Mathematical Statistics
Stockholm University

Improving divergence time estimation in phylogenetics: more taxa vs. longer sequences

Bodil Svennblad
Tom Britton

Research Report 2007:21

ISSN 1650-0377

Postal address:

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

Internet:

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

Improving divergence time estimation in phylogenetics: more taxa vs. longer sequences

Bodil Svennblad, Uppsala University[*]
Tom Britton, Stockholm University[†]

November 28, 2007

Running head: Divergence time estimation
Keywords: Phylogeny, divergence time estimation, Maximum Likelihood, Mean Path Length

Abstract

Maximum Likelihood (ML) is used as a standard method for estimating divergence times in phylogenetic trees. The method is consistent and hence the precision can be improved by analysing longer sequences. In this paper, we show that the estimates also can be improved by including more taxa to the existing tree. It is a theoretical study, complemented with simulations, showing that the gain in precision is faster with increasing sequence length than with increasing number of taxa, using symmetric trees.

We further compare the results of estimating divergence time using Maximum Likelihood with the much faster and less complex estimation method of Mean Path Length (MPL), which works with the evoultion model of Jukes-Cantor (1969). It is shown that MPL is as good as ML in estimating the divergence times of nodes that are located near the root in the tree, but ML is better in the divergence times of nodes lower down.

[*] Department of Mathematics, Uppsala University, Box 480, SE-751 06 Uppsala, Sweden *E-mail*: bodil.svennblad@math.uu.se

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

1 Introduction

Maximum Likelihood (ML) is a well known, well founded method for consistently estimating divergence times in a rooted phylogenetic tree as in many other areas. The method was introduced into phylogenetics by Edwards and Cavalli-Sforza (1964) but for gene frequency data. Felsenstein (1981) brought Maximum Likelihood to phylogenetic inference based on nucleotide sequences. It is now used as a standard method in softwares like PAUP* (Swofford, 2002) and PHYLIP (Felsenstein, 2005).

In this paper we investigate if divergence time estimation of a node gain in precision if more taxa are added to that part of an existing tree. (It was initiated when TB attended a phylogenetic meeting organized by NESCENT in September 2006, where the question was posed.)

Since Maximum Likelihood is a consistent method, the ML estimate of the divergence time of a node can be improved by analysing longer sequences. In general the variance is proportional to the inverse of the sequence length. The variance is hence divided by 2 if the sequence length is doubled. Under some conditions on the tree almost the same improvement can be achieved by instead squaring the number of taxa. The precision is hence improved by adding taxa, but not as fast as when sequence length is increased. To simplify the analytical calculations we have used the simplest possible model of evolution, but we believe similar results hold for more realistic models.

We further compare the results of estimating divergence time with Maximum Likelihood with the much faster and less complex estimation method of Mean Path Length. This method assumes the evolution model of Jukes-Cantor (1969), though a generalization of the method exists (PATHd8, Britton et. al. 2007) allowing different substitution rates in different parts of the tree. Under some conditions on the tree topology, the Mean Path Length is as good as Maximum Likelihood in estimating the divergence time of the nodes located high in the tree (close to the root). The divergence time of internal nodes lower down though, are always estimated with better precision using the method of Maximum Likelihood than of Mean Path Length.

The paper is organized as follows: the model used, tree topologies considered and observed data are described in the next section. Thereafter the two methods are explained and how to compare their precisions. A section of theoretical study of the improvement in precision as taxa are added is followed by a simulation section, examplifying the theoretical findings.

2 Model and data

In this paper we assume we have a binary rooted tree with $k = 2^l$ taxa. Where not otherwise stated it is assumed that the tree is symmetric in the sense that the two subtrees of a node have equally many terminal nodes. This makes formulae for variances simpler and it is possible to compare what happens if the number of taxa is e.g. doubled. If we would consider other trees, doubling the number of taxa would result in many possible tree topologies, all with different estimates of divergence time and precision thereof.

The nodes are denoted $\{1, 2, 3, \ldots, 2k - 1\}$, where $\{1, \ldots, k - 1\}$ are the internal nodes and $\{k, \ldots, 2k - 1\}$ are the terminal nodes (see Figure 1). Let the root be node 1 with daugther nodes 2 and 3. At the level beneath are the nodes 4 to 7 etc. A node is at level j if it has $j - 1$ nodes on the path from the root to the node. The number of levels is $\log_2 k = l$. Denote the divergence time of node i with t_i.

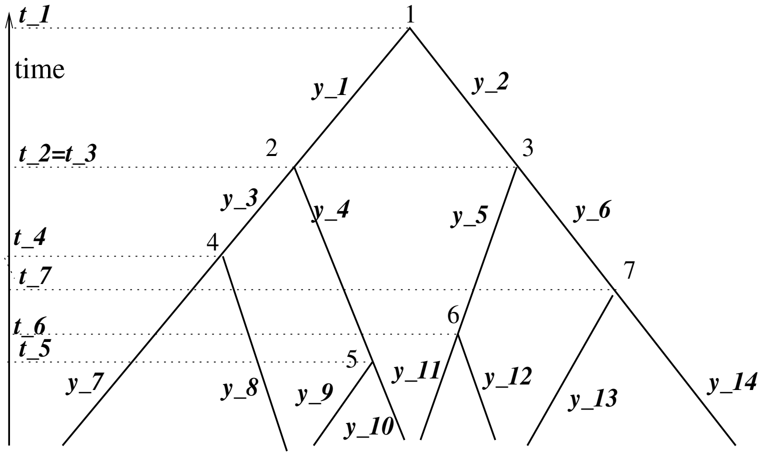

Figure 1: *A rooted binary tree with $k = 8$ taxa and $l = \log_2 k = 3$ levels. A node is at level i if it has $i - 1$ nodes on the path from the root to the node. Hence node 1 is the root at level 1 with divergence time t_1. The nodes $\{2, 3\}$ are at level 2, both with divergence time t_2. At the third level are the nodes $\{4, \ldots, 7\}$ with corresponding divergence times $\{t_4, \ldots, t_7\}$.*

Denote the data to analyse with $\mathbf{y}$, where y_i is the observed number of substitutions in a DNA sequence with n sites along the branch between node $[(i + 1)/2]$ and node $i + 1$, where $[\cdot]$ denotes the integer part. We admit it is an unrealistic situation to observe the number of substitutions along every branch, normally we only observe the DNA sequences at the terminal nodes and estimate the number of substitutions along the branches, but this is to make our investigation of the properties of the methods possible.

We further assume that substitutions occur according to the Jukes-Cantor model of evolution (1969). In this model, at each site j in the DNA sequence, substitutions occur independently in time with change rate r_j. When a change occur all bases are equally probable. Along a branch of length t, the number of substitutions at a site j is then Poisson distributed with mean tr_j. We further assume the sites to evolve independently and hence the total number of substitutions along the branch for a DNA sequence of n sites is $Y = Y_1 + \cdots + Y_n$ where all $Y_j \in Po(tr_j)$. Hence $Y \in Po(rt)$ where $r = \sum_{j=1}^{n} r_j$. To show the dependence on the number of sites in the sequences, define $\bar{r} = r/n$ and let $Y \in Po(n\bar{r}t)$. We assume that the mean rate $\bar{r}$ is constant over all branches since it otherwise may not be possible to estimate the divergence time consistently (Britton, 2005). Further the mean rate is assumed to be known as the the estimates otherwise will be on a relative scale.

