ACADEMIC PRESS

Molecular Phylogenetics and Evolution 24 (2002) 58–65

MOLECULAR PHYLOGENETICS AND EVOLUTION

www.academicpress.com

Phylogenetic dating with confidence intervals using mean path lengths

Tom Britton,[a,*] Bengt Oxelman,[b] Annika Vinnersten,[b] and Kåre Bremer[b]

[a] *Department of Mathematics, Uppsala University, P.O. Box 480, SE-751 06 Uppsala, Sweden*
[b] *Department of Systematic Botany, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, SE-752 36 Uppsala, Sweden*

Received 11 June 2001; received in revised form 7 November 2001

Abstract

The mean path length (MPL) method, a simple method for dating nodes in a phylogenetic tree, is presented. For small trees the age estimates and corresponding confidence intervals, calibrated with fossil data, can be calculated by hand, and for larger trees a computer program gives the results instantaneously (a Pascal program is available upon request). Necessary input data are a rooted phylogenetic tree with edge lengths (internode lengths) approximately corresponding to the number of substitutions between the nodes. Given this, the MPL method produces relative age estimates with confidence intervals for all nodes of the tree. With the age of one or several nodes of the tree being known from reference fossils, the relative age estimates induce absolute age estimates and confidence intervals of the nodes of the tree. The MPL method relies on the assumptions that substitutions occur randomly and independently in different sites in the DNA sequence and that the substitution rates are approximately constant in time, i.e., assuming a molecular clock. A method is presented for identification of the nodes in the tree at which significant deviations from the clock assumption occur, such that dating may be done using different rates in different parts of the tree. The MPL method is illustrated with the Liliales, a group of monocot flowering plants.

Keywords: Confidence interval; Liliales; Molecular clock; Phylogenetic dating; Phylogenetic tree; Poisson process; Reference fossil

1. Introduction

Phylogenetic trees of various groups of organisms are today common components of many evolutionary research projects. With the increasing amount of available DNA sequence data, phylogenetic trees are also becoming more and more robust and reliable. Given sufficient data, the production of large and well-supported phylogenetic trees is now a more or less standard procedure (e.g., Hillis et al., 1996). A logical next step in phylogenetic reconstruction is dating of the nodes in the trees. This is, however, far from a standard procedure. A host of methods have been developed (reviewed by Sanderson, 1998) but there are few simple and straightforward methods for dating the nodes of large phylogenetic trees involving perhaps several hundred terminals. One such method is the mean branch length method of Bremer and Gustafsson (1997). Herein, we discuss this method and describe a new approach for adding confidence intervals to the estimates. It provides a simple procedure for obtaining age estimates with confidence intervals for the nodes in a large phylogenetic tree. We adopt the mathematical terminology, call a single segment internode in the tree an *edge* and a sequence of edges from a node up to a terminal a *path*, and, as a consequence, denote the method of Bremer and Gustafsson (1997) the mean path length (MPL) method.

2. Methods

2.1. The MPL method

A data matrix of aligned DNA sequences for the terminals can be translated into a rooted dichotomously branching tree using a variety of methods (e.g., parsimony, distance, and likelihood methods). The branching tree is the basis for phylogenetic dating in the MPL method. The length of an edge in the tree is associated

* Corresponding author. Fax: +46-18-471-3201.
E-mail address: tom.britton@math.uu.se (T. Britton).

1055-7903/02/$ - see front matter
PII: S1055-7903(02)00268-3

with the dissimilarity between the pair of connected nodes, e.g., in number of substitutions. The accumulation of changes in gene sequences over time is thus manifested as increasing edge lengths, making these correlated with time. To make edge lengths directly proportional to time durations, a distance transformation accounting for superimposed events is often performed. There are various transformations suggested, for example the Jukes–Cantor model, depending on the specific application; see Swofford (1998). The edge lengths are, however, affected by stochastic variation, differences in substitution rates, and systematic biases introduced by the phylogenetic methods themselves. To reduce randomness when estimating relative ages for nodes in a phylogenetic tree, Bremer and Gustafsson (1997) suggested that the relative age of a node could be estimated by the mean path length from the node to the terminals descended from that node. Using one or more reference fossils that may be attached to specific nodes of the tree, an observed change rate may then be calculated by dividing the MPL estimate of a reference node with the age of the fossil attaching at that node. Ages for other nodes are then estimated by dividing their mean path lengths with the observed change rate. If more than one reference fossil is available, either one can use the average change rate from the fossils or, if there are reasons to assume different change rates in different parts of the tree, the tree may be divided into disjoint parts having different change rates in each part.

