Mathematical Statistics
Stockholm University

Divergence times in phylogenetic trees without a molecular clock cannot be estimated consistently

Tom Britton

Research Report 2003:16

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2003:16**,
http://www.math.su.se/matstat

Divergence times in phylogenetic trees without a molecular clock cannot be estimated consistently

Tom Britton[1]

October 2003

Abstract

It is generally accepted that rates of evolution often vary between lineages in a phylogenetic tree, or, equivalently, that the molecular clock assumption is not valid. The present paper is concerned with estimation methods for relative divergence times without assuming a molecular clock, where inference is based on DNA-sequences from the terminals of interest. Several methods, parametric and non-parametric, have been proposed for this estimation problem. In the present paper we show that consistent estimation of the divergence times is impossible, for the class of evolutionary models considered, in the sense that no set of estimators can surely converge to the true set of divergence times as the length of the DNA-sequence increases. The paper also discusses why estimation of divergence times could still be worthwhile, and what alternative types of data or model that may allow consistent estimation.

Keywords: Phylogeny, molecular clock, divergence times, estimation, consistency

[1]Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: tomb@math.su.se.

Introduction

The present paper is concerned with estimation of divergence times in phylogenetic trees, where estimation is based on aligned DNA (or amino acid or protein) sequences of 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 else real time. In the latter case some fossil dating relating the relative times to calendar time must also be available. Deriving divergence times estimates of a phylogenetic tree from sequence data involve several consecutive steps. The first step is that of aligning the sequences – a problem not treated in the present paper: all sequences are assumed aligned without gaps. A second step often involves estimating a rooted tree topology, sometimes also specifying the sequences of internal nodes, from the aligned terminal sequences. This is done using some explicit or implicit (parametric or nonparametric) assumption regarding the evolution of species and of evolution of sequences along a species tree. The last step is then to convert this rooted phylogenetic tree into a tree where the length unit of the edges are proportional to time making the tree ultrametric. The present paper is mainly concerned with the last step, i.e. to study methods for estimating the phylogenetic tree having edge-lengths proportional to time, from now on called the time-tree. When fossil datings are available, the last step also includes calibrating the divergence time estimates to calendar time.

Zuckerlandl and Pauling (1965) were the first to state the assumption of a molecular clock (MC), i.e. that substitution rates remain constant for different lineages in the tree, and also over time. A simple model for the evolution of sequences that obeys the MC-assumption is the model of Jukes and Cantor (1969). This model assumes that different sites in the genome evolve independently, that the per-site substitution rate is constant (independent of time and the present nucleotide, and identical between different sites) and that all nucleotides are equally likely once a substitution occurs. There are many more sophisticated models still obeying a molecular clock but allowing for rate differences between sites (both random and systematic), different substitution rates for different nucleotides, and unequal substitution probabilities once a mutation occurs. For example, Yang and Rannala (1997) work under a Bayesian setup where they, beside having a general substitution model, independent and identically distributed between sites, also model the speciation process. Felsenstein and Churchill (1996) use a hidden Markov model approach to allow rates to vary between sites in an unobserved fashion, whereas Rogers (2001) applies the general time reversible substitution model and allows the relative rate between sites to vary randomly according to independent identically distributed random variables. All these extensions of the simpler model of Jukes-Cantor are within the framework of a constant molecular clock (MC) assumption.

Langley and Fitch (1974) consider a Jukes-Cantor type model but allowing the clock-rate to differ between proteins. They assume that the tree topology is known and that data consist of the number of substitutions along each edge of the tree for the analysed proteins. For this model they derive maximum likelihood estimates for the relative divergence times. In the same paper they also derive a test of the MC-

assumption which was rejected for data on vertebrate evolution of four proteins. The MC-assumption has since then been tested on various types of data using many different methods, and most of the time the MC-assumption has been rejected (e.g. Britten, 1986; Li, 1997; Muse, 2000, Britton *et al.*, 2002).

The present paper is concerned with estimation methods for relative divergence times without assuming the MC-assumption, meaning that the substitution rate is allowed to vary over different lineages of the tree and perhaps also over time. To keep things simple we illustrate our findings on a fairly simple model where the substitution rate is constant, independent for different sites, and where the different nucleotides are equally likely after substitution. When this is the case a commonly used branch-length unit 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 molecular clock but concentrates on estimating the b-tree. A simplifying consequence of assuming a molecular clock is that the b-tree and the time-tree coincide up to proportionality, which implies that estimating the time-tree is equivalent to estimating the b-tree.