For a given tree, $\mathbf{y}$ denotes the number of substitutions along each branch. According to the model described above $Y_i \in Po(n\bar{r}(t_{[(i+1)/2]} - t_{i+1}))$ for internal branches ($i = 1, \ldots, k-2$) and $Y_i \in Po(n\bar{r}t_{[(i+1)/2]})$ for branches ending in a terminal node ($i = k-1, \ldots, 2k-2$).

3 Methods

3.1 Likelihood and Score function

The joint probability function of $\mathbf{y}$ can, as we assume that branches evolve independently, be written as the product of the probability functions for the observations at each branch. For a phylogenetic tree of the kind we consider, with k taxa, $\mathbf{t} = (t_1, \ldots, t_{k-1})$ and $\mathbf{y} = (y_1, \ldots, y_{2k-2})$ the joint probability function equals

$$\begin{aligned}
f(\mathbf{y}|\mathbf{t}) &= \prod_{i=1}^{k-2} \frac{1}{y_i!} e^{-n\bar{r}(t_{[(i+1)/2]}-t_{i+1})} (n\bar{r}(t_{[(i+1)/2]}-t_{i+1}))^{y_i} \times \\
& \prod_{i=k-1}^{2k-2} \frac{1}{y_i!} e^{-n\bar{r}t_{[(i+1)/2]}} (n\bar{r}t_{[(i+1)/2]})^{y_i} \\
&= \frac{1}{\prod_{i=1}^{2k-2} y_i!} \exp\Bigg\{ -n\bar{r}t_1 - \sum_{i=1}^{k-1} n\bar{r}t_i + \sum_{i=1}^{k-2} y_i \log(n\bar{r}(t_{[(i+1)/2]}-t_{i+1})) \\
& + \sum_{i=k-1}^{2k-2} y_i \log n\bar{r}t_{[(i+1)/2]} \Bigg\}. \qquad (1)
\end{aligned}$$

Let $h(y) = 1/(\prod y!)$, $T(\mathbf{y}) = \mathbf{y}$, $a(\mathbf{t}) = -n\bar{r}t_1 - \sum_{i=1}^{k-1} n\bar{r}t_i$ and

$$\eta(\mathbf{t}) = \begin{cases} \log(n\bar{r}(t_{[(i+1)/2]}-t_{i+1})) & i = 1, \ldots, k-2 \\ \log(n\bar{r}t_{[(i+1)/2]}) & i = k-1, \ldots, 2k-2. \end{cases}$$

Then (1) can be written on the form

$$f(\mathbf{y}|\mathbf{t}) = h(\mathbf{y}) \exp\{\eta(\mathbf{t})T(\mathbf{y}) - a(\mathbf{t})\}. \qquad (2)$$

A joint probability function that can be written on the form (2) is said to be in the exponential family. Hence (1) is in the exponential family.

The likelihood, $L(\mathbf{t})$ is the probability of data given the parameter $\mathbf{t}$, but seen as a function of $\mathbf{t}$, that is $L(\mathbf{t}) = f(\mathbf{y}|\mathbf{t})$. The log likelihood $l(\mathbf{t}) = \log L(\mathbf{t})$ is hence, using (2), $l(\mathbf{t}) = \log L(\mathbf{t}) = \log(h(\mathbf{y})) + \eta(\mathbf{t})T(\mathbf{y}) - a(\mathbf{t})$.

The score function $U_i(\mathbf{t}) := \frac{\partial l(\mathbf{t})}{\partial t_i}$ is in this case

$$\begin{aligned}
U_i(\mathbf{t}) &= \frac{\partial \eta(\mathbf{t})}{\partial t_i} T(y) - \frac{\partial a(\mathbf{t})}{\partial t_i} \\
&= \begin{cases} -2n\bar{r} + \frac{y_1}{t_1-t_2} + \frac{y_2}{t_1-t_3} & i = 1 \\ -n\bar{r} - \frac{y_{i-1}}{t_{[i/2]}-t_i} + \frac{y_{2i-1}}{t_i-t_{2i}} + \frac{y_{2i}}{t_i-t_{2i+1}} & i = 2, \ldots, \frac{k}{2}-1 \\ -n\bar{r} - \frac{y_{i-1}}{t_{[i/2]}-t_i} + \frac{y_{2i-1}+y_{2i}}{t_i} & i = \frac{k}{2}, \ldots, k-1. \end{cases} \qquad (3)
\end{aligned}$$

To obtain the Maximum Likelihood (ML) estimate of $\mathbf{t}$, the score function is set to 0. The solution is then found by solving the equations numerically. In the exponential family, under mild regularity conditions, the Maximum Likelihood estimate $\hat{\mathbf{t}}^{(ML)}$ is consistent, that is $\hat{\mathbf{t}}^{(ML)} \to \mathbf{t}$ when $n \to \infty$. The asymptotic variance of $\hat{\mathbf{t}}^{(ML)}$ equals the inverse of the expected information matrix, which will be defined later.

3.2 Mean Path Length

The algorithm of Mean Path Length (MPL) was first introduced by Bremer and Gustafsson (1997) and further developed by Britton et. al. (2002), who also showed it to be consistent for the Jukes-Cantor model. We focus on the divergence time of one node and call that node the root. The divergence time of the root is denoted t_1. Assume that a path from the root to a taxa is defined by the branches with corresponding observations $\{y_{i_1}, \ldots, y_{i_l}\}$. The total number of substitutions along that path is hence $x_i = y_{i_1} + \cdots + y_{i_l}$. With the assumptions of the tree made earlier in this paper we have that $X_i \in Po(n\bar{r}t_1)$. The Mean Path Length estimate of the divergence time of the root is the mean of the total number of substitutions of all paths from the root to the taxa divided by $n\bar{r}$.

The MPL estimate can be rewritten as a weighted sum of observations, where each observation y_i is weighted with a constant c_i, which equals the proportion of all paths that traverses the branch which y_i belongs to. This will make the estimate a sum of independent variables and the variance will then be easy to calculate. Since we assume a symmetric tree $c_i = \frac{1}{2^j}$, where j is the level which y_i belongs to. With the notations we use, y_1 and y_2 are at the first level, $y_3, \ldots, y_7$ at level 2 and level j consists of the observations $\{2^j - 1, \ldots, 2^{j+1} - 2\}$, $j = 1, \ldots, \log_2 k$. Hence the MPL estimate of the divergence time of the root will be

$$\hat{t}_1^{(MPL)} = \frac{1}{n\bar{r}} \sum_{j=1}^{\log_2 k} 2^{-j} \left(\sum_{i=2^j-1}^{2^{j+1}-2} y_i \right). \tag{4}$$

3.3 Precision of the estimates

Assuming the Jukes-Cantor model both methods consistently estimate the divergence time of the root, t_1. That is, as the sequence length n increases, the closer to the true value t_1 the estimates $\hat{t}_1^{(ML)}$ and $\hat{t}_1^{(MPL)}$ will be.

