Factors Affecting the Errors in the Estimation of Evolutionary Distances Between Sequences

D. C. Hoyle and P. G. Higgs
School of Biological Sciences, University of Manchester

Phylogenetic methods that use matrices of pairwise distances between sequences (e.g., neighbor joining) will only give accurate results when the initial estimates of the pairwise distances are accurate. For many different models of sequence evolution, analytical formulae are known that give estimates of the distance between two sequences as a function of the observed numbers of substitutions of various classes. These are often of a form that we call "log transform formulae". Errors in these distance estimates become larger as the time $t$ since divergence of the two sequences increases. For long times, the log transform formulae can sometimes give divergent distance estimates when applied to finite sequences. We show that these errors become significant when $t \sim 1/2 |\lambda_{\text{max}}|^{-1} \log N$, where $\lambda_{\text{max}}$ is the eigenvalue of the substitution rate matrix with the largest absolute value and $N$ is the sequence length. Various likelihood-based methods have been proposed to estimate the values of parameters in rate matrices. If rate matrix parameters are known with reasonable accuracy, it is possible to use the maximum likelihood method to estimate evolutionary distances while keeping the rate parameters fixed. We show that errors in distances estimated in this way only become significant when $t \sim 1/2 |\lambda_{\text{max}}|^{-1} \log N$, where $\lambda_{\text{max}}$ is the eigenvalue of the substitution rate matrix with the smallest nonzero absolute value. The accuracy of likelihood-based distance estimates is therefore much higher than those based on log transform formulae, particularly in cases where there is a large range of timescales involved in the rate matrix (e.g., when the ratio of transition to transversion rates is large). We discuss several practical ways of estimating the rate matrix parameters before distance calculation and hence of increasing the accuracy of distance estimates.

Introduction

There are now a large number of methods available for the construction of phylogenetic trees from molecular sequences. Where relatively small numbers of sequences are involved and large amounts of computer time is available, maximum likelihood (ML) inference of phylogenies can be used (Felsenstein 1981; Li and Gouy 1991). The ML method is based on sound statistical principles and allows tests to be made for comparing alternative tree topologies and alternative models of sequence evolution. However, when large numbers of taxa are involved, the ML method becomes very slow, and practical methods of dealing with large sets of sequences are still required. The fastest phylogeny methods usually use matrices of evolutionary distances. The neighbor joining (NJ) method (Saitou and Nei 1987) is one of the most popular of these, and tests have shown that it is relatively accurate in comparison with other heuristic clustering methods (Sourdis and Nei 1988; Saitou and Imanishi 1989; Hasegawa, Kishino, and Saitou 1991). If the distance matrix corresponds to an exactly additive tree, NJ reproduces the correct tree topology, and it is robust against small errors in the distance estimates between sequences (Atteson 1997). If errors in distance estimates are large, however, then NJ and other distance matrix methods will give incorrect tree topologies. In this article, we investigate the way in which the errors in the pairwise distances between sequences depend on the sequence length, the evolutionary model, and the degree of divergence between sequences.

The problems with distance estimation are already apparent in the simplest Jukes-Cantor (JC) (Jukes and Cantor 1969) model of sequence evolution. Evolutionary distance is usually measured in terms of the average number of substitutions per site, which we denote as $d$. For two aligned sequences of length $N$, the estimated evolutionary distance $\hat{d}$ is given within the JC model by

$$\hat{d} = -\frac{3}{4} \log \left(1 - \frac{4}{3}D\right)$$

where $D$ is the observed fraction of sites that differ. We will refer to equations such as equation (1) as "log transform" formulae because they transform observed quantities (in this case $D$) to estimates of distances that are not themselves directly observable quantities. A general method of deriving log transform formulae is given in Error Analysis of Distance Estimates for a General Rate Model, equations (4)–(9). For finite sequences, $n$ has a binomial distribution, and hence $D$ fluctuates about its mean value. Because $\hat{d}$ is a nonlinear function of $D$, the expectation value of $\hat{d}$ is not exactly equal to $d$; hence, there is a small systematic error of order $1/N$. In practice, this is dominated by the statistical error, which is of the order $1/N^{1/2}$ (Kimura and Ohta 1972; Nei, Stephens, and Saitou 1985).

