Molecular Biology and Evolution

Approximate Likelihood Calculation on a Phylogeny for Bayesian Estimation of Divergence Times

Mario dos Reis[1] and Ziheng Yang*[,1,2]

[1]Department of Biology, University College London, Darwin Building, Gower Street, London, United Kingdom
[2]Center for Computational and Evolutionary Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

***Corresponding author:** z.yang@ucl.ac.uk

Associate editor: Oliver Pybus

Abstract

The molecular clock provides a powerful way to estimate species divergence times. If information on some species divergence times is available from the fossil or geological record, it can be used to calibrate a phylogeny and estimate divergence times for all nodes in the tree. The Bayesian method provides a natural framework to incorporate different sources of information concerning divergence times, such as information in the fossil and molecular data. Current models of sequence evolution are intractable in a Bayesian setting, and Markov chain Monte Carlo (MCMC) is used to generate the posterior distribution of divergence times and evolutionary rates. This method is computationally expensive, as it involves the repeated calculation of the likelihood function. Here, we explore the use of Taylor expansion to approximate the likelihood during MCMC iteration. The approximation is much faster than conventional likelihood calculation. However, the approximation is expected to be poor when the proposed parameters are far from the likelihood peak. We explore the use of parameter transforms (square root, logarithm, and arcsine) to improve the approximation to the likelihood curve. We found that the new methods, particularly the arcsine-based transform, provided very good approximations under relaxed clock models and also under the global clock model when the global clock is not seriously violated. The approximation is poorer for analysis under the global clock when the global clock is seriously wrong and should thus not be used. The results suggest that the approximate method may be useful for Bayesian dating analysis using large data sets.

Key words: Bayesian method, MCMC, divergence time, likelihood function, molecular clock.

Research article

Introduction

The molecular clock assumption provides a powerful way to estimate species divergence times from molecular sequence data (Zuckerkandl and Pauling 1965). If protein and nucleic acid sequences accumulate substitutions at a uniform rate, the degree of divergence between two homologous sequences will grow linearly with the divergence time. If information on the time of divergence is available (for example, from the fossil record) for one or more pairs of sequences in a phylogenetic tree, the substitution rate can be calculated and used to obtain times of divergence for all the nodes in the tree.

Although intuitively appealing, this approach has two limitations. First, the molecular clock may not hold, and the rate may vary with time or over lineages (Kumar 2005). Second, information from the fossil record is uncertain, and this uncertainty needs to be incorporated in the computation of substitution rates (Thorne et al. 1998; Yang and Rannala 2006; Benton and Donoghue 2007). Recently, much effort has been taken to overcome these limitations. Several works have studied variation in molecular rates among lineages, and models that consider autocorrelated rates along the branches of a tree (Thorne et al. 1998; Rannala and Yang 2007) or independent rates following a specified statistical distribution (Drummond et al. 2006; Rannala and Yang 2007) have been developed. The uncertainties in the fossil record can be dealt with by the Bayesian method, during the specification of the prior for divergence times (Thorne et al. 1998; Drummond et al. 2006; Yang and Rannala 2006).

Due to the complexity of the model, Markov chain Monte Carlo (MCMC) methods are used to obtain numerical approximations to the posterior distribution. The MCMC method involves repeated evaluation of the likelihood function on the phylogeny to determine whether a proposed move should be accepted or rejected. Computation of the likelihood function is expensive, and a typical Bayesian analysis for a phylogeny of <50 species might take several days.

Thorne et al. (1998) proposed the use of Taylor expansion to approximate the likelihood function in the MCMC algorithm. This approximation is fast and has been used with success in several studies (Seo et al. 2004; Inoue et al. 2010; Guindon 2010). However, a rigorous assessment of the approximate versus exact likelihood calculations for various phylogenies under different clock models has not been carried out. Here, we examine the accuracy of the approximate methods in the estimation of divergence times and rates, using the program MCMCtree in the PAML package (Yang 2007), which implements both the exact and the approximate methods. First, we develop the approximation theory for the case of two species under the Jukes and Cantor (1969) model. This simple case inspired us to explore

several parameter transforms that may improve the approximation. We then compare the old and new approximations using a data set of mitochondrial protein-coding genes for 36 mammalian species and another data set of 18S ribosomal RNA and ATP1 mitochondrial protein-coding genes for 50 plant species.

A number of Bayesian molecular clock dating algorithms have been developed, with different rate–drift models developed to relax the molecular clock assumption (e.g., Thorne et al. 1998; Drummond et al. 2006; Yang and Rannala 2006; Rannala and Yang 2007). Inference under any of these models requires the calculation of the likelihood, that is, the probability of the sequence data given a set of branch lengths. Our approximate methods are not specific to any particular relaxed clock model and should be useful in many Bayesian dating algorithms. The approximation cannot be used when the topology changes, so joint inference of tree and divergence times (as in Drummond et al. 2006) will require extensions of this framework.

Theory and Methods

Taylor Expansion of the Log-Likelihood

The second-order Taylor expansion of the log-likelihood function around the maximum likelihood estimates (MLEs) is

$$\ell(\theta) \approx \ell(\hat{\theta}) + \mathrm{g}^T \Delta\theta + \frac{1}{2}\Delta\theta^T \mathrm{H} \Delta\theta \qquad (1)$$

or

$$\Delta\ell(\theta) = \ell(\theta) - \ell(\hat{\theta}) \approx \mathrm{g}^T \Delta\theta + \frac{1}{2}\Delta\theta^T \mathrm{H} \Delta\theta, \qquad (2)$$

where $\theta = \{\theta_i\}$ are the model parameters, $\Delta\theta = \theta - \hat{\theta}$, $\hat{\theta} = \{\hat{\theta}_i\}$ are the MLEs, and $\mathrm{g} = \{g_i\}$ and $\mathrm{H} = \{H_{ij}\}$ are the gradient and the Hessian matrix, respectively, that is, the vector of first derivatives and matrix of second derivatives, both evaluated at the MLEs. As the likelihood function is defined up to a constant, use of $\ell(\theta)$ or $\Delta\ell(\theta)$ leads to the same inference.

We apply the Taylor expansion to the $2s - 3$ branch length parameters on the unrooted tree for s species without assuming the clock. Commonly used phylogenetic models include, in addition to the branch lengths, parameters describing the evolutionary process, such as the transition/transversion rate ratio κ and the gamma shape parameter α for variable rates among sites. In theory, those substitution parameters can be treated in the same way as the branch lengths in equations (1) or (2). In particular, parameter α is known to be negatively correlated with branch lengths, and ignoring the uncertainties in the MLE of α may lead to too narrow posterior credibility intervals (CIs) for divergence times. However, in our test using the two data sets analyzed in this paper, we found that the CIs generated from the exact and approximate methods are nearly identical, possibly because the data sets are large so that α is estimated reliably by maximum likelihood (ML). Thus, for simplicity in this paper, θ includes the $2s - 3$ branch lengths only.

When the MLEs are inside the parameter space (i.e., if all MLEs of branch lengths are strictly positive), $\mathrm{g} = 0$, so that the linear (second) term in equation (1) is zero. In this case, the likelihood is $L = \exp(\ell) \approx L(\hat{\theta}) \times \exp(\frac{1}{2}\Delta\theta^T \mathrm{H} \Delta\theta)$, proportional to the density function for the multivariate normal distribution with mean vector $\hat{\theta}$ and variance–covariance matrix $-\mathrm{H}^{-1}$. This is the approximation used in the multidivtime program of Thorne et al. (1998). When the MLEs of some branch lengths are zero, multidivtime does not use the linear term but instead overestimates the variances of those branch lengths as a compensation.

