

Accuracy of Rate Estimation Using Relaxed-Clock Models with a Critical Focus on the Early Metazoan Radiation

Simon Y. W. Ho,* Matthew J. Phillips,* Alexei J. Drummond,† and Alan Cooper*

*Henry Wellcome Ancient Biomolecules Centre and †Evolutionary Biology Group, Department of Zoology, University of Oxford, Oxford, United Kingdom

In recent years, a number of phylogenetic methods have been developed for estimating molecular rates and divergence dates under models that relax the molecular clock constraint by allowing rate change throughout the tree. These methods are being used with increasing frequency, but there have been few studies into their accuracy. We tested the accuracy of several relaxed-clock methods (penalized likelihood and Bayesian inference using various models of rate change) using nucleotide sequences simulated on a nine-taxon tree. When the sequences evolved with a constant rate, the methods were able to infer rates accurately, but estimates were more precise when a molecular clock was assumed. When the sequences evolved under a model of autocorrelated rate change, rates were accurately estimated using penalized likelihood and by Bayesian inference using lognormal and exponential models of rate change, while other models did not perform as well. When the sequences evolved under a model of uncorrelated rate change, only Bayesian inference using an exponential rate model performed well. Collectively, the results provide a strong recommendation for using the exponential model of rate change if a conservative approach to divergence time estimation is required. A case study is presented in which we use a simulation-based approach to examine the hypothesis of elevated rates in the Cambrian period, and it is found that these high rate estimates might be an artifact of the rate estimation method. If this bias is present, then the ages of metazoan divergences would be systematically underestimated. The results of this study have implications for studies of molecular rates and divergence dates.

Introduction

The rate of molecular evolution can vary considerably among different organisms, challenging the concept of the “molecular clock.” If the assumption of rate constancy does not hold across a tree, then phylogenetic inference and divergence date estimation can be seriously misled unless this rate heterogeneity among lineages is adequately taken into account (Yoder and Yang 2000). In recent years, a number of methods have been developed to relax the clock assumption to varying degrees (e.g., Takezaki, Rzhetsky, and Nei 1995; Cooper and Penny 1997; Sanderson 1997, 2002; Rambaut and Bromham 1998; Thorne, Kishino, and Painter 1998; Cutler 2000; Huelsenbeck, Larget, and Swofford 2000; Aris-Brosou and Yang 2002; Thorne and Kishino 2002). These methods have been applied to a range of evolutionary questions, and their use is becoming more widespread with increasing awareness of the effects of ignoring violation of the molecular clock assumption.

One method of addressing the problem of non-clock-like data is to dispose of the molecular clock assumption altogether and to assign a unique rate parameter to each branch in the tree, thereby allowing rates to vary in an unrestricted manner. This method has been implemented for the purpose of inferring phylogenetic trees (e.g., Yang 1997; Huelsenbeck and Ronquist 2001).

Unfortunately, such an approach cannot be taken when estimating absolute rates or divergence dates because rate and time are inextricably linked, unless the sequences are noncontemporaneous and have known ages (Rambaut 2000; Drummond et al. 2002). Instead, we can implement a method that allows the rate to vary across the tree in a relatively constrained manner. For example, some methods

employ local molecular clock models, which permit a different rate in each region of the tree. These regions can be specified by the user (Yang 1997; Yoder and Yang 2000), assigned objectively by the algorithm (Foster 2004), or determined on the basis of phylogenetic relationships (Cooper and Penny 1997; Bromham et al. 1998). However, implementing local clock models for large trees can be difficult because of the number of possible rate configurations (Sanderson 1998).

An alternative to using local clock models is to allow the molecular rate to vary throughout the tree, but with closely related species sharing similar rates. This follows the idea that molecular rates are heritable because they are tied to physiological, biochemical, and life-history characteristics of the species in question. Adopting this philosophy, Sanderson (1997, 2002) proposed a nonparametric rate smoothing method and a semiparametric penalized likelihood approach, both of which operate by minimizing rate changes throughout the tree.

In a Bayesian framework, a rate can be assigned to each branch of a tree and a prior can be imposed that allows the rate to “evolve” throughout the tree in an autocorrelated manner. Thorne, Kishino, and Painter (1998) and Aris-Brosou and Yang (2002) implemented parametric models in which the rate in a given branch (the descendant rate) is a priori related to the rate in the branch immediately ancestral to it (the parent rate). The model they adopt for the prior is one in which the descendant rate is drawn from a chosen distribution, such as a lognormal, exponential, or gamma distribution, whose mean is the parent rate (Thorne, Kishino, and Painter 1998; Kishino, Thorne, and Bruno 2001; Aris-Brosou and Yang 2002). Aris-Brosou and Yang (2002) also employed a model in which the rate was assumed to evolve according to the Ornstein-Uhlenbeck process (OUP), a stationary Gaussian process that includes a rate-slowing friction term. The variance of each of these distributions corresponds to the degree of departure from the molecular clock. The hyperparameters (the parameters

Key words: molecular rates, relaxed clock, penalized likelihood, Bayesian inference, Cambrian explosion.

E-mail: simon.ho@zoo.ox.ac.uk.

Mol. Biol. Evol. 22(5):1355–1363. 2005

doi:10.1093/molbev/msi125

Advance Access publication March 9, 2005

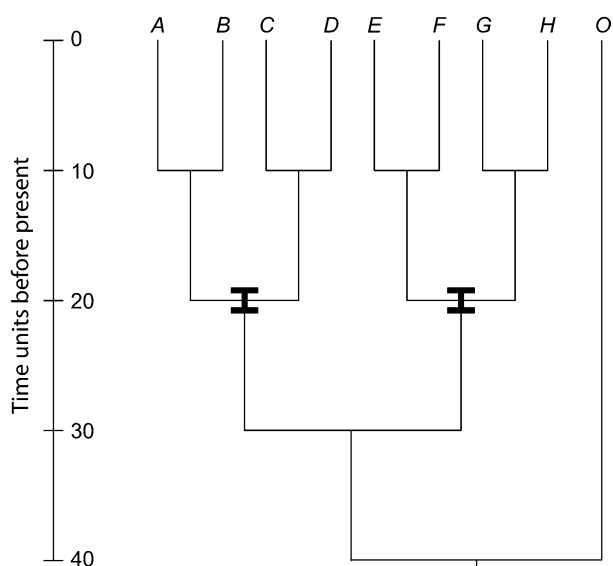


FIG. 1.—The rooted binary tree used for simulating sequence evolution. The timescale is drawn in arbitrary time units. Apart from the branch leading to the out-group, sequence *O*, all branches are 10 time units in length. Bold lines show the calibration ranges used for estimating rates.

that govern the prior distributions) can be optimized by comparing the posterior likelihoods of each rate prior given the sequence data.

