J Mol Evol (2002) 55:684–695
DOI: 10.1007/s00239-002-2364-7

JOURNAL OF MOLECULAR EVOLUTION

Estimation of Evolutionary Parameters with Phylogenetic Trees

Qiang Wang,[1] Laura A. Salter,[2] Dennis K. Pearl[1]

[1] Department of Statistics, The Ohio State University, Columbus, OH 43210, USA
[2] Department of Mathematics and Statistics, 415 Humanities Building, University of New Mexico, Albuquerque, NM 87131, USA

Received: 13 December 2001/Accepted: 20 June 2002

Abstract. An important issue in the phylogenetic analysis of nucleotide sequence data using the maximum likelihood (ML) method is the underlying evolutionary model employed. We consider the problem of simultaneously estimating the tree topology and the parameters in the underlying substitution model and of obtaining estimates of the standard errors of these parameter estimates. Given a fixed tree topology and corresponding set of branch lengths, the ML estimates of standard evolutionary model parameters are asymptotically efficient, in the sense that their joint distribution is asymptotically normal with the variance–covariance matrix given by the inverse of the Fisher information matrix. We propose a new estimate of this conditional variance based on estimation of the expected information using a Monte Carlo sampling (MCS) method. Simulations are used to compare this conditional variance estimate to the standard technique of using the observed information under a variety of experimental conditions. In the case in which one wishes to estimate simultaneously the tree and parameters, we provide a bootstrapping approach that can be used in conjunction with the MCS method to estimate the unconditional standard error. The methods developed are applied to a real data set consisting of 30 papillomavirus sequences. This overall method is easily incorporated into standard bootstrapping procedures to allow for proper variance estimation.

Correspondence to: Laura Salter; *email*: salter@stat.unm.edu

Key words: Maximum likelihood — Phylogenetics — Evolutionary model parameters — Bootstrapping — Monte Carlo sampling — Fisher information

Introduction

The appropriateness of the phylogenetic analysis of a particular data set is dependent upon the assumptions made by the method of analysis employed. An important issue in the analysis of nucleotide sequence data using maximum likelihood (ML) as the criterion for phylogenetic reconstruction is the underlying evolutionary model that is assumed. One advantage of the ML method is that it has been shown to be somewhat robust to misspecification of the parameters in this underlying substitution model (see, e.g., Gaut and Lewis 1995; Fukami-Kobayashi and Tateno 1991; Kuhner and Felsenstein 1994; Yang et al. 1994).

One particular aspect that has been examined in these studies is the effect of misspecification of the transition/transversion ratio on the ability of the ML method to recover the ''true'' phylogenetic tree. Estimation of this parameter is of interest in itself since it provides an idea of the rate and mode of evolution among the nucleotide sequences under consideration. Also of interest is an estimate of the mean instantaneous substitution rate, often abbreviated μ in the substitution models in the literature. For the common substitution models employed [e.g., the F84 model (Felsenstein 1984), the HKY85 model (Hasegawa et al. 1985), and all of their associated submodels], it is impossible to estimate the parameter μ indepen-

dently of the branch lengths, unless one is willing to assume a molecular clock model for the evolution of the species under consideration.

Many methods for estimating the transition/transversion ratio have been proposed, and several of these are reviewed by Wakeley (1996). We provide a short summary of the available methods here. The classical method of estimating the transition/transversion ratio, R, is to consider the ratio of the observed number of transitions to the observed number of transversions, with the idea that these observed numbers of transitions and transversions should reflect the true rates of transitional and transversional changes. Various methods can be used to count the numbers of transitions and transversions. One possibility is to consider pairs of sequences and to calculate the ratio of the numbers of transitions to transversions for each pair. However, it is not clear how to combine each of these pairwise estimates into an overall estimate of R for the entire data set under consideration. An alternative method of counting the numbers of transitions and transversions is to use parsimony to reconstruct the states at the internal nodes for a particular tree. For each site in the data set, the number of transitions and transversions is then the average of the numbers observed for each most parsimonious reconstruction of the states at the internal nodes for that site. We note that this method requires that the tree topology be specified prior to the estimation of R. See Wakeley (1996) for examples of these two methods.

Wakeley (1994, 1996) points out that when the sequences under consideration are substantially diverged from one another, both the pairwise method and the parsimony method give underestimates of the transition/transversion ratio. Even when recently diverged sequences are considered, the estimate of the transition/transversion ratio can be biased and quite variable (Wakeley 1996, 1994). Furthermore, failure to take rate variation among sites into consideration can also result in underestimation of the true transition/transversion ratio (Wakeley 1996, 1994; Yang 1994).

Several authors (Pollock and Goldstein 1995; Yang and Kumar 1996; Purvis and Bromham 1997; Ina 1998) have proposed alternative methods of estimating the transition/transversion ratio. Pollock and Goldstein's (1995) method is a modification of the pairwise method mentioned above which involves taking a particular weighted average of the ratios calculated for all pairs of sequences in the data set. Ina (1998) also modified the pairwise method, by using Haldane's correction for each of the pairwise ratios obtained from the data. Variances for these proposed estimates are also given. Purvis and Bromham (1997) used iterative-weighted least-squares regression to estimate the instantaneous transversion rate for a particular data set, assuming that the topology and branch lengths of the tree representing the evolutionary relationships among the sequences are known. Yang and Kumar (1996) proposed a modification of the parsimony method described above to correct the observed numbers of transitions and transversions for multiple substitutions.

ML has also been used previously (Yang 2000, 1994; Rambaut and Grassly 1996; Swofford 1998) to estimate the transition/transversion ratio. Swofford's program PAUP*, Yang's program PAML, and Rambaut and Grassly's program SPOT will calculate ML estimates (MLEs) of the transition/transversion ratio and the branch lengths when the tree topology is fixed. PAUP* and PAML then incorporate estimation of the transition/transversion ratio into the tree search procedure by periodically updating the estimate of this parameter for the current tree. Among the advantages of using the ML method for estimation of evolutionary model parameters such as the transition/transversion ratio is that the results of Chang (1996) and Rogers (2001) imply that the MLEs of the tree topology, branch lengths, and parameters are all consistent under many commonly used substitution models.

Here we consider the problem of estimating the parameters in the underlying evolutionary model jointly with the tree and of obtaining estimates of their standard errors. We begin with the case in which the tree and the branch lengths are known and obtain the MLEs of the parameters through a Newton–Raphson method. We then show that the MLEs are asymptotically efficient, by verifying that Birch's conditions (Birch 1964; Bishop et al. 1975) hold in the case in which the tree and branch lengths are known. By asymptotic efficiency, we mean that the joint distribution of the parameters is asymptotically normal with the variance–covariance matrix given by the inverse of the Fisher information matrix.