Furthermore we have assumed that we are able to observe the number of substitutions along the branches. In a practical situation this is rarely the case. Often, only DNA sequences of terminal taxa are observed and the number of substitutions along branches have to be estimated. This can be done by a number of algorithms, non-probabilistic or probabilistic (e.g. Maximum Likelihood). With this method it can be shown that the expected number of substitutions per site can be estimated consistently. By letting the observations of our tree be the estimated branch length multiplied by the sequence length we can consider the observations to be outcomes of the Poisson distribution with mean $n\bar{r}t$. The results of the analytical calculations of the precision of the two methods that follow are hence optimistic bounds when it comes to real data, but with large n the true variances should be close to the analytical ones obtained here.

Since the model implies that $Y_i \in Po(n\bar{r}(t_{[(i+1)/2]} - t_{i+1}))$, $i = 1, \ldots, k-2$ and $Y_i \in Po(n\bar{r}t_{[(i+1)/2]})$, $i = k-1, \ldots, 2k-2$, and $Y_1, \ldots, Y_{2k-2}$ are independent, the variance of $\hat{t}_1^{(MPL)}$ can be calculated as

$$\begin{aligned} V(\hat{t}_1^{(MPL)}) &= V\biggl(\frac{1}{n\bar{r}} \sum_{j=1}^{\log_2 k} 2^{-j} \biggl[\sum_{i=2^j-1}^{2^{j+1}-2} y_i \biggr]\biggr) \\ &= \frac{1}{n^2\bar{r}^2} \sum_{j=1}^{\log_2 k} 2^{-2j} \sum_{i=2^j-1}^{2^{j+1}-2} V(y_i) \\ &= \frac{1}{n^2\bar{r}^2} \sum_{j=1}^{\log_2 k-1} 2^{-2j} \sum_{i=2^j-1}^{2^{j+1}-2} n\bar{r}(t_{[\frac{i+1}{2}]} - t_{i+1}) + \frac{1}{n^2\bar{r}^2k^2} \sum_{i=k-1}^{2k-2} n\bar{r}t_{[\frac{i+1}{2}]} \\ &= \frac{1}{n\bar{r}} \sum_{j=1}^{k-1} 2^{-2j} \sum_{i=2^j-1}^{2^{j+1}-2} (t_{[\frac{i+1}{2}]} - t_{i+1}) + \frac{1}{n\bar{r}k^2} \sum_{i=k-1}^{2k-2} t_{[\frac{i+1}{2}]}. \qquad (5) \end{aligned}$$

For any estimate, $\hat{\mathbf{t}}$, Cramér-Rao's inequality states that $Var(\hat{\mathbf{t}}) \geq I^{-1}(\mathbf{t})$, where $I^{-1}(\mathbf{t})$ is the inverse of the Fisher information matrix (see e.g. Lindsey, 2001). Under mild regularity conditions (which are satisfied in our case the model being from the exponential family) the ML-estimate, $\hat{t}_1^{(ML)}$, achieves the Cramér-Rao lower bound, at least asymptotically as $n \to \infty$. The variance of $\hat{t}_1^{(ML)}$ will therefore be $I_{11}^{-1}(\mathbf{t})$ which is the first element of the inverse of the information matrix. The Fisher information matrix is defined as

$$\begin{aligned} I_{ij}(\mathbf{t}) &= E(J_{ij}(\mathbf{t})) \\ &= E\biggl(-\frac{\partial U_i(\mathbf{t})}{\partial t_j}\biggr), \end{aligned}$$

which can be shown to equal

$$I(\mathbf{t}) = \begin{cases} \frac{n\bar{r}}{t_1-t_2} + \frac{n\bar{r}}{t_1-t_3} & i=j=1 \\ \frac{n\bar{r}}{t_{[i/2]}-t_i} + \frac{n\bar{r}}{t_i-t_{2i}} + \frac{n\bar{r}}{t_i-t_{2i+1}} & i=j, i=2,\ldots,\frac{k}{2}-1 \\ \frac{n\bar{r}}{t_{[i/2]}-t_i} + \frac{2n\bar{r}}{t_i} & i=j, i=\frac{k}{2},\ldots,k-1 \\ -\frac{n\bar{r}}{t_j-t_i} & j=[i/2] \\ -\frac{n\bar{r}}{t_j-t_i} & j=(2i,2i+1), \end{cases} \tag{6}$$

where $i = 2, \ldots, k-1$ in the last row.

4 Increasing the number of taxa

Suppose as before that we are interested in the divergence time t_1 of two taxa A and B (the node is hence on referred to as the root). As data, we have the rooted tree to the left in Figure 2 as well as the number of substitutions (which we here denote x_1 and x_2) in a DNA sequence of length n for each branch. The observations are both outcomes from Poisson distributed random variables with mean (and variance) $n\bar{r}t_1$. The ML estimate, as well as the MPL estimate of t_1 will be

$$\hat{t}_1^{(ML)} = \hat{t}_1^{(MPL)} = \tfrac{x_1+x_2}{2n\bar{r}}.$$

The variance of the estimates is $V(\hat{t}_1^{(ML)}) = V(\hat{t}_1^{(MPL)}) = \frac{1}{(2n\bar{r})^2}(V(X_1)+V(X_2)) = \frac{t_1}{2n\bar{r}}$. Increasing the length n of the sequences decreases the variance as can be seen from the formula. One way of improving the estimate is hence to analyse longer sequences. We will now show that a different way to improve the estimate is by increasing the number of taxa. We will also investigate how much this improves the estimate.

At the right in Figure 2 two more taxa, C and D, are considered, with divergence times more recent than t_1, which is the time we want to estimate. We assume that the number of substitutions along each branch are observable and the observations are now $(y_1, \ldots, y_6)$, with $y_1 + y_3 = x_1$ and $y_2 + y_6 = x_2$. With real data those observations are not really observed but estimated, consistently as the sequence length increases. The number of levels is $l = \log_2 k = 2$ and the MPL estimate is then, according to (4)

$$\hat{t}_1^{(MPL)} = \frac{1}{n\bar{r}}\Big(2^{-1}(y_1+y_2) + 2^{-2}(y_3+\cdots+y_6)\Big),$$

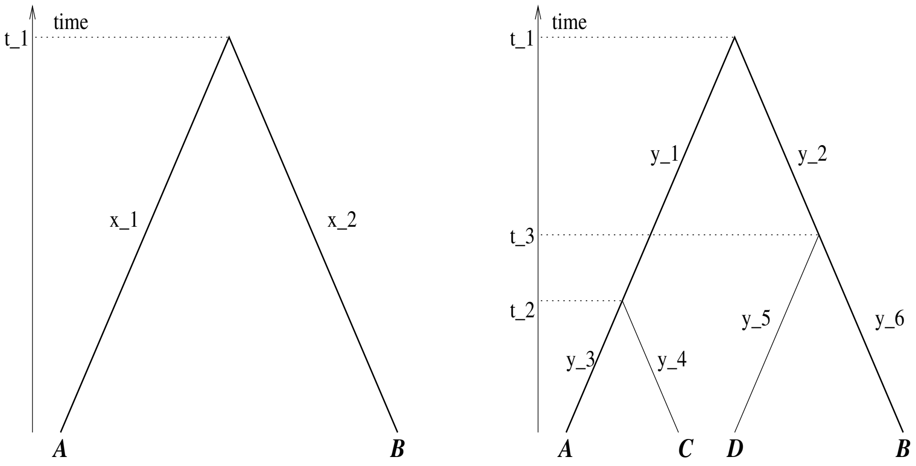

Figure 2: *To the left is the topology of the tree representing the divergence time we are interested in. To estimate the time t_1 we have the observations x_1 and x_2. To the right two more taxa are added, with divergence times t_2 and t_3 that we are not primarily interested in. The observations in the right hand tree are $\mathbf{y} = (y_1, \ldots, y_6)$.*