If the likelihood function is very asymmetrical around the MLEs, equation (1) may not provide a good approximation. Inclusion of the third-order term in equation (1) may be computationally expensive and numerically unstable. Instead, a transform of θ (i.e., a reparameterization) may be a better approach to improving the accuracy of the approximation. Suppose we apply the transform $\mathrm{u} = \mathrm{h}(\theta)$, in which θ and u form a multidimensional one-to-one mapping. The Taylor expansion of the log-likelihood function on the transformed parameters is

$$\Delta\ell(\mathrm{u}) = \ell(\mathrm{u}) - \ell(\hat{\mathrm{u}}) \approx \Delta\mathrm{u}^T \mathrm{g}_u + \frac{1}{2}\Delta\mathrm{u}^T \mathrm{H}_u \Delta\mathrm{u}, \qquad (3)$$

where $\Delta\mathrm{u} = \mathrm{u} - \hat{\mathrm{u}}$, $\mathrm{g}_u = \{g_{u,i}\}$, and $\mathrm{H}_u = \{H_{u,ij}\}$ are the gradient and Hessian of the log-likelihood function on the transformed parameters evaluated at the MLEs $\hat{\mathrm{u}}$. Note that $\hat{\mathrm{u}} = \mathrm{h}(\hat{\theta})$ as MLEs are invariant to reparameterization. Although the theory applies to quite general one-to-one transforms from θ to u, we consider in this paper only the element-wise transforms of the type $u_i = h(\theta_i)$. Then the gradient and Hessian for the transformed variables are given by

$$g_{u,i} = \frac{\partial\ell}{\partial\theta_i}\frac{\partial\theta_i}{\partial u_i} = g_i \frac{\partial\theta_i}{\partial u_i}, \qquad (4)$$

$$H_{u,ij} = \frac{\partial^2\ell}{\partial u_i \partial u_j} = \frac{\partial g_{u,j}}{\partial u_j} = \begin{cases} g_i \frac{\partial^2\theta_i}{\partial u_i^2} + H_{ii}\left(\frac{\partial\theta_i}{\partial u_i}\right)^2 & \text{if } i = j, \\ H_{ij}\frac{\partial\theta_i}{\partial u_i}\frac{\partial\theta_j}{\partial u_j} & \text{if } i \neq j. \end{cases} \qquad (5)$$

Transforms and Their Application to the Case of Two Sequences

In this section, we describe several transforms and test their performance in approximate likelihood calculation in the case of comparing two sequences to estimate the evolutionary distance (the branch length) b. We will describe the application of those transforms in divergence time estimation on a phylogeny later. Although the results obtained for two sequences do not apply exactly to phylogenetic analysis of many sequences, the general pattern appears to hold and offers important insight to approximate likelihood calculation on a phylogeny.

Consider two aligned nucleotide sequences with n sites and x differences. We use the Jukes and Cantor (1969)

model to estimate b. The log-likelihood function is

$$\begin{aligned}\ell(b) &= x\log p + (n-x)\log(1-p)\\ &= x\log\left(\frac{3}{4}-\frac{3}{4}e^{-4b/3}\right)\\ &\quad + (n-x)\log\left(\frac{1}{4}+\frac{3}{4}e^{-4b/3}\right), \qquad (6)\end{aligned}$$

where $p = \frac{3}{4} - \frac{3}{4}e^{-4b/3}$ is the expected proportion of differences between the two sequences. The value of b that maximizes ℓ is

$$\hat{b} = -\frac{3}{4}\log\left(1-\frac{4}{3}\hat{p}\right), \qquad (7)$$

where $\hat{p} = x/n$ is the observed proportion of differences in the alignment. The gradient and Hessian of equation (6) are

$$g = \frac{d\ell}{db} = \left(\frac{x}{p}-\frac{n-x}{1-p}\right)e^{-4b/3}, \qquad (8)$$

$$\begin{aligned}H = \frac{d^2\ell}{db^2} &= \left(-\frac{x}{p^2}-\frac{n-x}{(1-p)^2}\right)e^{-8b/3}\\ &\quad -\frac{4}{3}\left(\frac{x}{p}-\frac{n-x}{1-p}\right)e^{-4b/3}. \qquad (9)\end{aligned}$$

Both g and H are evaluated at the MLEs $\hat{p}$ and $\hat{b}$. The approximation by the second-order Taylor expansion is thus

$$\Delta\ell(b) \approx g\cdot(b-\hat{b}) + \frac{1}{2}H\cdot(b-\hat{b})^2. \qquad (10)$$

We call this the untransformed (NT) approximation.

In molecular phylogenetics, the log-likelihood as a function of branch lengths (e.g., eq. 6) has the feature that the curve drops steeply on the left (i.e., when $b < \hat{b}$) and decreases more slowly on the right (when $b > \hat{b}$). In other words, large branch length estimates tend to have large sampling errors. Thus, variance-stabilizing transforms are expected to improve the accuracy of the approximation. We consider three transforms of the branch length:

1. Square-root transform (SQRT): $u = \sqrt{b}$ and $b = u^2$. Then $db/du = 2u = 2\sqrt{b}$ and $d^2b/du^2 = 2$.
2. Log transform (LOG): $u = \log(b+\varepsilon)$ and $b = e^u - \varepsilon$. We use $\varepsilon = 0.1$ if $\hat{b} < 10^{-4}$ or $\varepsilon = 0$ otherwise. The use of ε here is to deal with the case of $\hat{b} = 0$, where the simple transform $u = \log(b)$ breaks down. Then $db/du = e^u = b+\varepsilon$ and $d^2b/du^2 = b+\varepsilon$.
3. Arcsine transform (ARCSIN): $u = 2\arcsin\sqrt{\frac{3}{4}-\frac{3}{4}e^{-4b/3}}$ and $b = -\frac{3}{4}\log\left(1-\frac{4}{3}\sin^2\left(\frac{u}{2}\right)\right)$. Then

$$\frac{db}{du} = \frac{\cos\left(\frac{u}{2}\right)\sin\left(\frac{u}{2}\right)}{1-\frac{4}{3}\sin^2\left(\frac{u}{2}\right)}, \qquad (11)$$

$$\begin{aligned}\frac{d^2b}{du^2} &= \frac{\frac{1}{2}\cos^2\left(\frac{u}{2}\right)-\frac{1}{2}\sin^2\left(\frac{u}{2}\right)}{1-\frac{4}{3}\sin^2\left(\frac{u}{2}\right)}\\ &\quad + \frac{\frac{4}{3}\cos^2\left(\frac{u}{2}\right)\sin^2\left(\frac{u}{2}\right)}{\left(1-\frac{4}{3}\sin^2\left(\frac{u}{2}\right)\right)^2}. \qquad (12)\end{aligned}$$

This transform is based on the following reasoning. The number of differences in the alignment, x, follows a binomial distribution $\mathrm{bi}(n,p)$, so that $E(\hat{p}) = p$ and $\mathrm{Var}(\hat{p}) = p(1-p)/n$. This dependency of the variance on the mean is undesirable, so we wish to find a transform $h(p)$ so that $\mathrm{Var}(\hat{h}(p)) = \mathrm{Var}(h(\hat{p})) \approx c$, a constant. Such a variance-stabilizing transform should lead to a more symmetrical likelihood in the transformed parameter space. By the delta technique (e.g., Yang 2006: p. 314), the asymptotic variance of the transform is

$$\begin{aligned}\mathrm{Var}(h(\hat{p})) &\approx \mathrm{Var}(\hat{p})\left[\frac{dh(p)}{dp}\right]^2\\ &= p(1-p)/n\left[\frac{dh(p)}{dp}\right]^2. \qquad (13)\end{aligned}$$

Equating this to constant c leads to the differential equation

$$\begin{aligned}c &= \frac{p(1-p)}{n}\left[\frac{dh(p)}{dp}\right]^2,\\ \frac{dh(p)}{dp} &= \frac{\sqrt{cn}}{\sqrt{p(1-p)}}. \qquad (14)\end{aligned}$$

We get

$$\begin{aligned}u = h(p) &\propto \int\frac{dp}{\sqrt{p(1-p)}} = 2\arcsin\sqrt{p}\\ &= 2\arcsin\sqrt{\frac{3}{4}-\frac{3}{4}e^{-4b/3}}. \qquad (15)\end{aligned}$$

We note that $\lim_{b\downarrow 0} 2\arcsin\sqrt{\frac{3}{4}-\frac{3}{4}e^{-4b/3}} = 2\sqrt{b}$ so the ARCSIN converges to the SQRT for small branch lengths. For amino acid or codon alignments, the likelihood formula (eq. 6) needs to be modified. We use $p = \frac{19}{20} - \frac{19}{20}e^{-20b/19}$ or $p = \frac{60}{61} - \frac{60}{61}e^{-61b/60}$ for those data types, respectively.