The asymptotic efficiency result suggests two possible estimators of the conditional variance of the parameter estimates when the tree and branch lengths are fixed. The first is to compute the observed Fisher information, which has been proposed previously by Yang (1994, 2000) and is implemented in his program PAML (Yang 2000) (Yang calls this the curvature method). We propose a second method, which entails estimating the expected Fisher information using a Monte Carlo sampling (MCS) approach. Simulations based on a 14-sequence model tree are used to compare the two methods.

We next consider the parameter estimation problem in the more realistic case where the tree and branch lengths are not known in advance. In this case, both Yang's method and our MCS method will underestimate the unconditional variance, since both

methods assume that the tree and branch lengths are known and thus fail to account for the additional variability due to simultaneously estimating these parameters. To overcome this, we consider several bootstrapping approaches and use simulations to compare them in terms of performance and in terms of computational requirements. Finally, we apply the methods developed here to a real data set consisting of 30 papillomavirus sequences.

Asymptotic Efficiency When the Tree Is Known

We begin by supposing that the phylogenetic tree and the branch lengths are known in advance and that it is desired to obtain estimates of the parameters in the underlying evolutionary model and their standard errors for this known tree. For the purpose of theoretical derivation, we use the following representation of the data. Consider the 4^N unique patterns of nucleotide configurations at a site that are possible for N sequences, and note that the data can be represented by the counts of these 4^N site patterns under the assumption that the sites evolve independently. The data, **X**, then have a multinomial distribution. The probability associated with the ith category, p_i, is the probability of the ith site pattern, which is determined by the evolutionary model and the phylogenetic tree.

As an example, we first consider Kimura's (1980) two-parameter model of evolution (K2P), in which the probabilities of change along branches are functions of only the mean instantaneous substitution rate, μ, and the transition/transversion parameter, K. We assume a molecular clock so that both of these parameters may be estimated. Note that for a given tree topology and set of branch lengths, the probability associated with any site pattern can be found by the pruning algorithm (Felsenstein 1981) and, thus, can be written as a function of μ and K. The above representation of the data then enables one to apply classical multinomial sampling theory. This theory states that if the function of the multinomial cell probabilities of the parameters satisfies certain conditions, it follows that the MLEs of the parameters are asymptotically efficient; that is, the joint limiting distribution (as the number of sites goes to ∞) of the MLEs is multivariate normal with the variance–covariance matrix given by the inverse of the Fisher information matrix. One such set of conditions that can be checked are Birch's conditions (Birch 1964; Bishop et al. 1975), which include conditions of smoothness for the likelihood function, identifiability of the parameters, and some other regularity conditions. These conditions are indeed satisfied for many standard site-independent models of evolution, and so, by applying Birch's theorem, we have

$$\sqrt{L}(\hat{\theta} - \theta) \rightarrow \mathcal{N}(0, (A'A)^{-1}) \tag{1}$$

where L is the number of sites in the sequences, $\hat{\theta}$ is the vector of MLEs of the evolutionary parameters, and θ is the vector of true values of the parameters. For example, in the K2P model, $\theta = (\mu, K)$. The matrix $A'A$ is the Fisher information matrix, with the following elements:

$$\begin{pmatrix} \sum_{k=1}^{4^N} \left(\frac{\partial \log(f_k(\mu,K))}{\partial \mu}\right)^2 f_k(\mu, K) & \sum_{k=1}^{4^N} \left(\frac{\partial \log(f_k(\mu,K))}{\partial \mu}\right)\left(\frac{\partial \log(f_k(\mu,K))}{\partial K}\right) f_k(\mu, K) \\ \sum_{k=1}^{4^N} \left(\frac{\partial \log(f_k(\mu,K))}{\partial \mu}\right)\left(\frac{\partial \log(f_k(\mu,K))}{\partial K}\right) f_k(\mu, K) & \sum_{k=1}^{4^N} \left(\frac{\partial \log(f_k(\mu,K))}{\partial K}\right)^2 f_k(\mu, K) \end{pmatrix} \tag{2}$$

where $f_k(\mu, K)$ is the likelihood of the data for site pattern k as a function of μ and K, given the tree topology and branch lengths.

To obtain variance estimates based on the observed information, one simply computes the entries of the above matrix at the MLEs of μ and K for those site patterns observed in the data. The variance estimates are thus given by

$$\hat{var}(\hat{\mu}) = \frac{1}{D}\left(\sum_{k=1}^{L}\left(\frac{\partial \log(f_k(\mu, K))}{\partial K}\right)^2\right) \tag{3}$$

and

$$\hat{var}(\hat{K}) = \frac{1}{D}\left(\sum_{k=1}^{L}\left(\frac{\partial \log(f_k(\mu, K))}{\partial \mu}\right)^2\right) \tag{4}$$

where D is the determinant of the matrix in Eq. (2),

$$\begin{aligned} D = &\left(\sum_{k=1}^{L}\left(\frac{\partial \log(f_k(\mu, K))}{\partial \mu}\right)^2\right) \\ &\times \left(\sum_{k=1}^{L}\left(\frac{\partial \log(f_k(\mu, K))}{\partial K}\right)^2\right) \\ &- \left(\sum_{k=1}^{L}\left(\frac{\partial \log(f_k(\mu, K))}{\partial \mu}\right)\left(\frac{\partial \log(f_k(\mu, K))}{\partial K}\right)\right)^2 \end{aligned} \tag{5}$$

We substitute $\hat{\mu}$ for μ and $\hat{K}$ for K in expressions (3), (4), and (5), and the sums in these equations are taken over the L observed sites in the data set.

We propose an alternative method of estimating the variances, which involves the use of a MCS procedure to estimate the matrix in expression (2). The method works by generating a large number, Q, of

site patterns according to the tree, branch lengths, and evolutionary model with parameters set at the MLEs. For each generated site, derivatives are evaluated at the MLEs of the parameters, and the sums in matrix (2) are formed. Variances are then estimated by using the same expressions as in (3), (4), and (5), except that the sums are taken over the Q sites generated by the MCS procedure, rather than over the L observed sites.

One technical point is that the true underlying variance of the parameters when the tree is known is infinite. This is because there is always some positive probability, albeit low, that the observed data will result in an estimate of the parameter that is infinite. Thus, the methods proposed here estimate an approximate asymptotic variance, namely, the variance of the limiting distribution of the parameter estimates. In what follows, when we refer to the variance or the asymptotic variance, we are referring to this approximate variance. Later, when we no longer condition on a known tree, we are referring to the analogous approximate asymptotic variance.

To evaluate our MCS estimator, we performed several simulation studies. We have chosen as the model tree for these studies the ML tree estimated from a real data set of mtDNA sequences for 14 species containing 231 sites (Hayasaka et al. 1988). The model tree is shown in Fig. 1. We next considered two values for each of the parameters of interest, resulting in four possible parameter combinations. The values selected were $\mu = 0.1$ and 0.5 and $K = 0.5$ and 10.0. For all of the simulations described here, MCS samples were generated using the program Seq-Gen (Rambaut and Grassly 1997).