with variance

$$\begin{aligned} V(\hat{t}_1^{(MPL)}) &= \frac{1}{n^2\bar{r}^2}\Big(2^{-2}(n\bar{r}(t_1 - t_2) + n\bar{r}(t_1 - t_3)) + 2^{-4}(2n\bar{r}t_2 + 2n\bar{r}t_3)\Big) \\ &= \frac{t_1}{2n\bar{r}} - \frac{t_2}{8n\bar{r}} - \frac{t_3}{8n\bar{r}}, \end{aligned} \tag{7}$$

which is smaller than the variance of the first estimate. Hence, by adding more information in the form of new taxa, corresponding unknown divergence times that we are not primarily interested in and observed substitutions along the branches, we have improved the estimate of the divergence time of the root.

For the method of Maximum Likelihood the information matrix $I(\mathbf{t})$ corresponding to the right tree in Figure 2 will be, according to (6),

$$I(t) = n\bar{r}\begin{pmatrix} \frac{1}{t_1-t_2} + \frac{1}{t_1-t_3} & -\frac{1}{t_1-t_2} & -\frac{1}{t_1-t_3} \\ -\frac{1}{t_1-t_2} & \frac{1}{t_1-t_2} + \frac{2}{t_2} & 0 \\ -\frac{1}{t_1-t_3} & 0 & \frac{1}{t_1-t_3} + \frac{2}{t_3} \end{pmatrix},$$

with inverse

$$I(\mathbf{t})^{-1} = c\begin{pmatrix} (2t_1-t_3)(2t_1-t_2) & t_2(2t_1-t_3) & t_3(2t_1-t_2) \\ t_2(2t_1-t_3) & t_2(4t_1-t_3-2t_2) & t_2t_3 \\ t_3(2t_1-t_2) & t_2t_3 & t_3(4t_1-2t_3-t_2) \end{pmatrix}, \tag{8}$$

where $c = \frac{1}{2n\bar{r}(4t_1-t_2-t_3)}$. It can be shown that the first element of (8) satisifies $\{I(\mathbf{t})^{-1}\}_{11} < \frac{t_1}{2n\bar{r}}$ and hence also $\hat{t}_1^{(ML)}$ is improved by adding more taxa.

From (7) and (8) we see that the variance is smaller the closer the nodes 2 and 3 are the root. To optimize the extra information taxa should hence be chosen so that the speciation did not happen recently.

If the unknown divergence times t_2 and t_3 in fact were the same (i.e. $t_2 = t_3$), then both (7) and the top left element of (8) would equal $((2t_1-t_2)/4n\bar{r})$. Adding more taxa, assuming the tree to be complete symmetric, by which we mean that all nodes at a level have the same divergence time, it can be verified that

$$V(\hat{t}_1^{(MPL)}) = V(\hat{t}_1^{(ML)}) = I_{11}^{-1}(\mathbf{t}), \tag{9}$$

for $k = 4$ (shown above), $k = 8$ and $k = 16$, but we have not managed to prove this for larger k. However, simulations indicate that it holds in general also for larger k. In the appendix it is shown analytically that (9) holds if we further require the branches to be of the same length. This type of tree we call equidistant complete symmetric. Then (9) becomes

$$V(\hat{t}_1^{(ML)}) = V(\hat{t}_1^{(MPL)}) = \frac{t_1}{n\bar{r}\log_2 k}\frac{k-1}{k}. \tag{10}$$

To divide the variance of the estimate of the divergence time of the root in an equidistant complete symmetric tree with 2, the sequence length should be doubled according to (10). Since $\frac{k-1}{k} \simeq 1$, at least for large k, almost the same improvement can be achieved by instead squaring the number of taxa. Adding more taxa hence improves the precision, but not as fast as increasing the sequence length.

For a symmetric tree where all nodes have individual divergence times, simulations indicate that ML estimates the divergence time of the root with slightly higher precision than the method of Mean Path Length. The difference between the estimates of the two methods and corresponding variances are small though, as long as it is the divergence time of the root that is of interest.

For other internal nodes, e.g. the most recent common ancestor of A and C in Figure 2 with divergence time t_2, are always estimated with better precision with Maximum Likelihood than with Mean Path Length. For the example in Figure 2 we have

$$\hat{t}_2^{(MPL)} = \tfrac{y_3+y_4}{2n\bar{r}}, \qquad V(\hat{t}_2^{(MPL)}) = \tfrac{t_2}{2n\bar{r}}.$$

The ML estimate is obtained as $U(\mathbf{t}) = 0$ is solved numerically, giving $\hat{\mathbf{t}}^{(ML)}$. The variance of $\hat{t}_2^{(ML)}$ is $\{I^{-1}(\mathbf{t})\}_{22}$, and by (8) this equals

$$V(\hat{t}_2^{(ML)}) = \tfrac{t_2}{2n\bar{r}}\left(1 - \tfrac{t_2}{4t_1-t_2-t_3}\right).$$

It can be shown that for this particular situation $\frac{1}{2}V(\hat{t}_2^{(MPL)}) \leq V(\hat{t}_2^{(ML)}) \leq V(\hat{t}_2^{(MPL)})$.

When estimating the divergence time for internal nodes, the method of Maximum Likelihood uses all observations of the entire tree. The method of Mean Path Length only takes the observations on the paths from the node to descending taxa, so the Mean Path Length uses less information than Maximum Likelihood. For a node located high in the tree the two methods use almost the same amount of information, for the root exactly the same. The estimates, as well as the precisions thereof should therefore be close. For a node lower down the tree though, MPL only uses part of the information and the precision is then lower than for Maximum Likelihood.

We have only considered adding taxa to the subtree defined by the node whose divergence time we are interested in and which will be the root of the subtree. Adding taxa to other parts of the tree will not affect the MPL estimate, since it only uses observations in the subtree. The ML estimate uses information from all parts of the tree and will gain in precision wherever the taxa are added. How much the variance of the estimate decreases is however hard to specify. It will depend on the size of the tree and subtree, and also on how far away from the node of interest the extra information is added.

5 Simulations

In the simulation part we considered three types of trees: (1) equidistant complete symmetric, (2) complete symmetric and (3) symmetric trees (see Figure 3). In the complete symmetric trees all nodes at the same level diverged at the same time. In the equidistant case we further require the times between speciation to be equal, that is all branches are of the same length. In the symmetric trees the two subtrees of a node have equally many nodes, but the divergence times may be individual for all nodes. When doing the inference, the type of tree is of course assumed to be unknown.

