Syst. Biol. 51(5):689–702, 2002
DOI: 10.1080/10635150290102456

Divergence Time and Evolutionary Rate Estimation with Multilocus Data

JEFFREY L. THORNE[1] AND HIROHISA KISHINO[2]

[1]*Bioinformatics Research Center, Box 7566, North Carolina State University, Raleigh, North Carolina 27695-7566, USA*
[2]*Laboratory of Biometrics, Graduate School of Agriculture and Life Sciences, University of Tokyo, Yayoi 1-1-1, Bunkyo-ku, Tokyo 113-8657, Japan*

Abstract.—Bayesian methods for estimating evolutionary divergence times are extended to multigene data sets, and a technique is described for detecting correlated changes in evolutionary rates among genes. Simulations are employed to explore the effect of multigene data on divergence time estimation, and the methodology is illustrated with a previously published data set representing diverse plant taxa. The fact that evolutionary rates and times are confounded when sequence data are compared is emphasized and the importance of fossil information for disentangling rates and times is stressed. [Markov chain Monte Carlo; Metropolis–Hastings algorithm; molecular clock; phylogeny.]

Because of improved technology, molecular sequence data are becoming increasingly easy to collect. As a result, the pattern and process of evolution are being characterized in ever finer detail. In the past, it was typical to infer evolutionary divergence times by selecting a single gene and then sequencing that gene in the taxa of interest. In the future, these single–gene data sets will give way to multigene data sets. The advantage of multigene over single–gene data sets is clear—bigger data sets contain more information.

How to properly combine evolutionary information from multiple genes is less clear. Because of recombination, the genealogy relating the sequences of one gene can differ from the genealogies of other genes. Hence, divergence times can vary among genes. However, this variation should be negligible when the number of generations between speciation events on a tree is large relative to effective population size.

Even when variation of divergence times among genes can be safely neglected, the question of how to properly extract and combine evolutionary information from multiple genes is not trivial. Evolutionary rates differ over time and among genes. Although variation of rates among genes and over time is problematic for estimating divergence times, understanding this variation is one of the central goals of evolutionary biology. Unfortunately, this variation cannot be directly studied. Comparisons between DNA or protein sequences yield estimates of amounts of molecular evolution, but evolutionary rates and times are confounded when molecular sequences are compared. Without information external to the molecular sequence data, these rates and times cannot be separated.

This confounding of rates and times has hampered the study of evolution. For example, effective population size and substitution rate per generation are two of the most important quantities in population genetics. Because rates and times are confounded when molecular sequence data are compared, population geneticists generally resort to estimating the product of effective population size and substitution rate per generation (e.g., Watterson, 1975) rather than the more desirable option of estimating each of these factors separately.

Similarly, the confounding of rates and times has affected the study of macroevolution by preventing the estimation of absolute rates of evolution and fostering the less desirable alternative of estimating the ratio of evolutionary rates of different genes. For macroevolutionary studies, it has long been recognized that fossil data permit some disentangling of times and absolute rates of evolution (Zuckerkandl and Pauling, 1962). Although serially sampled sequence data can be exploited for quickly evolving organisms (Leitner and Albert, 1999; Drummond and Rodrigo, 2000; Korber et al., 2000; Rambaut, 2000), the usual approach to estimating absolute rates of molecular evolution is to assume both that rates are constant over time and that fossil data allow estimation without error of the time since a common ancestral sequence. Both of these assumptions may be seriously flawed. Fossil data are imperfect and have associated uncertainty. A specific fossil specimen is unlikely to represent the specific organism that harbored the

gene corresponding to an internal node on an evolutionary tree. Likewise, rates of molecular evolution are certainly not constant over time. The pertinent question is not whether rates are constant over time but how much do rates change over time.

Advances in estimating absolute rates of evolution are being made. The uncertainty of fossil and other nonmolecular information can be incorporated via constraints on node times (Sanderson, 1997) or with more detailed treatments (see Huelsenbeck and Rannala, 1997; Waddell et al., 1999), and the assumption of a constant rate of evolution over time can be relaxed (Sanderson, 1997, 2002; Thorne et al., 1998; Huelsenbeck et al., 2000; Drummond et al., 2001; Seo et al., 2002). These recent advances have the potential to allow patterns of rate evolution to be compared over time and among genes and they permit improved estimates of divergence times to be made.

Here, we extend previously proposed Bayesian techniques for estimating divergence times to the analysis of data sets consisting of multiple gene sequences for each taxon of interest. The extensions are also valuable for comparing rates of evolution among genes and over time. We illustrate our Bayesian approach with sequence data representing four genes from diverse plant taxa (Nickrent et al., 2000) and introduce a method for investigating whether evolutionary rates for two genes change in a correlated fashion.

Hierarchical Framework

Our strategy was designed for multigene data sets where all genes can be safely assumed to share a common set of divergence times. This strategy is based on an explicit model for rate evolution that has been described previously in the context of single–gene data sets (Kishino et al., 2001). Its key property is that instead of assuming constant evolutionary rates, rates at different times can vary but will be correlated. With the autocorrelation of rates that is built into this model, homologous genes from closely related lineages are expected to evolve at similar rates, and those from distantly related lineages are likely to evolve at more different rates.

Specifically, the average rate on a branch of a phylogenetic tree is assumed to be the mean of the rate at the nodes that begin and end the branch. Given the rate at the beginning of the branch and the time duration of the branch, the logarithm of the rate at the end of the branch is modeled with a normal distribution. The mean of this normal distribution is such that the rate at the end of the branch (rather than the logarithm of the rate at the end of the branch) has an expected value equal to the rate at the beginning of the branch. The variance of this distribution is equal to the product of the time duration of the branch and a parameter ν that determines the amount of rate autocorrelation over time. A value of zero for ν represents the assumption that rates are constant. As the value of ν rises above zero, the expected difference between the rates at the beginning and ending of a branch increases.

In this way, the joint probability density of the rates at all nodes on the tree is defined for a given value of the rate at the root node. Multiple genes can be incorporated into this framework by having the rate at the root node for each gene be an independent realization from some prior distribution. Given the rate of a gene at the root node and a set of divergence times that is shared by all genes, each gene is modeled as experiencing its own independent rate trajectory on the rooted phylogenetic tree, i.e., the distributions of evolutionary rates among genes are a priori uncorrelated. In our implementation, rates at the root node for individual genes are independent realizations from a gamma distribution.

