were nonstationary at first positions for the ingroup at $P = 0.01$ using the chi-square test in PAUP*. The chi-square test in PAUP* appears to do a good job of identifying the worst offenders and although crude, is effective for a large data set when analyzing individual genes. We certainly anticipate that as more refined tests of stationarity are implemented (e.g., Foster, 2004), discrimination of genes that deviate from the stationary condition will improve. We note finally that although deviations from stationarity were detected at third positions, we are not suggesting that all misleading base compositional signal is at third positions. Since these nucleotides are linked as codons, we might expect that strong deviations at third positions would influence first and second positions of codons.

Broadly speaking, accurate phylogenetic trees can be recovered from correctly aligned sequences when the inference model is consistent with the process that gave rise to the data. When processes are stationary over lineages and time, relatively straightforward models can be designed to yield accurate inferences, even from short sequences (Steel and Penny, 2000). When processes differ across or within lineages, models must explicitly accommodate the nonstationarity involved. This is generally not straightforward, and even if it could be done, would require many more parameters and associated error terms (but see Foster, 2004). As such, at a given data set size, stationary sequences will prove to be more effective for recovering phylogeny. Stationary sequences will be less prone to the grouping of taxa with convergent base compositions. Of course, when taxa share an atypical base composition in a gene sequence because of shared history, nonstationary sequences may outperform stationary sequences in recovering that branch when using models that assume stationarity. Such instances are cases of obtaining the right answer for the wrong reason (e.g. Swofford et al. 2001) and are a poor argument for use. The criterion of stationarity should prove useful in selecting genes for phylogenetic analysis from completely sequenced genomes, and to the extent that genes that tend to remain stationary can be identified, will be useful for de novo sequencing studies. In general, avoidance of genes with strong deviations from base compositional equilibrium should prove to be a useful strategy for efficient recovery of accurate phylogenetic estimates with markedly fewer genes.

ACKNOWLEDGEMENTS

We thank Thomas Buckley, David Kizirian, Matt Osentoski, Rod Page, Matt Phillips, Tim Rawlings, Barry Williams, and an anonymous reviewer for helpful comments on the manuscript, and Dave Swofford and Mark Holder for discussions about the chi-square test in PAUP*. GN acknowledges the support of NSF grant DEB-0415486.

REFERENCES

Baker, R. H., and R. DeSalle. 1997. Multiple sources of character information and the phylogeny of Hawaiian drosophilids. Syst. Biol. 46:654–673.

Foster, P. G. 2004. Modeling compositional heterogeneity. Syst. Biol. 53: 485–495.

Gee, H. 2003. Ending incongruence. Nature 425:782.

Phillips, M. J., F. Delsuc, and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21:1455–1458.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Saccone, C., G. Pesole, and G. Preparata. 1989. DNA microenvironments and the molecular clock. J. Mol. Evol. 29:407–411.

Steel, M., and D. Penny. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17:839–850.

Swofford, D. L. 2002. PAUP* Phylogenetic analysis using parsimony (*and other methods), version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.

Swofford, D. L., P. J. Waddell, J. P. Huelsenbeck, P. G. Foster, P. O. Lewis, and J. S. Rogers. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. 50:525–539.

Taylor, D. J., and Piel, W. H. 2004. An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data. Mol. Biol. Evol. 21:1534–1537.

First submitted 10 June 2004; reviews returned 31 August 2004; final acceptance 8 December 2004

Associate Editor: Thomas Buckley

Syst. Biol. 54(3):500–507, 2005

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150590947311

Estimating Divergence Times in Phylogenetic Trees Without a Molecular Clock

TOM BRITTON

Department of Mathematics, Stockholm University, SE-106 91 Stockholm, Sweden; E-mail: tom.britton@math.su.se

Rates of evolution often tend to vary between lineages in a phylogenetic tree, implying that the molecular clock assumption is not valid. In this article, we are therefore concerned with estimation of divergence times without assuming a constant molecular clock, where inference is based on DNA (or amino acid or protein) sequences from the species of interest. "Time" could here either be relative time, i.e., all divergence times are relative to the unknown age of the root of the tree, or absolute time if some fossil dating(s) relating the relative times to absolute time are available. Here we focus on relative times, but in either case such a tree is ultrametric and will be denoted the time-tree.

By a molecular clock (denoted "clock" below) we mean that the average (or mean) substitution rate at a given site, and given the present nucleotide, is the same in all parts of the phylogenetic tree, i.e., during the whole evolution and for all species. On the other hand, the average substitution rate is allowed to differ (systematically and/or randomly) between different sites and it may also depend on the present state of the nucleotide without interfering with the clock-assumption. With this definition of a molecular clock the models of, e.g., Jukes and Cantor (1969), Felsenstein and Churchill (1996), Yang and Rannala (1997), and Rogers (2001), all obey the clock-assumption. Tests of the clock-assumption have been derived by several authors (e.g., Langley and Fitch, 1974; Britten, 1986; Li, 1997; Muse, 2000; Britton et al., 2002), and when applied to data, the clock-assumption is almost always rejected.

There are several general approaches for estimating time-trees without assuming a clock (see also Sanderson, 2002). One approach involves pruning taxa that depart from a tree-wide mutation rate (e.g., Takezaki et al., 1995). The local molecular clock method divides the tree into distinct parts assuming a constant rate in each part (e.g., Rambaut and Bromham, 1998). Sanderson (1997) adopts a nonparametric approach that aims at minimizing a certain quadratic function of the rate changes between adjacent edges, thus keeping rate changes small. In Sanderson (2002), he explores a semiparametric approach in which he penalizes a model likelihood according to how much the rates change over the tree. By specifying models for species evolution, substitutions, and rate changes, as well as priors for model parameters, a Bayesian framework can be used and an approximation of the posterior distribution of the time-tree and other parameters of interest can be obtained using Markov chain Monte Carlo methods (e.g., Thorne et al., 1998; Kishino et al., 2001).