2.2. Statistical properties of the MPL method

The MPL method described above aims to reduce the uncertainty of the estimated length from a node in the tree to its descendant terminals. This is achieved by using all the paths to terminals from the node in question, the estimated length from the node being the average length of these paths. The sample variance s^2 of these lengths should not be used when constructing confidence intervals, since s^2 is only suitable when the sample consists of independent observations. This is not the case in the present context. In fact, two terminals with a recent common ancestor (i.e., closely linked in the tree) share most of the path from a node lower down in the tree, thus making the two path lengths highly dependent.

The MPL method estimates the relative age of a node by its mean path length assuming the phylogenetic tree to be given, where input edge lengths correspond to the number of substitutions along the edge, adjusted for superimposed substitutions. To obtain a confidence interval for the estimate some assumptions about how data were generated have to be made. The MPL method has two assumptions. The first assumption is that substitutions occur randomly and independently in time and independently in different sites in the DNA sequence. However, it is not assumed that the substitution rates are the same for different sites. The second assumption is that the substitution rates (the rates at which new substitutions are expected) are approximately constant in time and hence the same in different lineages, meaning that there is a molecular clock running. It is hence not suited to a very non-clock-like tree with observed substitution differences among branches not attributable to stochastic variation.

A consequence of the two assumptions is that substitutions, along any lineage, occur randomly according to a Poisson process with an unknown constant rate r being the sum of the rates for all sites (e.g., Ross, 1997). The number of substitutions m along an edge (or sequence of edges) of time duration t is thus an outcome of a Poisson distribution with mean rt. Below we will make use of this and the fact that the mean and variance are identical for the Poisson distribution. Even if the assumptions are not entirely true, for example if there is some dependence between nearby sites or if the substitution rate at a site depends on its nucleotide, the fact that substitutions at many sites are considered simultaneously implies that they occur approximately according to a Poisson process. The crucial, and somewhat restrictive assumption, is that the substitution rate r remains approximately constant over time and hence for different species, the molecular clock assumption.

Consider a fixed node in the tree having k terminals and let $x_1, \ldots, x_k$ denote the accumulated edge lengths (i.e., the path lengths) from the node to the different terminals. The MPL estimate for that node is then

$$x = (x_1 + \cdots + x_k)/k.$$

As an example, consider node 7 of Fig. 1 (labels are given in the right tree and edge lengths in the left tree). There are $k = 3$ terminals descending from that node: *Uvularia perfoliata*, *Uvularia pudica*, and *Disporum*. The path lengths from node 7 to these nodes are $x_1 = 15$, $x_2 = 3 + 8 = 11$, and $x_3 = 3 + 20 = 23$, respectively. The mean path length is hence $x = (15 + 11 + 23)/3 = 16.3$.

Because each of the paths has a Poisson-distributed number of substitutions, each with mean ra, where a is the age of the node, the MPL estimate is an unbiased estimator of the relative age ra. To obtain an expression for the variance of estimate we write x in a different way. This is done using the separate input edge lengths of the subtree having the node of interest as the root. Label the s edges of the subtree $1, \ldots, s$, and denote the input it edge lengths (the number of substitutions) by $b_1, \ldots, b_s$. Let $n_1, \ldots, n_s$ denote the number of paths traversing the corresponding edges. Then it is easily verified that a different way of writing x is

$$x = (n_1 b_1 + \cdots + n_s b_s)/k.$$

We illustrate this with our example from Fig. 1. The edges from node 7 have lengths $b_1 = 15$, $b_2 = 3$, $b_3 = 8$,