We have explored two approaches for incorporating the autocorrelation parameter. The more general is to independently assign each gene its own autocorrelation parameter. By having autocorrelation parameters vary among genes, some genes are allowed to be more "clocklike" than others. We do this by having the value of the autocorrelation parameters for different genes be independent realizations from a gamma distribution. This gamma distribution can be interpreted as summarizing the variability of the tendency for rates to change over time among genes in the genome.

A less general approach is to force all genes to share a common value of the autocorrelation parameter. A gamma distribution is specified to serve as the prior for this value. Although a limitation of this less general approach is that it does not build in a tendency for the amount of rate variation to vary

among genes, the fact that all genes are governed by the same autocorrelation parameter value can be an advantage when the selected prior for the autocorrelation parameter is a poor summary of the values that are typical of actual genes. Such a situation is simulated below.

We incorporate fossil and other information external to the molecular data in the form of constraints on node times. Constraints that force a node age to exceed a specific value or that restrict a node age to less than a specific value are both allowed. Our experience (Kishino et al., 2001, unpubl. data) indicates that at least one of each of these kinds of constraints should be incorporated into an analysis to obtain reasonably narrow posterior distributions for node times. The presence of at least one of each kind of constraint is evidently important for allowing evolutionary rates and times to be somewhat decoupled in the posterior distribution.

As described elsewhere (Kishino et al., 2001), convergence of the Markov chain Monte Carlo (MCMC) approach to the posterior distribution is enhanced if one of the branches on the evolutionary tree is forced to have the same rate at its beginning and end rather than having the ending rate be determined by the model of rate evolution described above. In our implementation, a particular branch that emanates from the root is selected. The same branch is selected for each gene. Although the rate that is both at the root node and at the end of this branch is permitted to vary among genes, the rate of a gene at the root node and the rate of the same gene at the end of the selected branch are forced to be identical. Except for this selected branch, the rates at the beginning and ends of branches are governed by our model of rate evolution.

Our prior distribution for divergence times has two components (see Kishino et al., 2001). A gamma distribution describes the time duration separating the ingroup root and the tips. Conditional upon this time duration, the prior density for other internal node times depends on the proportion that is obtained by dividing the time separating these nodes and the tips by the time separating the root and the tips. We refer to these proportions as relative node times. Our implementation employs a generalization of the Dirichlet distribution to rooted tree structures to serve as the prior for these relative node times.

In the absence of fossil constraints or serially sampled data, the prior distribution for divergence times can then be expressed in terms of the gamma distribution for the ingroup root time and the generalized Dirichlet distribution for relative node times. This prior distribution is altered by conditioning upon node time constraints. Although conditioning upon node time constraints makes the analytical expression of the resulting divergence time prior much more complicated, this more complicated prior can be approximated via MCMC analysis (Kishino et al., 2001). Because serially sampled data result in internal node times being constrained, the prior conditional upon sequence sampling times can also be approximated via MCMC analysis. With serially sampled data, we determine a relative node time by calculating the time separating the most recently isolated tip and the node of interest divided by the time separating the most recently isolated tip and the root.

The Metropolis–Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) that we have employed for single gene analysis (Kishino et al., 2001) can be extended in a straightforward fashion to multigene data. To assess convergence of the Markov chain, our main strategy is to perform multiple MCMC runs from different initial states on the same data. We can then determine whether different runs yield similar approximations of the posterior distribution. Because long MCMC runs appear necessary to achieve convergence, computational tractability is attained by approximating the probability density of the sequence data given the branch lengths with a multivariate normal distribution (Thorne et al., 1998).

SIMULATIONS

Via simulation, we have explored the effect on divergence time estimates of the number of genes in the data set. Simulations based on two tree topologies are described here. Both topologies relate 16 ingroup and 1 outgroup taxa. Both have a true ingroup root time of 0.5 time units. Also, the total time represented on the path from the ingroup root back to the common ancestor of all 17 sequences and then forward to the outgroup taxon is 0.625 time units for both trees. With our MCMC implementation, the outgroup only serves to root the ingroup and only the

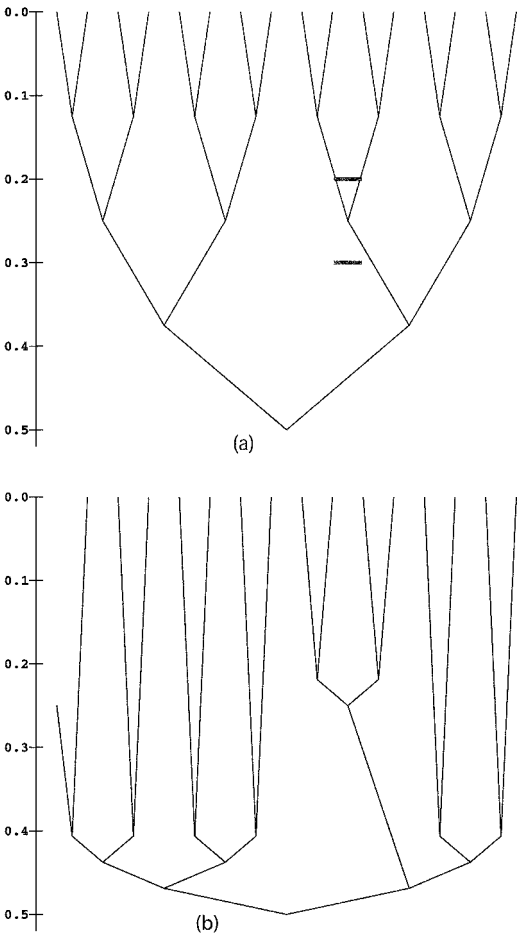

FIGURE 1. Ingroup topologies used in the simulations. (a) Contemporaneously isolated taxa. Constraints are depicted with horizontal lines above and below the node that they constrain. (b) Noncontemporaneously isolated taxa.

times of the ingroup nodes are estimated by our software. The true ingroup node times for the two trees are depicted in Figure 1.

For each gene that is simulated, the rate at the ingroup root node is sampled from a gamma distribution with a mean of 1.0 and an SD of 0.5. The evolutionary rate units are changes per sequence position per time unit. In the analyses, the prior distribution for the rate at the ingroup root node of a gene is in all cases also a gamma distribution with a mean of 1.0 and an SD of 0.5. We matched the gamma distributions of the simulations and analyses because more realistic situations where the true values of parameters are not always characteristic of the priors are unlikely to yield satisfactory results if satisfactory results cannot be obtained even for this ideal case.