We can use equations (4), (5), (8), and (9) to construct the gradient and Hessian for any of the three transforms suggested above. For example, for the SQRT, we have $g_u = g\frac{db}{du} = 2g\sqrt{b}$ and $H_u = g\frac{d^2b}{du^2} + H\left(\frac{db}{du}\right)^2 = 2g + 4Hb$, all evaluated at the MLE. Substituting these in equation (3), we obtain the approximate likelihood function:

$$\begin{aligned}\Delta\ell(b) &\approx 2g\sqrt{\hat{b}}\left(\sqrt{b}-\sqrt{\hat{b}}\right)\\ &\quad + (g+2H\hat{b})\left(\sqrt{b}-\sqrt{\hat{b}}\right)^2 \qquad (16)\end{aligned}$$

(cf: eq. (10)).

We are interested in how good the transformed approximations are for the simple two-species JC69 model. There are three cases of interest: 1) $\hat{p} = 0$. In this case, $\hat{b} = 0$ is at the lower boundary of the parameter space. 2) $0 < \hat{p} < 0.75$. This is the most common case with $0 < \hat{b} < \infty$ inside the parameter space. As $n \to \infty$, the likelihood tends asymptotically to the normal distribution and the Taylor expansion provides an increasingly better approximation. 3) $\hat{p} \geqslant 0.75$. In this case, the likelihood increases asymptotically with b, and $\hat{b} = \infty$ is at the upper boundary of the parameter space.

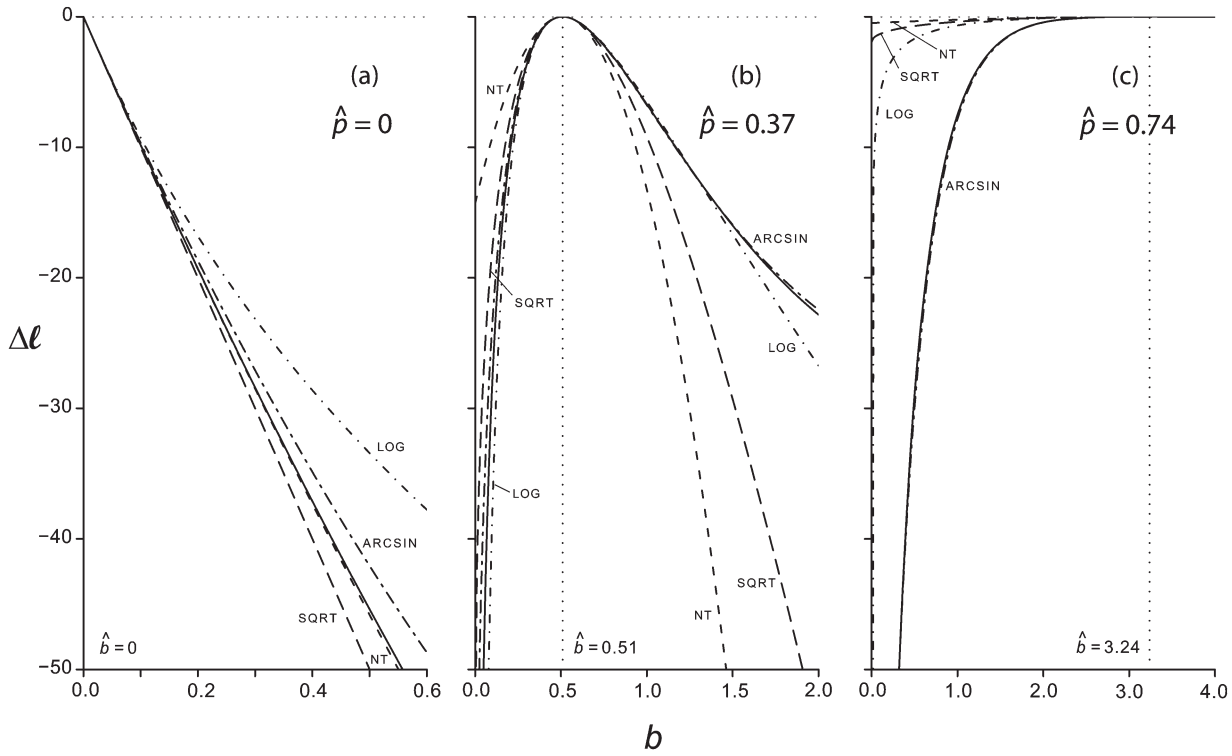

FIG. 1. Log-likelihood curves for the distance (b) between two sequences under the JC69 model. Three types of data are considered in which the number of differences between the two sequences is $x = 0$ (a), $x = 37$ (b), and $x = 74$ (c), respectively. The number of sites is $n = 100$. The log-likelihood ($\Delta\ell$) is calculated using the exact method (equation 6) (solid line) as well as four approximate methods: NT, SQRT, LOG, and ARCSIN.

Figure 1 shows the exact and approximate log-likelihood curves as a function of b when $n = 100$ and $x = 0$, $x = 37$, and $x = 74$. For the first case ($x = 0$, fig. 1*a*), all methods provide an adequate approximation to the true likelihood function. We note that in this case, $\lim_{\hat{b}\downarrow 0} g = -n$ and $\lim_{\hat{b}\downarrow 0} H = \frac{1}{3}n$. Therefore, as $\Delta b \to 0$, Δb^2 tends to zero much faster, and the linear term in equation 10 dominates the approximation. Figure 1*a* clearly shows this, as all methods but LOG approximate the likelihood by a fairly straight line with slope $\approx -n$.

The second case ($x = 37$, fig. 1*b*) is the classical case where the estimate $\hat{b}$ is inside the parameter space (i.e., $\hat{b}$ is neither 0 nor ∞) and the gradient at the MLE is zero. Because the gradient is zero, the Hessian entirely determines the approximation. However, the (exact) log-likelihood curve is steeper on the left side of the MLE and flatter on the right side (fig. 1*b*). Indeed, on the left side $\ell \to -\infty$ when $b \to 0$, but on the right side ℓ approaches a constant, $\ell \to x \log \frac{3}{4} + (n - x) \log \frac{1}{4}$ when $b \to \infty$, instead of $-\infty$ as expected from the normal approximation. The NT method approximates $\Delta\ell$ by a parabola on the untransformed branch length, which is always symmetrical around the MLE. Thus, NT overestimates the likelihood for branch lengths smaller than the MLE and underestimates the likelihood for branch lengths larger than the MLE. The SQRT corrects the problem to some extent but not enough. LOG does a much better job, but it slightly overcorrects on the left and undercorrects on the far right. Like the NT, the SQRT and ARCSIN also overestimate the likelihood on the left side of the MLE but to a lesser degree. The LOG method is the only one that underestimates the likelihood for the shorter branch lengths. The ARCSIN method provides the closest approximation to the true likelihood curve.

In the third case ($x = 74$, fig. 1*c*), substitutions are close to saturation, although there is a shallow peak in the likelihood function with $\hat{b} = 3.24$. In this case, the ARCSIN method is the only one that can approximate the likelihood. When the MLE of a branch length is infinity, the likelihood curve increases asymptotically and there is no maximum. As $b \to \infty$, the gradient and Hessian both tend to zero. This is a very pathological case and none of the transforms can approximate the likelihood curve appropriately. For example, when $x = 75$, equation (6) has no maximum and all transforms break down. One should not use such data in which the sequences are more divergent than random sequences.

Implementation of the Approximate Likelihood Calculation in the Dating Program MCMCtree

The approximate likelihood method with its various transforms is implemented in the program MCMCtree of the PAML package (Yang 2007). Estimation of divergence times with the approximate method follows two steps. First, the branch lengths are estimated by ML without assuming the clock using the BASEML or CODEML programs (Yang 2007). The gradient and Hessian for the branch lengths (g and H)

are estimated at this step. The gradient is calculated by the difference method, and the Hessian is calculated using the outer product of scores estimator (OPS, Porter 2002; Seo et al. 2004). This estimator of the Hessian is generally more stable than the difference approximation to second derivatives (see Appendix). In the second step, an MCMC algorithm is used to estimate the posterior distribution of divergence times and substitution rates. The likelihood function is approximated using the appropriate Taylor expansion (Yang 2006: fig. 7.10*b*). For the NT method, the likelihood is calculated using equation (2). As noted above, this differs from the approximation used in multidivtime (Thorne et al. 1998) if the MLEs of some branch lengths are zero. For SQRT, LOG, and ARCSIN, the likelihood is calculated using equation (3), with the transform applied to each branch length on the tree, and with the gradient g_u and the Hessian H_u for the transformed parameters calculated using equations (4) and (5).