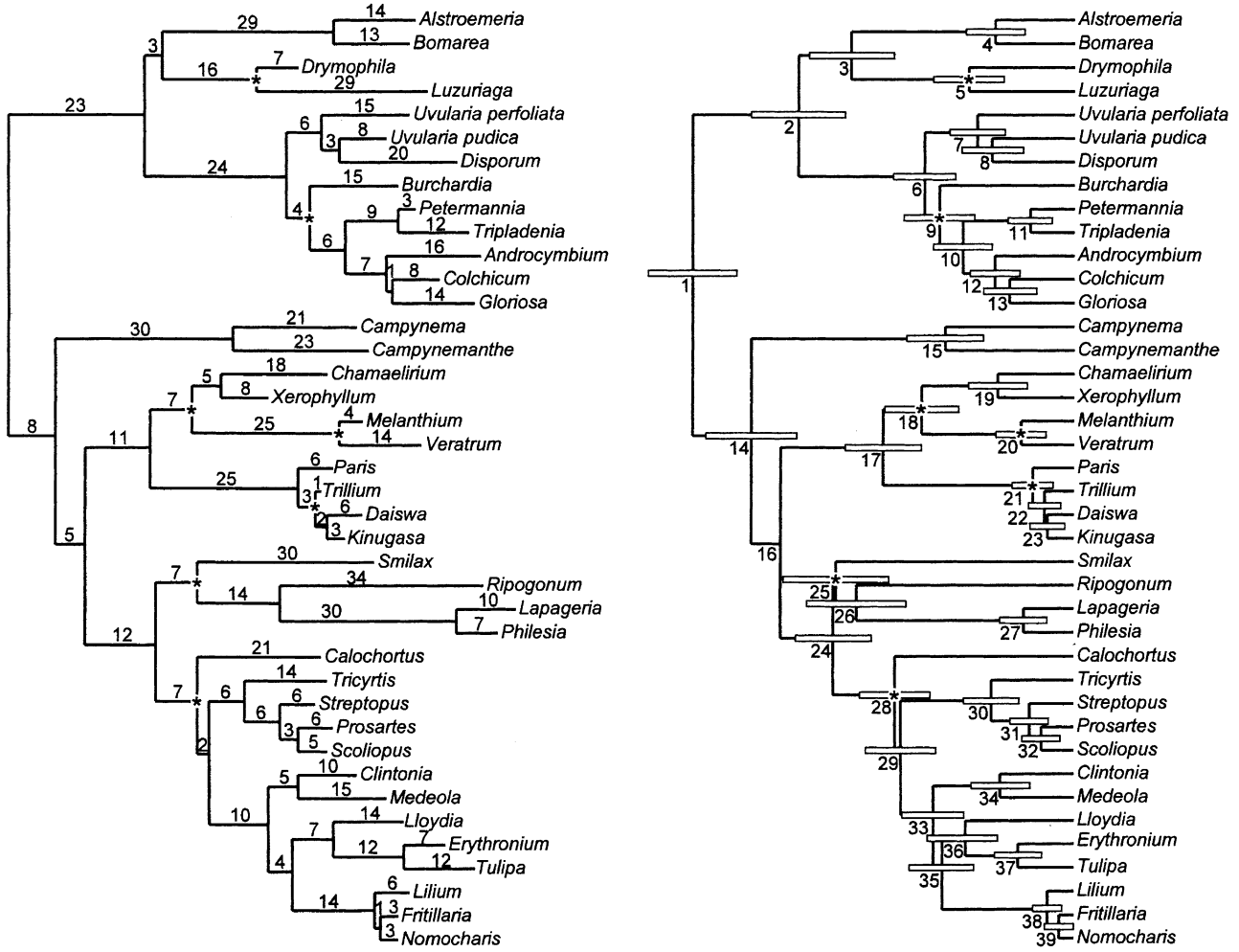

Fig. 1. (Left) Input phylogenetic tree of the monocot flowering plant group Liliales from Vinnersten and Bremer (2001) with edge lengths obtained from parsimony optimization of plastid DNA *rbcL* sequences. (Right) MPL tree with confidence intervals shown as white bars; node numbers refer to Table 1 where mean path lengths and confidence intervals for the nodes are reported; 1 cm = 15 Myr. Nodes marked with asterisks have subtrees with significantly different path lengths, indicating deviations from a molecular clock. When multiple testing is taken into account, however, it is only in node 28 that the subtrees have significantly different molecular clocks (see text for details).

and $b_4 = 20$ and the number of paths traversing the edges are $n_1 = 1$, $n_2 = 2$, $n_3 = 1$, and $n_4 = 1$, respectively. Clearly, $(1 \times 15 + 2 \times 3 + 1 \times 8 + 1 \times 20)\ /3 = 16.3 = x$, illustrating that the two ways of writing the mean path lengths are equivalent.