With a similar rationale, all genes were evolved according to the Jukes–Cantor model (Jukes and Cantor, 1969), and all genes were analyzed with this nucleotide substitution model. All genes consisted of exactly 1,000 sequence positions. To generate branch lengths for simulating according to the Jukes–Cantor model, the rate at the ingroup root was used to determine the rate at other nodes throughout the tree. For the tree depicted in Figure 1a, two cases were explored. One case was a constant rate of evolution over time, and the other allowed evolutionary rates to change over time. To permit rate change, each gene was assigned its own value of the autocorrelation parameter ν by sampling from a gamma distribution with a mean of 1.0 and an SD of 0.5. The model of rate evolution employed to generate data was that implemented in our MCMC approach except that no branch emanating from the ingroup root node was forced to have identical rates at its beginning and ending nodes. The rate of the tip corresponding to the outgroup taxon was also randomly evolved by treating the path from the ingroup root to the outgroup taxon to be a single branch of length 0.625 time units. For all simulations with the tree depicted in Figure 1b, only a constant rate of evolution was simulated.

Three varieties of MCMC analysis were performed. The first assumed that rates were constant throughout evolution. For the other two, unless otherwise noted, the prior for ν was a gamma distribution with a mean of 1.0 and an SD of 0.5. One of these two varieties assumed all genes shared a common value of ν, and the other assumed that each gene had a separate value of ν. We emphasize that the gamma prior with a mean of 1.0 and an SD of 0.5 was selected to assess the statistical behavior of our method. This gamma distribution has little mass near zero. Because the constant rate hypothesis of $\nu = 0$ is of general biological interest, a prior distribution for ν with more probability mass near zero may often be more worthwhile for analyses of actual data.

All posterior distributions were approximated with the MCMC approach by burning in the Markov chain for 100,000 proposal cycles, where each cycle consisted of a variety

of proposal steps (see Thorne et al., 1998; Kishino et al., 2001). Thereafter, the Markov chain was sampled after every 100 cycles until 10,000 total samples had been stored. Because prior distributions tend to be more diffuse than posterior distributions, the MCMC analyses employed more proposal cycles to approximate the prior distributions. The burn-in period for prior distribution approximations was 1,000,000 proposal cycles, and the Markov chain was then sampled every 1,000 cycles until 10,000 samples had been obtained.

The tree shown in Figure 1a is one of the best possible cases for our divergence time prior because the time intervals between nodes on the tree are evenly spaced. The tree shown in Figure 1b has time intervals between nodes that are not as evenly spaced. Also, one of the 16 ingroup taxa shown in Figure 1b is sampled at an earlier time than are the other 15 ingroup taxa. This lack of contemporaneous sampling produces evolutionary rate information even in the absence of fossil data. Analyses of data generated according to the tree depicted in Figure 1b were not provided with any external information except the relative times at which the 16 ingroup taxa were sampled. In contrast, all analyses of data generated on the tree depicted in Figure 1a rely on the information that one particular node time is constrained to be in the interval from 0.2 to 0.3 time units. The true time of this constrained node is 0.25 time units, and the two constraints placed upon this node time are depicted in Figure 1a.

Parameter Identifiability for Serial and Contemporaneous Sampling

The widths of confidence intervals are generally expected to be halved if the amount of data is quadrupled. This relationship does not apply to credibility intervals generated by Bayesian analyses because these intervals are influenced both by the data and by the prior distribution. However, when the information in the data greatly exceeds the prior information, then the width of a credibility interval might typically be expected to be about halved when the amount of data is quadrupled because the prior typically has little influence when data are abundant. This pattern relating the width of the credibility interval and the amount of data is not observed in row A and row B of Table 1.

TABLE 1. Widths of 95% credibility intervals for the ingroup root time from data sets with different numbers of genes. Entries in the columns for 1, 2, 4, 8, 16, 32, and 64 genes are respectively the medians of 8, 8, 8, 8, 4, 2, and 1 different set of simulated data. For each row, the prior distribution for ν is the distribution from which the true value of ν for each simulated gene is sampled. Row A represents the tree shown in Figure 1a and a constant rate of evolution. Row B also represents the tree shown in Figure 1a, but here the value of ν for each gene is sampled from a gamma distribution with a mean of 1.0 and an SD of 0.5. Row C represents the tree depicted in Figure 1b and a constant rate of evolution.

		No. genes						
	Prior	1	2	4	8	16	32	64
A	1.66	0.221	0.197	0.188	0.177	0.165	0.138	0.114
B	1.71	0.326	0.237	0.216	0.179	0.157	0.137	0.103
C	4.20	0.133	0.114	0.072	0.054	0.033	0.024	0.017

The unusual behavior in Table 1 can be understood by remembering the fundamental confounding of rates and times that arises when sequences are compared. The only information that the MCMC analysis is provided to separate rates and times stems from the prior distribution on rates and times. Especially important for separating rates and times in these analyses is the fact that a node with true time 0.25 is constrained to have a time that is between 0.2 and 0.3 (Fig. 1a). When a constant rate of evolution is both true and assumed, the ratio of the time from the tips to the constrained node relative to the time from the tips to the ingroup root (i.e., the relative time of the constrained node) will be increasingly well estimated as the number of genes employed to estimate this ratio increases. The true value of the relative time of the constrained node is 0.5, but even a perfectly estimated relative time for the constrained node does not lead to exact estimation of the root time. Knowledge that the relative time of the constrained node is exactly 0.5 allows the interval from 0.2 to 0.3 for the actual time of the constrained node to be converted to an interval from 0.4 to 0.6 for the time of the ingroup root, but more accurate estimation of the ingroup root time could only occur through other information in the prior. Without the prior, no amount of sequence data will allow perfect separation of rates and times here. In other words, even an entire genome (or an infinite number of genes) could not overcome the fossil

uncertainty and lead to exact estimation of divergence times here.

This example illustrates the crucial nature of fossil information for estimating divergence times with molecular sequence data. Sequence data cannot surmount uncertainty stemming from the inability of the fossil record to perfectly date an internal node. Too often when dating divergence times, the importance of collecting large molecular data sets is heavily stressed and the collection and summary of fossil information is an afterthought. Channeling extra effort into fossil information may frequently be much more worthwhile for estimating divergence times than is acquiring more sequence data.

A notable exception to the rule that molecular data cannot eliminate divergence time uncertainty without fossil information is the case of serially sampled data. With serially sampled data, the isolation dates of sequences provide the means for calibration of the rate of evolution. Row C of Table 1 enables a contrast between the results for the contemporaneously isolated taxa of the tree shown in Figure 1a and the results for serially sampled taxa of the tree shown in Figure 1b. The width of the 95% credibility interval for the ingroup root time for row C does seem to be roughly halving as the amount of data quadruples. With an infinite number of genes, the width of the credibility interval corresponding to row C should approach zero. Noncontemporaneous sample information is akin to a fossil that could perfectly date one internal node on a tree. For this reason, the asymptotic behavior of divergence time estimation from serially sampled data should be superior to that from contemporaneous taxon sampling with uncertain fossil evidence.