It is important to evaluate the performance of the different relaxed-clock methods that are available. In spite of their increasingly widespread application, there have been few studies into the accuracy of relaxed-clock models (for exceptions, see Kishino, Thorne, and Bruno 2001; Yang and Yoder 2003; Perez-Losada, Hoeg, and Crandall 2004). In this paper, we assess the accuracy of rate estimates produced by rate smoothing (penalized likelihood) and by Bayesian inference using various models of autocorrelated rate change. We tested the methods using sequence data generated by Monte Carlo simulation on a known tree, for which the true rate of substitution in each branch was known. The simulation study is followed by a critical simulation-based investigation of molecular rates in the Cambrian period.

Materials and Methods

The accuracy of methods for rate estimation can only be tested when the true phylogeny (including the ages of internal nodes) relating the sequences is known, which is rarely the case for real data. Instead, sequence evolution can be simulated on a known tree, and substitution rates can be recorded during the simulation process for comparison with rate estimates subsequently made from the data. We generated alignments of nine nucleotide sequences, each 1,000 nt in length, on the rooted tree in figure 1. The out-group sequence is only used to root the tree. Simulations were performed using the Java program Rate-Evolver v1.0 (available upon request from S.Y.W.H.), which can generate sequences under models of nucleotide substitution ranging from a Jukes-Cantor model (Jukes and Cantor 1969) to the HKY85 model (Hasegawa, Kishino, and Yano 1985). The program can also implement various

processes of autocorrelated rate change, including rates following lognormal, exponential, and gamma distributions, as well as independently sampled rates, where rates are drawn from a uniform distribution (fig. 2). Under these rate change models, with the exception of the independently sampled rates, each new rate is sampled from the chosen distribution with the parent rate as the mean. The user can also set the probability of a new rate being drawn at the beginning of each branch, which is set to 1.0 for the models studied in this paper.

Whenever it was allowed by the computer program, rate estimates were calibrated by constraining two internal nodes to the range 19–21 time units, which is an uncertainty range of 5% on either side of their actual age of 20 time units (fig. 1). In analyses using PhyBayes, calibration uncertainty could not be specified, so point values of 20 time units were used instead. We chose to calibrate rate estimates using internal nodes rather than the root node because in most applications of the relaxed clock, the ages can be ascribed to internal nodes with greater confidence than to the root node.

Case 1—Constant Rates Among Lineages

Ten sequence alignments were generated on the tree in figure 1 under a molecular clock model, in which the rate was held constant at 0.01 average substitutions per site per time unit throughout the tree. These data sets were analyzed using several methods: maximum likelihood, implemented by the program baseml (Yang 1997); penalized likelihood, implemented by r8s (Sanderson 2003); and Bayesian inference under various relaxed-clock models, implemented by multidivtime (Thorne, Kishino, and Painter 1998) and PhyBayes (Aris-Brosou and Yang 2002, 2003). In the Bayesian analyses, posterior distributions were approximated by sampling values at every 100 steps over 1,000,000 Markov chain Monte Carlo (MCMC) steps, after discarding a burn-in of 100,000 steps. Preliminary investigations showed that this burn-in was sufficient for the chain to reach the stationary distribution, from which subsequent samples were drawn. This was confirmed by a posteriori inspection of the MCMC samples from each analysis using the program Tracer (Rambaut and Drummond 2004), which also showed that the set of samples for each of the estimated parameters had an effective sample size of over 100. The effective sample size is influenced not only by the number of samples that are drawn from the MCMC but also by the degree of autocorrelation among the samples.

In all analyses using penalized likelihood, the (fixed) input tree was inferred using maximum likelihood, as implemented by the program baseml (Yang 1997), without the assumption of a molecular clock. In the process of tree inference, rate homogeneity among sites was assumed, matching the simulation conditions, but the transition/transversion ratio was estimated. In the estimation of rates using penalized likelihood, the smoothing parameter was optimized for each data set using the cross-validation procedure available in the program by comparing sums of squared errors for a wide range of smoothing values of 10^x ($x = 0.0, 0.2, \dots, 4.0$). Each data set was also analyzed using a smoothing parameter of 1.0, but the resulting rate estimates were extremely poor and are not presented here.