The reason for writing x in terms of edge lengths instead of path lengths is that the different edge lengths, $b_1, \ldots, b_s$, are independent by assumption, whereas the path lengths $x_1, \ldots, x_k$ are not. As a consequence we get $\mathrm{Var}(x) = (n_1\mathrm{Var}(b_1) + \cdots + n_s^2\mathrm{Var}(b_2))/k^2$. Because b_i is Poisson distributed, having variance equal to the mean, we can estimate the variances by the observed values:

$$s^2(x) = (n_1^2 b_1 + \cdots + n_s^2 b_s)/k^2.$$

The standard error for x is thus $s(x) = \sqrt{s^2(x)}$. Further, because x is a weighted sum of several (Poisson) random variables it is approximately normally distributed. The approximation is satisfactory whenever the total number of substitutions in the subtree exceeds 20. A 95% confidence interval for the MPL estimate is hence

$$x \pm 1.96 s(x).$$

For example, the standard error for the mean path length of node 7 in Fig. 1 is $s(x) = \sqrt{[(1^2 \times 15 + 2^2 \times 3 + 1^2 \times 8 + 1^2 \times 20)/3^2]} = 2.4$, and the 95% confidence interval is 16.3 ± 4.8. There are, in general, two possible ways to reduce the standard error, thus increasing precision. One way is to have more taxa in the tree so that the mean (path length) is taken over a larger number of paths. Another way is to include more genes in the underlying DNA sequences with the effect that edge lengths (the number of substitutions) become larger.

Under the assumptions of the model the MPL method gives an unbiased estimate of the relative ages of nodes as mentioned above. There are of course other

possible unbiased methods to estimate the relative age of a node, for example by simply picking one path length from the node at random as an estimate of the relative age of the node. A relevant question therefore whether the MPL method is an efficient way to estimate ages or whether there are other methods that give tighter confidence bounds. The answer is that the MPL method is not optimal (i.e., most efficient among all unbiased methods) but that the optimal method is very complicated to derive in a large tree, as the following small tree example aims at illustrating. Consider a completely clock-like tree having four terminals. The two edges from the root have lengths a and b. The node at length a from the root has two terminal edges of equal length c, and the node at length b from the root has two terminal edges of equal length $a+c-b$, making the tree completely clock-like (see Fig. 2). The MPL method gives equal weight ($=1/2$) to the edges of length a and b from the root. This seems natural if a and b are approximately equal (left tree in Fig. 2), but what if a is large and b and c are very small (right tree in Fig. 2)? Then the tree looks almost like a tree having three edges from the root, suggesting that the edge of length a should only be given weight 1/3. It can be shown that, for any a, b, and c, the optimal method gives weight $(a+b+c)/(3a+b+2c)$ to the edge having length a and weight $(2a+c)/(3a+b+2c)$ to the edge with length b in this situation. This weight will always lie between 1/3 and 2/3 and indeed gives weight 1/2 if $a=b$ and weight 1/3 if $b=c=0$ to the edge of length a. To derive optimal weights in a large tree is a very cumbersome task. The MPL method, on the other hand, is much simpler and easier to understand, thus being worth the price of a small loss in efficiency.

2.3. Identification of clock deviations

A useful corollary of the MPL method is that it provides a way to test whether the molecular clock assumption is reasonable not only over the whole tree (as in most other methods) but at each particular node. Deviations from the molecular clock may be identified, such that the tree may be divided into separate parts with different clocks. One simple such test, easily performed simultaneously with the confidence interval construction, is now described.

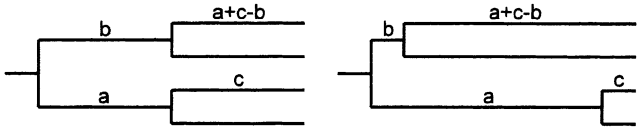

Fig. 2. Two small phylogenetic trees. The MPL method gives equal weight to the upper and lower subtrees in both examples, whereas the optimal method gives equal weight to the upper and lower subtrees in the left tree, but higher weight to the upper subtree in the right tree.

For a fixed node with k terminals the test compares the mean path lengths of the two subtrees rooted at the node and tests whether these are significantly different from each other, indicating whether the clocks are different in the two parts. Suppose the edges $1,\ldots,s$ of the subtree rooted at the node in question are labeled such that edges $1,\ldots,r$ belong to the first subtree (having j terminals) and edges $r+1,\ldots,s$ belong to the second (having k–j terminals). The mean path length from the node to the terminals of the first tree is then $x_1=(n_1b_1+\cdots+n_rb_r)/j$ and similarly $x_2=(n_{r+1}b_{r+1}+\cdots+n_sb_s)/(k-j)$ for the second subtree. These estimates are approximately normally distributed with estimated variances, $s^2(x_1)$ and $s^2(x_2)$, obtained using the same methods as above.