Our first simulation involves determining the number of MCS samples that need to be taken to obtain good estimates of the Fisher information matrix. For the model tree with known branch lengths and with each parameter set at one of the four combinations described above, we generated 40,000 sites in sets of 2000. For each of the sets of 2000, variances were estimated and compared to the overall estimate obtained from all 40,000 sites. To assess the accuracy of the estimates based on 2000 sites, we computed absolute differences between these estimates and the overall estimate. In estimating the parameter μ, the mean absolute difference ranged from 1.9×10^{-4} for $\mu = 0.1$ and $K = 0.5$ to 3.7×10^{-4} for $\mu = 0.5$ and $K = 0.5$ (the median ranged from 1.8×10^{-4} to 2.8×10^{-4} for the same parameter combinations). For estimation of the transition/transversion parameter, K, the mean absolute difference ranged from 0.001 for $\mu = 0.5$ and $K = 0.5$ to 0.019 for $\mu = 0.1$ and $K = 10.0$ (the median ranged from 0.001 to 0.017 for the same parameter combinations). Figure 2 shows each of the estimates based on 2000 sites, with a horizontal line drawn at the value of the estimate based on 40,000 sites for the two-parameter combinations described above. Results are similar for the other two-parameter combinations. Based on these plots and our examination of the errors above, we concluded that using 2000 sites would provide a good approximation of the conditional variance in this example. We give a recommendation for choosing the number of randomly generated sites in the general case in the Discussion.

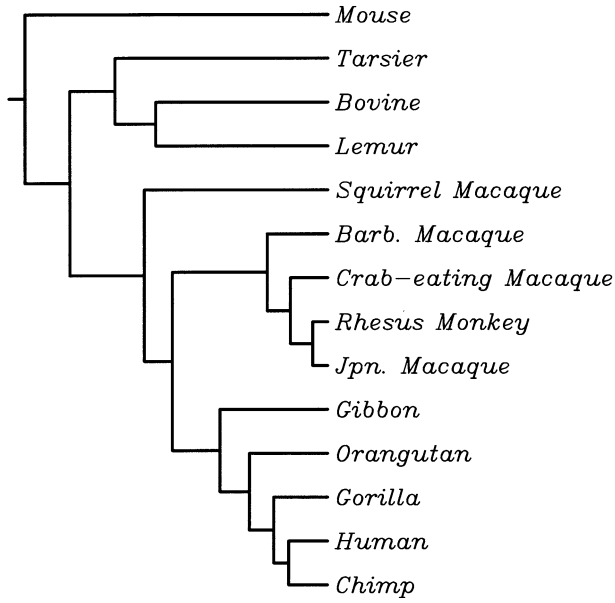

Fig. 1. Model tree used to perform the simulations.

Our next simulation involves assessing the variability in estimating the parameters and the variances when the tree and branch lengths are known. We generated simulated data sets containing 231 sites for the model tree in Fig. 1 for each of the four parameter combinations. For each simulated data set, we obtained the MLEs of μ and K and the estimates of their variances using both Yang's method and our MCS method. For each combination of parameter settings, 500 such simulations were performed. To assess the accuracy of the two methods of estimating the variance, we compared them to the results of the first simulation and to the empirical standard deviation of the 500 parameter estimates. The results are shown in the middle columns in Table 1.

Examination of these results shows that for the case in which the tree and branch lengths are assumed to be known (simulation 2), Yang's method based on the observed information generally gives underestimates of the standard errors. The problem is most severe for the lower value of μ. The MCS method, on the other hand, appears to provide reasonable estimates of the standard error for the case in which the tree and branch lengths are known. The only exception appears to be for the estimation of the standard error of K for the case in which both μ and K are at their "high" settings,

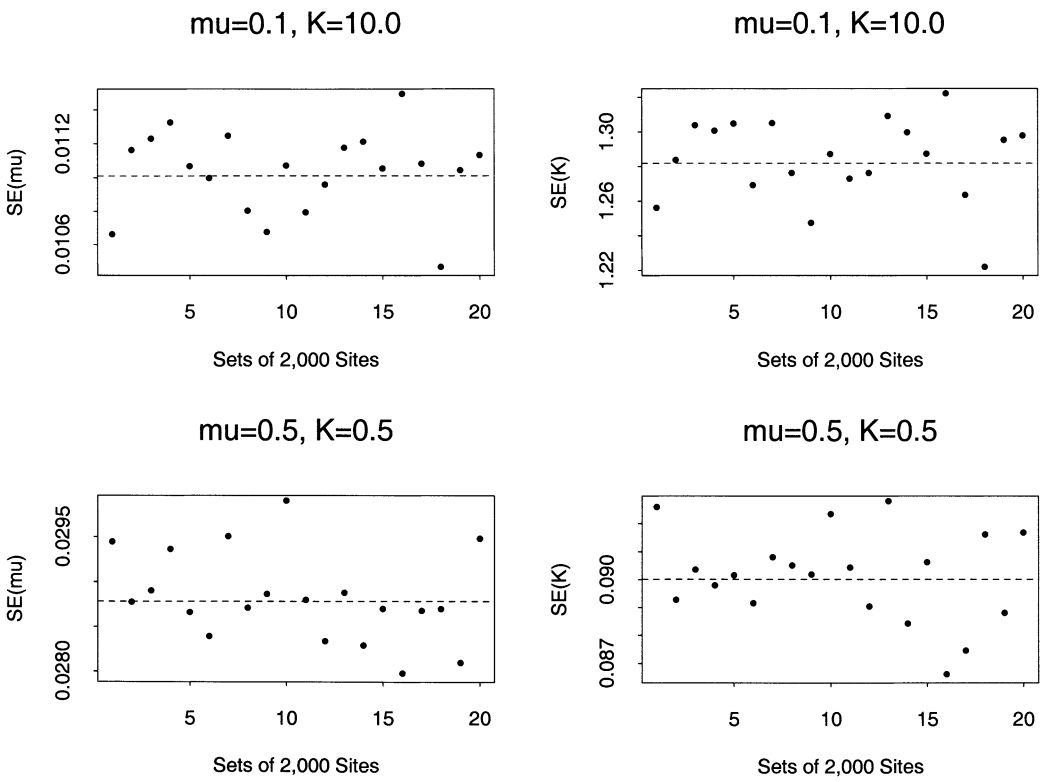

Fig. 2. Results of the first simulation. The *dashed line* in each plot indicates the standard error (SE) estimate based on 40,000 sites. Each point represents an individual estimate based on 2000 sites.

where both estimators provide estimates that are lower than the observed standard deviation (note, however, that both estimates are near the Simulation 1 value). Histograms of the parameter estimates from the 500 data sets verify that the distribution of the parameter estimates is approximately normal (results not shown).