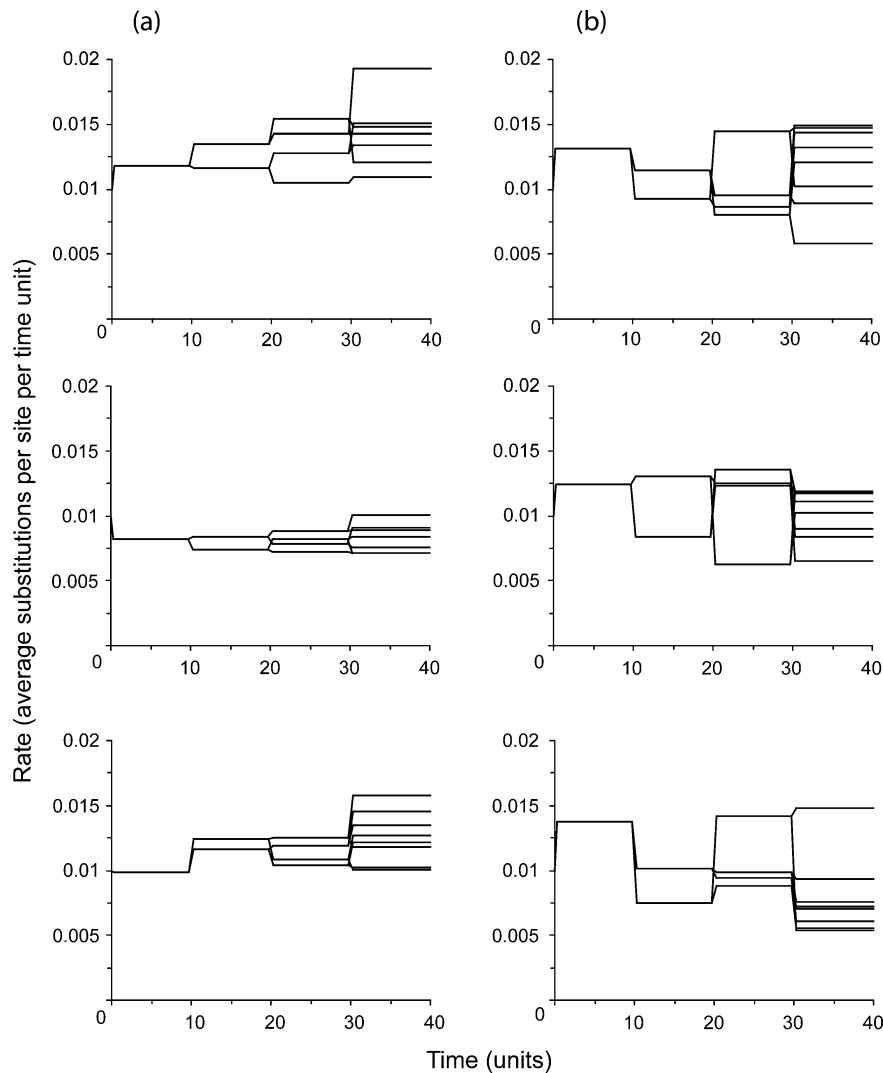


FIG. 2.—Profiles of the rates generated and used in the stochastic simulation process. Rate changes occur instantaneously but are depicted otherwise to facilitate interpretation. (a) Rates were autocorrelated, with the new rate being sampled from an LND with mean equal to the rate in its parent branch and a constant variance parameter of 0.025. (b) The rate in each branch was drawn independently from a uniform distribution over the interval [0.005,0.015].

In the first set of analyses, a molecular clock constraint was enforced in each of three methods used to estimate rates from the data (maximum likelihood, the Langley-Fitch method used in r8s, and Bayesian inference). The data were then analyzed a second time, without a molecular clock constraint, using penalized likelihood and relaxed-clock Bayesian inference. In the Bayesian analyses, four different inference models of rate change were tested: lognormally distributed (LND) rates (multidivtime and PhyBayes), gamma-distributed (GD) rates (PhyBayes), exponentially distributed (ED) rates (PhyBayes), and OUP (PhyBayes). The hyperparameters of the lognormal distribution were assumed to be 0.01 (mean) and 0.025 (variance parameter). The drift parameter for the OUP prior was optimized by comparing the posterior likelihoods when it was set at each of four different values (10^{-2} , 10^{-1} , 10^0 , and 10^1).

For all nonclock analyses performed with PhyBayes, the default values were used for the hyperparameters of the birth-death process prior on divergence times: the birth (speciation) rate, death (extinction) rate, and sampling fraction were drawn

from uniform distributions over [0,15], [0,5], and [0,0.001], respectively. The ages of internal nodes are specified relative to the age of the root node, which is arbitrarily set at 1.0. This prior is described in detail by Aris-Brosou and Yang (2002). In all nonclock analyses using multidivtime, the default value of 1.0 was used for the hyperparameter minab, which governs the prior distribution of divergence times. The distribution of internal node ages is conditional on the age of the root node, which is in turn described by its own gamma distribution prior. The generalized Dirichlet prior on divergence times used in multidivtime is described in detail by Kishino, Thorne, and Bruno (2001). In contrast to the two Bayesian relaxed-clock programs, penalized likelihood does not require the specification of any priors on divergence times or rates.

Case 2—Lognormally Autocorrelated Rates Among Lineages

Ten sequence alignments were generated on the tree in figure 1 under a model in which each descendant rate was

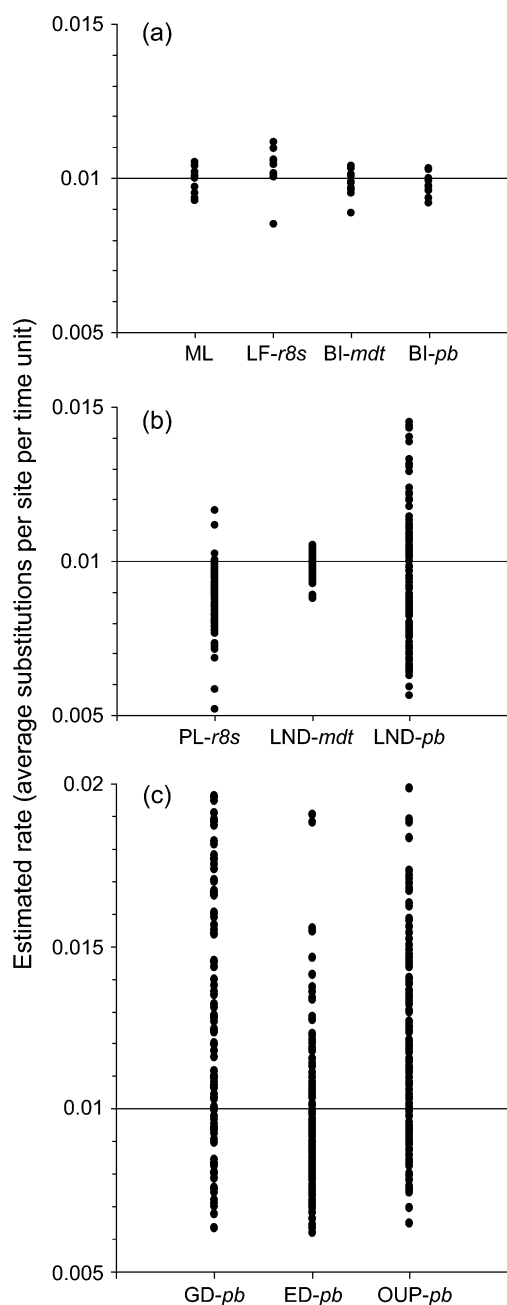


FIG. 3.—Branch-specific rate estimates made from 10 sets of sequence data generated under a constant-rate (molecular clock) model. (a) Estimates were made under the assumption of a molecular clock using maximum likelihood, Langley-Fitch method (LF-r8s), and Bayesian inference using the programs multidivtime (BI-mdt) and PhyBayes (BI-pb). Estimates were then made without the assumption of a molecular clock using (b) penalized likelihood (PL-r8s) and Bayesian inference with LND rate priors (LND-mdt and LND-pb) and (c) Bayesian inference using GD (GD-pb), ED (ED-pb), and OUP (OUP-pb) models of rate change.