This superiority of serially sampled data over uncertain fossil evidence is predicated on having exact isolation dates for serially sampled data. For viral data, these isolation dates are typically known. In other situations, DNA may be isolated from biological material with an unknown date of origin. In these situations, the uncertainty regarding the date of origin of the material will make the asymptotic behavior of divergence time estimation from serially sampled data qualitatively similar to estimation from contemporaneous sampling with uncertain fossil evidence.

Effects of Prior Specification on Divergence Time Estimates

Table 2 illustrates that performance of the multigene divergence time estimation procedures is relatively good when the analysis assumptions match the process that generated the data. When rates are constant and when constancy is assumed, the posterior means of the divergence time estimates for the ingroup root are close to their true value of 0.5 (see row A and row C, Table 2). Although the true value is outside the 95% credibility for the constant rate analysis of row C, the true and estimated values are still very close for this case.

Row B of Table 2 shows that the constant rate assumption produced a good estimate of the true ingroup root time when rates actually did change over time. Our previous study of divergence time estimates based on single genes revealed that the constant rate

TABLE 2. Posterior means (95% credibility intervals) for ingroup root times with data sets of 64 genes. The true ingroup root time is 0.5. Columns correspond to different analysis assumptions. Rows A and B had true times shown in the tree of Figure 1a. Row A data were simulated with constant rates of evolution, whereas row B data were simulated by independently sampling a value of ν for each gene from the gamma distribution with mean = 1.0 and SD = 0.5. Row C had true times shown in the tree of Figure 1b and data simulated with a constant rate of evolution.

		Individual ν[b]		
	Constant rate[a]	SD = 0.5	SD = 1.0	Shared ν[c]
A	0.527 (0.467, 0.581)	0.388 (0.378, 0.399)	0.431 (0.394, 0.479)	0.520 (0.463, 0.578)
B	0.485 (0.472, 0.510)	0.500 (0.451, 0.554)	0.491 (0.437, 0.550)	0.466 (0.414, 0.529)
C	0.487 (0.479, 0.496)	0.478 (0.468, 0.489)	0.492 (0.483, 0.502)	0.487 (0.479, 0.496)

[a] The prior distribution did not allow rates to change over time.

[b] Each of the 64 genes have a separate value of ν. Entries in the SD = 1.0 column correspond to gamma prior distributions for ν with mean = 1.0 and SD = 1.0.

[c] All 64 genes share a common value of ν. Entries in this column and the SD = 0.5 column correspond to gamma prior distributions for ν with mean = 1.0 and SD = 0.5.

assumption often produced poor estimates of divergence times (Kishino et al., 2001). The constant rate estimate was good here for multilocus data possibly because rate variation was not simulated so as to occur in a correlated fashion among genes. Lineage effects that cause all rates to tend to increase or to decrease on certain branches of the tree may prove more problematic for the constant rate assumption with multigene data. The 95% credibility interval for the ingroup root time when rates actually varied over time (i.e., row B) is more narrow when rates are forced to be constant than when rate variation assumptions match the truth. This coincides with expectations that the constant rate analyses result in divergence time uncertainty being underestimated when rates actually do vary. When rate variation over time is more extreme, the assumption of a constant rate of evolution does not yield good estimates of divergence times even when multiple genes are employed to estimate divergence times (Fig. 2).

Beyond the fact that rates of evolution do change, there is little knowledge of the nature of this rate variation. Consequently, an unrealistic prior distribution for the autocorrelation parameter is a serious possibility. Because posterior distributions are compromises between prior distributions and data, the more general approach of having a separate prior distribution for the autocorrelation parameter of each gene is liable to lead to posterior distributions that are heavily weighted toward the possibly unrealistic prior distributions. In contrast, the posterior distribution for an autocorrelation parameter that is shared among the genes is expected to be more heavily weighted toward the data than toward an unrealistic prior.

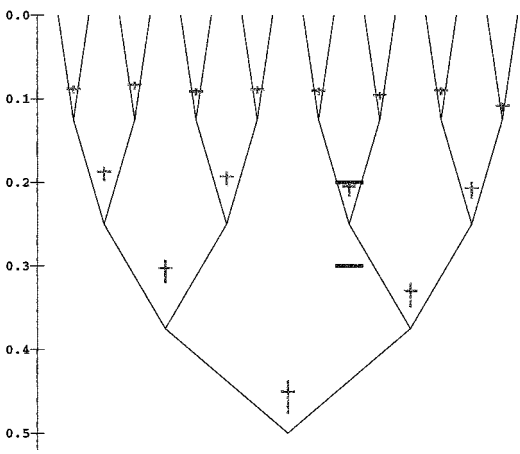

FIGURE 2. Divergence time estimates when rates vary over time but constant rates are assumed. A total of 64 data sets were simulated according to the tree of Figure 1a. Each of the 64 genes had a separate value of ν that was sampled from a gamma distribution with mean = 2 and SD = 2. The MCMC analysis was done with a prior forcing $\nu = 0$ for all genes. Thin lines show the true tree topology and node times. Constraints are depicted with heavy horizontal lines above and below the node that they constrain. Posterior mean estimates of node times are indicated with shaded horizontal lines that are centered above or below the relevant nodes. Shaded vertical lines represent the 95% credibility intervals for node times.

We believe this effect explains the entry seen in row A of Table 2, where the truth was a constant rate but the assumption was that each of the 64 genes possessed its own value of ν and where the prior for ν had a mean of 1.0 and an SD of 0.5. This scenario yielded a narrow 95% credibility interval for the ingroup root time that was centered far from the true value of 0.5. The problem seems to be that each gene had a value of ν with a prior that was mainly concentrated around large amounts of expected change. Because node times are shared by all 64 genes and because less rate variation is expected during short periods of evolution than during long periods of evolution, branch length estimates that do not deviate much (or at all) from a molecular clock can be explained either with a high rate of evolution at the root node and a short time duration of evolution or with the unlikely event (according to the prior) that all 64 genes happened to have values of ν that were close to zero and therefore far from their prior. The posterior mean of the ingroup root time and the narrow credibility interval for this scenario are close to 0.4 because an internal node was constrained to have a time exceeding 0.2, and this constrained node happens to represent half the time from the tips to the root. Without this constraint, the posterior mean for the ingroup root time is likely to have been even smaller. Having a prior for ν with a large standard deviation produced a better but still overly low estimate for the ingroup root time for the case of row A in Table 2. In actual data analyses, little prior information regarding rate change for a gene may exist, and therefore a large standard deviation in the prior for ν may be advisable.