It also is apparent from equation (1) that the estimated distance is infinite or undefined, if $D \geq 3/4$. When the true $d$ is large, $D$ approaches 3/4 for an infinite sequence; therefore, there is a significant chance that the observed $D$ in a finite sequence will be larger than 3/4 and that the distance estimate will be undefined. For small evolutionary distances, this will rarely occur. However, for any finite length $N$ and for any true distance $d$, there is a nonzero probability that $\hat{d}$ will be undefined.
When using the JC model, the problems described above will only become important when $d \gg 1$, in which case there is saturation of mutations. In this case there is very little phylogenetic information left in the sequences, and we probably should not be using these sequences for tree construction in the first place. However, we will show below that these problems tend to be more serious with more complex models. The effect of increasing the model complexity is illustrated by the next most simple model of evolution: the Kimura 2-parameter model (K2P) (Kimura 1980). In this model the base frequencies are equal, but a distinction is made between transitions (occurring at rate $1/4\alpha$) and transversions (occurring at rate $1/4\beta$). The ratio $\alpha/\beta$ is usually denoted by $\kappa$. The estimated evolutionary distance between two sequences is the sum of the estimated numbers of transversions and transitions per site and also can be written as a log transform formula, this time with two log terms,

$$d = -\frac{1}{4} \log(1 - 2V) - \frac{1}{2} \log(1 - 2S - V) \quad (2)$$

where $S$ and $V$ are the observed fractions of sites that differ by transitions and transversions between the two sequences. There are now two timescales, one associated with transitions and one with transversions. Transitions are usually faster ($\kappa > 1$). The distance estimate can diverge because of divergences in either of the two log terms. The quantity $1 - 2S - V$ tends to zero in a time governed by the rapid transition rate, whereas the term $1 - 2V$ tends to zero on the slow transversion timescale. Divergences, therefore, tend to arise because of saturation of the transitions, even when the overall evolutionary distance is small. The errors in this formula, therefore, become significant at smaller evolutionary distances than for the JC model.

In addition to the JC and K2P models, many other rate matrix models have been proposed. The model of Hasegawa, Kishino, and Yano (HKY) (Hasegawa, Kishino, and Yano 1985) and the model of Tamura and Nei (TN) (Tamura and Nei 1993) have rate matrices for which the eigenvectors and eigenvalues are analytically soluble. Hence, an explicit log transform formula can be written. A distance formula for general reversible models has been derived in various forms by Lanave et al. (1984), Barry and Hartigan (1987), Rodriguez et al. (1990), and Li and Gu (1996). Other distance measures such as the paralinear distance (Lake 1994) and the LogDet distance (Lockhart et al. 1994; Steel 1994) have been introduced to address nonstationarity, i.e., heterogeneity in the base composition of the sequences studied.

The observation that increasing model complexity can increase both the variance of distance estimates and the frequency with which undefined distance estimates are encountered has been made before (Zharkikh 1994; Li 1997, p. 84 for a discussion of this point). However, we wish to consider what generic features of a model affect the accuracy of distance estimates and the frequency of undefined distances.

We will begin by showing a few problems with distance estimates in real data. We will then give a quantitative theory to expand on the above arguments concerning log-transform formulae. We also consider likelihood-based methods of estimating distances where we fix the estimates of the rate parameters before estimation of the evolutionary distance. We show that these methods have errors associated with the slowest timescale in the model, and hence that these methods are more accurate.

Examples of Distance Estimates in Real Sequence Data

As an example with real data, we consider a set of large subunit (LSU) rRNA sequences obtained from the database of De Rijk et al. (2000). A set of 90 mitochondrial LSU sequences was used, including representatives from all groups of mitochondria-containing eukaryotes for which sequence data were available. The alignment in the database was taken without alteration. To exclude variable regions and poorly aligned regions from the analysis, we eliminated from the alignment those sites which contained a gap or unknown nucleotide for 10% or more of the sequences. In figure 1a, the evolutionary distance estimates for the K2P model (eq. 2) are plotted against the sequence dissimilarity $D$ for each pair of sequences. This is a well-behaved data set for which none of the distance estimates diverge and for which there is a clear relationship between $D$ and $\hat{d}$ so that all the points appear to lie on a smooth curve.

Figure 1b shows the same data with distance estimates calculated with the TN model, using the log transform formula given in Tamura and Nei (1993). In this case, the distance estimate diverges for 1.43% of sequence pairs. Pairs that diverge are plotted as points at $d = -0.25$ in the lower portion of the graph. Divergence occurs only for widely separated pairs with $D$ between 0.6 and 0.7. Also apparent from figure 1b is that the scatter of distance estimates is greater for the TN model than for the K2P model, and the relationship between $D$
and \( \hat{d} \) does not seem to be so well defined. In particular, at large \( D \) there are a number of outlier points, with \( \hat{d} \) well above the main curve. In any case, the TN distances could not be used without some finite value being assigned to the pairs that diverge.