There are several general approaches for estimating time-trees without assuming a molecular clock (see also Sanderson, 2002). One approach involves pruning taxa that depart from a tree-wide mutation rate (e.g. Takezaki, Rhetsky and Nei, 1995). Another one, the local molecular clock method, is to divide the tree into distinct parts assuming a constant rate in each part (e.g. Rambaut and Bromham, 1998). Sanderson (1997) adopts a non-parametric approach which aims at minimizing a certain quadratic function of the rate changes between adjacent edges. Another parametric approach is to use the Bayesian framework. By specifying models for species evolution, substitutions and rate changes, as well as parameter priors, it is possible to obtain the (approximate) posterior distribution of the time-tree and other parameters of interest by using Markov chain Monte Carlo methods (e.g. Thorne, Kishino and Painter, 1998; Kishino, Thorne and Bruno, 2001). A semi-parametric approach has been suggested by Sanderson (2002) in which he penalizes a model likelihood according to how much the rates change over the tree. Most methods adopt several simplifying assumptions, for example that the rooted tree topology is known, that rates do not change over sites and sometimes that the occurred number of substitutions along internal edges are observed.

Methods

Model and notation

Let X be the $k \times n$ matrix of aligned sequences of length n from k terminals. Let τ denote a rooted binary tree topology of the k terminals which hence has $2k-2$ edges. 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 normed by defining the aggregated time from the root to the terminals (leaves) to equal 1.

The labeling of the edges depend 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 were observing the evolution 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'.

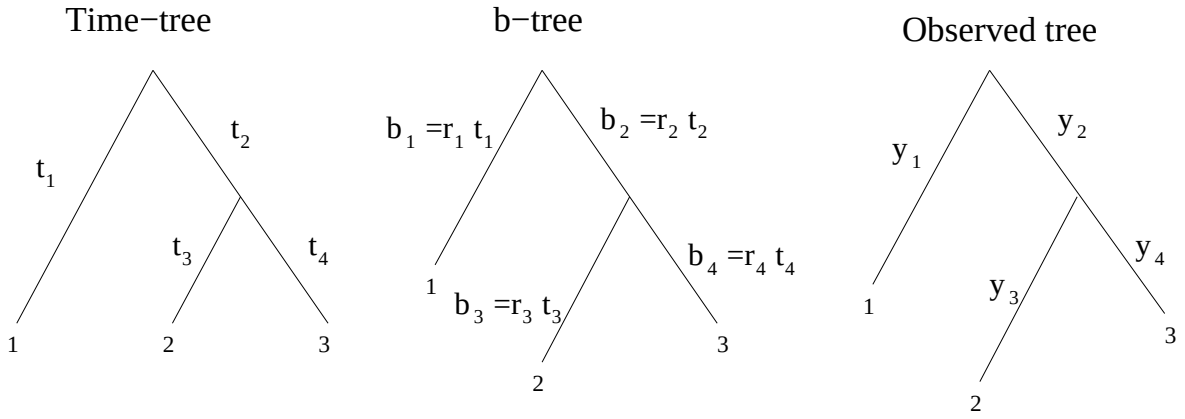

Figure 1: The three types of trees. The time-tree, which is ultrametric, is a result of the speciation process. The b-tree has time multiplied by the substitution rates in each lineage and is no longer ultrametric. The observed tree is a random perturbation of the b-tree where the number of substitutions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

We illustrate our analysis with the model of Thorne, Kishino and Painter (1998) for species evolution (the time-tree) and variation of substitution rates over the edges. For sequence evolution on the time-tree, also given the substitution rates, we use the Jukes-Cantor model in our illustration. More specifically, the evolution of species, i.e. the time-tree, is modelled by a binary-splitting branching process with constant splitting rate, which is evolved up until the time just before the first splitting-time resulting in one more terminal 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. Given the time-tree the substitution rates vary between edges according to a Markov model as follows. The logarithm of the rate r_d of a daughter edge of length t_d with mother edge having rate r_m and time duration t_m is assumed to be normally distributed with mean equal to the logarithm of r_m and variance $\nu(t_m + t_d)/2$. 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 modelled using the Jukes-Cantor model which means that substitutions occur randomly, independent and identically distributed between sites, with a constant mean rate which is independent of the present nucleotide, and all remaining nucleotides are equally likely once a substitution has occurred. Since 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. Then all variances are 0 implying that all rates will be identical. The case $\nu = 0$ is hence equivalent to saying that the MC-assumption is fulfilled. That there is only one parameter is true because k, the number of terminals, is given and the time from the root to the terminals is defined to equal unity.