Our third simulation evaluates the performance of the conditional variance estimators when the underlying tree is unknown and must be simultaneously estimated. Again, we simulated data sets of the same size as the real data for each of the four parameter combinations. For each simulated data set, we obtained simultaneous estimates of the tree and parameters using a modification of the SSA method (Salter and Pearl 2001; Salter 1999). SSA is an algorithm for simultaneous estimation of the ML tree and substitution model parameters that utilizes a stochastic search strategy based on a simulated annealing-type algorithm. SSA has previously been shown to be computationally more efficient than existing methods for ML tree estimation and is less likely to provide estimates that are only locally optimal (Salter and Pearl 2001). The program is available for download at: www.stat.unm.edu/~salter/software/ssa/ssa.html

Following estimation of the tree and parameters using SSA, estimates of the variances using both Yang's method and our MCS method were obtained. We again compared the two methods to the results obtained in the first simulation and to the empirical variance of the 500 parameter estimates. The results are shown in the right-hand columns in Table 1. In the case in which the tree and branch lengths are unknown, the observed standard deviation of the parameter estimates increases, as is expected due to the additional variability associated with simultaneously estimating the phylogenetic tree. Both Yang's method and our MCS method give underestimates of the standard error in this case, since both fail to account for this additional variability. In the next section, we propose a bootstrapping method to deal with this issue.

Note that the results of simulations 2 and 3 show that there is additional variability contributed by estimating the branch lengths and topology, without examining the relative proportion that each contributes to this variability. Of course, the two components are heavily intertwined, and indeed changes in topology can be viewed as changes in the lengths of the branches which lead to placement of a branch on the opposite side of an internal node.

Bootstrap Estimates of the Variance When the Tree Is Unknown

The results of the third simulation in the last section demonstrate that both Yang's method and the MCS method give underestimates of the unconditional variance of the parameter estimates when the tree must

Table 1. Results of the first three simulations[a]

		Simulation 1	Simulation 2			Simulation 3		
				SE(μ)			SE(μ)	
μ	K	SE(μ)	SD(μ)	Observed info	MCS	SD(μ)	Observed info	MCS
0.1	0.5	0.0110	0.0108	0.0087 ± 0.0007	0.0110 ± 0.0007	0.0145	0.0086 ± 0.0011	0.0109 ± 0.0013
0.1	10.0	0.0110	0.0109	0.0083 ± 0.0011	0.0110 ± 0.0007	0.0148	0.0079 ± 0.0013	0.0110 ± 0.0012
0.5	0.5	0.0288	0.0290	0.0268 ± 0.0016	0.0288 ± 0.0012	0.0492	0.0262 ± 0.0026	0.0283 ± 0.0026
0.5	10.0	0.0289	0.0290	0.0289 ± 0.0012	0.0289 ± 0.0012	0.0573	0.0289 ± 0.0032	0.0289 ± 0.0032
0.1	0.5	0.1591	0.1591	0.1269 ± 0.0169	0.1617 ± 0.0263	0.1702	0.1320 ± 0.0182	0.1688 ± 0.0284
0.1	10.0	1.2818	1.2885	0.9897 ± 0.1724	1.3210 ± 0.2134	1.4143	0.9584 ± 0.1748	1.3370 ± 0.2392
0.5	0.5	0.0900	0.0874	0.0848 ± 0.0067	0.0910 ± 0.0074	0.0885	0.0829 ± 0.0067	0.0895 ± 0.0074
0.5	10.0	1.2188	1.3204	1.2500 ± 0.2281	1.2500 ± 0.2201	2.5078	1.5320 ± 0.5646	1.5340 ± 0.5474

[a] Simulation 1 investigates the number of MCS samples that should be taken for estimation of the SE. The results shown are the standard errors estimated using all 40,000 MCS sites. Simulation 2 compares the SE estimates obtained using both the observed information and the MCS method when the tree and branch lengths are assumed known. This is compared with the simulation 1 result and with the standard deviation (SD) of the parameter estimates from each trial in simulation 2. The design of simulation 3 is the same as that of simulation 2, except that the tree and branch lengths are not assumed to be known and are estimated simultaneously with the tree. For simulations 2 and 3, the numbers listed are the means ± standard deviations (for the SE estimates) of the 500 independent simulations.

be simultaneously estimated. Bootstrapping has been widely used in estimating the variance of statistics whose distribution is not tractable (Efron 1982; Efron and Tibshirani 1993). Bootstrap estimates are appropriate for many common statistics, including smooth functions of solutions to smooth estimating equations, including most ML estimators (Davison and Hinkley 1997). The bootstrap was introduced into phylogenetic analysis by Felsenstein (1985) as a means of assessing the confidence at internal nodes of the tree. Since its introduction, bootstrapping has become widely used in phylogenetic analysis, and the proper interpretation of bootstrap values on trees has been discussed (Efron et al. 1996; Hillis and Bull 1993). This motivates a direct bootstrap (DB) method for our example as follows: suppose that the data set has N sequences, each of which is L nucleotides long. We sample columns of the data matrix x with replacement L times to form a new data set, x_i^*. Using x_i^*, we find the MLEs of μ and K, which we denote $\hat{\mu}_i^*$ and $\hat{K}_i^*$, simultaneously with the tree topology and branch lengths. We repeat the above procedure B times and then use the empirical variances of the $\hat{\mu}_i^*$ and $\hat{K}_i^*$ to give the bootstrap estimates of the variance of the parameters. The method is easily adapted to parameters in other substitution models by simply estimating all parameters of interest simultaneously with the tree for each bootstrap data set and then computing the variance of the bootstrap estimates of each parameter.

We note that this procedure corresponds to a nonparametric bootstrapping approach, since it involves simply resampling from the columns of the original data matrix. It is also possible to use a parametric bootstrapping approach by using the MLEs of the tree, branch lengths, and parameters along with the substitution model to simulate B data sets each consisting of L sites [parametric bootstrapping in the context of phylogenetic trees has been discussed by Hillis et al. (1996), among others]. Determination of the variance estimates then proceeds in exactly the same manner as described above. We note that parametric bootstrapping is similar to drawing MCS samples, as discussed in the previous section. The primary distinction between the two is that in bootstrapping, the size of each resampled data set is equal to L, the number of observations in the original data set, since our goal is to use the empirical c.d.f. as an estimate of the underlying c.d.f. that gives rise to the data. In contrast, MCS involves generating a large sample of sites from the assumed distribution with the goal of estimating an expectation, and hence the number of sites generated in our MCS procedure is generally much larger than L.

The direct bootstrap approach proposed above is very computationally expensive, since the procedure of finding the MLEs of the substitution model parameters entails simultaneously optimizing the tree topology, branch lengths, and parameters. We observe that the time required to find the ML tree given the parameters and to find the MLEs of the parameters given the tree is substantially less than the time required for simultaneously optimizing all of these parameters. This leads to the proposal of another bootstrapping approach.