The main conclusion from the present article is that the precision in the divergence times estimates cannot become arbitrary high by collecting sufficiently long DNA sequences for a fixed number of species. In other words, without the clock assumption, no method, likelihood based, Bayesian, or other, for estimating the time-tree can be consistent as the sequence lengths are increased. To keep things simple we illustrate our findings on a Jukes-Cantor type model of sequence evolution (Jukes and Cantor, 1969) but where different edges in the tree may have different substitution rates. A commonly used edge- or branch-length unit for this model is the "expected number of substitutions" along the branches. A tree measured in this length unit is from now on denoted the b-tree. Felsenstein (1981), for example, does not assume a clock but concentrates on estimating the b-tree, which can be estimated consistently.

Model, Data, and Notation

Let the data X denote the $k \times n$ matrix of aligned sequences of length n from k species. Let τ denote a rooted binary tree topology of the k species and label the $2k-2$ edges $1, \ldots, 2k-2$. Further, let $\mathbf{t}^{(\tau)} = (t_1^{(\tau)}, t_2^{(\tau)}, \ldots, t_{2k-2}^{(\tau)})$ denote the vector of relative time durations of the edges of the tree, let $\mathbf{r}^{(\tau)}$ denote the corresponding vector of relative substitution rates, and define $\mathbf{b}^{(\tau)}$ by $\mathbf{b}^{(\tau)} = \mathbf{r}^{(\tau)} \cdot \mathbf{t}^{(\tau)} = (r_1^{(\tau)} t_1^{(\tau)}, \ldots, r_{2k-2}^{(\tau)} t_{2k-2}^{(\tau)})$, the vector of expected number of substitutions. The vector $\mathbf{t}^{(\tau)}$ of relative time durations is normalized by defining the aggregated time from the root to the terminals/species to equal 1. (There are several other constraints and, in fact, only $k-2$ "free" time durations. For example, the two time durations from a final speciation to present time must be identical.) The labeling of the edges depends on the specific topology τ, which is shown explicitly. From now on we will let the time-tree be specified by $(\tau, \mathbf{t}^{(\tau)})$, the topology, and the time durations of the edges, and the b-tree by $(\tau, \mathbf{b}^{(\tau)})$, the topology, and the expected number of substitutions along the branches. Neither of these trees are ever observed: even if we observed the substitutions continuously over a set of sites, the number of substitutions on the observed set of sites that occurred along the different edges would make up a randomly perturbed version of the b-tree. The different trees are illustrated in Figure 1 where the last tree is denoted "Observed tree."

We illustrate our analysis with the same model for evolution (i.e., time-tree) and variation of substitution

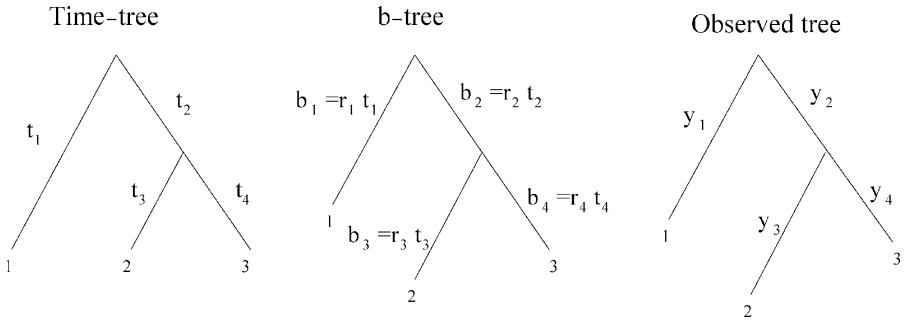

FIGURE 1. The three types of trees. The time-tree, which is ultrametric, is a result of the speciation process. The b-tree has time multiplied by the substitution rates in each lineage and is no longer ultrametric. The observed tree is a random perturbation of the b-tree where the number of substitutions y_i is a random outcome having b_i as its mean.

rates over edges as in Thorne et al. (1998). For sequence evolution on the time-tree, also given the substitution rates, we use the Jukes-Cantor model in our illustration. More specifically, the splitting times of the time-tree are modeled by a Yule process that is evolved up until the time just before the first splitting-time, resulting in one more species than in the data set of interest. In Figure 1, for example, each branch splits into two at constant rate and independently between branches as time evolves, i.e., as one moves downward in the time-tree. In this particular tree, only one split occurred (the next occurred just below the leaves). Given the time-tree, the substitution rates vary between edges, as in Thorne et al. (1998), as follows. The substitution rate r_d of a daughter edge of length t_d with mother edge having rate r_m and time duration t_m is an observation of a random variable R_d with distribution $\ln(R_d) \sim N(\ln(r_m), \nu(t_m + t_d)/2)$, where the parameter ν reflects the degree of correlation between adjacent edges. For example, the logarithm of the rate r_3 in Figure 1 is drawn from a normal distribution with mean $\log(r_2)$ and variance $\nu(t_2 + t_3)/2$. Daughter edges have independent rates conditional on the mother rate. Additional to this, the two rates for the edges stemming from the root have to be defined. Assume that one of them is exponentially distributed, with mean equal to some plausible number (e.g., 0.01 substitutions per site per unit time) and the other lognormal as for the remaining edges, but now relating to the sister edge. Finally, given the time-tree and the substitution rates of all edges, substitutions are modeled using the Jukes-Cantor model along each edge. This means that substitutions occur randomly, independent, and identically distributed between sites, with a constant mean rate that equals the substitution rate of the edge, and each of the remaining nucleotides is equally likely after the substitution. Because sites as well as nucleotides are interchangeable under this model, it is sufficient to keep track of the total number of substitutions along each edge. Further, given the time duration t of an edge and its per-site substitution rate r, the total number of substitutions along the edge will be an outcome of a Poisson random variable with mean nrt (n is the sequence length).