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. In general θ could be a vector of parameters for the model of species evolution and the model for sequence evolution (θ could for example contain parameters specifying the substitution rate matrix, variation between sites and so forth). Further, η can denote the parameters of the model of the evolution of substitution rates over the time-tree. The crucial assumption for the class of models considered in the paper 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 rate over the whole genome in that lineage. The absolute substitution rates may however vary randomly and/or systematically over the genome but this variation should be the same over the whole phylogenetic tree.

For the model specified above there is no parameter for the species evolution nor for the evolution of sequences given the substitution rates, but the model of substitution rate evolution has one parameter ν. We hence have $\theta = \emptyset$ (no parameter) and $\eta = \nu$ in our specific example, but for other models both θ and η can be multi-dimensional.

Probability distribution and likelihood

We want to make inference about the time-tree implying that we should study the the probability distribution of the terminal 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 $\mathbf{r}^{(\tau)}$ and

summing over all possible sequences $X_I^{(\tau)}$ of the internal nodes including the root. If $f(\cdot)$ denotes a generic probability distribution we have:

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

In the last row of (1) we have removed θ in the distribution of $\mathbf{r}^{(\tau)}$ and η in the distribution of X because they don't affect the distributions. More importantly, the distribution of X, given the topology τ and parameter θ, depends only on $\mathbf{r}^{(\tau)}$ and $\mathbf{t}^{(\tau)}$ through their product $\mathbf{r}^{(\tau)} \cdot \mathbf{t}^{(\tau)} = \mathbf{b}^{(\tau)}$, the expected number of substitutions along the different branches. This distribution is in turn obtained by summing over possible values of the extended data set also containing internal node sequences:

$$f(X\ ;\ \tau, \mathbf{r}^{(\tau)} \cdot \mathbf{t}^{(\tau)}, \theta) = \sum_{X_I^{(\tau)}} f(X, X_I^{(\tau)}\ ;\ \tau, \mathbf{r}^{(\tau)} \cdot \mathbf{t}^{(\tau)}, \theta). \tag{2}$$

In Figure 1 for example, $X_I^{(\tau)}$ would be the (unknown) sequences of the root and the branching point that splits into terminal 2 and 3. The observed sequences X are for the terminals 1, 2 and 3.

For the specific model defined in the previous section in which $\eta = \nu$, the distribution of the rates, $f(\mathbf{r}^{(\tau)}\ ;\ \tau, \mathbf{t}^{(\tau)}, \nu)$, splits up into a product, one factor for each rate $r_i^{(\tau)}$, as specified by the Markov model. With the exception of the edges stemming from the root, the logarithms of observed edge-rates are normally distributed with mean equal to the logarithm of the mother rate and variance equal to ν times the average of the mother and daughter edge-times.

For our simple model (with $\theta = \emptyset$), the summands on the right hand side of (2) can also be written explicitly. Let $b_i^{(\tau)} = r_i^{(\tau)} t_i^{(\tau)}$ denote the expected number of substitutions per site along branch i. Then the probability of having the same nucleotide in both ends of this edge at a given site equals $p(b_i^{(\tau)}) = (1 + 3e^{-3b_i^{(\tau)}/4})/4$ (this is shown by conditioning on the number of substitutions). Because of independence between sites, and between edges given the branch lengths $\mathbf{b}^{(\tau)}$ $(= \mathbf{r}^{(\tau)} \cdot \mathbf{t}^{(\tau)})$ it follows that

$$f(X, X_I^{(\tau)}\ ;\ \tau, \mathbf{r}^{(\tau)} \cdot \mathbf{t}^{(\tau)}) \propto f(m_1, \ldots, m_{2k-2};\ \tau, \mathbf{b}^{(\tau)}) = \prod_{i=1}^{2k-2} p(b_i^{(\tau)})^{m_i} \left(1 - p(b_i^{(\tau)})\right)^{n-m_i},$$

where m_i is the observed number of sites at which edge i have the same nucleotide in both ends of the edge, and $n - m_i$ is the number of sites where the nucleotides differ. In the middle expression we have left out some combinatorial terms specifying how many nucleotide combinations that allow a given site to have the same/different nucleotide at the end of the edges why we use "$\propto$", denoting "proportional to", rather than equality.