We also can use likelihood ratio tests (Goldman 1993) for model selection purposes, and this gives results that are contradictory to those above. Because both the K2P and TN model distance estimates can be derived by maximizing the likelihood in terms of the observed numbers of tranversions and transitions and the K2P can be considered as a special case of the TN model, we can apply a likelihood ratio test to each sequence pair from the this data set. It is straightforward to calculate the log-likelihood values for the two models for any pair of sequences and, from this, to determine the difference \( \Delta \) in the log-likelihoods. Because the two models are nested, we know that \( 2\Delta \) is distributed according to a \( \chi^2 \) distribution under the hypothesis that the simpler K2P model is correct (Goldman 1993). On performing such a calculation for each possible pair of sequences, we find that the TN model provides a significantly better fit to the data (\( P < 0.05 \)) for 99.8% of the sequence pairs for which the distance estimate does not diverge. From this, we would conclude that the TN model provides an overwhelmingly better explanation of the observed sequence data than does the K2P model. (We note that the likelihood ratio test is usually applied to the likelihood of whole trees, whereas here we are interested in the likelihood of each sequence pair.)

Our second example consists of 455 small subunit (SSU) rRNA sequences taken from the database of Van De Peer et al. (2000), chosen so that there is one example of a sequence from each genus of Eubacteria. This data set has been analyzed previously by Higgs (2000) and Savill, Hoyle, and Higgs (2001) because it is an example of sequence evolution under the constraint of a conserved RNA secondary structure. These articles (and references therein) discuss a range of models for evolution of the paired stem regions of RNA sequences. Stem regions evolve through compensatory substitutions that maintain the pairing pattern; hence, the substitutions in the sites on either side of a pair are strongly correlated, and it is necessary to consider the evolution of the pair as a single unit rather than as two separate sites. One appropriate model is that of Tillier and Collins (1995), which considers six allowed paired states, AU, GU, GC, UA, UG, and CG, (occurring with frequencies \( \pi_{AU}, \ldots, \pi_{CG} \)) and has rate parameters \( \alpha \), \( \beta \), and \( \gamma \) quantifying the rates of double transitions, double transversions, and single transitions, respectively. The structure of the model is simple enough for an explicit log transform formula to be calculated (Tillier and Collins 1995), namely \( \hat{d} = \hat{K}_V + \hat{K}_{S_1} + \hat{K}_{S_2} + \hat{K}_V + \hat{K}_{S_1} + \hat{K}_{S_2} \) are estimators of \( K_V \), \( K_{S_1} \), and \( K_{S_2} \), the expected numbers of transversions, single transitions, and double transitions, respectively. The problem with the pairwise distance estimates is that the values of parameters like the base frequencies and rate matrix parameters \( \alpha \), \( \beta \), etc. are estimated separately for each pair, using only information from that pair, i.e., the parameters have different values for each pair, i.e., the parameters have different values for each

\[
\hat{K}_V = -\frac{1}{2} \log(1 - 2V)
\]

\[
\hat{K}_{S_1} = -4\pi_3(\pi_1 + \pi_3) \left[ 2 \log \left( 1 - V - \frac{S_1}{8\pi_2(\pi_1 + \pi_3)} \right) \right]
- \log(1 - 2V)
\]

\[
\hat{K}_{S_2} = -\frac{4\pi_1\pi_3}{\pi_1 + \pi_3} \log \left[ 1 - V - \frac{S_1}{4(\pi_1 + \pi_3)} - \frac{S_2(\pi_1 + \pi_3)}{4\pi_1\pi_3} \right]
+ \frac{8\pi_1\pi_2\pi_3}{\pi_1 + \pi_3} \log \left( 1 - V - \frac{S_1}{8\pi_2(\pi_1 + \pi_3)} \right)
+ 4\pi_1\pi_3\log(1 - 2V)
\]

where \( V \), \( S_1 \), and \( S_2 \) are the observed fractions of transversions, single transitions, and double transitions, respectively. Clearly, the ratios of the rate parameters also can be estimated from \( \hat{K}_V \), \( \hat{K}_{S_1} \), and \( \hat{K}_{S_2} \), i.e., \( \kappa_1 = \alpha/\beta \) is estimated by \( \hat{\kappa}_1 = \hat{K}_{S_1}/8\pi_2\pi_3 \hat{K}_V \), and \( \kappa_2 = \gamma/\beta \) is estimated by \( \hat{\kappa}_2 = \hat{K}_{S_2}/8\pi_2(\pi_1 + \pi_3) \hat{K}_V \). Figure 2a shows the estimated distances using this formula. A divergent distance estimate occurs for 0.91% of the sequence pairs. These divergent pairs (shown in the lower portion of the graph at \( d = -0.25 \)) occur over a range of dissimilarity that begins with a value of \( D \) as small as 0.25. There also is considerable scatter in the distance estimates for the points that do not diverge. The problems of distance estimation are therefore quite serious in this example.