Analysis of Real Data Sets

We use two real data sets to assess the accuracy of the approximate method. In both data sets, the global clock is seriously violated. Although we use the global clock to test the performance of the approximate methods in comparison with the exact calculation of likelihood, we do not recommend its use in divergence time estimation in such data sets as use of the global clock when it is seriously wrong is known to generate unreasonable time estimates. We test all the transforms assuming the global clock (Yang and Rannala 2006) or assuming a relaxed clock with autocorrelated rates following a log-normal distribution (Rannala and Yang 2007). The log-normal distribution of rates is specified by the overall rate μ and by the rate–drift parameter σ^2 (Thorne et al. 1998; Rannala and Yang 2007). Large σ^2 indicates large variation in rates along the branches of the tree, whereas σ^2 close to zero indicates that the tree is clock-like. When the global clock is assumed, the proposed branch lengths during the MCMC iteration are expected to be far from the likelihood peak. The global clock is thus a stern test of the suitability of the approximate method.

The first data set analyzed is an alignment of the first and second codon positions from the 11 protein genes (>90 codons) on the *H* strand of the mitochondrial genome of 36 mammalian species, compiled by Jun Inoue. The alignment has 7,260 sites. The prior on divergence times is specified using fossil calibrations with soft bounds (Yang and Rannala 2006; Inoue et al. 2010). We use 24 minimum and 14 maximum constraints based on the fossil record (Benton et al. 2009). We use a gamma prior $G(1, 1)$ for the mean substitution rate μ (for both the global clock and the correlated rates models). This is a diffuse prior with the mean at one change per site per 100 My. We also use a diffuse gamma prior $G(1, 1)$ for the rate–drift parameter σ^2 (for the correlated rates model only). The tree and fossil calibration are shown in figure 2*a*. Examination of the MLEs of branch lengths under the no-clock model suggests that the global clock is seriously violated in this data set: the likelihood ratio test of the clock under the Hasegawa-Kishino-Yano (HKY)+Γ_5 model rejects the clock at $p = 9.9 \times 10^{-324}$ (with $2\Delta\ell = 1641.1$, degree of freedom [df] = 34).

The second data set is an alignment of two slowly evolving genes: the 18S ribosomal RNA gene (nuclear-encoded) and the ATPase alpha subunit (ATP1) gene (mitochondrial-encoded) from 50 land plant species, kindly provided by Joseph Brown and Yin-Long Qiu. The alignment has 1,974 sites. This is a subset of the seven-gene 192-species alignment of Qiu et al. (2007). We use seven minimum and three skew-*t* constraints based on the fossil record. We use a gamma prior $G(2, 0.04)$ for the mean substitution rate μ (for both the global clock and the correlated rates models), and a gamma prior $G(1, 10)$ for the rate–drift parameter σ^2 (for the correlated rates model). The tree and the fossil constraints are shown in figure 2*b*. As in the mammal data set, the global clock is seriously violated: the likelihood ratio test of the clock under the HKY+Γ_5 model rejects the global clock at $p = 1.8 \times 10^{-182}$ (with $2\Delta\ell = 1,020.6$, df = 48). Although the test statistic is smaller than for the mammal data set, note that there is far more data in the mammal data set (36 sequences each of 7,260 sites compared with 50 sequences each of 1,974 sites for the plant data set). Thus, the global clock is violated more seriously in the plant data set than in the mammal data set.

For each data set, we estimated the branch lengths, the gradient g and the Hessian H with BASEML using the HKY85+Γ_5 substitution model (Hasegawa et al. 1985; Yang 1994). We then used the program MCMCtree to estimate the divergence times under both the global clock (clock = 1 in the MCMCtree control file) and the correlated rates (clock = 3 for MCMCtree) models, using each one of the three transforms (SQRT, LOG, ARCSIN), the plain approximation (NT), and the exact method. We also used the independent rates model (Drummond et al. 2006; Rannala and Yang 2007) (clock = 2 for MCMCtree) to analyze the two data sets. The results are very similar to those for the correlated rates model and are not presented. We ran each MCMC setup twice from different random starting values to check convergence to the posterior distribution.

Results and Discussion

Mammal Data Set

When the relaxed clock was assumed, all four approximations gave essentially the same posterior mean times as the exact method (fig. 3). The 95% posterior CIs were virtually the same as well (results not shown). The posterior distribution of the root age, and the posterior distribution of the mean rate μ were also the same as from the exact method (fig. 4). The mean posterior rate was $\mu = 0.16 \times 10^{-8}$ per site per year $(0.091, 0.28)$ and the drift parameter $\sigma^2 = 0.61\,(0.37, 0.97)$ by the exact method. There is substantial rate variation among lineages, consistent with the rejection of the global clock by the likelihood ratio test.

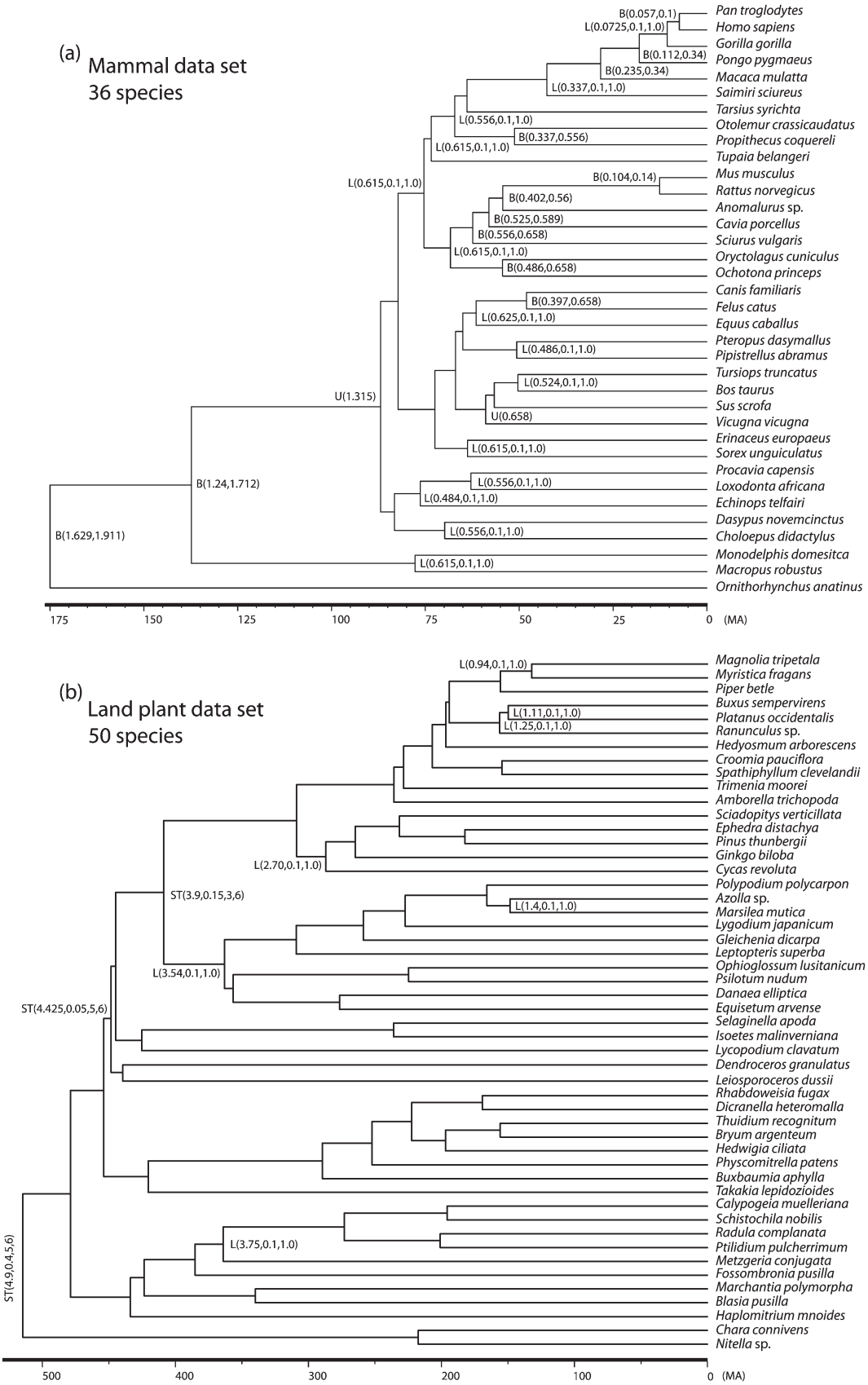

FIG. 2. Phylogenetic trees used for divergence time estimation for the mammal and plant data sets. $L(a,p,c)$: lower (minimum) age bound a with distribution parameters p and c; $B(a,b)$: joint bounds with the minimum at a, and maximum at b; $U(b)$: upper (maximum) age bound b; and $ST(\xi,\omega,\alpha,\mathrm{df})$: Skew-$t$ distribution for the node age with distribution parameters ξ, ω, α, and df. For details about the various distribution parameters and specification of the fossil calibrations, see Inoue et al. (2010) and the PAML documentation.