The goal of the simulations were to estimate the divergence time of the root for a number of datasets and see how well the methods meet the theoretical variances by calculating the sample variances. We also wanted to see how the variances decreased when the number of taxa, k, was increased. To obtain the ML estimate of the divergence time of the root, $\hat{t}_1^{(ML)}$, for a given tree and dataset, the log likelihood function $l(\mathbf{t})$ should be maximized with respect to $\mathbf{t}$ (or equivalently $U(\mathbf{t}) = 0$ should be solved). We have to do this under the "complex" constraints $t_1 > 0$, $0 < t_i < t_{[i/2]}$, since otherwise the time of a daughter node can be older than the mother node. Several numerical algorithms exist that are able to optimize a function under "simple" constraints. Let $\mathbf{x} = \mathbf{x}(\mathbf{t}) = (t_1, \frac{t_2}{t_1}, \frac{t_3}{t_1}, \ldots, \frac{t_i}{t_{[i/2]}}, \ldots, \frac{t_{k-1}}{t_{[(k-1)/2]}})$, that is x_i is the time t_i divided by the time of the node above it. Consider now $l(\mathbf{x})$ with the simple constraints $x_1 > 0$, $0 < x_i < 1$, $i = 2, \ldots, k-1$. Reparameterizing back to $\mathbf{t}(\mathbf{x}) = (x_1, \ldots, \prod_{j=0}^{[\log_2 i]} x_{[i/2^j]}, \ldots, \prod_{j=0}^{[\log_2 (k-1)]} x_{[(k-1)/2^j]})$ with $\hat{\mathbf{t}} = \mathbf{t}(\hat{\mathbf{x}})$ will give the ML estimate of $\mathbf{t}$.

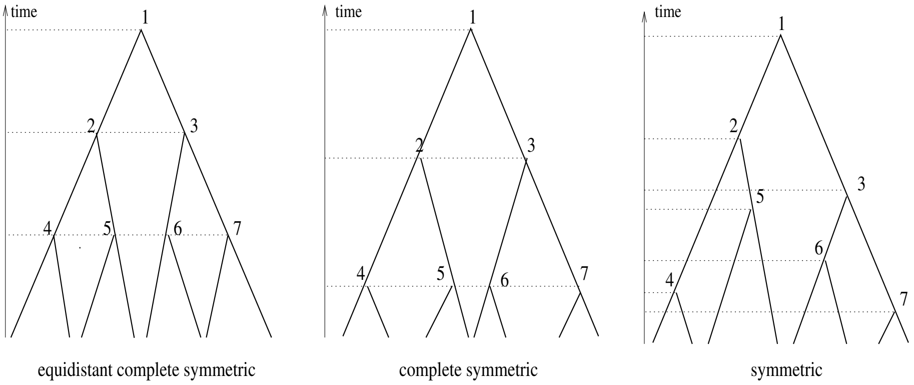

Figure 3: *The three different types of tree considered in the simulations. To the left is the equidistant complete symmetric case where all nodes at a level diverged at the same time and the times between the levels are equal. In the middle is the complete symmetric case where the times between the levels do not need to be equal. At the right is the third type which is symmetric in the sense that the two subtrees of a node have equally many nodes, but the times of divergence may differ between nodes.*

For different types of trees a starting tree with $k = 128$ taxa was created by simulating the time of divergences $\mathbf{t} = (t_1, \ldots, t_{127})$. The divergence time of the root, t_1, was always set to 1.0, implying relative times. The divergence times of the other nodes were simulated according to Table 1. To be able to compare the results for different k, $n\bar{r}$ should be held constant and was always set to 30. We then expect about 30 substitutions on a path from the root to a taxa.

type of tree	simulation algorithm
equidistant complete symmetric	(i) $l_i = 1 - \frac{i}{\log_2 k}, i = 1, \log_2 k - 1$ (ii) $t_j = l_i, j = 2^i, \ldots, 2^{i+1} - 1$
complete symmetric	(i) $u_1, \ldots u_{\log_2 k - 1} \in U(0, 1)$ (ii) $\mathbf{u}$ sorted from max to min (ii) $t_j = u_{(i)}, j = 2^i, \ldots, 2^{i+1} - 1$
symmetric	a) Start from the root. For the two branches: (i) $u_1, \ldots, u_{\log_2 k - 1} \in U(0, 1)$ (ii) each $u_{(i)}$ represents a divergence node b) For each divergence node, add a branch. (i) $u_1, \ldots, u_{l'} \in U(0, 1)$, where $l' = log_2 k - j$, where j is the level the node belongs to (ii) each $u_{(i)}$ represents a divergence node on the added branch. For each new divergence node that is not on the last level, goto b)

Table 1: *A tree was created by setting* $t_1 = 1.0$ *and simulating the divergence times of the other nodes according to the type of tree wanted. The number of taxa in the tree is* k.

For the tree created, the theoretical variances were calculated using formula (5) and by solving the inverse of (6) numerically. Then 500 data sets were simulated. For each dataset $\hat{t}_1^{(ML)}$ was calculated using the optimization routine of Nelder and Mead (see e.g. Mathews, 1987), and $\hat{t}_1^{(MPL)}$ was calculated using formula (4). The sample variances were then calculated from the 500 estimates.

The number of taxa were then divided by 2 by deleting a level. The level to be deleted was chosen randomly between the levels $\{2, \ldots, \log_2 k\}$. For each node at the chosen level, the left or the right subtree was deleted with probability 0.5 each. The resulting tree was then of the same type as the original tree, the divergence times of the nodes that exist in both trees maintained, but with half as many taxa as the original tree. New data sets were created and all calculations were done for the new tree. The number of taxa, k, were then divided again by deleting a level of nodes, the calculations were redone etc. until $k = 4$.

The equidistant complete symmetric tree was an exception of the this procedure as deleting a level in such a tree would result in a tree not being equidistant. For this type of tree we started with $k = 4$ taxa, calculated the theoretical variances, created 500 data sets for which the estimates were calculated. Finally the sample variances were calculated from the estimates. Then a new equidistant complete symmetric tree was created by squaring the number of taxa. This could be interpreted as adding two levels of nodes to the existing tree. New data sets were created, the calculations redone etc.

Table 2 summarizes the results from the equidistant complete symmetric case. The first box is where the tree started with 4 taxa, the second box where the starting tree had 8 taxa. As is shown in the appendix, the theoretical variances are equal for the two methods. Formula (10) showed that the variances should almost decrease with a factor 2 when the number of taxa were squared, a result that holds in the simulations.

	mean		sample variance		Theoretical variance	
k	$\hat{t}_1^{(ML)}$	$\hat{t}_1^{(MPL)}$	$\hat{t}_1^{(ML)}$	$\hat{t}_1^{(MPL)}$	I_{11}^{-1}	$V(\hat{t}_1^{(MPL)})$
4	0.99834	0.99820	0.01332	0.01328	0.0125	0.0125
16	1.00016	1.00057	0.00761	0.00756	0.0078	0.0078
256	1.00295	0.99895	0.00350	0.00400	0.0041	0.0041
8	0.99666	0.99686	0.01030	0.01029	0.0097	0.0097
64	1.00468	1.00391	0.00489	0.00477	0.0055	0.0055

Table 2: *Results from simulations of 500 datset for each tree in the equidistant complete symmetric case. Each box represents a tree where the number of taxa has been squared in each step.*

In the complete symmetric case the theoretical variances of $\hat{t}_1^{(ML)}$ and $\hat{t}_1^{(MPL)}$ are equal, but the numerical values depend on the simulated $\mathbf{t}$, and can therefore differ even if the same k and $n\bar{r}$ have been used. In Table 3 we present the results from two of the many simulated trees to visualize the dependence on $\mathbf{t}$. Each box represents a unique starting tree. In the last row of the box, where $k = 128$ the original starting tree has been used to create datasets. In the row with $k = 64$ a level of nodes has been deleted as described earlier and the resulting tree has been used to simulate data. The results within a box can therefore be compared, but not between boxes as another starting tree has been used in box 2.