In frequentist statistics, trade-offs between the number of parameters in a model and fit to the data are necessary to avoid overparameterization. The Bayesian framework is not immune to this sort of trade-off. The Bayesian compromise is between the fit of the model and how highly dimensional and diffuse is the prior. For each of the middle two columns in Row A of Table 2, examination reveals that the 64 posterior means for ν were all much less than the prior means (data not shown). The unrealistic nature of the prior distributions for ν could have been detected via their posteriors, and this detection would have pointed to potential problems in divergence time estimates.

Although the absolute times for these entries in row A of Table 2 were poorly estimated, the ratio of the time from a tip to the node of interest and the time from the tip to the ingroup root can also be considered. We refer to such ratios as relative times. In our experience, an unrealistic prior distribution for ν is likely to yield poor posterior estimates of absolute times, but estimates of relative times are more robust (e.g., see Fig. 3). Relative rates can be well estimated even when absolute rates are poorly estimated because of unrealistic prior distributions.

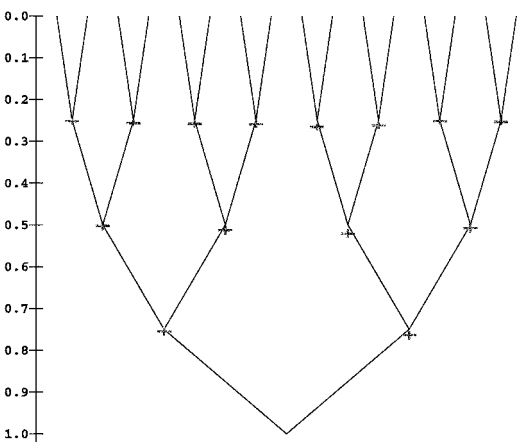

FIGURE 3. Estimated relative times when the truth is a constant rate of evolution but genes have separate ν values with a gamma prior that has mean = 1.0 and SD = 0.5. Relative times for the relevant entry of row A in Table 2 are depicted. Thin lines show the true tree topology and node times. Posterior mean estimates of relative times are indicated with shaded horizontal lines. Shaded vertical lines show the estimated 95% credibility intervals for relative node times.

Detecting Correlated Changes in Evolutionary Rates Among Genes

Estimation of divergence times is not the only application of our model of rate evolution to multigene data sets. Because posterior distributions for evolutionary rates among genes and over time can be approximated, our approach and future modifications can be employed in comparative genomic studies. One important question is how much rate variation can be attributed to locus effects, lineage effects, and locus by lineage interactions (see Muse and Gaut, 1997). The posterior distribution of rates over time and among genes provides a basis for addressing this question. To date, we have mainly concentrated on investigating whether evolutionary rate changes of two genes are positively correlated over time. A positive correlation implies that the portions of an evolutionary tree on which one gene is slowly evolving tend to be the portions on which the other gene is slowly evolving. A reasonable and robust summary statistic for measuring this correlation could be calculated by ranking the posterior means of node rates from low to high for one gene and then determining the correlation between these ranks and the corresponding ranks for the other gene. When calculating this rank correlation statistic, we do not include the ingroup root node because our implementation forces it to have the same rate as one of the nodes that it is connected to by a branch. To evaluate the null hypothesis of uncorrelated rate changes over time among the two genes, the distribution of this rank correlation statistic under this null hypothesis needs to be determined. Because of the autocorrelation of rates in our model, the rate for a gene at one node will be dependent on the rate for the gene at other nodes on the tree. For this reason, conventional tests of association that employ the rank correlation coefficient must be avoided here.

To approximate the distribution of the rank correlation coefficient under the null hypothesis, we exploit the symmetry inherent in our model of rate evolution. Specifically, the model assumes that given the rate at the beginning of a branch, the logarithm of the rate at the end of the branch will have a normal distribution. Because normal distributions are symmetric about their mean, a value exceeding the mean by a certain amount has the

same probability density as a value below the mean by the same amount. Therefore, the difference between the logarithm of the rate at the end of the branch and its expected value has the same probability density as it would if it had the same magnitude but opposite sign. We refer to this difference as the deviation for a branch. We estimate the deviations for branches from the posterior means of the node rates and times. For node i, R_i is the estimated posterior mean rate and T_i is the estimated posterior mean time. The parental node of node i is p_i. The deviation for the branch that ends at node i is then equal to $log(R_i) - E[log(R_i) \mid log(R_{p_i}), T_i, T_{p_i}]$. For simplicity, other parameters that are conditioned upon in the above expectation are omitted.

Our approach relies on the fact that the estimated deviation for a branch and a deviation of the same magnitude but opposite sign are equally likely a priori. We use this fact to generate random assignments of rates to nodes. These random assignments are made in such a way as to be independent among genes but to preserve the autocorrelation of rates at different nodes for the same gene. After each random assignment is made, the rank correlation statistic is calculated. By making random assignments of rates to nodes many times, the distribution of our test statistic under the null hypothesis can be approximated.

The rate assigned by our randomization procedure to node i for a gene of interest is denoted R_i^*. The root node is referred to as node 0, and its rate is R_0. The process for determining the R_i^* values begins at the ingroup root node by setting $R_0^* = R_0$ and proceeds toward the tip nodes.

As mentioned earlier, our implementation analyzes data by constraining one branch that emanates from the ingroup root to have the same rate at its beginning and ending nodes. Likewise, our randomization procedure assigns the same rate to the beginning and ending nodes of this branch. For other nodes, the assigned rate of a node depends on the rate assigned to its parental node and on the deviations estimated from the posterior means of the node rates and times. With probability 0.5, a node is assigned $E[log(R_i^*)|log(R_{p_i}^*), T_i, T_{p_i}]$ plus the estimated deviation for the branch separating the node and its parental node. Otherwise, the node is assigned $E[log(R_i^*)|log(R_{p_i}^*), T_i, T_{p_i}]$ minus the estimated deviation for the branch. This means that Node i ($i \neq 0$) has a rate assigned by

$$\begin{aligned} log(R_i^*) &= E[log(R_i^*)|log(R_{p_i}^*), T_i, T_{p_i}] + H_i(log(R_i) \\ &\quad - E[log(R_i)|log(R_{p_i}), T_i, T_{p_i}]), \end{aligned}$$

where H_i is a random variable that takes the value 1 with probability 0.5 and the value -1 with probability 0.5.