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

$$L(\tau, \mathbf{t}^{(\tau)}, \theta, \eta) = f(X\ ;\ \tau, \mathbf{t}^{(\tau)}, \theta, \eta). \tag{3}$$

Table 1
Substitution rates used in simulation study

Rate	Distribution	Numerical value
r_1	$r_1 \sim Exp(100)$	0.0100
r_2	$\log(r_2) \sim N(\log(r_1),\ \nu(t_1+t_2)/2)$	0.0098
r_3	$\log(r_3) \sim N(\log(r_2),\ \nu(t_2+t_3)/2)$	0.0120
r_4	$\log(r_4) \sim N(\log(r_2),\ \nu(t_2+t_4)/2)$	0.0108

Note – The t_i's are taken from the time-tree of Figure 2, and $\nu = 0.01$.

For example, the maximum likelihood estimates $(\hat{\tau}, \hat{\mathbf{t}}^{(\hat{\tau})}, \hat{\theta}, \hat{\eta})$ is the set of parameter values that maximize the likelihood function. In our specific model we would hence estimate the topology, time-tree and ν by numerically maximizing $L(\tau, \mathbf{t}^{(\tau)}, \nu)$ with respect to τ, $\mathbf{t}^{(\tau)}$ and ν.

In the simpler case, treated in the simulation study, where the actual number of substitutions in each site along each edge is observed, the probability distribution becomes simpler. With this more detailed data we also observe multiple substitutions. A given site will have the same nucleotide in both ends of an edge if there were no substitutions, but also if there were at least two substitutions and if the last substitution was to the original nucleotide. With this more detailed data and our specific model, the number of substitutions from different sites can be aggregated without loss of information, so if we let $y_1, \ldots, y_{2k-2}$ denote the total number of substitutions along the different edges the probability distribution is given by

$$f(y_1, \ldots, y_{2k-2}\ ;\ \tau, \mathbf{r}^{(\tau)} \cdot \mathbf{t}^{(\tau)}) = \prod_{i=1}^{2k-2} \frac{(nr_it_i)^{y_i} e^{-nr_it_i}}{y_i} = \prod_{i=1}^{2k-2} \frac{(nb_i)^{y_i} e^{-nb_i}}{y_i},$$

a product of Poisson probabilities. The likelihood for this more detailed data is $L(\tau, \mathbf{t}^{(\tau)}, \nu) = \int f(\mathbf{r}^{(\tau)}\ ;\ \tau, \mathbf{t}^{(\tau)}, \nu) f(y_1, \ldots, y_{2k-2}\ ;\ \tau, \mathbf{r}^{(\tau)} \cdot \mathbf{t}^{(\tau)}) d\mathbf{r}^{(\tau)}$.

Simulation study

Maximum likelihood estimates for the model described above were derived numerically for a given 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 Figure 2). Substitution edge-rates were generated once according to the model of Thorne, Kishino and Painter (1998) with $\nu = 0.01$ (however, since 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.

The resulting b-tree is shown in Figure 2. 'Data' was generated for three different sequence lengths: $n = 1000$, $n = 10000$, and $n = 100000$. For each n the data $(y_1, ..., y_4)$, 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 $n = 10000$ have $y_2 = nb_2 = nr_2t_2 = 49$ observed substitutions (within the 10000 sites) as input data.