**Distance Estimates with Fixed Rate Parameters**

The problem with the pairwise distance estimates is that the values of parameters like the base frequencies and rate matrix parameters \( \alpha \), \( \beta \), etc. are estimated separately for each pair, using only information from that pair, i.e., the parameters have different values for each
sequence pair. Usually when building a tree, we make the assumption that the rate matrix parameters are constant during evolution of the whole tree. It, therefore, makes sense to obtain a single estimate of the parameters using information from all the sequences and then to estimate ML distances for each sequence pair while keeping the parameters fixed at their previously estimated values. The accuracy of pairwise distance estimates obtained in this way is analyzed in ML Distance Estimates with Fixed Rate Parameters and Appendix B.

The most natural way of estimating the parameters would be simply to do an ML calculation of the optimum tree and parameter values for the whole set of sequences. However, this is not practical for large data sets of sequences without spending huge amounts of computer time. Indeed, if we were able to carry out ML for the complete set of sequences, we would not be particularly interested in using distance matrix methods anyway.

One way around the problem is to take a subset of sequences and do the ML calculation and then to use the ML parameters estimated from this subset when estimating pairwise distances for the whole set. Figure 1c shows the distances obtained in this way for the LSU rRNA sequences with the TN model. The estimates of the rate parameters have been obtained using TREE-PUZZLE (Strimmer and von Haeseler 1996; Strimmer 1997; Strimmer, Goldman, and von Haeseler 1997) on 10 sequences selected at random from the data set. The pairwise distances are then obtained for each sequence pair by maximizing the likelihood while keeping the rate parameters fixed at the values estimated by TREE-PUZZLE. There are now no divergent distance estimates, and the scatter of points at high D is substantially reduced in comparison with figure 1b. It is now practical to use the TN model on this data, which we know was the preferred model according to the likelihood ratio tests.

The model of Tillier and Collins (1995) for paired sites in RNAs is not implemented in TREE-PUZZLE or other currently available programs. We are currently in the process of developing a package that will use ML and Monte Carlo Markov Chain methods to construct phylogenies of RNAs using a variety of paired-site models. This will be available shortly (Jow et al. 2002).

However, for the present article, we used a much simpler method to estimate the rate parameters for the model of Tillier and Collins (1995) in the example with the SSU rRNA sequences.

For each sequence pair of the SSU rRNA data set, the parameter ratios $\kappa_1$ and $\kappa_2$ can be estimated as outlined in the previous section. We obtain final estimates of $\kappa_1$ and $\kappa_2$ by averaging finite estimates over all the 455 × 454/2 sequence pairs of the SSU rRNA data set. $\kappa_1$ and $\kappa_2$ are then held fixed at these values. Fixing $\kappa_1$ and $\kappa_2$ (along with the usual timescale normalization condition that the expected number of substitutions per site per unit time is 1) completely specifies all three rate parameters, $\alpha$, $\beta$, and $\gamma$. Specifically, $\beta = (1 + 8\pi_i\pi_j\kappa_1 + 8\pi_i\pi_j\kappa_2)^{-1}$, $\alpha = \kappa_2\beta$, and $\gamma = \kappa_2\beta$. Figure 2b shows ML distances calculated from the paired states in the SSU rRNA sequences and using the six-state model of Tillier and Collins (1995) with fixed rate parameters estimated in this fashion. There are now no divergent distance estimates and the scatter of points is reduced, indicating that distance estimates are more reliable than those with the log transform formula.

The key point of this section is that if good estimates of the rate parameters are known by ML estimation on a tree or by averaging the estimates of parameter ratios from each pair as described above, then accurate estimates of the pairwise distances can be obtained by maximizing the likelihoods for each pair with the fixed values of the rate parameters. These estimates suffer much less from problems of divergence than do estimates using the log transform formulae. In the next two sections, we consider in detail why this is so. Those readers not interested in the mathematical detail can proceed to Discussion and Conclusions, where we discuss the implications of these findings.

**Error Analysis of Distance Estimates for a General Rate Model**

Consider a rate matrix $\mathbf{R}$, whose elements $R_{ij}$ ($R_{ij} > 0 \forall i \neq j$) define the rate of substitution from state $i$ to state $j$. Thus, the probability $P_{ij}(t)$ of finding a particular nucleotide in state $j$ at time $t$, given that it was initially in state $i$, obeys the simple evolution equation

$$\frac{dP_{ij}}{dt} = P_{ij}\mathbf{R}.$$  \hfill (4)

$\mathbf{R}$ is defined such that the diagonal elements $R_{ii} = -\sum_{j \neq i} R_{ij}$ and also so that the substitution process is reversible, i.e., $\pi_i R_{ij} = \pi_j R_{ji} \forall i, j$. At long times, $P_{ij}(t)$ tends to $\pi_j$, the equilibrium frequency of state $j$.