autocorrelated with its parent rate (see fig. 2a). The log of each descendant rate, r_i , was drawn from a normal distribution with a mean equal to the log of the parent rate, r_{i-1} : $\log(r_i) \sim N(\log(r_{i-1}), \sigma^2 t)$, where σ^2 is the variance parameter and t is the length (in time units) of the descendant branch. For the simulations in this study, σ was set at 0.5 and t was 10.0 time units for all in-group branches; con-

sequently, the variance of the distribution ($s^2 = \sigma^2 t$) was fixed at a constant value of 0.25 throughout the tree (disregarding the out-group taxon). This value was chosen after preliminary investigations showed that it was sufficiently large to cause appreciable rate change throughout the tree but small enough for clear autocorrelation to remain (see fig. 2a).

Rates were estimated from the simulated sequences using the same methods as those described in Case 1, but without a molecular clock constraint.

Case 3—Independent Rates Among Lineages

Ten sequence alignments were generated on the tree in figure 1 under a model in which a descendant rate was independent of its parent rate (see fig. 2b). Each rate was drawn from a uniform distribution, $r_i \sim U(0.005, 0.015)$.

Rates were estimated from the simulated sequences using the same methods as those described in Case 1, but without a molecular clock constraint.

Results

Case 1

Three of the four methods performed well when the sequence data evolved according to a molecular clock constraint and were analyzed under a clock assumption (fig. 3a). The Langley-Fitch method implemented in r8s overestimated the rate by about 5% on average, which suggests that there might be an intrinsic bias (in either the method itself or its current implementation). In practice, it is probably better to use penalized likelihood with a high value for the smoothing parameter to analyze clocklike data.

Rates estimated by the other methods were close to the true rate of 0.01 average substitutions per site per time unit, with the observed variation probably resulting from the stochastic nature of Poisson-distributed nucleotide substitutions. There were small discrepancies in the spread of the estimates made using different programs, which are likely to be due to differences in priors, algorithms, and software architecture and because of stochastic variation in the MCMC realizations. Most importantly, multidivtime and PhyBayes employ different priors for divergence times, while r8s does not incorporate prior information on divergence times (unless the user specifies constraints on node ages). The multidivtime program uses the Dirichlet distribution for internal nodes and a gamma distribution for the age of the root, whereas PhyBayes uses a generalized birth-death process with species sampling (Yang and Rannala 1997). In view of these differences, it is not surprising that the programs produce different rate estimates, even in the most basic comparisons.

When the clock assumption was relaxed, penalized likelihood tended to underestimate rates. Penalized likelihood will tend to underestimate the smoothing parameter when there are random fluctuations in the realized substitution rate, which would be more pronounced when sequences are short and substitution rates are low. An underestimation bias was also seen, to a much lesser extent, in the rate estimates produced under an LND model of rate change (multidivtime and PhyBayes) (fig. 3b). Under other models of rate change (GD, ED, and OUP), the rate of substitution was

Table 1
Different Models of Rate Change Compared to the OUP

| Model of Rate Change | Average $\log(p(X M))^a$ | | |
|--------------------------|--------------------------|----------|----------|
| | Case 1 | Case 2 | Case 3 |
| OUP | -7,654.2 | -7,782.4 | -7,560.6 |
| Exponential distribution | -2.9 | -2.6 | -3.7 |
| Lognormal distribution | -2.3 | -2.2 | -2.3 |
| Gamma distribution | -5.1 | -5.9 | -4.8 |

^a Marginal posterior probability of the data given the model. Numbers in bold denote absolute values, with other numbers giving values relative to these.

often overestimated (fig. 3c). In all cases, the spread of rate estimates was much greater than under a molecular clock assumption. This suggests that if sequence data have evolved under clocklike (or approximately clocklike) conditions, it may be better to assume a molecular clock during phylogenetic analysis of the sequences to improve the precision, and perhaps accuracy, of substitution rate estimates (and, consequently, of inferred divergence dates). The particular conditions leading to a balance between obtaining better models of the data and the loss of precision will vary among different data sets, and so it is important to make an informed choice in this respect.

There was a large discrepancy between the levels of estimated uncertainty associated with rate estimates made by multidivtime and PhyBayes using a lognormal model

of rate change. The narrower spread of rate estimates made by multidivtime is probably because the program requires the user to specify a prior for the tree height. Another difference between the two programs, which could affect the levels of estimated uncertainty, is that PhyBayes calculates the full likelihood, whereas multidivtime uses a multivariate normal approximation based on maximum likelihood estimates of branch lengths in the tree.

The average (across data sets) of the sampled marginal probabilities of the data given the model was the highest when the OUP prior was used (table 1), which might simply reflect the fact that the use of this prior involves two hyperparameters. However, the OUP model still yields the highest marginal posterior probability even when each method is penalized 1 log unit for each hyperparameter. Nevertheless, there does not appear to be a clear relationship between the marginal posterior probability and the accuracy of the rate estimates. Unfortunately, a formal comparison using Bayes factors is difficult to perform without reversible-jump MCMC.