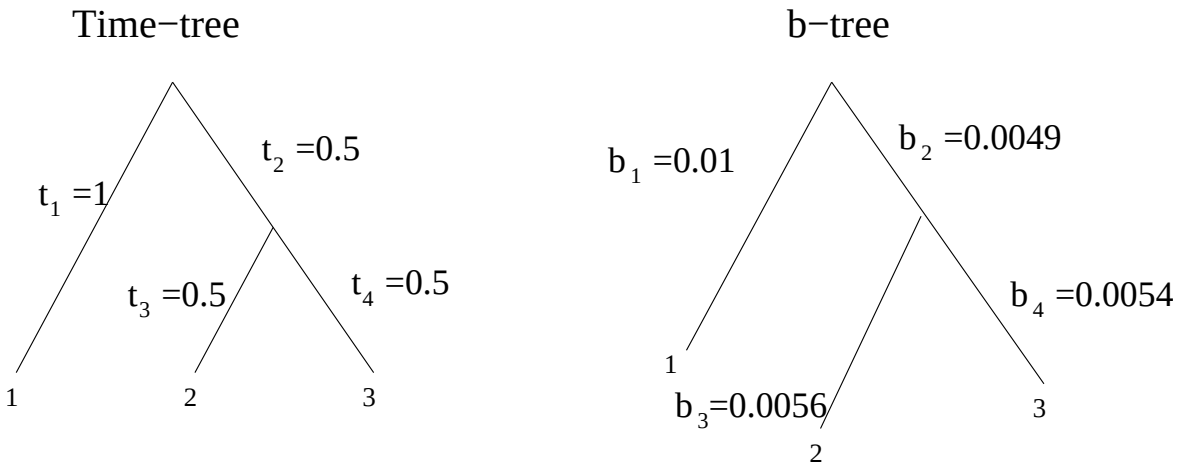

Figure 2: The time-tree and b-tree used in the simulation study. The substitution rates are taken from Table 1. For example, $b_2 = r_2t_2 = 0.0098 \times 0.5 = 0.0049$. The changes between the time-tree and the b-tree are exagerated in the figure.

To simplify inference the rooted tree topology τ and the variance parameter $\nu = 0.01$ in the rate variation model, are assumed known (τ and ν are hence omitted in the notation). Further, as mentioned in the previous section we assume that data consists of observing the actual number of substitutions that have occurred along each edge, also counting multiple substitutions. 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 Figure 2). The likelihood in (3) hence only depends on the parameter t_2 why we drop the index and write $L(t)$. For the different sequence lengths n, $L(t)$ 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 $f(y_1, \ldots, y_4\ ;\ \mathbf{r}\cdot\mathbf{t})$ 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

$$L(t) = f(y_1, \ldots, y_4; t) = \int f(y_1, \ldots, y_4\ ;\ \mathbf{r}\cdot\mathbf{t}) f(\mathbf{r}\ ;\ \mathbf{t}) d\mathbf{r}.$$

The (approximate) ML-estimate is then defined as the grid-point having highest likelihood. Simulations and figures were obtained using Matlab version 6.

Results

The main result of the present paper is that relative divergence times cannot be estimated consistently with increasing sequence lengths when the MC-assumption is not valid. This means that, even if the adopted model describes reality perfectly, one cannot find the relative divergence times with arbitrary high precision by collecting sufficiently long DNA sequences from the terminals of interest. This negative conclusion is true whatever estimator is used, non-parametric, likelihood based or other.

The conclusion follows from the observation already mentioned in the Methods section; the fact that data only affect 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. This is most easily seen from Equation (1) where it was shown that

$$L(\tau, \mathbf{t}^{(\tau)}, \theta, \eta) = f(X\ ;\ \tau, \mathbf{t}^{(\tau)}, \theta, \eta) = \int f(\mathbf{r}^{(\tau)}\ ;\ \tau, \mathbf{t}^{(\tau)}, \eta) f(X\ ;\ \tau, \mathbf{r}^{(\tau)} \cdot \mathbf{t}^{(\tau)}, \theta) d\mathbf{r}^{(\tau)}.$$

As more and more data (i.e. longer sequences) is collected $f(X;\ \tau, \mathbf{r}^{(\tau)} \cdot \mathbf{t}^{(\tau)}, \theta)$ becomes more and more peaked around $f(X;\ \tau, \hat{\mathbf{b}}, \theta)$, where $\hat{\mathbf{b}}$ is the maximum likelihood estimator for $\mathbf{b}$ which also approaches the true parameter value. From this it follows that, as n gets large,

$$L(\tau, \mathbf{t}^{(\tau)}, \theta, \eta) \approx f_{\mathbf{r}^{(\tau)}}(\hat{\mathbf{b}}^{(\tau)}/\mathbf{t}^{(\tau)};\ \tau, \mathbf{t}^{(\tau)}, \eta) f(X\ ;\ \tau, \hat{\mathbf{b}}^{(\tau)}, \theta), \tag{4}$$

where the first function on the right hand side, as indicated by the sub-index, is the probability density for $\mathbf{r}^{(\tau)}$ evaluated in $\hat{\mathbf{b}}^{(\tau)}/\mathbf{t}^{(\tau)} = (b_1^{(\tau)}/t_1^{(\tau)}, \ldots, b_{2k-2}^{(\tau)}/t_{2k-2}^{(\tau)})$. As a function of $\mathbf{t}^{(\tau)}$ the second factor on the right hand side is constant and is therefore irrelevant for inference on $\mathbf{t}^{(\tau)}$. The first factor does not get more and more peaked as longer sequences are collected which implies that the information about the substitution rates $\mathbf{r}^{(\tau)}$ does not tend to infinity. It will have a maximal value, but the density value of other points $\mathbf{r}$ are comparable in size and their relation is independent of the sequence length n (see Figure 3 for an illustration from the simulation study).