The theoretical variances always decreased with increasing number of taxa. Sometimes though they can be quite close as for $k = 64$ and

$k = 128$ in the first box of Table 3. This can be explained by the level chosen to be deleted being close to taxa or to another level. The impact on the estimates from the deleted branches were then small. As a result the estimate of the variance, i.e. the sample variance, could then increase.

	mean		sample variance		Theoretical variance	
k	$\hat{t}_1^{(ML)}$	$\hat{t}_1^{(MPL)}$	$\hat{t}_1^{(ML)}$	$\hat{t}_1^{(MPL)}$	I_{11}^{-1}	$V(\hat{t}_1^{(MPL)})$
4	0.99551	0.99552	0.01652	0.01651	0.0158	0.0158
8	1.00302	1.00310	0.00781	0.00776	0.0083	0.0083
16	0.99705	0.99761	0.00638	0.00637	0.0066	0.0066
32	1.00051	0.99938	0.00544	0.00548	0.0055	0.0055
64	1.00154	1.00023	0.00364	0.00372	0.0041	0.0041
128	0.99852	0.99592	0.00364	0.00396	0.0040	0.0040
4	1.00278	1.00290	0.00905	0.00909	0.0094	0.0094
8	1.00201	1.00207	0.01002	0.00999	0.0092	0.0092
16	0.99966	0.99958	0.00999	0.00993	0.0090	0.0090
32	1.00216	1.00139	0.00701	0.00697	0.0078	0.0078
64	1.00643	1.00231	0.00635	0.00604	0.0060	0.0060
128	1.00383	0.99606	0.00509	0.00550	0.0057	0.0057

Table 3: *Results from simulations of 500 datset for each tree in the complete symmetric case. The theoretical variances, as well as the sample variances, depend on the* **t** *vector simulated for the unique tree represented by a box in the table.*

The table for the symmetric case (Table 4) should be read in the same manner as for the complete symmetric case (Table 3). The theoretical variances have been calculated according to the formulae (5) and (6) and depend on the simulated **t**. As we can see in the table, $V(\hat{t}_1^{(ML)}) < V(\hat{t}_1^{(MPL)})$, but the ratio between them are always less than 1.15. Only once, for all simulations we did with a symmetric tree was the ratio greater than 1.5 which happened for 128 taxa.

For the complete symmetric and the symmetric cases we have simulated many starting trees. For all trees, the theoretical variances decreased with increasing number of taxa. The sample variances were close to the theoretical ones, but for large k the sample variances for the method of Maximum Likelihood tend to be smaller than the theoretical ones. A reason could be that with large k, the number of nuicance parameters is large and the log likelihood function more complex. The numerical optimization algorithm may then have problems

	mean		sample variance		Theoretical variance	
k	$\hat{t}_1^{(ML)}$	$\hat{t}_1^{(MPL)}$	$\hat{t}_1^{(ML)}$	$\hat{t}_1^{(MPL)}$	I_{11}^{-1}	$V(\hat{t}_1^{(MPL)})$
4	1.00323	1.00225	0.00991	0.01008	0.0112	0.0113
8	0.99922	0.99895	0.00934	0.00950	0.0106	0.0107
16	1.00289	1.00347	0.00714	0.00763	0.0069	0.0072
32	1.00698	1.00531	0.00699	0.00747	0.0063	0.0067
64	1.00019	0.99698	0.00346	0.00377	0.0032	0.0035
128	1.00829	0.99946	0.00196	0.00301	0.0027	0.0031
4	0.99923	0.99910	0.01180	0.01206	0.0132	0.0132
8	1.00098	1.00181	0.01009	0.01033	0.0104	0.0107
16	0.99980	1.00037	0.00577	0.00587	0.0057	0.0057
32	1.00416	1.00436	0.00378	0.00379	0.0035	0.0036
64	1.00890	1.00457	0.00341	0.00348	0.0034	0.0035
128	1.00953	0.99962	0.00256	0.00334	0.0028	0.0030

Table 4: *Results from simulations of 500 datset for each tree in the symmetric case. The theoretical variances, as well as the sample variances, depend on the* $\mathbf{t}$ *vector simulated for the unique tree represented by a box.*

finding the global maximum and is stuck in one of the local maxima. For the same reason the much less complex estimation method of Mean Path Length seem to be closer to the true value for $k \geq 64$.

6 Concluding Remarks

In this paper we have investigated the method of Maximum Likelihood for estimating divergence time of a root. We have compared two ways of improving the estimate - either by analyzing longer sequences, that is adding more information by increasing the sequence length n, or by adding more information in form of new taxa with more recent divergence times than the root. The first approach is the usual one when investigating consistency and asymptotics. The number of parameters is fixed, here the divergence times of the internal nodes $(t_1, \ldots, t_{k-1})$ where k is the number of taxa, but n increases, theoretically $n \to \infty$. We have shown that in general, the variance of the estimate of the divergence time of a node decreases with a factor 2 if the sequence length is doubled. In the second approach n is large but fixed and the number of taxa k is increased implying that the number of parameters is increased. Since we only have finite sequences we cannot, even theoretically, let $k \to \infty$. If e.g. $k > n$ there is not enough data to estimate the number of substitutions along branches. Instead we consider the situtation where k is increased in a nice way and not too much so that the sequence length always is much larger than the number of taxa. If possible, to gain as much as possible, the taxa added ought to be as close to the root as possible. If the taxa added diverged recently, the precision is still improved, but not with the same amount. In a situation "in between" the two extremes, e.g. the equidistant case where all branches are of the same length, we have shown that the same reduction of the variance of the ML estimate can almost be achieved if the number of taxa is squared as if the sequence length is doubled, assuming the Jukes-Cantor model of evolution.

Throughout the paper we have assumed the number of substitutions along branches to be observable. Admittedly, this is rarely the case in practice, usually only DNA sequences at terminal nodes are observed and the number of substitutions along branches are estimated. We performed a simulation of DNA sequences of terminal taxa with equidistant complete symmetric trees under the Jukes-Cantor model and used PAUP* to analyze the sequences with the Maximum Likelihood method, Jukes-Cantor model and a global molecular clock. For 100 data sets the first 500 sites where analyzed as well as all 1000 sites. The results are summarized in Table 5 where the sample variances of the estimates of the divergence time of the root are calculated. The results verify our theoretical findings, the sample variance is approximately divided by two if the sequence length is doubled, and in the equidistant case the variance is almost divided by 2 when the number of taxa is squared from $k = 4$ to $k = 16$.

In the study we have focused on the divergence time of one given node, which we have denoted the root and restricted the study to symmetric trees. Such trees maintain the topology when taxa are added to the tree, making the precisions for different number of taxa comparable. The conclusion that precision is improved by adding more taxa holds for nonsymmetric trees too, but it is then harder to express how fast it is improved.