Case 2

Rates estimated from data that had evolved under a process of LND rate change were recovered with reasonable accuracy when an LND model of rate change was assumed (fig. 4). This is not surprising because the model

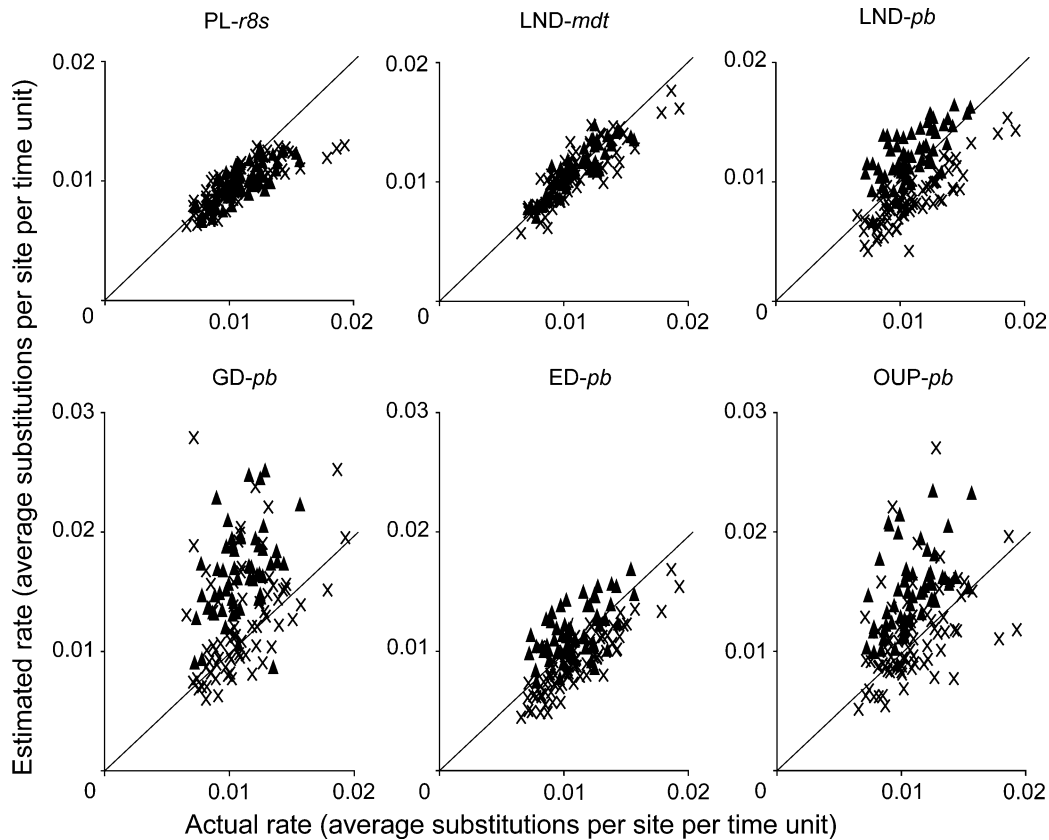


FIG. 4.—Branch-specific rate estimates made from 10 sets of sequence data generated under an autocorrelated LND rate model. Crosses and filled triangles denote estimates of rates in terminal and internal branches, respectively. Estimates were made without the assumption of a molecular clock, using (a) penalized likelihood (PL-r8s) and Bayesian inference with LND rate priors using the programs multidivtime (LND-mdt) and PhyBayes (LND-pb), and (b) Bayesian inference using GD (GD-pb), ED (ED-pb), and OUP (OUP-pb) rate priors.

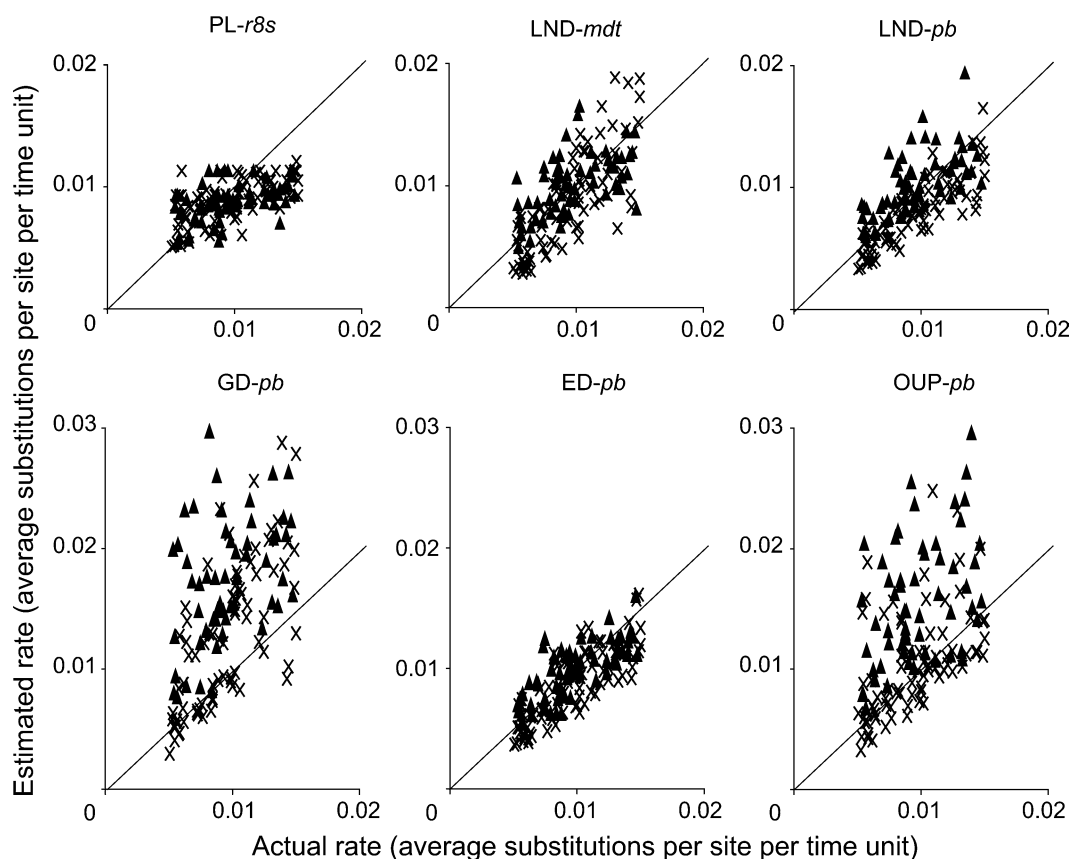


FIG. 5.—Branch-specific rate estimates made from 10 sets of sequence data generated under a model in which rates were independently drawn from a uniform distribution. Figure details follow those of figure 4.

used for generating the data matched that used to analyze the data. Use of the exponential model also yielded rate estimates that were quite accurate. However, rate estimates produced by penalized likelihood were inaccurate when the actual rate was high, due to the effect of the smoothing parameter. Even if rate estimates were calibrated by placing constraints on the age of the root node, rather than on other internal nodes, it is unlikely that the rate estimates would show considerable change because they are already close to the average simulation rate.