We have labeled the eigenvalues of $\mathbf{R}$ as $\lambda_i$, where $i = 0, \ldots, N_{\text{state}} - 1$. There is one eigenvalue $\lambda_0$ which is zero, whereas the other eigenvalues are negative. We have assumed the eigenvalues are ordered so that $|\lambda_1| \leq |\lambda_2| \leq \cdots \leq |\lambda_{N_{\text{rate}}}|$. We also label $\lambda_{N_{\text{rate}} - 1}$ as $\lambda_{\text{max}}$ because it is the eigenvalue of the largest absolute value. We also can define the matrices $\mathbf{U}$ and $\mathbf{V}$ whose columns are formed from the eigenvectors of $\mathbf{R}$ and $\mathbf{R}^T$, respectively. The transition probability matrix $\mathbf{P}$ can then be written in terms of this decomposition as

$$P_{ij}(t) = \sum_{k=0}^{N_{\text{rate}}-1} A_{ijk} \exp(\lambda_k t), \text{ where}$$

$$A_{ijk} = U_{ik} V_{jk}. $$ \hfill (5)

The evolutionary distance $d$ between two sequences is defined as the expected number of nucleotide changes per site from which one finds $d = -t \sum_i \pi_i R_{ip}$. This can be written as a function of the eigenvalues as

$$d = -\sum_i \pi_i \sum_{k=0}^{N_{\text{rate}}-1} A_{ik} \lambda_k t.$$ \hfill (6)

To derive the log transform formula, the $P_{ij}(t)$ in equation (5) is replaced by the observed values in the pair of sequences in question, i.e., equating the expectation value $P_{ij}(t)$ with the observed number of substitutions of
that type. Thus, the derivation essentially uses the “Methods of Moments” (see for example, Casella and Berger 2002, p. 312). This gives a set of simultaneous equations for the unknown quantities \(\exp(\lambda_i t)\). The number of unknowns is equal to the number of distinct non-zero eigenvalues of \(R\), and we call this \(N_c\). Solving these equations for \(\lambda_i t\) and substituting into equation (6) gives an estimate of the evolutionary distance as a log transform formula

\[
\hat{d} = -\sum_{i=1}^{N_c} M_i \log(Q_i).
\]  

(7)

The quantities \(M_i\) depend on the frequencies of the states \(\pi_t\) but not on the rate parameters. Therefore, they can easily be estimated from the observed state frequencies. The \(Q_i\) can be taken to be of the form

\[
Q_i = Q_{i0} + \sum_{a=1}^{M} B_{a0} S_{a}.
\]  

(8)

Here, the quantities \(S_{a}\), \(\alpha = 1, \ldots, M\), represent observed fractions of various substitution classes, e.g., \(S\) and \(V\) in the K2P model. The coefficients \(B_{a0}\) will be explicit from the particular log transform formula used. The log transform formulae given above (eqs. 1 and 2) are both of this form, as are those for the TN model and the RNA model used in Distance Estimates with Fixed Rate Parameters.

In Appendix A, we carry out an analysis of the errors in a log transform formula of the general form given in equation (7). The probability that the \(i\)th individual logarithmic term in equation (7) diverges is

\[
\frac{1}{2} \left[ 1 - \text{erf} \left( \frac{\sqrt{N}}{2} \exp(\lambda_i t)/C_{\lambda i}^{1/2} \right) \right]
\]  

(9)

where

\[
C_{\lambda i} = \left( \sum_a (S_{a}) B_{a0}^2 - \left( \sum_a (S_{a}) B_{a0} \right)^2 \right)^{1/2}.
\]  

(10)

Here, the notation \(\langle \rangle\) represents the expectation value. From equation (9), we can see that the individual logarithmic term in equation (7) that is estimating the fastest eigenvalue \(\lambda_{\text{max}}\) is the most likely to produce a divergence, and so \(\lambda_{\text{max}}\) predominantly determines the likelihood of obtaining a divergent evolutionary distance estimate from a log transform formula.