In our example using the K2P model, let $\hat{\theta}$ and $\hat{t}$ denote the MLEs of the parameters, $\theta = (\mu, K)$, and the tree, t (both topology and branch lengths), for the observed data set x. Let $\hat{\theta}_{\hat{t}}(x)$ denote the MLE of θ for data set x given tree $\hat{t}$. By the conditioning principle,

$$var(\hat{\theta}) = E(var(\hat{\theta}_{\hat{t}}(x)|\hat{t})) + var(E(\hat{\theta}_{\hat{t}}(x)|\hat{t})) \quad (6)$$

Applying standard bootstrap techniques, we obtain an estimate of the variance of $\hat{\theta}$ by using

$$var_{\hat{F}}(\hat{\theta}) = E_{\hat{F}}(var(\hat{\theta}_{\hat{t}^*}(x^*)|\hat{t}^*)) + var_{\hat{F}}(E(\hat{\theta}_{\hat{t}^*}(x^*)|\hat{t}^*)) \quad (7)$$

where $\hat{F}$ denotes the empirical c.d.f. of the multinomial distribution of the data x, x^* denotes a random bootstrap sample (formed by drawing from the columns of x with replacement L times or by applying a parametric resampling method as described above), and $\hat{t}^*$ denotes the ML tree estimated from the bootstrap sample x^*.

To obtain a practical algorithm, we use the MCS method for the variance of the MLEs of the parameters described in the last section to approximate the conditional variance in the first term on the right-hand side of Eq. (7) for each bootstrap data set and then average over all B of the bootstrap samples to estimate the expectation. We approximate the conditional expectation in the second term on the right-hand side of Eq. (7) for each bootstrap data set using a Monte Carlo estimate of the mean, and the variance is taken over all B of the bootstrap samples. The details of the algorithm are as follows.

1. From the data set x, obtain the global MLEs of the tree, $\hat{t}$, and the parameters, $\hat{\theta}$.
2. Generate B nonparametric or parametric bootstrap samples, either by resampling the columns of x with replacement L times or by generating L sites according to the tree, parameter estimates, and substitution model from Step 1.
3. For each of the bootstrap samples x_i^* obtained in Step 2, find the ML tree $\hat{t}_i^*$ with the parameters fixed at their ML values from Step 1, as is common in bootstrap analysis.
4. For each ML tree $\hat{t}_i^*$ in Step 3, generate nL sites (n an integer) using the MCS procedure.

   Using these nL sites, two quantities will be estimated.

   (a) Conditional on the tree $\hat{t}_i^*$, estimates of μ and K are found for each group of L of the nL sites. The averages of these n estimates of μ and K are then an approximation to the conditional expectation in the second term of Eq. (7). These averages are denoted $\tilde{\mu}_i^*$ and $\tilde{K}_i^*$.

   (b) For all of the nL sites combined, the conditional variances in the first term of Eq. (7) are estimated using Eqs. (3) and (4) (where the sums are taken over the nL MCS samples, as described in the previous section). These conditional variances are denoted $\tilde{v}_i^*(\mu)$ and $\tilde{v}_i^*(K)$.
5. The averages of the $\tilde{v}_i^*(\mu)$ and the $\tilde{v}_i^*(K)$ are computed to get the first term in Eq. (7), and the sample variances of the $\tilde{\mu}_i^*$ and the $\tilde{K}_i^*$ are computed to get the second term in Eq. (7). The estimate of the variance of $\hat{\mu}$ is then the sum of these two quantities for μ, and the estimate of the variance of $\hat{K}$ is the sum of these two quantities for K.

Implementation of the bootstrap procedure proposed above for other site-independent substitution models is straightforward, as the steps above are simply modified to deal with all parameters of interest in the substitution model under consideration.

Further simulations were used to evaluate each of these two bootstrap approaches [the direct approach (DB) and the conditional variance approach (CVB)] using both nonparametric and parametric bootstrapping. The simulations involved first generating 10 data sets for the model tree shown in Fig. 1 with the four combinations of parameter settings for μ and K. For each simulated data set, variances were estimated using both the DB and CVB approaches. In each case, 50 bootstrap replicates were used. Results using the parametric bootstrap are shown in Table 2 and results using the nonparametric bootstrap are shown in Table 3.

We note that when generating the MCS samples in step 4 above, there are several choices for the parameter values used. For real examples, we could either let the parameters take on the values of their global MLEs, $\hat{\theta}$, or use the parameter estimates specifically obtained for bootstrap replicate i with tree $\hat{t}_i^*$. In our simulated examples, we have the additional choice of using the true parameter values that were used to generate the initial simulated data set. We denote these three options for choosing parameter values SP (simulated data parameters), BP (bootstrap parameters), and TP (true parameters) in Tables 2 and 3. We note that when using either the parameters from the simulated data or the true parameters, fixed values of the parameters are used in generating the MCS data in step 4. In contrast, different values of the parameters are used for each bootstrap replicate when the bootstrap parameters are used. For this reason, we expect that using the bootstrap parameters will lead to larger estimates of the variance, since the second term in Eq. (7) is expected to be larger.

In examining Tables 2 and 3, we first note that there is reasonably close agreement between the parametric and the nonparametric results, with no clear trend toward over- or underestimation of variances for either method, although the mean squared error is more often lower for the parametric method (results not shown). In comparing the direct bootstrapping method (DB) with the conditional variance bootstrap approaches (CVB), we note that the DB method generally gives the values closest to the values from simulation 3 but that the DB method sometimes underestimates the variance. The CVB method which uses the parameter values from the bootstrap sample

Table 2. Results of the parametric bootstrapping simulations[a]

μ	K	Simulation 3 SD	DB	CVB(BP)	CVB(SP)	CVB(TP)
			Results for μ			
0.1	0.5	0.0145	0.0146 ± 0.0037	0.0192 ± 0.0037	0.0123 ± 0.0015	0.0121 ± 0.0005
0.1	10.0	0.0148	0.0148 ± 0.0030	0.0194 ± 0.0026	0.0128 ± 0.0015	0.0119 ± 0.0005
0.5	0.5	0.0492	0.0421 ± 0.0120	0.0501 ± 0.0094	0.0305 ± 0.0027	0.0304 ± 0.0011
0.5	10.0	0.0573	0.0530 ± 0.0129	0.0593 ± 0.0093	0.0388 ± 0.0166	0.0355 ± 0.0099
			Results for K			
0.1	0.5	0.1702	0.1644 ± 0.0242	0.2495 ± 0.0364	0.1813 ± 0.0309	0.1781 ± 0.0097
0.1	10.0	1.4143	1.2937 ± 0.3344	1.9272 ± 0.4162	1.4363 ± 0.2606	1.4596 ± 0.0777
0.5	0.5	0.0885	0.0901 ± 0.0097	0.1309 ± 0.0134	0.0934 ± 0.0091	0.0945 ± 0.0028
0.5	10.0	2.5078	2.2132 ± 1.0082	2.9827 ± 1.2708	1.7749 ± 0.9557	1.5780 ± 0.1722

[a] Ten independent simulations were performed, with 50 bootstrap replicates in each simulation. The entries are the means ± standard deviations of the estimates obtained in the 10 simulations. The DB column gives the results of the direct bootstrap method, while the CVB columns give the bootstrapping results using the conditional variance bootstrap. The three CVB bootstrap methods differ in the manner in which the parameters are set in generating the MCS samples: BP refers to using the parameters from the particular bootstrap replicate, SP refers to using the parameters from the simulated data set that generated the bootstrap samples, and TP refers to using the true values of the parameters that generated the simulated data sets.