This model for speciation, substitution-rate evolution, and sequence evolution only has one parameter $\nu \geq 0$, a measure of how correlated the substitution rates are: the smaller ν the more correlated are the substitution rates. The case $\nu = 0$ is special: that all variances are 0 implying that all rates will be identical and the clock-assumption is fulfilled. Our results in the next section apply to more general models of species evolution, evolution of substitution rates over edges, and evolution of nucleotide sequences. The crucial assumption for our result to hold true is that the same evolution of (relative) substitution rates apply to the whole sequence meaning that a high substitution rate at a given lineage is reflected in high substitution rates over the whole genome in that lineage. The absolute substitution rate is allowed to vary, randomly and/or systematically, over the genome, but this variation should be the same over the whole phylogenetic tree.

Probability Distribution and Likelihood

We want to make inference about the time-tree implying that we should study the probability distribution of the species sequences X as a function of the rooted tree topology, time durations, and model parameters. This distribution can, at least in principle, be obtained by integrating over all possible substitution rates $\mathbf{r}^{(\tau)}$. If $f(\cdot)$ denotes a generic probability distribution we have

$$\begin{aligned} f(X \mid \tau, \mathbf{t}^{(\tau)}, \nu) &= \int f(X \mid \tau, \mathbf{t}^{(\tau)}, \nu, \mathbf{r}^{(\tau)}) f(\mathbf{r}^{(\tau)} \mid \tau, \mathbf{t}^{(\tau)}, \nu) d\mathbf{r}^{(\tau)} \\ &= \int f(X \mid \tau, \mathbf{r}^{(\tau)} \cdot \mathbf{t}^{(\tau)}) f(\mathbf{r}^{(\tau)} \mid \tau, \mathbf{t}^{(\tau)}, \nu) d\mathbf{r}^{(\tau)}. \end{aligned} \tag{1}$$

In the last row of (1) ν has been removed and $\mathbf{t}^{(\tau)}$ and $\mathbf{r}^{(\tau)}$ have been replaced by the product $\mathbf{r}^{(\tau)} \cdot \mathbf{t}^{(\tau)}$ in the conditional argument for X. This can be done since, by the definition of the model, the distribution of X, given the topology τ, depends only on ν, $\mathbf{r}^{(\tau)}$, and $\mathbf{t}^{(\tau)}$ through the product $\mathbf{r}^{(\tau)} \cdot \mathbf{t}^{(\tau)} = \mathbf{b}^{(\tau)}$, the expected number of substitutions along the different branches.

The two distributions appearing in Equation 1 can be computed numerically implying that also $f(X| \tau, \mathbf{t}^{(\tau)}, \nu)$ can, at least in principle. The factor $f(\mathbf{r}^{(\tau)}| \tau, \mathbf{t}^{(\tau)}, \nu)$ splits up into a product of log-normal densities, one factor for each rate $r_i^{(\tau)}$, and $f(X| \tau, \mathbf{r}^{(\tau)} \cdot \mathbf{t}^{(\tau)})$ can be computed by summing over all possible internal node sequences, the distribution being explicit once the internal sequences (including the root) are also specified.

The likelihood function for the data is simply the probability distribution but viewed as a function of the parameters rather than of the data:

$$\begin{aligned} L_X(\tau, \mathbf{t}^{(\tau)}, \nu) &= f(X \mid \tau, \mathbf{t}^{(\tau)}, \nu) \\ &= \int f(\mathbf{r}^{(\tau)} \mid \tau, \mathbf{t}^{(\tau)}, \nu) f(X \mid \tau, \mathbf{r}^{(\tau)} \cdot \mathbf{t}^{(\tau)}) d\mathbf{r}^{(\tau)}. \end{aligned} \tag{2}$$

This is the function to make (likelihood-based) inference from. For example, the maximum likelihood estimates $(\hat{\tau}, \hat{\mathbf{t}}^{(\hat{\tau})}, \hat{\nu})$ is the set of parameter values that maximize the likelihood function, and these estimates can be obtained by numerically maximizing Equation 2 with respect to τ, $\mathbf{t}^{(\tau)}$ and ν.

RESULTS

The main result of the present article is that relative divergence times cannot be estimated consistently by increasing the analyzed sequence length n from the k species of interest, when the clock-assumption is not valid. More specifically, even if the adopted model describes reality perfectly and the phylogeny τ as well as ν are known (a best case scenario), one cannot estimate the relative divergence times $\mathbf{t}^{(\tau)}$ with arbitrary high

precision by collecting sufficiently long DNA sequences from the species of interest. This negative conclusion is true whatever estimator is used, nonparametric, likelihood based, Bayesian, or other. The result is also true if fossil datings are available for one or several of the internal nodes and divergence times are estimated using absolute time, unless of course there are fossil datings for all internal nodes when there is nothing to estimate.