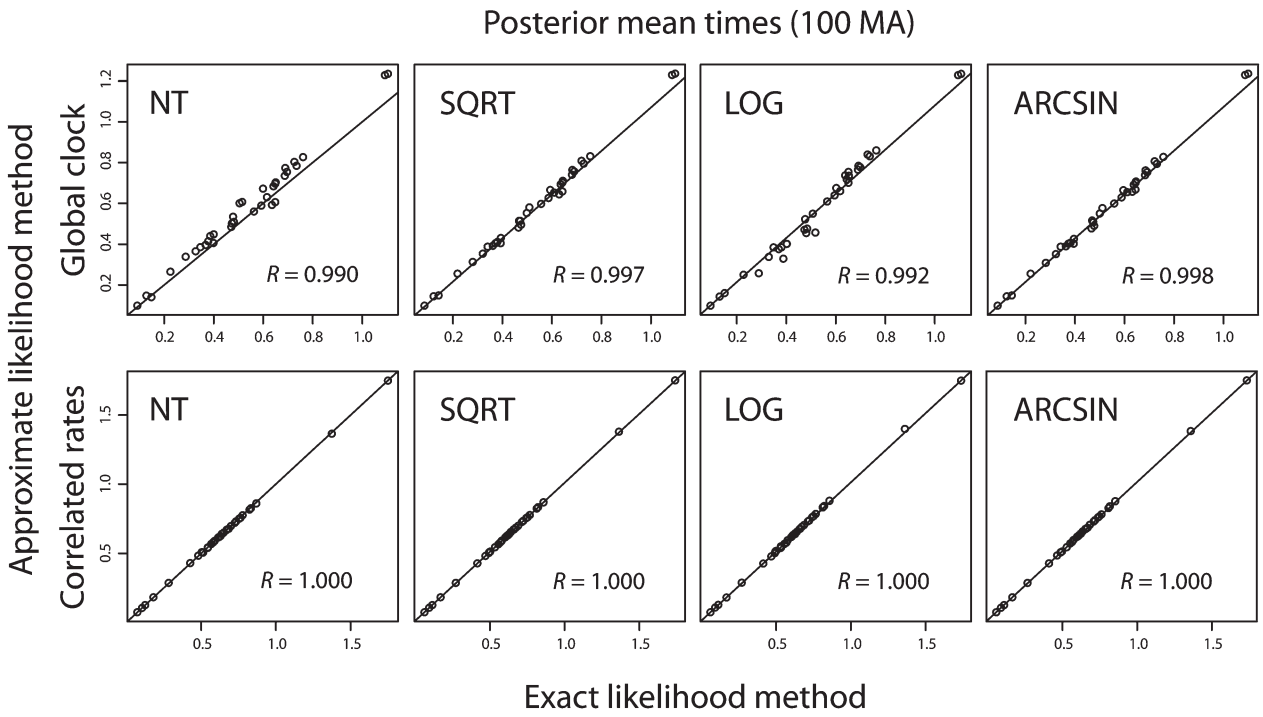

FIG. 3. The posterior means of divergence times obtained using the approximate methods of likelihood calculation (NT, SQRT, LOG, and ARCSIN) plotted against those obtained using the exact method of likelihood calculation. The mammal data set was analyzed, and the posterior means of the 35 node ages in the tree of figure 2*a* are used in the scatterplots. Either the global clock (top) or the correlated clock (bottom) was assumed. See text for the specification of priors and other details of the analysis.

When the global clock was assumed, substantial differences in the mean posterior times among the transforms were observed (fig. 3). The SQRT and ARCSIN outperformed the LOG, and all these three clearly outperformed the plain NT approximation. There were noticeable differences in the posterior distribution of the root age and the substitution rate (fig. 4). The age of the root for the approximations was between 1.03 and 1.12 times the age estimated with the exact method. Furthermore, the NT, SQRT, and ARCSIN underestimated the substitution rate by 25%, 12%, 10%, respectively, whereas the LOG slightly overestimated it by 3% (fig. 4). Overall, for the worst performing approximation, NT, node age estimates were between 0.93 and 1.19 times those from the exact method. For the best performing approximation, ARCSIN, node age estimates ranged between 0.95 and 1.09 times those from the exact method. Therefore, node ages were estimated more reliably with the approximations than the substitution rate.

Figure 5 shows the approximate $\Delta\ell$ values for parameter values proposed during the stationary phase of the MCMC for each transform plotted against $\Delta\ell$ calculated using the exact method. The discrepancies between the exact and the approximate methods are more apparent when the global clock is considered. In this case, all approximate methods overestimated the likelihood. With the relaxed clock, $\Delta\ell$ values sampled during the stationary phase of the MCMC ranged between 20 and 80 log-likelihood units away from the likelihood peak. For the global clock, $\Delta\ell$ values were considerably further away from the peak, by roughly between 900 and 1,300 units. This is because the molecular clock is seriously violated in the data, and it is impossible for the global clock model to fit the branch lengths estimated without the clock assumption. It is noteworthy that the ARCSIN can still achieve a reasonable approximation so far away from the likelihood peak.

Land Plant Data Set

When the relaxed clock was assumed, all approximations gave similar posterior mean times to the exact method, although some discrepancies in the age of some nodes were apparent (fig. 6). The LOG appears to be slightly worse than the other transforms. Likewise, the posterior distribution of the root age and mean rate μ were similar across methods, although the plain approximation (NT) underestimated the mean substitution rate by about 28% (fig. 7). The mean posterior rate was $\mu = 0.049 \times 10^{-8}$ per site per year $(0.021, 0.11)$ and the drift parameter $\sigma^2 = 0.93\ (0.65, 1.3)$ by the exact method. There is considerably more rate variation among lineages and the molecular clock is violated much more severely in the plant data set than in the mammal data set.

When the global clock was assumed, large discrepancies in node ages were observed between the approximations and the exact method (fig. 6). Also, the various approximations produced different posterior distributions for the root age and substitution rate (fig. 7). The NT, SQRT, and ARCSIN underestimated the substitution rate by 72%, 37%, and 35%, respectively, whereas LOG overestimated it by 30% (fig. 7).

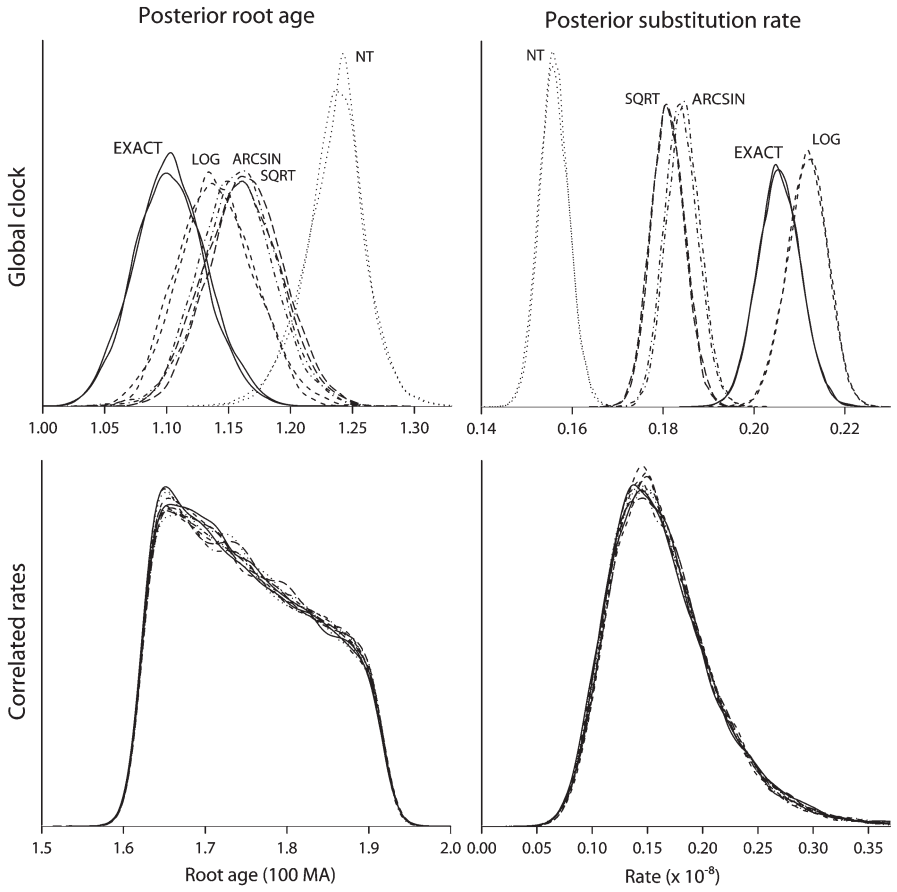

FIG. 4. Estimated posterior density for the age of the root (left) and the overall rate μ (right) under the global clock (top) and autocorrelated rates (bottom) models. The mammal data are used. Two curves are shown for each analysis, from two independent MCMC runs.