Rates were frequently overestimated under the GD model of rate change, with some rates being overestimated by a factor of three, while the performance of the OUP model was slightly better. The rate estimates produced under an ED model of rate change were good, even though the standard deviation is equal to the mean (by definition) and cannot be set to the value used in generating the data. This result is unusual because the true rates are homoscedastic because the variance of the rates depends on the temporal branch lengths and the variance parameter, both of which were held at constant values for all simulations. An additional point of note is that the estimated uncertainty under the ED rate model was smaller than that under either the GD or OUP models but comparable to that under the LND model.

In all the analyses performed using the program PhyBayes, rates in internal branches tended to be overestimated and rates in terminal branches tended to be under-

estimated. Estimates of terminal rates are expected to be more accurate because they are constrained by the calibration points at the internal nodes. Conversely, the age of the root node and rates in internal branches are relatively unconstrained, resulting in rate estimates that are less accurate and less precise. Estimates of rates deep in the tree are probably influenced by the birth-death process prior on the divergence times; a small sampling fraction will lead to short internal branches, which inflates the estimates of rates in these branches (Yang and Rannala 1997). Aris-Brosou and Yang (2002) found that increasing the upper boundary on the prior sampling interval for the sampling fraction to 0.5 caused divergence date estimates to increase by 36%, with a doubling of the associated credibility interval. The large difference between the prior range of the birth rate and the prior range of the extinction rate is also liable to produce trees with short internal branches.

Case 3

When sequences evolved under a model where the rate in each branch was independently drawn from a uniform distribution $U(0.005,0.015)$, few methods performed well (fig. 5). In particular, penalized likelihood produced rate estimates of around 0.01 regardless of the true rate in the branch, such that it overestimated lower rates and underestimated higher rates. Penalized likelihood was not expected to accurately recover the true rates because there was a

considerable violation of its fundamental assumption of minimal rate changes between branches. Use of the GD model also yielded poor rate estimates, with the majority of rates being considerably overestimated, sometimes by a factor of four. Furthermore, many rates were substantially overestimated under the OUP model of rate change, despite the hyperparameters being optimized by comparison of marginal posterior probabilities. This may be a consequence of the constraints on the friction term, a possibility requiring further investigation.

In this case, as with the previous two cases, the OUP model of rate change yielded the highest marginal posterior probabilities of the data given the model (table 1), even when penalties were imposed for the number of hyperparameters involved in each method.

The rate estimates made under an ED model of rate change appeared to be the most accurate of all the methods that were tested. The overall performance of the ED model, even in the analysis of data generated under independent rates, suggests that it should be used for estimating rates and divergence dates from sequences that show considerable departure from clocklike evolution. An additional benefit of using the ED model of rate change is that the variance of the prior distribution does not need to be specified, thus minimizing the number of hyperparameters in the analysis. If a conservative approach to divergence time estimation is desired, then the ED model seems to be a good choice.

Case Study—Elevated Rates in the Cambrian

Studies of rates in branches of the metazoan tree have revealed that molecular rates may have been elevated during the Cambrian period, which would partly or wholly explain the discrepancy between molecular and paleontologic estimates of metazoan divergence dates (Bromham and Hendy 2000; Aris-Brosou and Yang 2002, 2003). It has been suggested that rates of molecular change in the Cambrian may have been at least twice as high as in the periods preceding and following it. However, the consistent overestimation of rates seen in the cases above suggests that the high rate estimates may be a consequence of the method used. In order to examine this possibility, we reanalyzed the 18S metazoan data set used by Aris-Brosou and Yang (2002, 2003) using a simulation-based approach.

Ten sequence alignments were generated using the program Seq-Gen 1.3.1 (Rambaut and Grassly 1997), using model parameters estimated from the 18S alignment: a transition/transversion ratio of 1.73, gamma shape parameter of 0.373, and empirically derived base frequencies (A: 0.264, C: 0.216, G: 0.275, T: 0.245). Ungapped alignments of 40 sequences, each with a length of 1,710 nt, were generated on the metazoan tree used by Aris-Brosou and Yang (2003), using the relative divergence times that they estimated under an ED model of rate change.

The original data set, along with the 10 simulated data sets, was analyzed using PhyBayes with an ED model of rate change, using the same substitution model parameter values as those used to generate the data. The resulting rate estimates followed a pattern similar to that seen in the study by Aris-Brosou and Yang (2003) (fig. 6a), with elevated rates in certain branches, despite the sequence data having