The conclusion follows from the observation that distribution of the data (X) only depend on the time and rate vectors $\mathbf{t}^{(\tau)}$ and $\mathbf{r}^{(\tau)}$ through their product vector $\mathbf{b}^{(\tau)} = \mathbf{r}^{(\tau)} \cdot \mathbf{t}^{(\tau)}$. As a consequence, only $\mathbf{b}^{(\tau)}$ can be estimated consistently, and not $\mathbf{t}^{(\tau)}$ and $\mathbf{r}^{(\tau)}$ separately. We now explain this fact using Equation 2 by studying what happens with the integral factors to the right in Equation 2 for different $\mathbf{r}^{(\tau)}$ as the sequences get longer (i.e., n increases) for the k species. The first factor, $f(\mathbf{r}^{(\tau)} | \tau, \mathbf{t}^{(\tau)}, \nu)$ is unaffected by n and is continuous in $\mathbf{r}^{(\tau)}$. The second factor, the probability function $f(X| \tau, \mathbf{r}^{(\tau)} \cdot \mathbf{t}^{(\tau)})$ viewed as a function of $\mathbf{r}^{(\tau)}$, becomes more and more peaked as n increases and the maximum is obtained for $\mathbf{r}^{(\tau)}$ such that $\mathbf{r}^{(\tau)} \cdot \mathbf{t}^{(\tau)} = \hat{\mathbf{b}}^{(\tau)}$ where $\hat{\mathbf{b}}^{(\tau)}$ is the maximum likelihood estimator for $\mathbf{b}^{(\tau)}$. More explicitly, the maximum is obtained for $\mathbf{r}^{(\tau)} = \hat{\mathbf{b}}^{(\tau)}/\mathbf{t}^{(\tau)} = (b_1^{(\tau)}/t_1^{(\tau)}, \ldots, b_{2k-2}^{(\tau)}/t_{2k-2}^{(\tau)})$. As n increases the contribution from such $\mathbf{r}^{(\tau)}$ to the integral expression making up the likelihood Equation 2 becomes more and more dominating. It hence follows that

$$L_X(\tau, \mathbf{t}^{(\tau)}, \nu) \approx c_n f_{\mathbf{r}^{(\tau)}}(\hat{\mathbf{b}}^{(\tau)}/\mathbf{t}^{(\tau)} \mid \tau, \mathbf{t}^{(\tau)}, \nu) f(X \mid \tau, \hat{\mathbf{b}}^{(\tau)}), \quad (3)$$

for some constant c_n, where we made it explicit that the first function on the right hand side is the probability density for $\mathbf{r}^{(\tau)}$ evaluated in $\hat{\mathbf{b}}^{(\tau)}/\mathbf{t}^{(\tau)}$. As a function of $\mathbf{t}^{(\tau)}$, the constant c_n as well as $f(X \mid \tau, \hat{\mathbf{b}}^{(\tau)})$ in Equation 3 are constant and hence irrelevant for inference on $\mathbf{t}^{(\tau)}$. The middle factor on the right hand side of Equation 3 clearly depends on $\mathbf{t}^{(\tau)}$, but the maximal value, having argument $\hat{\mathbf{r}}^{(\tau)}$ say, will not go to infinity as n does. From the model for substitution rates, the density value for other arguments $\mathbf{r}^{(\tau)}$, or equivalently $\hat{\mathbf{b}}^{(\tau)}/\mathbf{t}^{(\tau)}$, are comparable in size and will not be negligible as n goes to infinity (see Fig. 3 for an illustration from the simulation study).

A more intuitive explanation to why the divergence times cannot be estimated consistently as the sequence length increases is that each of the $2k - 2$ substitution rates is generated only once: the edge-specific substitution rate for every site along the edge. Consistency, on the other hand, relies on the fact that more and more random quantities are observed, and that the average of these many random quantities becomes less and less random due to the law of large numbers. For example, the *average* number of substitutions among the different sites, along an edge having substitution rate r and time duration t will tend to the constant $b = rt$ even though each such, per-site, number of substitutions is an outcome of a Poisson random variable (with mean b). Consequently the parameter $\mathbf{b}^{(\tau)} = \mathbf{r}^{(\tau)} \cdot \mathbf{t}^{(\tau)}$ is possible to estimate consistently (see Fig. 4 for an illustration). This means that the b-tree can be estimated consistently by collecting longer and longer sequences. However, given the b-tree, data contain no additional information about the time-tree, and the b-tree does not allow the time-tree to be estimated without uncertainty.

How the maximum likelihood (ML) estimator of a specific edge time-length in a phylogenetic tree relates to the true time-length depends on the substitution rates of the edge and the surrounding edges. Typically a low substitution rate of an edge implies that the ML estimate of the time-length of this edge is smaller than its true time-length (cf. next section). However, because a substitution rate can either be large or small, there is no systematic bias in the estimators for the divergence times. This means that the divergence times can be estimated in an approximately unbiased way albeit not consistently.

The conclusion that divergence times cannot be estimated consistently holds true also if a Bayesian viewpoint is adopted. In the Bayesian framework this would be formulated by saying that the posterior distribution of the divergence times does not converge to a point mass at the true divergence times, as longer and longer sequences are collected. In the Bayesian framework, a prior distribution $\pi(\tau, \mathbf{t}^{(\tau)}, \nu)$ for the parameters has to be specified additional to the evolutionary model. The knowledge about the parameters, after the data X has been collected, is then expressed in the posterior distribution $\pi(\tau, \mathbf{t}^{(\tau)}, \nu | \ X)$. Using Bayes' formula the posterior distribution and prior distribution and likelihood are related by

$$\pi(\tau, \mathbf{t}^{(\tau)}, \nu \mid X) \propto \pi(\tau, \mathbf{t}^{(\tau)}, \nu) L_X(\tau, \mathbf{t}^{(\tau)}, \nu). \quad (4)$$

Because the likelihood will not get infinitely peaked around the true divergence times, neither will the posterior distribution. A consequence of this is that the choice of prior distribution will have a big impact irrespective of how long sequences are collected. This is in contrast to the usual situation where the choice of prior becomes less important as more data is collected.