Land Plant Data

The multigene analysis methodology is illustrated with a data set of land plant sequences that was collected and studied by Nickrent et al. (2000). Four genes were included in this data set: 1,351 aligned positions from the *rbc*L locus encoded by the chloroplast, 1,480 alignment positions of 16S ribosomal DNA (rDNA) from the chloroplast genome, 1,720 aligned positions from the 18S rDNA of the nuclear genome, and 1,544 aligned positions from the 19S rDNA of the mitochondrial genome. Thirty taxa, 2 outgroup green algal species and 28 ingroup land plant taxa, were represented in these data. All taxa were present for all genes, with the exception that the *Selaginella* sequence was absent from the 19S rDNA data.

We adopted the tree topology that was reconstructed by Nickrent et al. (2000) by combining the data from the four genes and then applying parsimony. This topology was inferred by ignoring transitions at the third-codon position of the *rbc*L locus. To estimate branch lengths of each gene, a discretized gamma distribution for rate heterogeneity of sites (Yang, 1994) was used in conjunction with the Felsenstein 1984 model of nucleotide substitution (see Felsenstein, 1989). The parameter that determines the amount of rate heterogeneity among sites was separately estimated with the PAML software (Yang, 1997) for each gene.

The prior distribution of divergence times along with the constraints on node times are depicted in Figure 4a. These constraints and times were based on those employed in previous work (Sanderson, 1997). After inspecting the branch length estimates from the four gene data sets, the evolutionary rate at the

root node was given a gamma prior distribution with mean and standard deviation both equal to 0.02 substitutions at the average site per 100 million years. The motivation for choosing this prior was to obtain a distribution for the root rate that was simultaneously reasonable and relatively diffuse. For those analyses where the rate of evolution was allowed to change over time, the autocorrelation parameter was given a gamma prior distribution with a mean of 0.5 and an SD of 0.5. Here, the units of ν represent the accumulated variance per 100 million years in the logarithm of the rate of evolution per 100 million years. This prior for ν was selected rather arbitrarily to be diffuse and to allow for the

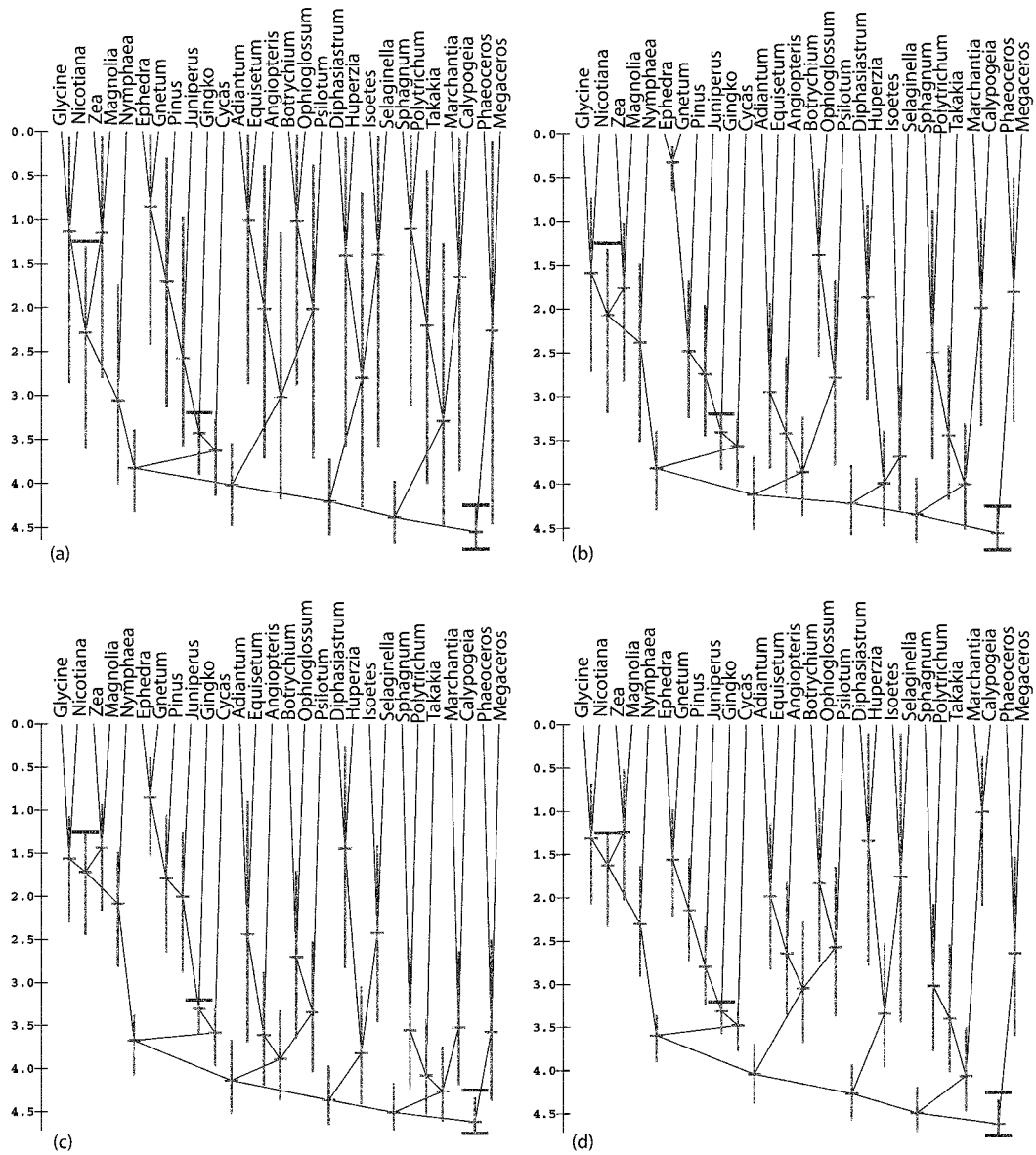

FIGURE 4. Divergence time estimates based on the land plant data. Time units are hundreds of millions of years. Constraints are heavy horizontal lines centered above or below the node that they constrain. Shaded vertical lines passing through a node represent 95% credibility intervals for the node. Posterior mean estimates are indicated with shaded horizontal lines. (a) Prior distribution. (b) 16S rDNA analysis. (c) 18S rDNA analysis. (d) 19S rDNA analysis. (e) *rbc*L analysis. (f) Multigene analysis with constant rate assumption. (g) Multigene analysis with each gene having its own value of ν.