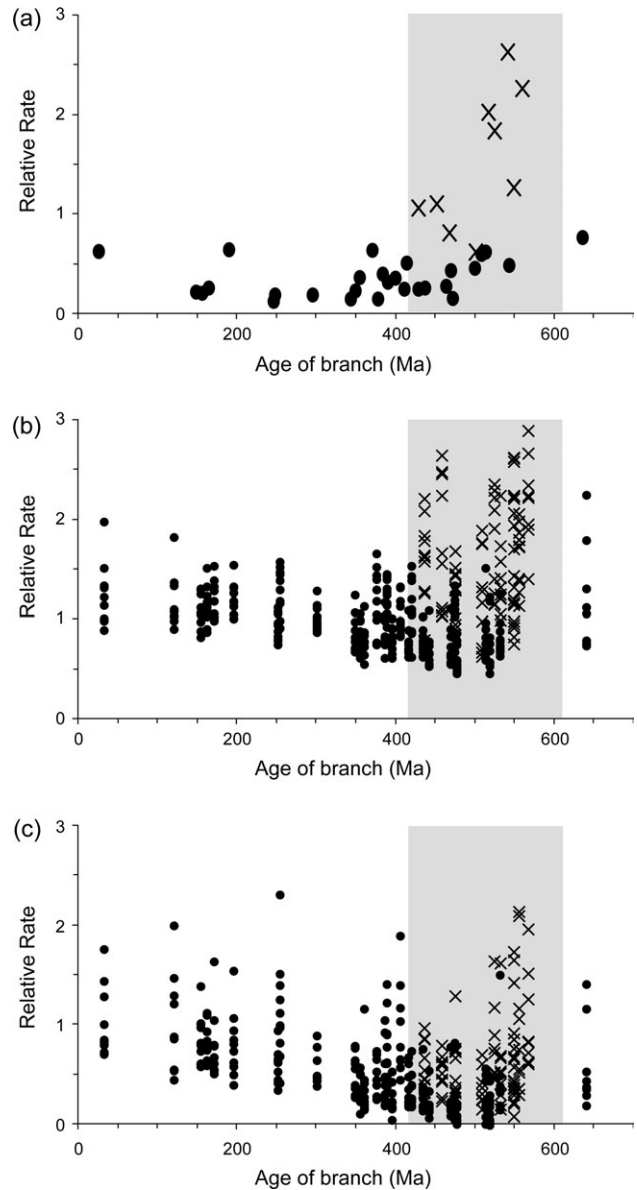


FIG. 6.—Rate estimates made from an alignment of 40 sequences, based on 18S ribosomal RNA (rRNA), under an ED model of rate change using the program PhyBayes. Crosses denote rate estimates for branches for which Aris-Brosou and Yang (2002) inferred elevated substitution rates from 18S data, while filled circles denote rate estimates for other branches. The shaded area marks the time period during which Aris-Brosou and Yang (2003) found the rate to be elevated. (a) Rates were estimated from the 18S data set analyzed by Aris-Brosou and Yang (2002). (b–c) Rates were estimated from 10 replicate data sets, each simulated under a molecular clock constraint on a 40-taxon metazoan tree, with the simulation parameters based on values estimated from 18S rRNA sequences. The hyperparameters for the birth-death prior were (b) birth rate of 15, death rate of 5, and sampling fraction of 0.001 and (c) birth rate of 15, death rate of 15, and sampling fraction of 0.01.

been generated under a molecular clock constraint (fig. 6b). Aris-Brosou and Yang (2003) found that the average relative rate in the period from 640 to 420 MYA was 1.37 per site per MYA, compared with 0.62 in the period from 420 MYA to the present. The values obtained from data simulated under a molecular clock, for these time intervals, were

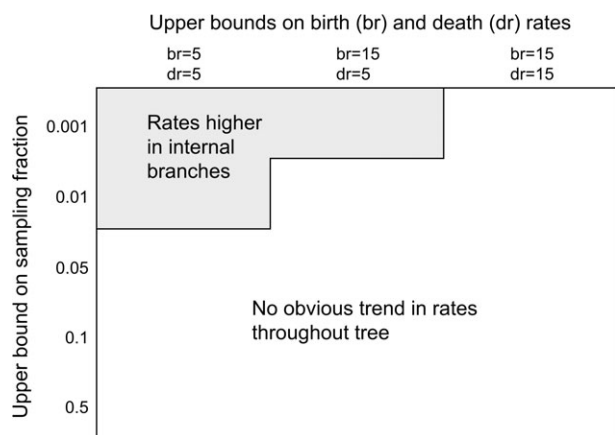


FIG. 7.—Distribution of relative rates when different values are used for the upper bounds on the hyperparameter distributions for the birth-death prior, for the data described in figure 6. Rates were estimated under an ED model of change using the program PhyBayes.

1.24 and 0.87, respectively. This implies that the evidence for elevated rates in the Cambrian was at least partly, if not entirely, an artifact of the estimation method. Rate estimates from nodes more recent than the Cambrian also fluctuated considerably (average standard deviation of 0.62 for the 10 simulated data sets), while rates in terminal branches were estimated with much better precision (average standard deviation of 0.18).

The ranges for the values of hyperparameters governing the birth-death process prior were varied to investigate their effects on rate estimates. Aris-Brosou and Yang (2002) only tested a small number of different prior ranges for the sampling fraction and failed to test for the effect of varying the birth and death rates of the birth-death prior on divergence times. High rates in internal branches were only produced for three of the parameter value sets tested, with the rates being fairly uniform when other sets were tested (figs. 6c and 7). However, the parameter values that produced the most accurate rate estimates were not the ones yielding the highest average (over each group of data sets) marginal probabilities of the data given the model (results not shown).

As seen in the case study, as well as in the simulation studies throughout this paper, popular methods for inferring rates from sequence data tend to overestimate rates in the internal branches close to the root of the tree. This has wide-ranging implications for studies of divergence dates: if the rate is overestimated in internal branches, then the ages of nodes older than the calibration points will be systematically underestimated. In view of this, previously proposed timescales for evolutionary events may need to be reassessed.

Discussion

Rates can be difficult to recover accurately, even when the tree topology and calibration points are accurately known, which is rarely the case. In practice, the inference of rates is even more difficult than is demonstrated in the present study, being further confounded by uncertainties in calibration points, by tree topology, and by asymmetric tree shapes. The distribution of node heights in the true tree can

also affect the accuracy of rate estimates, particularly if they digress markedly from those generated under the prior model.

Other factors that are known to affect phylogenetic analysis, such as rate heterogeneity among sites, compositional heterogeneity, invariant sites, and covariation of mutations at different sites, can also adversely affect the process of rate estimation. It is difficult to predict the impact of these factors because they are generally not well understood even in much simpler cases.

The consistent overestimation of rates in internal branches has important consequences for the estimation of divergence times. Its impact is likely to be dependent on whether rate and date estimates are calibrated using the root or using other internal nodes. In the case of the former, the ages of internal nodes would be overestimated because the overestimation of rates in branches between internal nodes and the root leads to a reduction in the temporal length of these branches. If nodes other than the root are used for calibration, however, then the age of the root might be underestimated. This is particularly relevant for studies of deep divergences, such as the evolution of metazoan phyla.

It is relatively straightforward to test whether data conform to a molecular clock (but see Bromham et al. 2000), even if the data set contains noncontemporaneous sequences (Rambaut 2000). If the clock is rejected, however, there is currently no method available to determine whether rates are sufficiently autocorrelated to warrant the application of relaxed-clock methods. It is unclear whether the currently available models can satisfactorily deal with such data sets by taking high values for the variance parameter.