The test rejects the hypothesis that the same molecular clock applies to the two subtrees at $\alpha=5\%$ significance level if the absolute value of $z=(x_1-x_2)/\sqrt{[s^2(x_1)+s^2(x_2)]}$ exceeds 1.96. The denominator in the expression is the standard error of x_1-x_2, and ±1.96 is the limit for a symmetric 95% confidence interval of the standard normal distribution. The P value corresponding to the test is obtained by computing the probability that a standard normal variate obtains a value as large or larger than the observed z value.

As an example, node 7 of Fig. 1 has two subtrees, the first only consisting of one edge with length $b_1=15$, implying that the mean path length of that subtree is $x_1=15$, having estimated variance $s^2(x_1)=15$. The second subtree has three edges, two paths, and mean path length $x_2=(2\times3+1\times8+1\times20)/2=17$. The variance is estimated by $s^2(x_2)=(2^2\times3+1^2\times8+1^2\times20)/2^2=7$. The z-value for the example is hence $z=(15-17)/\sqrt{[15+7]}=-0.426$. The absolute value 0.426 is less than 1.96 and hence the clock assumption is not rejected at this node. The probability (P) that $|z|>0.426$ is $P=0.668$, a P value far from significant. Since the P value is not significant the two subtrees may be analyzed with the same molecular clock. The test procedure just described is performed by the Pascal program which computes the MPL estimates and confidence intervals.

In a large tree many such tests, one for each node, will be performed and some P values can be small purely by chance. This complicated statistical problem is known as the mass-significance phenomenon. A procedure that can be applied when several tests are to be performed simultaneously is known as the sequentially rejective Bonferroni test (Holm, 1979). This method consists of ordering the P values from the various tests and sequentially comparing them with the significance level α ($=0.05$ for example) as follows. If the smallest of all P values is larger than the significance level α divided by the total number of tests ($=$number of nodes in the present context), then none of the tests are significant. Otherwise, the test corresponding to the smallest P value

is rejected, and the procedure is repeated for the second smallest P value, but this time the significance level α is divided by the number of nodes minus 1. The procedure is repeated sequentially until the first time that the P value is larger than what it is compared with. The method is illustrated on the example presented below.

It is worth noting that the sequentially rejective Bonferroni test is conservative in that it will not falsely reject a hypothesis with a probability larger than the chosen significance level α. However, it may very well falsely accept a hypothesis with an unnecessarily large probability.

2.4. Computer program

The calculations are simple enough to be done by hand but for larger trees a program is useful. We provide a Pascal program for calculation of mean path lengths with confidence intervals from all nodes to the terminals of an input tree (written in NEXUS format; Maddison et al., 1997). The program is available upon request from the first author. Since the method has best performance when the input tree is clock-like, meaning that input edge lengths are proportional to time except for random fluctuations, a distance transformation may be utilized to meet this criterion. The program reports the estimated ultrametric tree from the MPL method with recalculated edge lengths, mean path lengths with a 95% confidence interval for all nodes to their terminals, and z values for all nodes to determine whether the two subtrees of any node have significantly different clocks. The latter situation occurs when the absolute z value exceeds 1.96. However, as discussed above, it should be noted that due to the mass significance phenomenon, in large trees with many nodes the clock assumption will not be rejected for all nodes with $|z| > 1.96$.

In trees where input path lengths vary greatly for paths from the same node, thus violating the clock assumption, it is possible that a mean path length from one particular node turns out to be shorter than that from a subsequent node in the tree. This paradoxically implies that a group is younger than one of its subgroups. As a consequence the recalculated edge length attains a negative value. In such cases the program reports a zero value for negative recalculated edge lengths and reduces the next recalculated edge lengths in the tree with the negative value to preserve the ultrametricity of the MPL tree. The rationale for this modification is simply that variation in input edge length should not overrule the predefined branching sequence of the tree.