Simulation Study

We now illustrate our results using simulated data from a rooted 3-taxon tree. A rooted 3-taxon tree was chosen in order to keep numerical problems to a minimum. The internal node was chosen to have equal time-length (= 0.5) to the root as to the terminals (see Fig. 2). Substitution edge-rates were generated once according to the model of Thorne et al. (1998) with $\nu = 0.01$ (however, because time is only given in relative terms, the substitution rate r_1 was set to equal 0.01 rather being generated from the exponential distribution). All rates and their distributions are given in Table 1 and the resulting b-tree is shown in Figure 2.

We treat the simpler inference problem where the actual number of substitutions in each site along each edge is observed, which implies that also multiple substitutions are observed, and also that the rooted tree

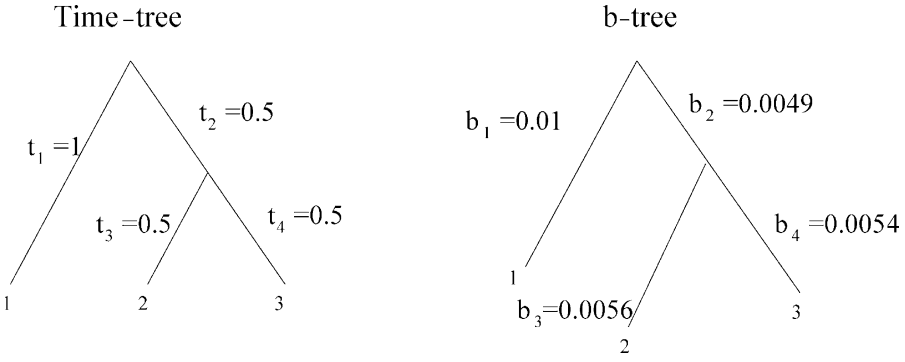

FIGURE 2. The time-tree and b-tree used in the simulation study. The substitution rates are taken from Table 1. For example, $b_2 = r_2 t_2 = 0.0098 \times 0.5 = 0.0049$. The changes between the time-tree and the b-tree are exaggerated in the figure.

topology τ and the variance parameter $\nu = 0.01$ in the rate variation model are assumed known. Under these assumptions, the number of substitutions from different sites can be aggregated without loss of information, so if we let $\mathbf{y} = (y_1, \ldots, y_4)$ denote the total number of substitutions along the four different edges, the probability distribution is given by

$$f(\mathbf{y} \mid \tau, \mathbf{r}^{(\tau)} \cdot \mathbf{t}^{(\tau)}) = \prod_{i=1}^{4} \frac{(nr_i t_i)^{y_i} e^{-nr_i t_i}}{y_i} = \prod_{i=1}^{4} \frac{(nb_i)^{y_i} e^{-nb_i}}{y_i} \quad (5)$$

a product of Poisson probabilities. The likelihood for this type of data is given by

$$L_{\mathbf{y}}(\tau, \mathbf{t}^{(\tau)}, \nu) = \int f(\mathbf{r}^{(\tau)} \mid \tau, \mathbf{t}^{(\tau)}, \nu) f(\mathbf{y} \mid \tau, \mathbf{r}^{(\tau)} \cdot \mathbf{t}^{(\tau)}) d\mathbf{r}^{(\tau)}. \quad (6)$$

Because τ and ν are assumed known and the time from the root to the leaves is defined to equal 1, there is only one remaining parameter: the time t_2 from the root to the internal node. This is true since $t_1 = 1$ and $t_3 = 1 - t_2 = t_4$ (see Fig. 2). The likelihood in Equation 5 hence only depends on the parameter t_2 so we drop the 2-index and write $L_{\mathbf{y}}(t)$.

"Data" was generated for three different sequence lengths: $n = 1000$, $n = 10{,}000$, and $n = 100{,}000$. For each n the data $(y_1, \ldots, y_4)$, the total number of substitutions along each edge, was simply set to equal the corresponding expected values, rounded to the nearest integer. So for example edge 2, with time length $t_2 = 0.5$ and substitution rate $r_2 = 0.0098$, will for $n = 10{,}000$ have $y_2 = nb_2 = nr_2 t_2 = 49$ observed substitutions (within the 10,000 sites) as input data.

For each of the three different sequence lengths, $L_{\mathbf{y}}(t)$ was computed for $t = 0.01,\ 0.02, \ldots,\ 0.99$. In each such grid point t the likelihood was computed numerically by Monte Carlo simulation. This was done by generating 10,000 independent rate vectors $\mathbf{r} = (r_1, \ldots, r_4)$ according to the model and given the time-tree, and for each such vector the probability function $f(y_1, \ldots, y_4 \mid \mathbf{r} \cdot \mathbf{t})$ was computed for the observed data (we associate $t = t_2$ with $\mathbf{t} = (1, t, 1 - t, 1 - t)$). Taking the mean of these probability functions gives a good approximation of

$$L_{\mathbf{y}}(t) = f(y_1, \ldots, y_4 \mid t) = \int f(y_1, \ldots, y_4 \mid \mathbf{r} \cdot \mathbf{t}) f(\mathbf{r} \mid \mathbf{t}) d\mathbf{r}.$$

In Figure 3, the likelihood plots are shown for the sequence lenghts $n = 1000$, $n = 10{,}000$, and $n = 100{,}000$ (in all figures the $\mathbf{y}$-index is dropped in the likelihood functions). We have also plotted the likelihood function for $n = \infty$ where we used Equation 3, which is an equality in the limit. Because the first three figures are obtained using Monte Carlo simulations, they are plotted using dashed lines as opposed to the exact likelihood of the last plot. Simulations and figures were obtained using Matlab version 6.