We have used the Jukes-Cantor model of evolution througout the paper when investigating the properties of Maximum Likelihood as a method of estimating the divergence time. It is a simple and unrealistic model, but it simplifies the complex calculations and make theoretical studies of the estimation method possible. In order to obtain the ML estimate of divergence times, the time vector $\mathbf{t}$ that maximizes the log likelihood function has to be computed numerically. The task to find the log likelihood function is harder with a more complex model, it will be harder to solve the score function numerically and to find the inverse of the information matrix. The method is time

α	n	k	$\hat{t}_1^{(ML)}$	$V(\hat{t}_1^{(ML)})$
0.02	500	2	1.032	0.0479
		4	1.021	0.0402
		16	0.994	0.0245
	1000	2	1.030	0.0241
		4	1.015	0.0219
		16	1.007	0.0129
0.03	500	2	0.967	0.0302
		4	0.979	0.0220
		16	0.985	0.0139
	1000	2	0.976	0.168
		4	0.978	0.0139
		16	0.987	0.0074

Table 5: *Results from simulations of 100 data sets for each tree, with substitution rate* $\alpha = 0.02$ *and* $\alpha = 0.03$ *respectively. The first 500 sites of the sequences as well as all 1000 sites have been analyzed with PAUP* using the Maximum Likelihood Method, Jukes-Cantor model and a global molecular clock. The estimates and variances given in the table are the means and the sample variances.*

consuming even for a moderate number of taxa and a simple model.

Another reason to use the model of Jukes-Cantor was to be able to compare the results of ML with the much less complex method of Mean Path Length which only works with that model. There is however a generalization of MPL, allowing different substitution rates in different segments of the tree in a software recently introduced called PATHd8 (Britton et. al. 2007).

The method of Mean Path Length is consistent for the Jukes-Cantor model. We have shown that for this method too the variance of the estimate decreases with increasing sequence length or increasing number of taxa.

If the tree is symmetric with the same divergence times across a level, both methods are efficient, they achieve the Cramér-Rao lower bound. This is shown analytically when all branches are of the same length and by simulations when the times between levels may differ. If the divergence time of the nodes at a level are individual, Maximum Likelihood has slightly better precision of estimating the divergence time of the root than Mean Path Length. MPL uses almost the same amount of information for nodes high in the tree, near the root, as the method of ML. The advantages of MPL for those nodes are that the algorithm is very fast and it does not use a numerical optimization routine that can get stuck in local maxima. For internal nodes lower down though, Maximum Likelihood is the better but slower method as it uses more information than the method of Mean Path Length.

Appendix

In this appendix we will show that the variances of the estimators of the divergence time of the root of a tree that in fact is an equidistant complete symmetric one are equal for the methods of Maximum Likelihood and Mean Path Length. We will also show that the estimators achieve the Cramér-Rao lower bound.

In the equidistant complete symmetric case with $k = 2^l$ taxa all nodes on a level have the same divergence time. Further the times between the levels are equal, $t_1/l = t_1/\log_2 k$, where t_1 is the time of the root. Figure 4 shows two equidistant trees with $k = 4$ and $k = 8$ taxa respectively.

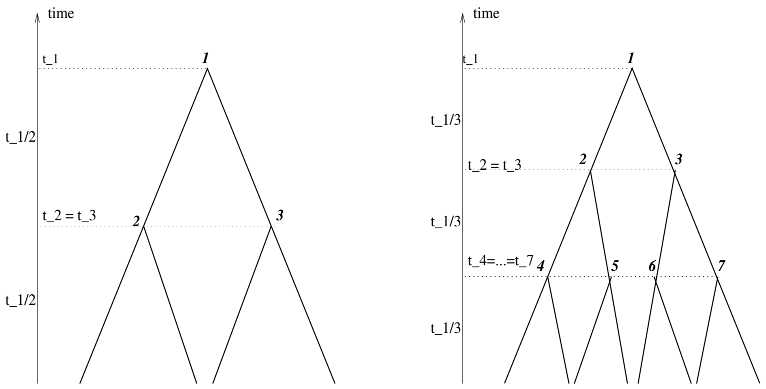

Figure 4: *To the left is a equidistant tree with k=4 taxa. The time between the divergences is then* $t_1/\log_2 k = t_1/2$*. To the right the tree has* $k = 8$ *taxa and the branches are then of length* $t_1/\log_2 8 = t_1/3$*.*

In such a tree, the estimator of t_1 using the method of Mean Path Length, $\hat{t}_1^{(MPL)}$, is as good as the estimator using Maximum Likelihood, $\hat{t}_1^{(ML)}$, as the next theorem will show. The methods of course do not assume the tree to be equidistant nor complete.

Theorem: If the number of taxa in a rooted phylogenetic tree is $k = 2^l$ and all branches are of the same length $(t_1/\log_2 k)$, where t_1 is the time of the root, then

$$V(\hat{t}_1^{(MPL)}) = V(\hat{t}_1^{(ML)}) = I_{11}^{-1} = \tfrac{t_1}{n\bar{r}\log_2 k}\tfrac{k-1}{k},$$

where I is the information matrix. Hence $V(\hat{t}_1^{(ML)}) = V(\hat{t}_1^{(MPL)})$ in the equidistant complete symmetric case and both estimators are efficient.

Proof:

Since all branches have the same length, $t_1/\log_2 k$, the variance of the time estimate of the root using the method of Mean Path Length in the equidistant complete symmetric case is, according to (5),

$$V(\hat{t}_1^{(MPL)}) \quad = \quad \frac{t_1}{n\bar{r}\log_2 k}\frac{k-1}{k}, \tag{11}$$

Using formula (6) the information matrix will have the elements

$$I_{t_i t_j}(\mathbf{t}) =$$
$$\begin{cases} \frac{n\bar{r}}{t_1-t_2} + \frac{n\bar{r}}{t_1-t_3} = \frac{2n\bar{r}\log_2 k}{t_1} & i = j = 1 \\ \frac{n\bar{r}}{t_{[i/2]}-t_i} + \frac{nr}{t_i-t_{2i}} + \frac{n\bar{r}}{t_i-t_{2i+1}} = \frac{3n\bar{r}\log_2 k}{t_1} & i = j, i = 2, \ldots, \frac{k}{2}-1 \\ \frac{n\bar{r}}{t_{[i/2]}-t_i} + \frac{2nr}{t_i} = \frac{3n\bar{r}\log_2 k}{t_1} & i = j, i = \frac{k}{2}, \ldots, k-1 \\ -\frac{n\bar{r}}{t_j-t_i} = -\frac{n\bar{r}\log_2 k}{t_1} & j = [i/2], \\ -\frac{n\bar{r}}{t_j-t_i} = -\frac{n\bar{r}\log_2 k}{t_1} & j = (2i, 2i+1), \end{cases}$$

where $i = 2, \ldots, k-1$ in the last two rows. The information matrix can hence be written as $I = n\bar{r}\log_2 kE$, where E is a matrix describing the tree topology. The diagonal elements of E will tell how many branches that are connected to the nodes (2 for the root and 3 for the rest). For $i \neq j$ the elements will be 0 if node i is not connected to node j and -1 if it is.