The molecular clock was violated seriously, so the branch lengths proposed during the MCMC were far from their MLEs. Correspondingly, the $\Delta\ell$ values for the various transforms differ substantially from the exact method, specially under the global clock model (fig. 8). With the relaxed clock, $\Delta\ell$ values for proposals made during the stationary phase of the MCMC ranged between 50 and 200 log-likelihood units away from the likelihood peak. Under the global clock, $\Delta\ell$ values were further away from the peak, by roughly between 600 and >1, 600 units. All approximations overestimated the likelihood under both clock models, with the NT method providing the poorest approximation.

In comparison with the mammal data set, the clock is more seriously violated in the plant data set. As a result, the branch lengths visited by the MCMC are farther away from the likelihood peak and the approximations are in general poorer in the plant data set than in the mammal data set (cf., fig. 8 with fig. 5).

Proposal Space in the MCMC Algorithm and Adequacy of the Approximate Method

It is interesting to consider how large the range of parameter values should be within which we need to calculate the likelihood reliably. For the simple two-species example when $0 < \hat{b} < \infty$ (fig. 1*b*), the 99.9% "likelihood interval (region)" is constructed by lowering the log-likelihood from the peak by $\frac{1}{2}\chi^2_{\nu,0.1\%}$, where $\chi^2_{\nu,0.1\%}$ is the 0.1% critical point of the χ^2 distribution with ν df (e.g., Yang 2006: p. 25). In our current example with $\nu = 1$ parameter, the likelihood interval consists of all values of b at which $\Delta\ell > -\frac{1}{2} \times 10.83 = -5.41$. If the prior is diffuse and not in strong conflict with the likelihood, the Bayesian 99.9% CIs may roughly coincide with the 99.9% likelihood interval. Then in 99.9% of the samples taken in the MCMC, the log-likelihood will be within 5.41 units of the maximum at the MLE. Many rejected proposals, however, may have even lower likelihood values, so that we need to be able to calculate the log-likelihood reliably over a larger (but perhaps not much larger) region than the 99.9% likelihood interval, that is, the interval with $\Delta\ell > -\frac{1}{2}\chi^2_{\nu,0.1\%}$. Very poor proposals will be rejected whether their log-likelihood is reliably calculated: it makes virtually no difference to the acceptance or rejection of the proposal whether the logarithm of the likelihood ratio for the proposal is −1, 000 or −2, 000.

Divergence time estimation on a phylogeny under the clock and relaxed clock models is more complicated. Currently, for a phylogeny of s species, PAML estimates $2s - 3$ branch lengths on the unrooted tree without assuming the molecular clock in order to construct the Taylor approximation to the likelihood surface used in the MCMC. The global clock and the relaxed clock models are all special cases of this no-clock model and their likelihood cannot be

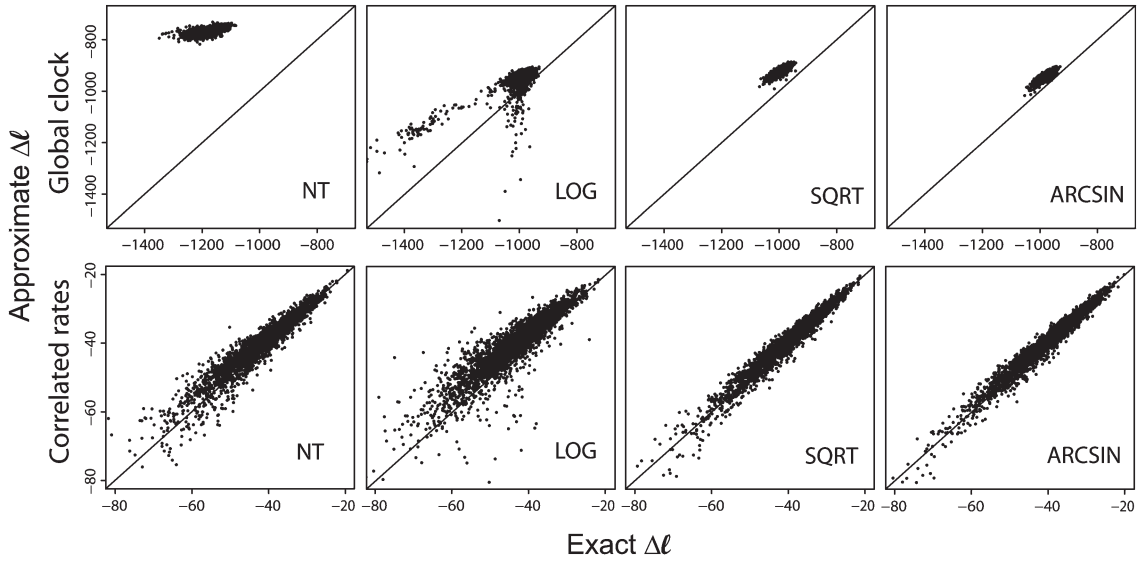

FIG. 5. The log-likelihood values ($\Delta\ell$) calculated using the approximate methods (NT, SQRT, LOG, and ARCSIN) for branch lengths visited during the MCMC plotted against the exact log-likelihood values. The mammal data set was analyzed under the global-clock (top) and correlated rates (bottom) models. See also figures 3 and 4 for results from the same analysis.

higher than the optimal likelihood achieved at the MLEs of branch lengths under the no-clock model. Because the assumed clock model (either the global clock or relaxed clock model) may impose unrealistic constraints on the branch lengths, proposals and samples taken in the MCMC may be consistently far away from the likelihood peak calculated under the no-clock model. In fact, branch lengths near the likelihood peak may not be achievable. For the plant and mammal data sets, the branch lengths proposed in the MCMC were from 20 (mammal data set, correlated rates model) to over 1,600 (plant data set, global clock) log-likelihood units away from the peak (figs. 5 and 8).

We note that the global clock should not be assumed to estimate divergence times if it is seriously violated by the data. Nevertheless, if time estimates under the global clock are desired, an alternative procedure is to obtain the MLEs, gradient and Hessian for parameters under the global clock, which are the $s - 1$ node distances (node ages measured