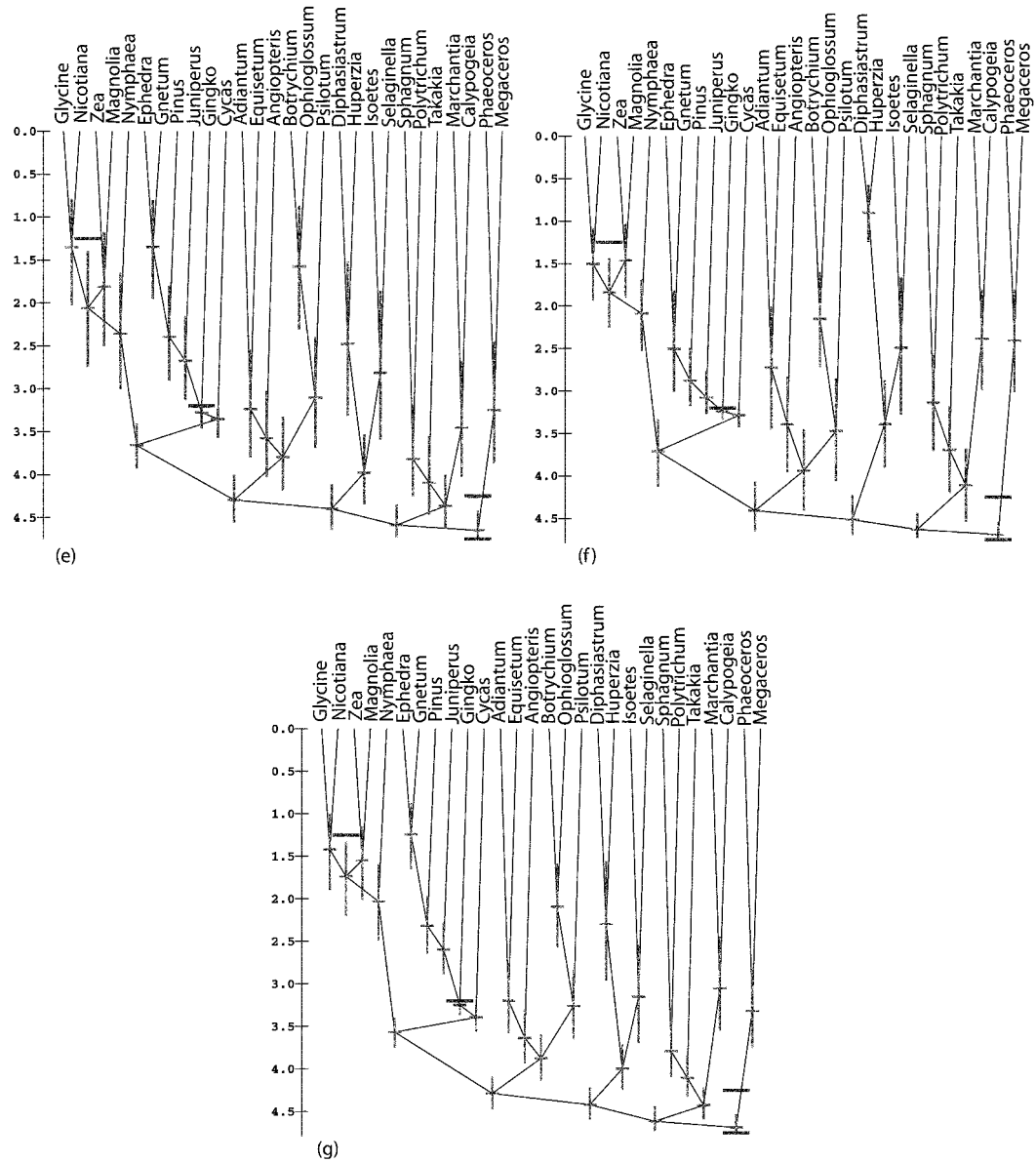

FIGURE 4. (Continued).

possibility that a gene may have a high probability of experiencing a very large amount of rate variation over time.

Divergence times were estimated separately for each of the four single gene data sets and by combining the four genes. Combined analyses were performed with each gene having a separate ν value, with all genes sharing a common value of ν, and with $\nu = 0$ for all genes. The cases where ν was not forced to be zero indicate that these four genes seem to be experiencing a large amount of rate variation over time (Table 3), but there does not seem to be enough information in the data to obtain an accurate estimate of ν.

Overall, the divergence time estimates from single gene analyses when rates were allowed to vary (Figs. 4b–e) were similar for most nodes, but there are some notable exceptions. The same pattern emerges when the divergence time estimates from the multigene analyses are compared for the scenarios with a constant rate (Fig. 4f) and with different ν values for different genes (Fig. 4g). The

TABLE 3. Posterior means and 95% credibility intervals of ν. In each analysis, the prior for ν was a gamma distribution with mean 0.5 and standard deviation 0.5.

	Combined gene	Single gene
Prior	0.48 (0.01, 1.76)	
16S	1.06 (0.54, 1.80)	0.84 (0.45, 1.41)
18S	0.78 (0.40, 1.35)	0.80 (0.36, 1.53)
19S	1.67 (0.96, 2.65)	0.96 (0.46, 1.73)
rbcL	0.82 (0.42, 1.46)	0.86 (0.40, 1.59)
Shared ν	1.23 (0.86, 1.69)	

The "Combined Gene" column represents the case where all 4 genes are analyzed together but individual genes are each given their own autocorrelation parameter. The "Single Gene" column represents the case where the genes are analyzed individually. The row labeled "Prior" shows the prior distribution for ν as estimated by MCMC analysis. The final row shows the situation where all 4 genes are analyzed together but are forced to share the same value of ν.

TABLE 4. Rank correlation coefficients between evolutionary rates of plant genes. Results are from the multigene analysis where each gene has a separate value of ν. The numbers in parentheses are the proportion of times that the observed rank correlation coefficient was equaled or exceeded when the distribution of the correlation statistic was approximated under the null hypothesis of independent rate changes among the two genes. The null distribution of this statistic was approximated by independently sampling 1,000 values with the method described in the text.

	18S	19S	*rbc*L
16S	0.211 (0.790)	0.487 (0.441)	0.625 (0.074)
18S		0.506 (0.246)	0.372 (0.504)
19S			0.530 (0.335)

estimated time since the common ancestor of the liverworts *Marchantia* and *Calypogeia* and the estimated time since the common ancestor of the gymnosperms *Ephedra* and *Gnetum* were especially variable among analyses. As expected, the credibility intervals are more narrow for the multigene analyses than for the single–gene analyses. Divergence times for the case where all genes share the same autocorrelation parameter ν are not depicted here because they are very similar to those in Figure 4g.