Table 3. Results of the nonparametric bootstrapping simulations[a]

μ	K	Simulation 3 SD	DB	CVB(BP)	CVB(SP)	CVB(TP)
			Results for μ			
0.1	0.5	0.0145	0.0153 ± 0.0021	0.0198 ± 0.0023	0.0124 ± 0.0014	0.0122 ± 0.0005
0.1	10.0	0.0148	0.0135 ± 0.0030	0.0182 ± 0.0033	0.0120 ± 0.0015	0.0120 ± 0.0005
0.5	0.5	0.0492	0.0454 ± 0.0114	0.0520 ± 0.0102	0.0315 ± 0.0030	0.0314 ± 0.0020
0.5	10.0	0.0573	0.0621 ± 0.0109	0.0640 ± 0.0076	0.0355 ± 0.0075	0.0372 ± 0.0083
			Results for K			
0.1	0.5	0.1702	0.1638 ± 0.0429	0.2478 ± 0.0511	0.1815 ± 0.0293	0.1790 ± 0.0088
0.1	10.0	1.4143	1.5019 ± 0.3145	2.1335 ± 0.4029	1.5444 ± 0.2946	1.4134 ± 0.0927
0.5	0.5	0.0885	0.1031 ± 0.0487	0.1514 ± 0.0697	0.0947 ± 0.0096	0.0967 ± 0.0056
0.5	10.0	2.5078	2.9305 ± 1.5593	3.5386 ± 1.5741	1.7453 ± 0.6363	1.6747 ± 0.3759

[a] Ten independent simulations were performed, with 50 bootstrap replicates in each simulation. The entries are the means ± standard deviations of the estimates obtained in the 10 simulations. The DB column gives the results of the direct bootstrap method, while the CVB columns give the bootstrapping results using the conditional variance bootstrap. The three CVB bootstrap methods differ in the manner in which the parameters are set in generating the MCS samples: BP refers to using the parameters from the particular bootstrap replicate, SP refers to using the parameters from the simulated data set that generated the bootstrap samples, and TP refers to using the true values of the parameters that generated the simulated data sets.

also appears to be a reasonably good estimator but always gives values larger than the simulation 3 values. The CVB methods which use either the simulated data parameters (SP) or the true parameters (TP) generally give values lower than the CVB(BP) values, as expected, and often underestimate the variance.

Based on these observations, we recommend using either the DB or the CVB(BP) approach for estimating the variance. Since estimation using the DB method is computationally intensive due to the need to estimate the tree and parameters simultaneously, we suggest the use of the CVB(BP) estimate as a conservative approximation of the standard error. In addition, the time savings experienced using the CVB method compared to the DB method will increase as the number of sequences increases, making CVB the preferable method for data sets with a large number of sequences. We next apply the methods developed here to a real data set.

Application to the HPV Data Set

Aligned DNA sequences for 30 papillomaviruses (28 human papillomaviruses (HPVs), a rhesus papillomavirus, and a pygmy chimpanzee papillomavirus) were obtained from the Los Alamos National Database web site (http://hpv-web.lanl.gov). A data set containing 1382 base pairs was obtained for the L1 gene after removal of all sites in which an insertion or deletion occurred in any of the sequences. The sequences used here have been considered in similar studies of the evolutionary relationships among papillomaviruses by Chan et al. (1992, 1995) and Ong et al. (1997). In particular, Ong et al. (1997) found that

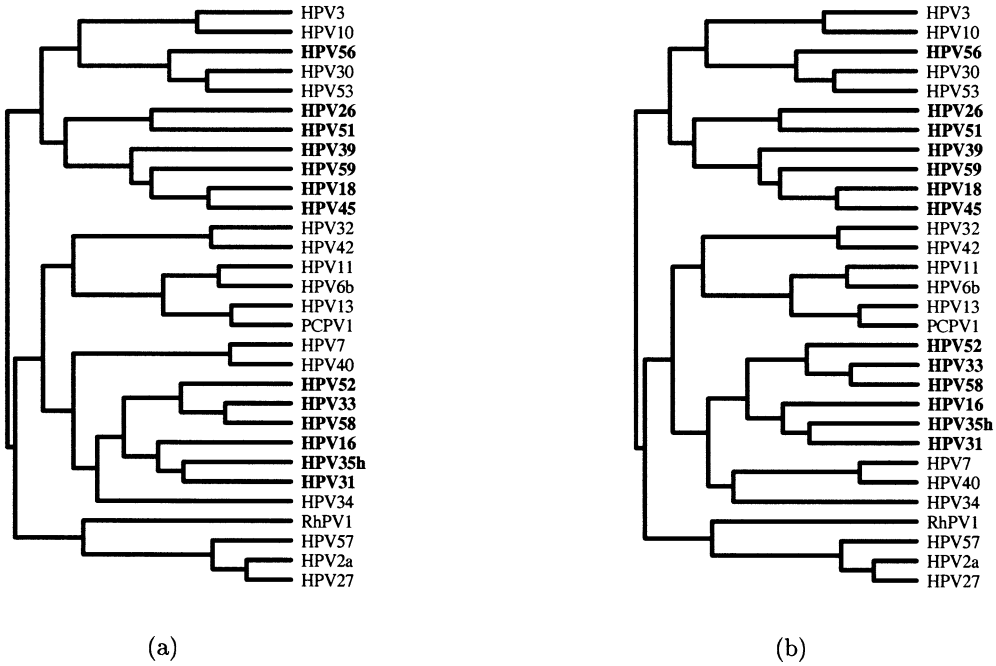

Fig. 3. Estimates of the phylogenetic tree for the papillomavirus data set. The tree in **a**, which has log likelihood −27,864.73 when $K = 0.6606$, is the MLE when the transition/transversion parameter is fixed at 1.5691 during the search for the ML tree. The tree in **b**, which has log likelihood −27,864.11 when $K = 0.6578$, is the MLE when the substitution model parameters and the tree are estimated simultaneously.

the molecular clock model is appropriate for these sequences. Here we consider the problem of simultaneously estimating the tree and the parameters μ and K in the context of the F84 model (Felsenstein 1984).