In Figure 3 it is seen that, as the sequence length n increases, the likelihood gets more peaked to start off but that this concentration then stops. Even for $n = \infty$ the likelihood is not negligible for values of t in the range (0.42, 0.52) say. This illustrates that the divergence time t $(= t_2)$ cannot be estimated consistently. The maximum likelihood estimate is $\hat{t} \approx 0.47$ for all data sets (i.e., sequence lengths). Note that this value differs from the true value $t = 0.5$. The reason for this difference is that, by chance, the corresponding substitution rate $r_2 = 0.0098$ was relatively small compared to the other substitution rates. This makes the corresponding branch length $b_2 = r_2 t_2 = 0.0049$ relatively smaller (compare the edges in the time-tree and the b-tree in Fig. 2). And, having a small branch length b implies that the estimated t-value will tend to be smaller than its true value.

As a comparison, we also show plots of the likelihood $L_{\mathbf{y}}(b)$ for the corresponding branch length $b = b_2$ for the same data sets (see Fig. 4). If these plots are compared with the plots of $L_{\mathbf{y}}(t)$ in Figure 3 it is seen that $L_{\mathbf{y}}(b)$ concentrates at a higher rate as n increases, and also that,

TABLE 1. Substitution rates used in simulation study. The t_is are taken from the time-tree of Figure 2, and $\nu = 0.01$.

Edge	Rate distribution	Obtained numerical value
1	$R_1 \sim \text{Exp}(100)$	$r_1 = 0.0100$
2	$\log(R_2) \sim N(\log(r_1), \nu(t_1 + t_2)/2)$	$r_2 = 0.0098$
3	$\log(R_3) \sim N(\log(r_2), \nu(t_2 + t_3)/2)$	$r_3 = 0.0120$
4	$\log(R_4) \sim N(\log(r_2),\ \nu(t_2 + t_4)/2)$	$r_4 = 0.0108$

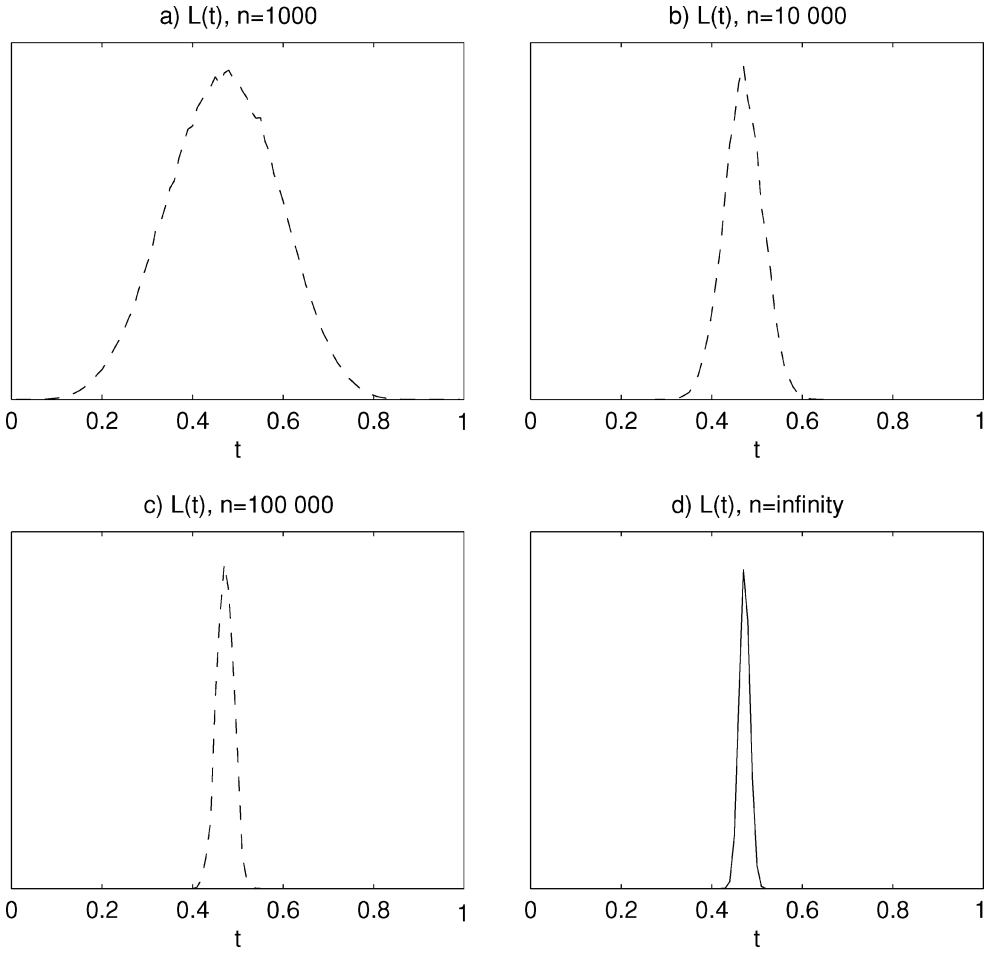

FIGURE 3. Likelihood plots of t for the simulation example, for different values of the sequence length n. The first three plots are obtained using Monte Carlo simulations and the last plot is plotted with the limiting likelihood. It is seen that information about the relative time t increases with n but the amount of information is limited making consistent estimation impossible.

in the limit as n tends to infinity, all mass concentrates at the true value $b_2 = 0.0049$. This illustrates that the branch lengths **b** can be estimated consistently, whereas the relative times **t** cannot.