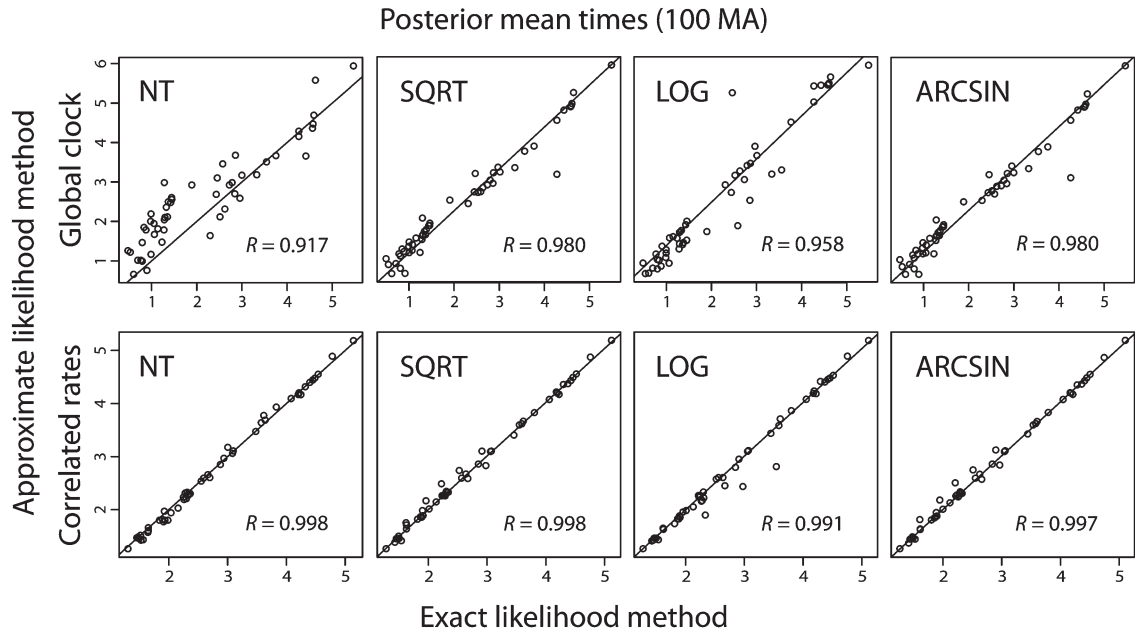

FIG. 6. The posterior means of divergence times obtained using the approximate methods of likelihood calculation (NT, SQRT, LOG, and ARCSIN) plotted against those obtained using the exact method. The plant data set was analyzed, and the posterior means of the 49 node ages in the tree of figure 2*b* are used in the scatterplots. Either the global clock (top) or the correlated clock (bottom) was assumed. See text for details of the analysis.

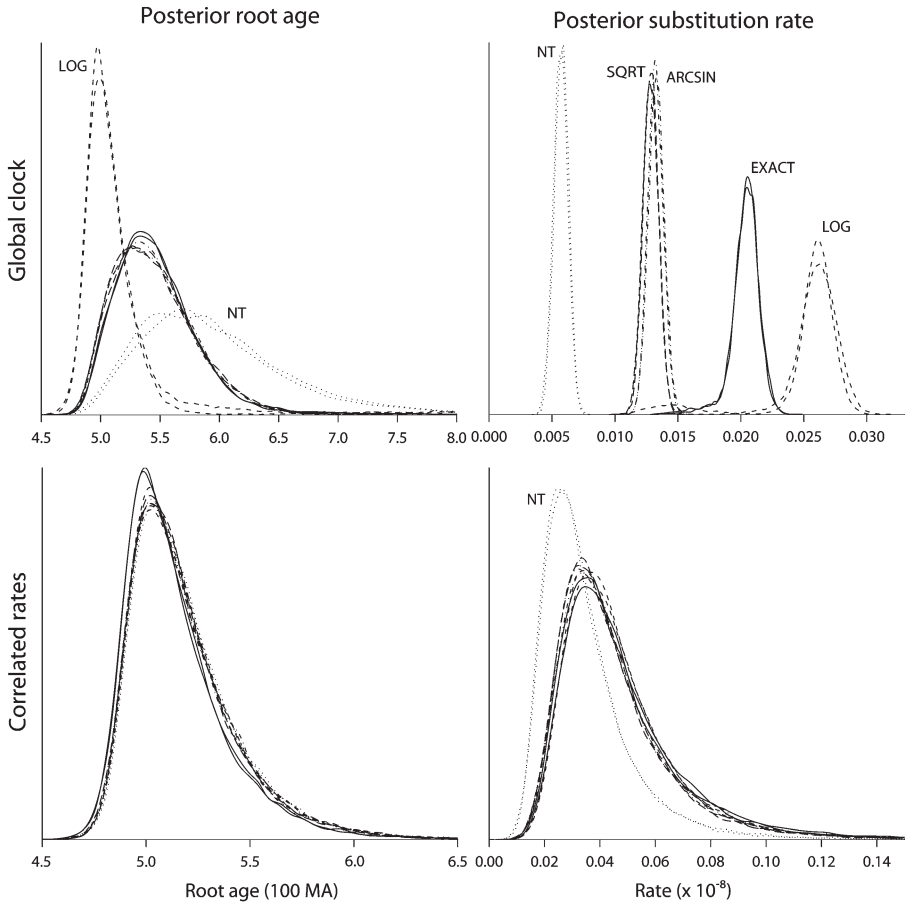

FIG. 7. Estimated posterior density for the age of the root (left) and overall rate μ (right) for the global clock (top) and autocorrelated rates (bottom) models. The plant data are used. Two curves are shown for each analysis, from two independent MCMC runs.

by the number of substitutions) on the rooted tree. The ML and Bayesian models would then be conducted under the same clock model and proposals in the MCMC would be expected to be close to the likelihood peak, thus improving the approximation for analysis under the global clock. We have not pursued this approach in this paper, partly because there seems to be little point in fitting the global clock when it is seriously violated.

To summarize the results obtained from the real data analysis, the approximation is very good when the branch lengths proposed during the MCMC are close to the MLEs. This is the case for the mammal data set under the correlated rates model, where the time and rate estimates from all approximate methods are virtually identical to those from the exact method. The approximation is quite good under the correlated rates model for the plant data set as well. In contrast, when the proposed branch lengths are far from the MLEs, the discrepancies between the true likelihood and the approximation become important. Because the NT approximation overestimates the likelihood on the left side of the MLE, there will be an excess of short branch lengths sampled during the MCMC. Overall, trees sampled when the likelihood is calculated with the NT approximation under the global clock are much shorter (in terms of substitutions per site) than if the likelihood is calculated exactly. The posterior substitution rate is therefore underestimated, as shorter trees require slower rates to accommodate the fossil constraints. Because the SQRT and ARCSIN approximations also overestimate the likelihood for the short branch lengths, they underestimate the posterior rate as well, although not as severely as the NT method. In contrast, the LOG overestimates the rates. This pattern was apparent in the posterior distribution of substitution rates in the global clock analysis for both data sets (figs. 4 and 6).

It is clear that the suitability of the approximate method depends on the data set being analyzed. Two factors seem important: the alignment length and the adequacy of the molecular clock. ML theory establishes that as the sample size increases the MLEs are asymptotically normally distributed around the true parameter values. If the alignment is too short, the asymptotic theory may not be reliable. Furthermore, as discussed above, the approximation may be poor for analysis under the global clock when the global clock is seriously violated. The plant data set represents such an extreme, where the clock is seriously violated, the alignment is relatively short, and the differences between the exact and approximate methods are large.

As an example of good approximation for the clock analysis when the clock is largely correct, we reanalyzed the cat data set of Johnson et al. (2006) and Rannala and Yang

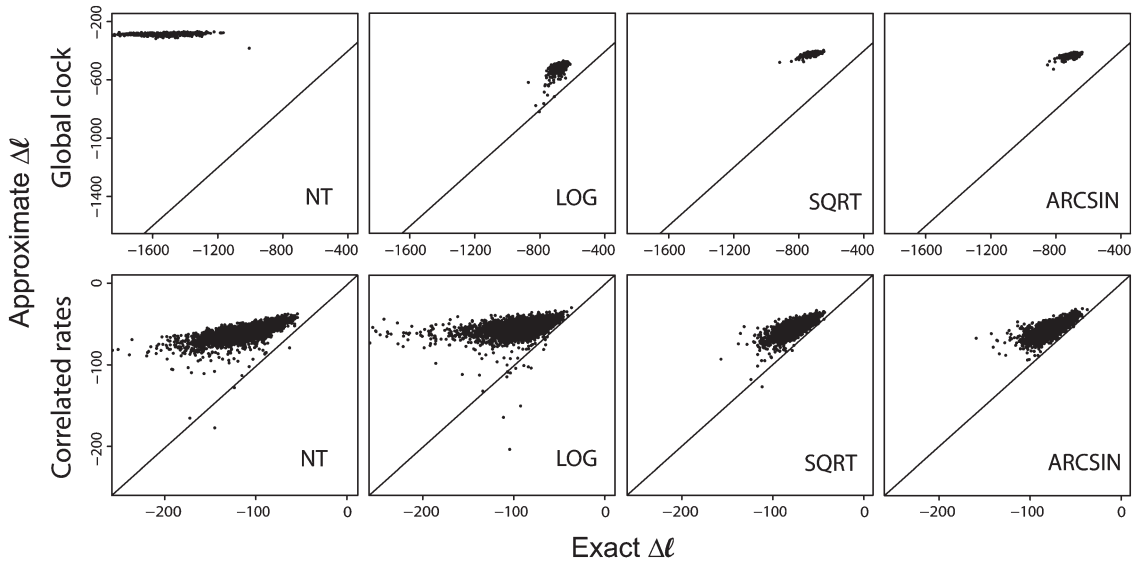

FIG. 8. The log-likelihood values ($\Delta\ell$) calculated using the approximate methods (NT, SQRT, LOG, and ARCSIN) for branch lengths visited during the MCMC plotted against the exact log-likelihood values. The plant data set was analyzed under the global-clock (top) and correlated rates (bottom) models. See also figures 6 and 7 for results from the same analysis.