We first suppose that the ML tree has been estimated assuming that the transition/transversion ratio R, which is a function of K, is 2.0, which is the default value in PHYLIP and PAUP*. In this example we consider the F84 model (Felsenstein 1984) with base frequencies fixed at their empirical values, and thus $R = 2.0$ corresponds to $K = 1.5691$. From the ML tree estimated for K fixed at 1.5691 (which is shown in Fig. 3a), we can find the MLE of K, $\hat{K} = 0.6606$. Using the methodology in the previous section, we can then obtain an estimate of the standard error of $\hat{K}$ and use this to construct a confidence interval for K. Fifteen parametric bootstrap samples were used to estimate the standard error. In the MCS procedure, parameters were set at their estimates for each bootstrap sample (BP), and $n = 10$ was used. In this case, we have $\mathrm{SE}(\hat{K}) = 0.0413$, and so the 95% confidence interval for K is (0.5797, 0.7415) (see Table 4). We note that this interval does not contain 1.5691, which suggests that estimation of the transition/transversion parameter simultaneously with the tree may be important in this case.

Therefore, we next consider the problem of simultaneous estimation of the parameters and the tree for this data set. The same procedure as described above was used to estimate the standard error, and the results are shown in Table 4. The resulting MLE of the phylogenetic tree is shown in Fig. 3b. We note that this tree, which was obtained by simultaneously estimating the tree and the parameters, differs from that in Fig. 3a, which was obtained by holding the transition/transversion parameter fixed, in the placement of the HPV34 sequence. In both trees, sequences shown in boldface are those which are known to be associated with human cancers (Burk 1999). Examination of the two trees shows that in the first case (transition/transversion ratio fixed; Fig. 3a), HPV34 clusters with a group of oncogenic HPV strains, while in the second case (transition/transversion ratio estimated simultaneously; Fig. 3b), HPV34 is more closely related to the two nononcogenic HPV sequences in that clade. However, the two trees are close to one another, in the sense that they are separated by only a single NNI (nearest-neighbor interchange) branch swap. The log likelihoods of the two trees with parameter values optimized are quite close (−27,864.73 for the tree in Fig. 3a vs −27,864.11 for the tree in Fig. 3b). The MLE of K is $\hat{K} = 0.6578$. The estimated standard error of $\hat{K}$ is 0.0577, which gives a 95% confidence interval of (0.5447, 0.7709).

It is also of interest to compare not only the ML trees in each of the two cases described above (K fixed at 1.5691 and K estimated simultaneously with the tree), but also other trees of high likelihood. More detailed study of the problem shows that when the transition/transversion parameter K is fixed at 1.5691, the trees in Fig. 3 have log likelihoods of −28,083.67 and −28,084.42, respectively. There exists a group of

Table 4. Estimates of μ and K and their standard errors for the papillomavirus data set

Tree estimated assuming	$\hat{\mu}$	SE($\hat{\mu}$)	$\hat{K}$	SE($\hat{K}$)
R fixed	0.3010	0.0125	0.6606	0.0413
R estimated	0.3014	0.0142	0.6578	0.0577

three trees having log likelihoods ranging from −28,086.23 to −28,086.26, which are distinct from the trees in Fig. 3 in that more than a single NNI rearrangement is required to convert either tree in Fig. 3 to one of these three. When the transition/transversion parameter K is simultaneously estimated, however, the distinction between these two groups of trees becomes more clear, since the log likelihoods of these other trees are then all near −27,871.83, which is much farther from the optimized log likelihoods for the trees in Fig. 3 (a difference in log likelihood between the two groups of trees of about 7, compared to a difference in log likelihood of about 3 or 4). Thus simultaneous estimation of the transition/transversion parameter is also beneficial in evaluating and comparing trees other than the ML tree.

Discussion

Here we have considered the problem of estimation of the standard errors of the parameter estimates in nucleotide substitution models used in phylogenetic analysis. As an example, this paper considered the transition/transversion parameter, K, and the mean instantaneous substitution rate parameter, μ, which can be estimated when a molecular clock is assumed. Although we initially assumed the K2P model, more general models are commonly used in practice. Perhaps the most commonly employed models are the HKY85 model and the F84 model, each of which allows the generalization of the K2P model to incorporate different base frequencies. We note that if the base frequencies are estimated empirically and then held fixed through the tree and parameter estimation process, as is commonly done in practice, the results shown here for K are directly applicable. This strategy was used in the analysis of the papillomavirus data above. In addition, the results derived here are general enough that they are easily applied to the situation in which a molecular clock is not assumed (though then the parameter μ cannot be estimated, so primary interest will be in estimating K) and that they may be extended to include additional parameters in more general substitution models [e.g., the TrN model of Tamura and Nei (1993)].

In the fortunate situation in which the tree and branch lengths can be considered known, parameters and estimates of their variability can be simply computed. We have compared two possible estimates of the standard error, one based on the observed information and one based on the expected information, and found in simulations that the expected information outperforms the observed information, in the sense that the observed information tends to underestimate the standard error. This result is in contrast to the work of Efron and Hinkley (1978), who found that the use of the observed information was superior to the use of the expected information in the simpler one-parameter setting. When the tree is unknown, both the observed information and the expected information provide underestimates of the variability, because both fail to account for the additional variability introduced by estimating the tree simultaneously.

To deal with this problem, we propose the use of the bootstrap to estimate the standard error and demonstrate through simulation that bootstrapping provides reasonable estimates of the standard error. However, bootstrap estimation is computationally intensive, since it involves simultaneous estimation of the tree and the parameters for each bootstrap data set. We have thus proposed a more computationally tractable bootstrap procedure through use of the "conditional variance formula." We demonstrate through simulation that this method provides reasonable estimates of the standard error. The time savings associated with this approach will increase as the number of taxa increases, since for larger problems simultaneous estimation of the tree and parameters becomes more computationally expensive.

To implement practically the methods proposed here, we recommend that each MCS sample contain at least 10 times the number of sites in the original data set (i.e., $n \geq 10$). We of course also recommend the use of as many bootstrap samples as is computationally feasible, though for large data sets it will be difficult to obtain a large number of bootstrap replicates. Since both parametric and nonparametric bootstrap sampling performed well, we suggest that the choice of method used to generate the bootstrap samples be left to the investigator, whose decision may be guided by other uses for the bootstrap samples generated.