In the simulation studies described here, the choice of rate prior has a considerable effect on posterior estimates of substitution rates. The effect of the prior was evident in spite of the relatively high levels of variation in the data, suggesting that the prior has a strong influence on the analysis. The present study, however, is based on the analysis of alignments comprising only a small number of relatively short sequences. Further investigation is required to assess whether or not the amount of information about rate variation contained in real sequence data is sufficient to overcome any biases caused by prior rate models.

Our results suggest that careful selection of rate priors is important, especially when the information content of the data set being analyzed is low. However, our results also question the reliability of simplistic (and often inappropriate) model comparison methods, such as using the Akaike Information Criterion, for choosing the model that will most accurately infer molecular rates and divergence dates. It would be important to develop valid formal procedures to compare the models, perhaps incorporating the Bayesian model comparison criteria described by Kass and Raftery (1995) and Spiegelhalter et al. (2002).

One of the major drawbacks of current relaxed-clock methods is their inability to simultaneously infer topology and implement relaxed-clock models. The development of such a method would offer a preferable alternative to the undesirable practice of having to specify a tree in order to use more sophisticated models of rate variation. Moreover, being able to infer trees in a relaxed-clock framework

would represent a significant advance in methods for phylogenetic analysis.

Relaxed-clock models present a biologically plausible and computationally tractable approach to relaxing the assumption of rate constancy among lineages. They are able to perform well under conditions of limited rate variation, but models of rate change must be carefully chosen when the data have evolved under more complex processes of rate change.

Acknowledgments

The authors thank R. A. Fortey for reading a draft of the manuscript. S.Y.W.H. was supported by a Commonwealth (Oxford) Scholarship from the Commonwealth Scholarship Commission and a Domus Research Studentship and Edward Penley Abraham Cephalosporin Scholarship from Linacre College, Oxford. A.J.D. was supported by the Wellcome Trust. A.C. was supported by the Biotechnology and Biological Sciences Research Council and the Wellcome and Leverhulme Trusts.

Literature Cited

- Aris-Brosou, S., and Z. Yang. 2002. Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. *Syst. Biol.* **51**:703–714.
- . 2003. Bayesian models of episodic evolution support a Precambrian explosive diversification of the Metazoa. *Mol. Biol. Evol.* **20**:1947–1954.
- Bromham, L., D. Penny, A. Rambaut, and M. D. Hendy. 2000. The power of relative rates tests depends on the data. *J. Mol. Evol.* **50**:296–301.
- Bromham, L., A. Rambaut, R. Fortey, A. Cooper, and D. Penny. 1998. Testing the Cambrian explosion hypothesis by using a molecular dating technique. *Proc. Natl. Acad. Sci. USA* **95**:12386–12389.
- Bromham, L. D., and M. D. Hendy. 2000. Can fast early rates reconcile molecular dates with the Cambrian explosion? *Proc. R. Soc. Lond. B Biol. Sci.* **267**:1041–1047.
- Cooper, A., and D. Penny. 1997. Mass survival of birds across the Cretaceous-Tertiary boundary: molecular evidence. *Science* **275**:1109–1113.
- Cutler, D. J. 2000. Estimating divergence times in the presence of an overdispersed molecular clock. *Mol. Biol. Evol.* **17**:1647–1660.
- Drummond, A. J., G. K. Nicholls, A. G. Rodrigo, and W. Solomon. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161**:1307–1320.
- Foster, P. G. 2004. Modelling compositional heterogeneity. *Syst. Biol.* **53**:485–495.
- Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**:160–174.
- Huelsenbeck, J. P., B. Larget, and D. L. Swofford. 2000. A compound Poisson process for relaxing the molecular clock. *Genetics* **154**:1879–1892.
- Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**:754–755.
- Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21–132 in H. N. Munro, ed. *Mammalian protein metabolism*. Academic Press, New York.
- Kass, R. E., and A. E. Raftery. 1995. Bayes factors. *J. Am. Stat. Assoc.* **90**:773–795.
- Kishino, H., J. L. Thorne, and W. J. Bruno. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. Evol.* **18**:352–361.
- Perez-Losada, M., J. T. Hoeg, and K. A. Crandall. 2004. Unraveling the evolutionary radiation of the thoracican barnacles using molecular and morphological evidence: a comparison of several divergence time estimation approaches. *Syst. Biol.* **53**:244–264.
- Rambaut, A. 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* **16**:395–399.
- Rambaut, A., and L. Bromham. 1998. Estimating divergence dates from molecular sequences. *Mol. Biol. Evol.* **15**:442–448.
- Rambaut, A., and A. J. Drummond. 2004. *Tracer*. University of Oxford, Oxford.
- Rambaut, A., and N. C. Grassly. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* **13**:235–238.
- Sanderson, M. J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.* **14**:1218–1231.
- . 1998. Estimating rate and time in molecular phylogenies: beyond the molecular clock? Pp. 242–264 in D. E. Soltis, P. S. Soltis, and J. J. Doyle, eds. *Molecular systematics of plants II: DNA sequencing*. Kluwer Academic Publishers, Norwell, Mass.
- . 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol.* **19**:101–109.
- . 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**:301–302.
- Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde. 2002. Bayesian measures of model complexity and fit (with discussion). *J. R. Stat. Soc. Ser. B* **64**:583–639.
- Takezaki, N., A. Rzhetsky, and M. Nei. 1995. Phylogenetic test of the molecular clock and linearized trees. *Mol. Biol. Evol.* **12**:823–833.
- Thorne, J. L., and H. Kishino. 2002. Divergence time and evolutionary rate estimation with multilocus data. *Syst. Biol.* **51**:689–702.
- Thorne, J. L., H. Kishino, and I. S. Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* **15**:1647–1657.
- Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**:555–556.
- Yang, Z., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo Method. *Mol. Biol. Evol.* **14**:717–724.
- Yang, Z., and A. D. Yoder. 2003. Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. *Syst. Biol.* **52**:705–716.
- Yoder, A. D., and Z. H. Yang. 2000. Estimation of primate speciation dates using local molecular clocks. *Mol. Biol. Evol.* **17**:1081–1090.

William Martin, Associate Editor

Accepted February 23, 2005