A more intuitive explanation to why the divergence times cannot be estimated consistently is that the $2k-2$ substitution rates are generated only once since the substitution rate of a specific edge is the substitution rate for every site. 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 instance, the average number of substitutions among the different sites, along an edge having substitution rate r and time duration t will tend to $b = rt$ eventhough 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 Figure 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 has no additional information about the time-tree, and the b-tree does not allow the time-tree to be estimated without uncertainty.

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 $\pi(\tau, \mathbf{t}^{(\tau)}, \theta, \eta)$ 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 $\pi(\tau, \mathbf{t}^{(\tau)}, \theta, \eta | \, X)$. The relation between the posterior distribution and prior distribution and likelihood satisfies

$$\pi(\tau, \mathbf{t}^{(\tau)}, \theta, \eta | \, X) \propto \pi(\tau, \mathbf{t}^{(\tau)}, \theta, \eta) L(\tau, \mathbf{t}^{(\tau)}, \theta, \eta). \tag{5}$$

Because the likelihood will not get more and more peaked around the true divergence times it hence follows that nor will the posterior distribution. A consequence from 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 results

In Figure 3 we show likelihood plots for $t = t_2$ for the 3-taxon example presented earlier, for sequence lengths equal to $n = 1000$, $n = 10000$ and $n = 100000$. We have also plotted the likelihood for $n = \infty$ where we used equation (4) 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. In the figure 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 negligeble for values of t in the range (0.42, 0.52) say – recall that the true t equals 0.5. This illustrates that the divergence time $t_2 = t$ 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 $t_2 = 0.5$. 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 Figure 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 likelihood plots for the corresponding branch length $b = b_2$ for the same set of sequence lengths (see Figure 4). If these plots are compared with the plots of $L(t)$ in Figure 3 it is seen that $L(b)$ concentrates at a higher rate as n increases, and also that, in the limit as n tends to infinity, all mass concentrates at the true value $b_2 = 0.0049$. This illustrates that the branch lenghts b_i can be estimated consistently whereas the relative times t_i 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