The E-matrix can be divided into blocks A, B, C and D such that

$$E = \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right),$$

where A describes the relationships between all internal nodes except the ones on the last level, right above the taxa, i.e. nodes $\{1, \ldots, \frac{k}{2} - 1\}$. D describes the relationships between the nodes right above the taxa, i.e. nodes $\{\frac{k}{2}, \ldots, k-1\}$. Since they are independent, D will be a diagonal matrix with the diagonal elements equal 3. B describes the relationships between the nodes on the last level and the rest of the nodes, $C = B^T$.

By multiplication it can be verified that the inverse of E, E^{-1}, can be expressed as

$$E^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} - D^{-1}C(A - BD^{-1}C)BD^{-1} \end{pmatrix}. \tag{12}$$

Since we are only interested in the first element of this inverse, we concentrate on the upper left part. As $C = B^T$ and $D = 3\mathcal{I}$, where $\mathcal{I}$ is the identity matrix, it follows that $BD^{-1}C = BD^{-1}B^T = \frac{1}{3}BB^T$. The elements of B are 0 except for the places corresponding to a branch between two nodes where the element is -1. BB^T will be a diagonal matrix with 0 or 2 at the diagonal as

$$\begin{aligned} (BB^T)_{ij} &= \sum_m b_{im} b^T_{mj} \\ &= \sum_m b_{im} b_{jm} \\ &= \begin{cases} 0 & i \neq j \\ 0 & i = j, i = 1, \ldots, \frac{k}{4} - 1 \\ 2 & i = j, i = \frac{k}{4}, \ldots, \frac{k}{2} - 1 \end{cases} \end{aligned}$$

The elements equal 2 will correspond to the nodes of the second last time level of the tree.

Let $E_{(1)} = (A - \frac{1}{3}BB^T)$. As noted earlier, the submatrix A describes the relationships between all nodes except the ones on the last level, right above the taxa. Finding the inverse of $E_{(1)}$ is then equivalent to finding the inverse of the corresponding E matrix for a equidistant tree where the nodes on the last level are deleted, i.e. for a tree with $k^* = k/2$ taxa, but with the diagonal elements of the submatrix D^* equal $3 - \frac{2}{3} = \frac{7}{3}$ instead of 3.

With the same arguments as above, finding the inverse of the upper left part of $E_{(1)}$ results in finding the inverse of $E_{(2)} = A_{(1)} - \frac{3}{7}B_{(1)}B^T_{(1)}$, which is equivalent with finding the inverse of $A_{(2)}$ where the elements of the lower right sub matrix have changed to $3 - 2 \cdot \frac{3}{7}$. This will go on, reducing the problem to finding the inverse of $E_{(j)}$ with change of the diagonal elements of the lower right sub matrix until we reach the matrix $E_{(\log_2 k - 1)}$. We then have the problem of finding the inverse to

$$E_{(\log_2 k-1)} = \left(\begin{array}{c|cc} 2 & -1 & -1 \\ \hline -1 & a & 0 \\ -1 & 0 & a \end{array}\right) \tag{13}$$

where a is 3 minus 2 times the inverse of the lower right part of the previous step. The first element of the inverse of $E_{(\log_2 k-1)}$ is then, according to (12) $(2-\frac{2}{a})^{-1} = (\frac{2a-2}{a})^{-1} = \frac{a}{2a-2}$. We can calculate the a element recursively like the following:

Starting with the matrix for the tree with $\log_2 k$ time levels, let a_j denote the diagonal element of the matrix A at step j, that is $a_0 = 3$, $a_1 = 3 - 2a_0^{-1} = \frac{7}{3}$, $a_2 = 3 - 2a_1^{-1} = \frac{15}{7}$ etc. Rewrite this as

$$\begin{aligned} a_1 = \frac{7}{3} &= \frac{2^{1+2}-1}{2^{1+1}-1} \\ a_2 = 3 - 2a_1^{-1} &= 3 - \frac{2(2^{1+1}-1)}{2^{1+2}-1} = \frac{2^{2+2}-1}{2^{1+2}-1} \end{aligned}$$

Assume that a_{j-1} can be written as $\frac{2^{(j-1)+2}-1}{2^{(j-1)+1}-1}$. Then

$$\begin{aligned} a_j &= 3 - 2a_{j-1}^{-1} = 3 - \frac{2(2^{(j-1)+1}-1)}{2^{(j-1)+2}-1} \\ &= 3 - \frac{2(2^j-1)}{2^{j+1}-1} \\ &= \frac{(3-1)2^{j+1}-1}{2^{j+1}-1} \\ &= \frac{2^{j+2}-1}{2^{j+1}-1}. \end{aligned}$$

Hence, by induction proof, it is shown that $a_j = \frac{2^{j+2}-1}{2^{j+1}-1}$, where $j = 1, \ldots \log_2 k - 2$. At this stage we have the situation in (12) and the first element of the inverse of $E_{(\log_2 k 1)}$ can be calculated as

$$\begin{aligned} \{E^{-1}_{(\log_2 k-1)}\}_{11} &= (2 - 2a_{\log_2 k-2})^{-1} \\ &= \left(2 - \frac{2(2^{\log_2 k-1}-1)}{2^{\log_2 k}-1}\right)^{-1} \\ &= \left(2 - \frac{2(\frac{k}{2}-1)}{k-1}\right)^{-1} \\ &= \frac{k-1}{k}. \end{aligned} \tag{14}$$

The first element of inverse of the information matrix $I = \frac{n\bar{r}\log_2 k}{t_1}E$ will then be $\frac{t_1}{n\bar{r}\log_2 k}\frac{k-1}{k}$, which is equal to (11). Hence the estimator of the Mean Path Length for the divergence time of the root achieves the Cramér Raos lower bound.

References

Bremer, K. and Gustafsson, M.H.G., 1997, East Gondwana ancestry of the sunflower alliance of families. plants., *Proc. Natl. Acad. Sci.* USA, **94, 9188-9190**

Britton, T., 2005, Estimating Divergence Times in Phylogenetic Trees Without a Molecular Clock, *Systematic Biology*, **54:3, 500-507**

Britton, T., Anderson, C.L., Jaquet, D., Lundquist, S. and Bremer, K., 2007, Estimating Divergence Times in Large Phylogenetic Trees, *Systematic Biology*, **56:5, 741-752**

Britton, T., Oxelman, B., Vinnersten, A. and Bremer, K., 2002, Phylogenetic dating with confidence intervals using mean path lengths, *Molecular Phylogenetics and Evolution*, **24: 58-65**

Edwards, A.W.F. and Cavalli-Sforza, 1964, Reconstruction of evolutionary trees, **p. 67-76** in *Phenetic and Phylogenetic Classification* ed. V.H. Heywood and J.MacNeill, Systematics Association Publ. No 6, London

Felsenstein, J., 1981, Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution, **17: 368-376**

Felsenstein, J., 2005, PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Jukes, T.H and Cantor, C.R. 1969. Evolution of protein molecules. In Munro, H.N. ed. *Mammalian Protein Metabolism*, **pp. 21-132**, Academic Press, New York

Lindsey, J.K., 2001, Parametric Statistical Inference, Oxford University Press, Inc., New York

Mathews, J.H., 1987, Numerical Methods for computer science, engineering and mathematics, Prentice-Hall, Inc.

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