The total probability that the estimate (eq. 7) diverges is

\[
P_{\text{div}} = 1 - P(Q_1 > 0, Q_2 > 0, \ldots, Q_{N_c} > 0).
\]

Obtaining a generic analytic form for \(P_{\text{div}}\) is not possible. However, we approximate

\[
P(Q_1 > 0, Q_2 > 0, \ldots, Q_{N_c} > 0) \approx P(Q_1 > 0)P(Q_2 > 0) \cdots P(Q_{N_c} > 0),
\]

i.e., we approximate the \(Q_i\) as being independent. This gives

\[
P_{\text{div}} \approx 1 - \prod_{k=1}^{N_c} \frac{1}{2} \text{erfc} \left( - \frac{\sqrt{N}}{2} \exp(\lambda_{k} t)/C_{\lambda k}^{1/2} \right)
\]  

(11)

where the complementary error function \(\text{erfc}(x) = 1 - \text{erf}(x)\). In the case of the K2P model, one has

\[
P_{\text{div}} = 1 - \frac{1}{2} \text{erfc} \left( - \frac{\sqrt{N}}{2} \exp(\lambda_{\text{max}} t)/C_{\lambda_{\text{max}}}^{1/2} \right)
\]  

(11)

where the complementary error function \(\text{erfc}(x) = 1 - \text{erf}(x)\). In the case of the K2P model, one has

\[
P_{\text{div}} = 1 - \frac{1}{2} \text{erfc} \left( - \frac{\sqrt{N}}{2} \exp(\lambda_{\text{max}} t)/C_{\lambda_{\text{max}}}^{1/2} \right)
\]  

(11)

where the complementary error function \(\text{erfc}(x) = 1 - \text{erf}(x)\). In the case of the K2P model, one has
duced by assuming that $Q$, in equation (7) is independent to be largely unimportant in comparison with the fact that at this stage greater than 50% of the distance estimates will be undefined. From equation (11), we see that for a general rate model $P_{dv}$ consists of $N_d$ step-like increases reaching a limiting value (for long sequences and large evolutionary distances) of $1 - (1/2)^{N_e}$.

**ML Distance Estimates with Fixed Rate Parameters**

Our calculations on real data sets revealed that no divergent pairwise evolutionary distance estimates were obtained if the rate parameters were held fixed and not simultaneously estimated from the sequence data of the pair. We now consider this point in more detail and determine what controls the probability of a divergent estimate and the accuracy of that estimate when the rate parameters are held fixed.

Let us consider methods of distance estimation based on ML. We suppose that estimates of all the parameters in the rate matrix are known and that we maximize the likelihood with respect to only the distance $t$. If the specified values of the rate parameters are equal to the true values, we find (see Appendix B) that the probability of obtaining a divergent distance estimate, for large $N$ and $t$, is now controlled by the slowest non-zero eigenvalue $\lambda_i$. The limiting value, as $t \to \infty$, for the probability of divergence is $1/2$ (to leading order in $N^{-1}$), irrespective of the rate matrix used. For the K2P model, we find

$$p_{dv} \approx \frac{1}{2} \left[ 1 - \text{erf} \left( \frac{N^{1/2}}{2} \exp(-\beta t) \right) \right].$$

(13)

The result from this equation is shown in figure 3 in comparison with estimates of $P_{dv}$ from simulation. The sequence data has been generated with $k_0 = 10.0$, and the same value has been specified in the ML calculation. It can be seen that the range of applicability of the ML distance estimate (with fixed rate parameters) is much larger because the divergence probability is a function of the slow timescale (transversions) only. The simulation results are in excellent agreement with the theory indicating that the probability of obtaining a divergent evolutionary distance estimate is predominantly controlled by the slowest nonzero eigenvalue $\lambda_i$.

In Appendix B, we derive a perturbative result for the RMS error of the ML estimate (with fixed rate parameters) of $d$. Again the theoretical result (eq. 26) is in excellent agreement with simulation (not shown). Consequently, we also can conclude that the error in the ML estimate of $d$ is indeed controlled by the slowest nonzero eigenvalue $\lambda_i$.

In general, we will not know a priori the true parameters in the rate matrix. If the sequence data have been generated by the same class of model that is used to estimate the pairwise evolutionary distances, then for the methods we have used to estimate the rate parameters, e.g., for ML on a fixed tree (as in TREE-PUZZLE), we expect the estimates of rate parameters to become more accurate as sequence length $N$ is increased. In this case, our analysis in Appendix B is unchanged to leading order in $N$. Consequently, for increasing sequence length, distance estimates are still controlled by the slowest eigenvalue $\lambda_i$. Obviously, for any fixed sequence length $N$, the exact value of $\langle (\delta t)^2 \rangle$ will depend on precisely how close the estimates of the rate matrix parameters are to the true values.

**Discussion and Conclusions**

The focus of this article has been on the accuracy and reliability of distance matrix methods for phylogeny reconstruction. Our analysis has shown that distance estimates from log transform formula are accurate only on the shortest distance scale, $|\lambda_{max}|^{-1}$, defined by the underlying rate matrix $R$ and can frequently give rise to undefined distances. Pairwise distance estimates can be improved using ML if the rate parameters are known with reasonable accuracy, for example, by estimation from more than just the two particular sequences under consideration. In such cases, pairwise distance estimates are accurate on the longest distance scale, $|\lambda_1|^{-1}$, defined by $R$. With improved pairwise distance estimates, better phylogenies can be constructed using NJ or variants such as Neighbor (Bruno, Socci, and Halpern 2000).