For example, we note that computation of standard error estimates as proposed here is easily incorporated into the standard bootstrap analysis that is commonly performed following ML estimation. In particular, either the DB or the CVB(BP) method of standard error estimation may be easily added to the traditional bootstrap analysis performed to assess support for the internal nodes of the ML tree in the following way. If bootstrap analysis is to be performed with parameter estimates fixed at their ML values from the original data set, then such an analysis will result in the generation of B bootstrap samples,

each of which will be used to estimate an ML tree with fixed parameter values. Once this tree has been estimated, little computational expense is added by performing step 4 of the CVB method for each bootstrap sample and then summarizing the results over all bootstrap samples as in step 5. Moreover, such a procedure is easily implemented with currently available software [for example, PAUP* (Swofford 1998) and PHYLIP (Felsenstein 1993) will generate bootstrap samples, the Seq-Gen program (Rambaut and Grassly 1997) can be used to generate the required MCS samples, and PAUP* can be used for parameter estimation for given trees]. If, on the other hand, parameters are to be estimated simultaneously with the tree for each of B bootstrap replicates, then it is straightforward to store the values of these parameter estimates for each bootstrap sample and compute their empirical standard deviation, which amounts to carrying out the DB method (this approach may be even more readily accessible with currently available software such as PAUP* since all that is required is output of the analysis of each bootstrap sample). Hence, the methods we propose may be easily implemented in a standard bootstrap analysis with little additional computational expense and may provide important information concerning the ML tree estimate and other trees of high likelihood.

We note, finally, that another potential approach to assessing variability in parameter estimates is to perform a Bayesian analysis using available MCMC methods (e.g., Mau et al. 1999; Larget and Simon 1999; Yang and Rannala 1997; Li et al. 2000, Huelsenbeck and Ronquist 2001). Such an analysis will yield the posterior distribution of the tree and parameters in the evolutionary model. This approach is more informative if the focus is not on only the MLE or posterior mode but, instead, on the entire posterior distribution.

Acknowledgments. We thank Jeff Thorne and an anonymous referee for helpful comments on an early draft of the manuscript.

References

Birch MW (1964) A new proof of the Pearson-Fisher theorem. Ann Math Stat 35:818–824

Bishop YM, Fienberg SE, Holland PW (1975) Discrete multivariate analysis: Theory and practice. MIT Press, Cambridge, MA

Burk R (1999) Human papillomaviruses and the risk of cervical cancer. Hosp Pract (www.hosppract.com/issues/1999/11/burk.htm)

Chan SY, Bernard HU, Ong CK, Chan SP, Hoffman B, Delius H (1992) Phylogenetic analysis of 48 papillomavirus types and 28 subtypes and variants: A showcase for the molecular evolution of DNA viruses. J Virol 66:5714–5725

Chan SY, Delius H, Halpern A, Bernard HU (1995) Analysis of genomic sequences of 95 papillomavirus types: Uniting typing, phylogeny, and taxonomy. J Virol 69:3074–3082

Chang J (1996) Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency. Math Biosci 137: 51–73

Davison AC, Hinkley DV (1997) Bootstrap methods and their application. Cambridge University Press, New York

Efron B (1982) The jackknife, the bootstrap, and other resampling plans. CBMS-NSF Regional Conference Series in Applied Mathematics, Monograph 38. Society for Industrial and Applied Mathematics, Philadelphia

Efron B, Hinkley D (1978) Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. Biometrika 65:457–487

Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman and Hall, New York

Efron B, Halloran E, Holmes S (1996) Bootstrap confidence levels for phylogenetic trees. Proc Natl Acad Sci USA 93:13429–13434

Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 17:368–376

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

Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791

Felsenstein J (1993) *Phylogenetic inference package (PHYLIP), version 3.5*. University of Washington, Seattle

Fukami-Kobayashi K, Tateno Y (1991) Robustness of maximum likelihood tree estimation against different patterns of base substitutions. J Mol Evol 32:79–91

Gaut BS, Lewis PO (1995) Success of maximum likelihood phylogeny inference in the four-taxon case. Mol Biol Evol 12:152–162

Hasegawa M, Kishino H, Yano TA (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 21:160–174

Hayasaka K, Gojobori T, Horai S (1988) Molecular phylogeny and evolution of primate mitochondrial DNA. Mol Biol Evol 5:626–644

Hillis D, Bull J (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst Biol 42:182–192

Hillis D, Mable B, Moritz C (1996) Applications of molecular systematics: The state of the field and a look to the future. In: Hillis D, Moritz C, Mable B (eds) Molecular systematics, 2nd ed. Sinauer Associater, Sunderland, MA, pp 515–543

Huelsenbeck J, Ronquist F (2001) MrBayes: Bayesian inference of phylogenetic trees, Bioinformatics 17:754–755

Ina Y (1998) Estimation of the transition/transversion ratio. J Mol Evol 46:521–533

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kuhner MK, Felsenstein J (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol 11:459–468

Larget B, Simon D (1999) Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol Biol Evol 16:750–759

Li S, Pearl D, Doss H (2000) Phylogenetic tree construction using Markov chain Monte Carlo. J Am Stat Assoc 95:493–508

Mau B, Newton M, Larget B (1999) Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics 55:1–12

Ong CK, Nee S, Rambaut A, Bernard HU, Harvey PH (1997) Elucidating the population histories and transmission dynamics of papillomaviruses using phylogenetic trees. J Mol Evol 44:199–206

Pollock DB, Goldstein DD (1995) A comparison of two methods for reconstructing evolutionary distances from a weighted contribution of transition and transversion differences. Mol Biol Evol 12:713–717

Purvis A, Bromham L (1997) Estimating the transition/transversion ratio from independent pairwise comparisons with an assumed phylogeny. J Mol Evol 44:112–119

Rambaut A, Grassly N (1996) *Sequence parameters of trees: SPOT*. Department of Zoology, University of Oxford, Oxford

Rambaut A, Grassly N (1997) Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci 13:235–238

Rogers J (2001) Maximum likelihood estimation of phylogenetic trees is consistent when substitution rates vary according to the invariable sites plus gamma distribution. Syst Biol 50:713–722

Salter L (1999) Simulation-based estimation of phylogenetic trees, PhD thesis. Ohio State University, Columbus

Salter L, Pearl D (2001) Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. Syst Biol 50:7–17

Swofford D (1998) PAUP*. Phylogenetic analysis using parsimony (* and other methods); version 4. Sinauer Associates, Sunderland, MA

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526

Wakeley J (1994) Substitution-rate variation among sites and the estimation of transition bias. Mol Biol Evol 11:436–442

Wakeley J (1996) The excess of transitions among nucleotide substitutions: New methods of estimating transition bias underscore its significance. TREE 11:158–163

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 (2000) Phylogenetic analysis by maximum likelihood (PAML), version 3.0. University College London, London

Yang Z, Kumar S (1996) New parsimony-based methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites and comparison with likelihood methods. Mol Biol Evol 13:650–659

Yang Z, Rannala B (1997) Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. Mol Biol Evol 14:717–724

Yang Z, Goldman N, Friday A (1994) Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. Mol Biol Evol 11:316–324