We stress that the substitution rates $(r_1, \ldots, r_4)$ are only generated once from the model. If a new set of substitution rates were generated we would get a different b-tree. The likelihood for $t = t_2$ would then look somewhat different, but it would still have non-negligible likelihood values for a range of t-values as the sequence length n grew large. The likelihood for $b = b_2$, on the other hand, would just like before tend to a point mass, but now around the new true value of $b_2 = r_2 t_2 = 0.5 r_2$.

DISCUSSION

This article shows that it is impossible to estimate the relative divergence times of a phylogenetic tree consistently, by taking longer and longer sequences for a fixed set of species, without assuming a constant molecular clock. The treated model contains several unrealistic simplifications. First, the Jukes-Cantor type substitution model is oversimplistic. However, the same qualitative result would still hold if a more general substitution rate model was used. Secondly, we assume the same substitution rate for each site. Relaxing this assumption to let the magnitude of the substitution rates vary over the sequence in a systematic and/or random way would not alter the result either. The crucial assumption for the result to remain true is that the (relative) evolution of the substitution rates over the tree is the same for different sites. We focus on maximum likelihood estimation of divergence times. However, it follows that no estimator for the divergence times can be consistent by increasing sequence length only. This holds also when adopting the Bayesian framework in which the posterior distribution reflects the uncertainty of the divergence times. Given that the prior distribution is correct—and its impact is non-negligible even when long sequences are collected—the posterior distribution summarizes the information about the divergence times correctly. However, the distribution does not converge to a point mass at the true set of divergence times, as one would hope, when longer sequences are collected.

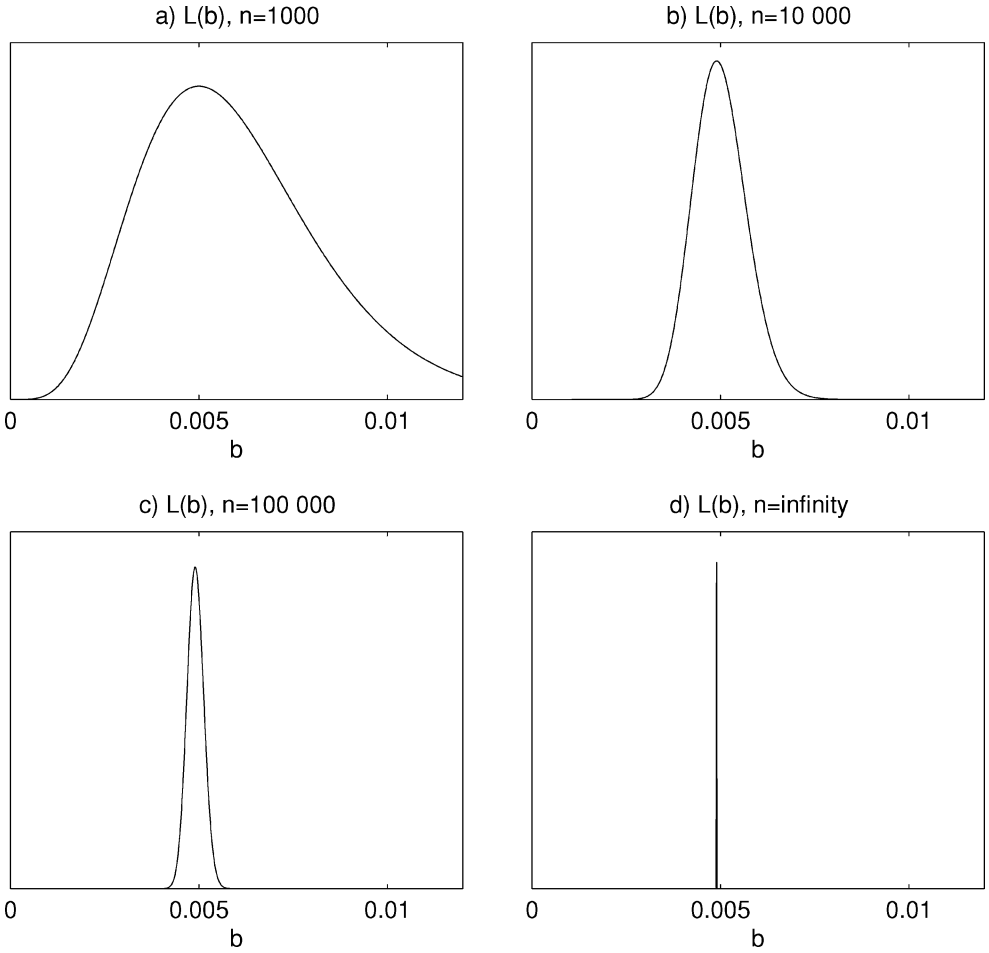

FIGURE 4. Likelihood plots of b for the simulation example, for different values of the sequence length n. The amount of information increases to infinity with n making consistent estimation possible.