Recently, a multigene approach assuming constant rates of evolution has been applied to estimating divergence times for fungi, plants, and animals (Heckman et al., 2001). This analysis yielded divergence time estimates for several nodes that are substantially earlier than previously believed; it produced an estimate of about 700 million years for the time since divergence of vascular plants and mosses. Because our analyses constrained the ingroup root to be no more than 475 million years old, our Bayesian approach could not possibly have estimated an age of 700 million years. As we have stressed, the constraints placed on node times play a central role in what divergence time estimates result. It would be interesting and probably computationally tractable to analyze the data of Heckman et al. with an approach such as ours that does not assume rate constancy and that does not treat fossil information as simply providing calibration points.

Table 4 shows that correlations of rate change are positive for all pairwise comparisons involving the four genes. An attractive future topic would be to investigate whether positive correlations between these genes are mainly due to certain gene regions or are distributed throughout the genes. However, no statistically significant correlations of rate change are detected between pairwise comparisons of the four genes. A substantial amount of rate variation over time was inferred for each of the four genes in this analysis. A consequence of these large amounts of rate variation is that our test statistic tends to be positive even when the null hypothesis of independent rate evolution among genes is true. Because we expect that lineage effects are ubiquitous in molecular evolution, we are predisposed to believe that the lack of significant correlations is attributable to a lack of power of our method to detect these correlations.

DISCUSSION AND FUTURE DIRECTIONS

The method introduced here is designed for multilocus data where all genes share a common set of divergence times. Because differences of divergence times among genes may be substantial for many data sets, alternative methods would be of great interest. For data sets with small numbers of closely related taxa, methods that explicitly incorporate population genetic theory would be promising. For other data sets, statistical approaches that allow divergence times to vary among genes but still have some correlation structure among genes could be explored.

Future implementations of divergence time estimation procedures could explicitly allow for a priori dependence of evolutionary rates among genes. This sort of dependence structure could capture the possibility that different genes might share a tendency to change rates in the same direction on certain

branches because of factors common to all genes (e.g., a change in generation time on these branches). It may also be desirable to have the prior distribution reflect the possibility that certain factors affecting evolutionary rates may be specific only to those genes that are encoded by the same genome (e.g., nucleus, mitochondria, or chloroplast), those genes with products that physically interact, those genes in the same metabolic pathway, or those genes that have similar expression patterns. Although these potential factors responsible for correlated rates can be studied via the posterior distribution even if the prior does not incorporate dependencies, more biologically reasonable priors will undoubtedly lead to more informative analyses.

Systematists may be faced with choosing whether to do a multigene analysis by having a constant rate, by having a single shared value of ν for all genes, or by having individual ν values for each gene. We do not recommend constant rate analyses because they are unable to appropriately represent uncertainty in divergence time estimation. The question of when divergence time estimates obtained through constant rate analyses are biased remains unanswered. Both options for incorporating rate variation warrant exploration, and careful comparison between the prior and posterior distributions can shed light on potential problems in the analysis. It may also be helpful to add another level to our hierarchical model for ν. In addition to the value of ν for each gene and the prior for these ν values, this extra level would be a hyperprior. The hyperprior would reflect uncertainty regarding the appropriate prior distribution for ν values.

Acknowledgments

We thank D. L. Nickrent for kindly providing the sequence data, M. Sanderson for advice on fossil constraints and plant phylogeny, and S. Muse for discussion. This manuscript also benefitted from comments by two anonymous reviewers. J.L.T. was supported by J.S.P.S. grant 13308013, NSF grant INT-990934, and NSF grant DEB-0089745. H.K. was supported by J.S.P.S. grants 12554037, 13308013, and BSAR-497. Software written in the C language that implements these techniques can be obtained from J.L.T. (thorne@statgen.ncsu.edu).

References

Drummond, A., R. Forsberg, and A. G. Rodrigo. 2001. The inference of stepwise changes in substitution rates using serial sequence samples. Mol. Biol. Evol. 18:1365–1371.

Drummond, A., and A. G. Rodrigo. 2000. Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial–sample UPGMA. Mol. Biol. Evol. 17:1807–1815.

Felsenstein, J. 1989. PHYLIP—Phylogeny inference package (version 3.2). Cladistics 5:164–166.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.

Heckman, D. S., D. M. Geiser, B. R. Eidell, R. L. Stauffer, N. L. Kardos, and S. B. Hedges. 2001. Molecular evidence for the early colonization of land by fungi and plants. Science 293:1129–1132.

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 B. Rannala. 1997. Maximum likelihood estimation of phylogeny using stratigraphic data. Paleobiology 23:174–180.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules, Pages 21–132 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, New York.

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.

Korber, B., M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B. H. Hahn, S. Wolinsky, and T. Bhattacharya. 2000. Timing the ancestor of the HIV-1 pandemic strains. Science 288:1789–1796.

Leitner, T., and J. Albert. 1999. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. Proc. Natl. Acad. Sci. USA 96:10752–10757.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equations of state calculations by fast computing machines. J. Chem. Phys. 21:1087–1092.

Muse, S. V., and B. S. Gaut. 1997. Comparing patterns of nucleotide substitution rates among chloroplast loci using the relative ratio test. Genetics 146:393–399.

Nickrent, D. L., C. L. Parkinson, J. D. Palmer, and R. J. Duff. 2000. Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. Mol. Biol. Evol. 17:1885–1895.

Rambaut, A. 2000. Estimating the rate of molecular evolution: Incorporating non–contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics 16:395–399.

Sanderson, M. J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol. Biol. Evol. 14:1218–1232.

Sanderson, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. Mol. Biol. Evol. 19:101–109.

Seo, T.-K., J. L. Thorne, M. Hasegawa, and H. Kishino. 2002. A viral sampling design for estimating evolutionary rates and divergence times. Bioinformatics 18:115–123.

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.

Waddell, P. J., Y. Cao, M. Hasegawa, and D. P. Mindell. 1999. Assessing the Cretaceous superordinal divergence times within birds and placental mammals by using whole mitochondrial protein

sequences and an extended statistical framework. Syst. Biol. 48:119–137.

WATTERSON, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7:256–276.

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. 1997. Phylogenetic analysis by maximum likelihood (PAML), version 2.0. Univ. California, Berkeley.

ZUCKERKANDL, E., AND L. PAULING. 1962. Molecular disease, evolution, and genic heterogeneity. Pages 189–225 *in* Horizons in biochemistry: Albert Szent-Györgyi dedicatory volume (M. Kasha and B. Pullman, eds.). Academic Press, New York.

First submitted 10 October 2001; reviews returned 9 November 2001; final acceptance 1 March 2002
Associate Editor: Rasmus Neilsen