Both the probability of divergence (eq. 11) and the estimate of the variance of $d$ (eq. 22 in Appendix A) provide a measure of how much sequence information is required to resolve a given evolutionary distance between two sequences using log transform formulae. To maintain a fixed probability of divergence $P_{dv}$ or fixed variance of the evolutionary distance estimate, we have the scaling (for large evolutionary distances)

$$\frac{N_{C_{N_eN_e}}}{C_{N_eN_e}} \exp(\lambda_{max} t) \sim \text{Constant}$$

$$\Rightarrow t \sim -\frac{1}{2} \lambda_1^{-1} \log N \sim \log C_{N_eN_e}. $$

(14)

Because $C_{N_eN_e}$ tends to a constant, independent of $N$, as $t \to \infty$, we can, for large sequence lengths $N$, consequently ignore it in the scaling relation above.

Models with higher $N_e$ tend to be applicable over a narrower range of $t$ and give more divergences because the more eigenvalues there are, the larger the one tends to be. This was already seen in the example with the LSU rRNA sequences in figure 1a and b. We also can see from equation (14) that the evolutionary distance between two sequences that can be resolved accurately depends logarithmically on $N$. To increase the evolutionary distance probed requires an exponential increase in sequence length.

What are the sequence lengths required to obtain a given level of accuracy if evolutionary distances are estimated using ML with fixed rate parameters? Again, one finds that to maintain a constant level of accuracy or constant probability of obtaining a divergent evolutionary distance estimate, the required scaling is

$$N^{1/2} \exp(\lambda_1 t) \sim \text{Constant} \Rightarrow t \sim -\frac{1}{2} \lambda_1^{-1} \log N.$$  

(15)

Thus, the evolutionary distances that one can probe still...
depends logarithmically on $N$. However, one should note that the prefactor to log $N$ is $\lambda_1^{-1}$ as opposed to $\lambda_{\text{max}}$ when using log transform formula. Thus, if one has a fixed reasonably accurate estimate of the underlying rate matrix, then the evolutionary distance estimate obtained by maximizing the likelihood with respect to $t$ can be considerably more accurate than that obtained by using the appropriate log transform formula. This is particularly the case when evolution of sequences is occurring through processes on vastly different timescales. For the K2P model, when $\kappa = 10$ (i.e., transitions occurring much more rapidly than transversions) the ratio $\lambda_{\text{max}}/\lambda_1 = 5$.

Appendix A

Probability of Divergence in Log Transform Formulae

We take as our starting point a log transform formula of the form

$$\hat{d} = -\sum_{i=1}^{N} M_i \log(Q_i)$$

(16)

which provides an estimate $\hat{d}$ of the true evolutionary distance $d$ between two sequences. The amplitudes $M_i$ are assumed to depend only on the state frequencies $\{\pi_i, i = 1, \ldots, N_{\text{state}}\}$. $N_e$ is the number of distinct nonzero eigenvalues of the rate matrix model to which the log transform formula (16) applies. The $Q_i$ can be taken to be of the form

$$Q_i = Q_{i0} + \sum_{\alpha=1}^{M} B_{i\alpha} S_{\alpha}.$$  

(17)

where $S_{\alpha}, \alpha = 1, \ldots, M$, represent observed fractions of various substitution classes. With the quantities $S_{\alpha}$ estimated from the sequence data, $Q_i$ provides estimates of $\exp(\lambda_i t)$ and in particular $\langle Q_i \rangle = \exp(\lambda_i t)$.

The distribution, $P(S_1, \ldots, S_M)$ is multinomial and so, as $N \to \infty$, is well approximated by a multivariate Gaussian,

$$P(S_1, \ldots, S_M) = \left(\frac{2\pi}{N}\right)^{N^2/2} \delta\left(1 - \sum_{\alpha=1}^{M} S_{\alpha}\right) \prod_{\alpha=1}^{M} \left(\frac{N}{2\pi \langle S_{\alpha} \rangle}\right)^{1/2} \times \exp\left(-\frac{N}{2} \sum_{\alpha=1}^{M} \frac{(S_{\alpha} - \langle S_{\alpha} \rangle)^2}{\langle S_{\alpha} \rangle}\right).$$

(18)