Although difficulties in estimating divergence times when substitution rates vary over the tree have been reported elsewhere (e.g., Sanderson, personal communication, and Thorne and Kishino, 2002) our result seems not to have been shown previously. Two natural questions arise from this rather negative result: (1). Can estimation still be worthwhile under the present situation? and (2). Are there other models or limiting scenarios when divergence times can be estimated consistently even if the clock-assumption is not valid? Fortunately, the answer is yes to both of these questions as we now explain.

We start with the first question. For a specific problem there is usually a fixed amount of data available on which to base inference, and then the asymptotic scenario is less important. In particular, data do contain information about the divergence times, and uncertainty measures, such as confidence intervals or Bayesian credibility intervals, of the estimated divergence times can describe this amount of information correctly. In other words, the divergence times are not unidentifiable (see, for example, Rannala, 2002, for a discussion on unidentifiable parameters in overparametrized models). The lack of consistency means that the amount of information is limited even as the sequence length increases (see Fig. 3), the reason being that divergence times are partially confounded with the substitution rates. The fact that data contain information about the divergence times also implies that the clock-assumption can be tested, and the power of this test can be made arbitrarily high by collecting sufficiently long sequences (see Langley and Fitch, 1974, and other references mentioned in the introduction for tests on the clock-assumption).

We now move to the second question concerning other models and/or limiting situations enabling consistent estimation without relying on the clock-assumption. Let us first look at alternative models for which consistent estimation is feasible. Recall that the reason for not obtaining consistency was that the (random) substitution rates were only generated once for each edge in the tree. If instead the variation of substitution rates over lineages is believed to differ for different groups of sites, for example, between genes, then this should make consistent estimation of the divergence times feasible. For example, if the variation of substitution rates over edges for a specific gene is modeled as in the present model, but assuming

that substitution rate variation between different genes are completely independent, then it is possible to estimate divergence times consistently by collecting DNA sequences from more and more genes (cf. Thorne and Kishino, 2002, who also consider uncertainty in fossil information). The same result should hold true even if there is some correlation between substitution rates of different genes, with the effect that a high substitution rate for one gene of a specific lineage makes high substitution rates for other genes on the same lineage somewhat more likely. Modeling such correlation can be done in different ways (see for example, Thorne and Kishino [2002]), and how correlated substitution rates can be, while still allowing consistent estimation, remains an open problem.

We now return to our original class of models for which the (relative) rate variation over edges was the same for all sites. Even under this class of models there is hope for consistent estimation of the divergence times, but under different asymptotic situations. In particular, if the number of fossil dates is increased this will improve precision in divergence time estimates if used correctly. (How to use several fossil datings, and to admit for uncertainty in the dating, when estimating divergence times is in itself important problems not treated in the present article (see, for example, Sanderson [1997] and Thorne and Kishino [2002]). However, this can on its own only lead to consistent estimation if fossil dates are available for *all* interior nodes of the tree, but then the inference problem is trivial. The more realistic and promising situation is where more and more fossil dates are collected, but in a tree containing more and more taxa. In other words, even if one is primarily interested in a given set of taxa, it can be worthwhile to collect longer sequences from these taxa but also from closely related taxa, especially if also the number of fossil dates increases. Exactly what the criteria are to allow for consistent estimation, and what are the optimal rates at which taxa, sequence length, and fossil dates should grow at, remain open and important questions. The underlying explanation why this may lead to consistent estimation is that, when the number of taxa increase, there will still not be complete information about each substitution rate in the tree, but the separate edge lengths will become shorter by including more and more closely related species, so the influence of each substitution rate will decrease, enabling consistent estimation of a sequence of edges in the larger tree, corresponding to one edge in the original tree of interest.

ACKNOWLEDGEMENTS

I want to thank Kåre Bremer for interesting discussions on phylogenetic inference and Michael Sanderson for posing the question of consistency. Financial support from the Swedish Research Council is gratefully acknowledged. Many helpful suggestions from the associate editor and two referees improved the presentation of the article considerably.

REFERENCES

Britten, R. J. 1986. Rates of DNA sequence evolution differ between taxonomic groups. Science 231:1393–1398.

Britton, T., B. Oxelman, A. Vinnersten, and K. Bremer. 2002. Phylogenetic dating with confidence intervals using mean pathlengths. Mol. Phyl. Evol. 2:58–65.

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

Felsenstein, J., and G. A. Churchill. 1996. A hidden Markov approach to variation among sites in rate of evolution. Mol. Biol. Evol. 13:93–104.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages. 21–32 *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.

Langley, C. H., and W. M. Fitch. 1974. An examination of the consistency of the rate of molecular evolution. J. Mol. Evol. 3:161–177.

Li, W.-H. 1997. Molecular evolution. Sinauer, Sunderland, Massachusetts.

Muse, S. V. 2000. Examining rates and patterns of nucleotide substitution in plants. Plant Mol. Biol. 42:25–43.

Rambaut, A., and L. Bromham. 1998. Estimating divergence data from molecular sequences. Mol. Biol. Evol. 15:442–448.

Rannala, B. 2002. Identifiability of parameters in MCMC Bayesian inference of phyologeny. Syst. Biol. 51:754–760.

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

Sanderson, M. J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol. Biol. Evol. 14:1218–1231.

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

Takezaki, N., A. Rhetsky, 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 melocular evolution. Mol. Biol. Evol. 15:1647–1657.

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

First submitted 27 April 2004; reviews returned 7 July 2004; final acceptance 23 November 2004
Associate Editor: Thomas Buckley