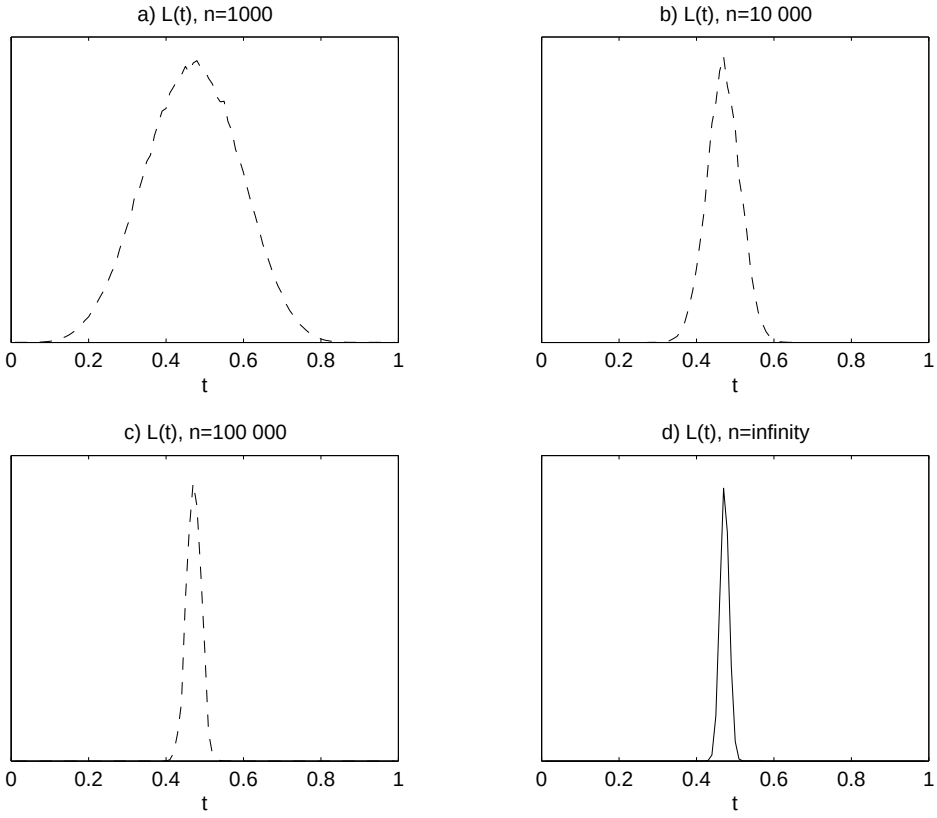

Figure 3: Likelihood plots of t for the simulation example, for different values of the sequence lenght n. The first three plots are obtained using Monte Carlo simulations and the last plot is plotted with the limiting likelihood. It is seen that information about the relative time t increases with n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

nonnegligible likelihood values for a range of t-values as the sequence lenght n grew. 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 $b_2 = r_2 t_2 = 0.5 r_2$.

The simulation example contains several unrealistic simplifications. First, the evolutionary model is very simple with the same substitution rate for each site and Jukes Cantor type substitution model. However, the same qualitative result would still hold if a more general substitution rate model would be used, and also if the magnitude of the substitution rates varied over the sequence in a systematic and/or random way. A crucial assumption for our result to remain true in the latter case is that the (relative) evolution of the substitution rates over the tree is the same for different sites. When this assumption fails we are in a different class of models which are discussed briefly in the next section. Another simplification in the example is that we assume that the tree topology is known, that $\nu = 0.01$ is known, and that we observe the actual occurred number of substitutions (counting also multiple substitutions) along each edge of the

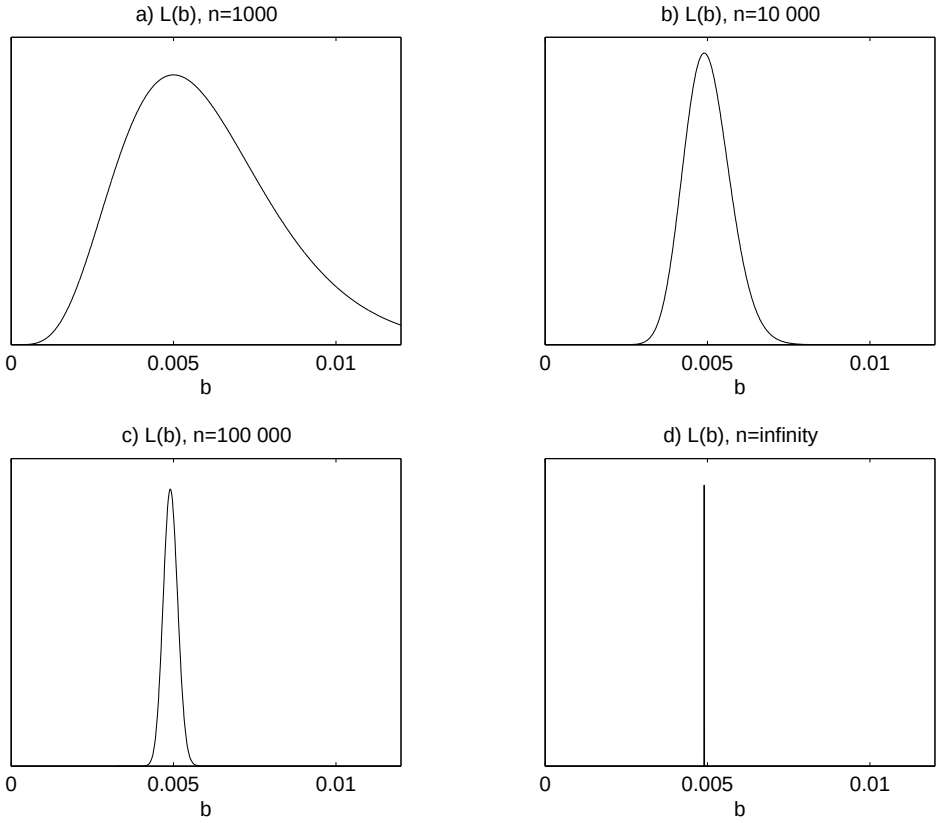

Figure 4: Likelihood plots of b for the simulation example, for different values of the sequence lenght n. The amount of information increases to infinity with n making consistent estimation possible.

tree. Even with the simple model treated in the example this is not very realistic, but by assuming this data rather than sequence data makes inference more precise, so the likelihood for t based on sequence data should be even less peaked. Finally, in the example we only consider a rooted tree for three terminals. The reason for this very small example is to keep numerical problems to a minimum. For more realistically sized examples it is a difficult numerical problem to obtain the likelihood function or even an approximation of it, but this problem is outside the scope of the present article.