(2007), using exactly the same fossil calibrations and rate prior as in Inoue et al. (2010) but assuming the global clock. The alignment has 19,984 sites. The divergences are within 15 My, and the substitution rate appears constant over lineages. Under these settings, the NT method and the exact method gave virtually the same posterior times (results not shown).

Computational Efficiency

The approximate method is substantially faster than conventional likelihood calculation. For the exact method, one MCMC run under correlated rates model for the 36 mammalian species alignment with 7,260 sites took ~ 2 days on a desktop computer, but only ~3 min. with the approximations. The plant data set, having more species (50) but a shorter alignment (1,974 sites), took ~20 hr for the exact method and about ~10 min with the approximations. Calculation of the MLEs, the gradient, and the Hessian took a negligible amount of time for those two data sets. Such speed performance is appealing as it opens up the possibility of analyzing large genomic alignments. We recommend the use of the approximate method, especially the ARCSIN, for analysis of large data sets under relaxed clock models such as the correlated rates and independent rates models. If results under the global clock are desirable, care should be taken to confirm that the molecular clock is not seriously violated, as otherwise the exact method is necessary.

Acknowledgments

We thank Stephane Guindon, Mark Holder, and Oliver Pybus for many constructive comments. We are grateful to Jun Inoue for preparing the alignment, tree, and fossil calibrations for the mammal data set, and Joseph Brown and Yin-Long Qiu for providing the alignment, tree, and fossil calibrations for the plant data set. We thank Jeff Thorne for clarifications concerning implementations of approximate likelihood calculation in multidivtime. This work was supported by a grant (BB/G006431/1) from the Biotechnological and Biological Sciences Research Council (United Kingdom) to Z.Y. Z.Y. gratefully acknowledges the support of K.C. Wong Education Foundation, Hong Kong.

Appendix

Calculation of the Gradient and Hessian

In BASEML and CODEML, the gradient for the branch lengths (g) is calculated using the central difference method. For a general multivariate function $f(\mathbf{x})$, with $\mathbf{x} = \{x_i\}$, the central difference method approximates the gradient by

$$g_i = \frac{\partial f(\mathbf{x})}{\partial x_i} \approx \frac{f(\mathbf{x} + h_i \mathbf{e}_i) - f(\mathbf{x} - h_i \mathbf{e}_i)}{2h_i}, \qquad (17)$$

where $\mathbf{e}_i$ is a vector with the ith element to be 1 and all other elements to be 0. The step length is set at $h_i = \epsilon(|x_i| + 1)$, with ϵ to be a small number, around 10^{-8}-10^{-5}. Equation (17) is stable and normally leads to a good approximation.

Similarly, the second-order difference method can be used to calculate the Hessian

$$H_{ii} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i^2} \approx \frac{f(\mathbf{x} + h_i \mathbf{e}_i) - 2f(\mathbf{x}_i) + f(\mathbf{x} - h_i \mathbf{e}_i)}{h_i^2}, \qquad (18)$$

$$H_{ij} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \approx \frac{f(\mathbf{x} + h_i \mathbf{e}_i + k_j \mathbf{e}_j) - f(\mathbf{x} + h_i \mathbf{e}_i - k_j \mathbf{e}_j) - f(\mathbf{x} - h_i \mathbf{e}_i + k_j \mathbf{e}_j) + f(\mathbf{x} - h_i \mathbf{e}_i - k_j \mathbf{e}_j)}{4h_i k_j}, \qquad (19)$$

where h_i and k_j are the step sizes. This was used in previous versions of PAML. However, equations (18) and (19) are unstable and highly sensitive to the step sizes. The current version of PAML (since version 4.3) uses the OPS estimator of the Hessian (Porter 2002; Seo et al. 2004), which is generally more stable. However, the OPS estimator is based on the assumption that the expectation of the gradient at the MLE is zero. This is not true when the MLEs are at the boundary of the parameter space (i.e., when they are zero). Thus, the current version of PAML does not provide accurate estimates of the Hessian for zero branch lengths.

One way to address this problem may be to use the difference approximation (equations (18) and (19)) to calculate the Hessian elements for zero branch lengths, and the OPS method to estimate the rest of the matrix. Additionally, the first derivatives and the diagonals of the Hessian could be computed exactly, as described by Yang (2000) and combined with the first-order difference method: that is, we compute g_i exactly and then apply the difference approximation

$$H_{ij} = \frac{g_i(\mathbf{x} + h_j \mathbf{e}_j) - g_i(\mathbf{x} - h_j \mathbf{e}_j)}{2h_j}. \quad (20)$$

This approach should lead to correct H_{ij} values even for branches of length zero. In any case, as the gradient of a zero branch length is typically not zero, the linear term of the Taylor expansion has a substantial effect on the approximation. The plant data set contains nine branches of length zero. The gradient for these nine branch lengths ranges from $-1,218$ to -53, so their contribution to the approximation is important. Therefore, the corresponding node age estimates are reasonably close to the estimates obtained by the exact method.

Compared with zero branch lengths, infinite branch lengths cause even bigger problems. In this case, the branch lengths reported by BASEML or CODEML are arbitrarily large and depend on the particular data set and on the upper bound set in the program. These branch lengths have Hessian elements that approach zero asymptotically. The approximate methods described here are not expected to perform well in this case, although the ARCSIN might provide the most robust approximation. Users who wish to use the approximate method should inspect the ML tree and observe whether unusually long branches are present. If they are, some species may be removed, a new alignment may be prepared or the exact method should be used.

References

Benton M, Donoghue P, Asher R. 2009. Calibrating and constraining molecular clocks. In Hedges S, Kumar S, editors. The timetree of life. Oxford: Oxford University Press. p. 35–86.

Benton MJ, Donoghue PC. 2007. Paleontological evidence to date the tree of life. *Mol Biol Evol.* 24:26–53.

Drummond AJ, Ho SY, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4:e88.

Guindon S. 2010. Bayesian estimation of divergence times from large sequence alignments. *Mol Biol Evol.* 27:1768–1781.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol.* 22:160–174.

Inoue J, Donoghue PC, Yang Z. 2010. The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. *Syst Biol.* 59:74–89.

Johnson WE, Eizirik E, Pecon-Slattery J, Murphy WJ, Antunes A, Teeling E, O'Brien SJ. 2006. The late Miocene radiation of modern Felidae: a genetic assessment. *Science* 311:73–77.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In Munro HN, editor. Mammalian protein metabolism. New York: Academic Press. p. 21–123.

Kumar S. 2005. Molecular clocks: four decades of evolution. *Nat Rev Genet.* 6:654–662.

Porter J. 2002. Efficiency of covariance matrix estimators for maximum likelihood estimation. *J Bus Econ Stat.* 20:431–440.

Qiu YQ, Li L, Wang B, et al. (13 co-authors) 2007. A non-flowering land plant phylogeny inferred from nucleotide sequences of seven chloroplast, mitochondrial, and nuclear genes. *Int J Plant Sci.* 168:691–708.

Rannala B, Yang Z. 2007. Inferring speciation times under an episodic molecular clock. *Syst Biol.* 56:453–466.

Seo TK, Kishino H, Thorne JL. 2004. Estimating absolute rates of synonymous and nonsynonymous nucleotide substitution in order to characterize natural selection and date species divergences. *Mol Biol Evol.* 21:1201–1213.

Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol.* 15:1647–1657.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol.* 39:306–314.

Yang Z. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *J Mol Evol.* 51:423–432.

Yang Z. 2006. Computational molecular evolution. Oxford: Oxford University Press.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.

Yang Z, Rannala B. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol Biol Evol.* 23:212–226.

Zuckerkandl E, Pauling L. 1965. Evolutionary divergence and convergence in proteins. In Bryson V, Vogel HJ, editors. Evolving genes and proteins. New York: Academic Press. p. 97–166.