For large $N$, we can consider the quantities $S_1, \ldots, S_M$ and $Q_1, \ldots, Q_N$ to be continuously distributed between $-\infty$ and $\infty$. Using the distribution (eq. 18), we calculate

$$P(Q) = \left(\frac{2\pi}{N}\right)^{N^2/2} (\det C)^{-1/2} \times \exp\left[-\frac{N}{2} (Q - \langle Q \rangle) \cdot C^{-1} (Q - \langle Q \rangle)\right].$$

(19)

The probability that the $i$th individual logarithmic term in equation (16) diverges is

$$\left(\frac{N}{2\pi \langle C \rangle_i}\right)^{1/2} \int_{-\infty}^{0} dQ_i \exp\left(-\frac{N}{2} \langle C \rangle_i (Q_i - \langle Q_i \rangle)^2\right) = \frac{1}{2} \left[1 - \text{erf}\left(\left(\frac{N}{2}\right)^{1/2} \exp(\lambda_i t)/\langle C \rangle_i^{1/2}\right)\right].$$

(21)

From this we can see that the individual logarithmic term in equation (16) that is estimating the fastest eigenvalue $\lambda_{\text{max}}$ is the most likely to produce a divergence, and so $\lambda_{\text{max}}$ predominately determines the likelihood of obtaining a divergent evolutionary distance estimate from a log transform formula.

Assuming that the dominant contribution to equation (16) comes from estimating $\lambda_{\text{max}}$, then the variance in such a distance estimate is equally easily derived by generalizing the analysis of Kimura and Ohta (1972)

$$\sigma^2(d) = N^{-1} \exp(2\lambda_{\text{max}} t) M_{k} C_{N_e} t \to \infty.$$  

(22)

Appendix B

Analysis of ML for a Sequence Pair with Fixed Rate Parameters

The likelihood of observing two sequences 1 and 2 is

$$L(t) = \prod_{k=1}^{N} \pi_{\alpha k} P_{s_1 \alpha s_2 k}(t)$$

(23)

where $s_1$ and $s_2$ label the states at the $k$th site for sequences 1 and 2, respectively.

We wish to derive the probability of obtaining an infinite value for the estimate of $t$ which maximizes the likelihood (eq. 23) when keeping the rate matrix $R$ fixed, i.e., the probability that the solution, $t_{ML}$, to $\partial \log L/\partial t = 0$ is infinite. For the two sequence problems, we assume that there exists at most one maxima in $t$ of the likelihood $L$ at a finite value of $t_{ML}$. Thus, a divergent solution $t_{ML}$ is equivalent to the asymptotic gradient of log $L$ approaching zero from above, i.e., $\partial \log L/\partial t \to 0^-$, $t \to \infty$. Thus, we will focus on the distribution of $\lim_{t \to \infty} \partial \log L/\partial t$.

We take two sequences separated by branch length $t_0$ and generated with rate matrix $R$. We consider our fixed estimate of $R$ in equation (23) to be exact. As $N \to \infty$, we can take the distribution of $x(t) = \partial \log L/\partial t$ to be Gaussian, i.e., a probability density $(2\pi B_1)^{-1/2} \exp(-[x - B_0]/2B_1)$, where $B_0 = \langle \partial \log L/\partial t \rangle$ and $B_1 = \text{Var} \langle \partial \log L/\partial t \rangle$. Consequently, the probability of obtaining a divergent estimate for $t_{ML}$ is

$$\lim_{t \to \infty} (2\pi B_1)^{-1/2} \int_{-\infty}^{\infty} \exp\left[-\frac{(x - B_0)^2}{2B_1}\right] = \frac{1}{2} \text{erf}(C) = 1 - \frac{1}{2} \text{erf}(C).$$

(24)

where $C = \lim_{t \to \infty} B_{t}/\sqrt{2B_1}$. We find
A perturbative analysis of the accuracy of the ML estimate $t_{ML}$ also can be obtained in a straightforward manner. The derivation is tedious, and so we quote only the final result here. Writing $t_{ML} = t_0 + \delta t_{ML}$ ($t_0$ again being the true separation between the two sequences), we find that the mean squared error is given by

$$< (\delta t_{ML})^2 > = N^{-1} \left[ \sum_{i,j} \pi_i \pi_j^{-1} (A_{ij})^2 \right] \lambda_1^{-1} \exp(-2\lambda_1 t_0).$$

Consequently, in contrast to equation (22), the accuracy of the ML (with fixed rate parameters) estimate $t_{ML}$ when the two sequences are well separated is on a distance scale determined by the slowest nonzero eigenvalue $\lambda_1$.

Acknowledgment

This work was supported by the U.K. Biotechnology and Biological Sciences Research Council.

Literature Cited


Edward Holmes, Associate Editor
Accepted July 17, 2002