Discussion

In the previous section it was shown that it is impossible to estimate the relative divergence times of a phylogenetic tree consistently, by taking longer and longer sequences, without assuming the molecular clock assumption. Although difficulties in estimating divergence times when substitution rates vary over the phylogenetic tree have been reported previously (e.g. Sanderson, personal communication, and Thorne and Kishino, 2002) this result has not been discovered previously. Since consistency is one of the

most important properties when making parameter inference, a relevant question is if estimation is completely useless in the situation treated. The answer to this question is no – there is information about the divegence times in the data, although the amount of information is limited even as the sequence length increases (see Figure 3). This implies that the divergence times are not unidentifiable; some sets of divergence times are more likely than others, and maximum likelihood estimates exists and are unique (see for example Rannala, 2002, for a discussion on unidentifiable parameters in overparametrised models). How the maximum likelihood (ML) estimate 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 small 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. the simulation study). 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 unbiased way albeit not consistently.

Divergence times can of course be estimated in several parametric or non-parametric (e.g. parsominy) ways. If one knows that the model is correct then ML-estimation is often the preferred method. Except for the case with very small trees and simple evolutionary models ML-estimates can be practically impossible to derive due to numerical problems; an issue not treated in the present paper. In such situations other estimators may be preferred. Another reason why other estimators can be preferred is if they are more robust in the sense that they are less sensitive to model assumptions. Unfortunately it follows from the paper that no other estimator for relative divergence times can be consistent either. When choosing estimator robustness and numerical tractability as well as statistical properties such as approximate unbiasedness and small standard deviation should be considered.

In the Bayesian framework the posterior distribution reflects the uncertainty of the divergence times. Given that the prior distribution is correct – and its impact is not negligible even when long sequences are collected! – the posterior distribution summarizes the information about the divergence times correctly. However, as pointed out before, the distribution does not get more and more peaked, as one would hope, when longer sequences are collected.

That data contains information about the divergence times also implies that the molecular clock (MC) assumption can be tested, and the power of this test can be made arbitrary powerful by collecting sufficiently long sequences (see Langley and Fitch, 1974 and other references mentioned in the introduction for tests on the MC-assumption). For most sets of data the MC assumption is rejected meaning that the time-tree is not proportional to the b-tree.

Having concluded that relative divergence times cannot be estimated consistently, under the type of models and data treated in the present paper, makes it relevant to ask under what alternative data sets and/or model assumptions divergence times indeed can be estimated consistently. If we first look at other, more informative, types of data, it is clear that precision increases the more fossil datings that are available. However, it seems as if precise fossil datings are necessary for all nodes where substitu-

tion rates change in the tree in order to estimate divergence times consistently, and if the substitution rate is believed to change in more or less all nodes this does not seem like realistic data. Further, fossil datings are usually not very precise in its location in a tree nor the actual dating. For the type of model considered here we hence have no obvious suggestion of complementary data to make consistent estimation possible, but more work is needed in this area.

If we instead look at alternative models for which consistent estimation is feasible we first 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 the variation of substitution rates over lineages is allowed 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 for a specific gene is modelled as in the present model, but assuming that substitution rate between different genes are completely independent, then it is possible to estimate divergence times consistently by collecting DNA sequences from more and more genes. The result should hold true even if some small correlation between substitution rates of different genes is allowed, with the effect that a high substitution rate for one gene of a specific lineage make high substitution rates for other genes on the same lineage somewhat more likely. To model such correlation can be done in different ways, see for example Thorne and Kishino (2002), and how correlated substitution rates can be, still allowing consistent estimation, remains an open problem.

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.

Literature cited

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 path-lenghts. Mol. Phyl. Evol. **24**: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. Pp. 21-32 *in* H. N. Munro, ed. Mammalian protein metabolism. 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, Mass.
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 phylogeny. 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.
ZUCKERLANDL, E., and L. PAULING. 1965. Evolutionary divergence and convergence in proteins. Pp. 97-166 *in* V. Bryson and H. J. Vogel, eds. Evolving genes and proteins. Academic press, New York.