3. Results

We exemplify with the Liliales, a group of monocot flowering plants including lilies (*Lilium*), tulips (*Tulipa*), and autumn crocus (*Colchicum*). Vinnersten and Bremer (2001) analyzed age and biogeography of the Liliales using a phylogeny based on 40 plastid DNA *rbcL* gene sequences, each sequence comprising 1428 positions (the first 26 being excluded from analysis). Their phylogenetic tree with edge lengths obtained from standard parsimony analysis of the *rbcL* data is shown in the left tree of Fig. 1. This is also the input tree for our MPL calculation with confidence intervals. The output tree with its mean path lengths and their 95% confidence intervals, obtained from the simple methods of the present paper, is shown in the right tree of Fig. 1. The MPL values are reported in Table 1. Vinnersten and Bremer also used the MPL method, however, without the calculation of confidence intervals presented here. Instead they obtained confidence intervals by bootstrap resampling and reanalysis of the data as suggested by Sanderson (1997) for his nonparametric rate smoothing method (also given in Table 1 as comparison).

The clock test revealed that 7 of the 39 nodes have subtrees with significantly different path lengths at the 5% level, i.e., with absolute P values < 0.05 and hence $|z|$ values > 1.96, indicating deviations from the clock. These P values are reported in Table 1 and the corresponding nodes are marked with asterisks in Fig. 1. If the sequentially rejective Bonferroni test (Holm, 1979) is applied to this multiple test problem the smallest P value ($= 0.000216$ for node 28) is first compared with $0.05/39 = 0.0012$ since there are 39 nodes in the tree. Because $0.000216 < 0.0012$ we conclude that the clocks of the two subtrees descending from node 28 have significantly different substitution rates. The second smallest P value ($= 0.00270$ for node 5) is then compared with $0.05/38 = 0.0013$, but since $0.00270 > 0.0013$ we cannot reject this nor any of the remaining tests of the molecular clock hypothesis. The conclusion is hence that only in node 28 of the tree is there clear evidence that the two subtrees descending from the node have significantly different molecular clocks.

To transform mean path lengths to ages Vinnersten and Bremer (2001) used a previous dating (Bremer, 2000) of the base node of Liliales to 82 Myr (a minimum age constraint). Since the mean path length from the base node is 65.2 steps (see Table 1) the observed (maximum) change rate is $65.2/82 = 0.79$ steps per Myr. Hence, the mean path length values including the confidence intervals reported in Table 1 may be transformed to (minimum) ages in Myr by dividing by 0.79. It should be noted that the dating is not fully justified in the nodes above (to the right of) node 28. This is because the multiple test procedure rejected the hypothesis that the same change rate applied to both subtrees above node 28. Ideally, individual change rates obtained from separate reference fossils should be used for both of these subtrees if dating in this part of the tree is important.

Table 1
Node numbers from right tree of Fig. 1, mean path lengths from node to terminals, 95% confidence intervals obtained by the MPL method, 95% confidence intervals obtained by bootstrap analysis (Vinnersten and Bremer, 2001), and *P* values for test of whether the two subtrees above the node have the same molecular clock

Node number	MPL	MPL conf. interval	Bootstrap conf. interval	*P* value
1	65.2	±7.6	±8.3	>0.05
2	47.0	±8.0	±9.7	>0.05
3	38.2	±7.6	±8.3	>0.05
4	13.5	±5.1	±7.2	>0.05
5	18.0	±5.9	±4.5	0.00270
6	25.6	±5.2	±6.4	>0.05
7	16.3	±4.8	±3.9	>0.05
8	14.0	±5.2	±3.9	>0.05
9	23.2	±5.9	±6.4	0.0214
10	18.8	±4.9	±5.2	>0.05
11	7.5	±3.8	±3.3	>0.05
12	13.3	±4.2	±3.7	>0.05
13	11.0	±4.6	±4.7	>0.05
14	54.9	±7.8	±7.4	>0.05
15	22.0	±6.5	±4.1	>0.05
16	50.1	±7.1	±7.4	>0.05
17	32.6	±6.6	±6.2	>0.05
18	26.0	±6.3	±5.4	0.00693
19	13.0	±5.0	±4.9	>0.05
20	9.0	±4.2	±3.4	0.0132
21	7.2	±3.5	±4.0	0.0357
22	4.7	±2.8	±2.6	>0.05
23	4.5	±2.9	±3.4	>0.05
24	41.1	±6.6	±7.6	>0.05
25	45.8	±8.9	±7.8	0.0357
26	37.0	±8.5	±6.6	>0.05
27	8.5	±4.0	±4.0	>0.05
28	30.5	±6.1	±6.9	0.000215
29	29.3	±5.9	±6.6	>0.05
30	13.8	±4.8	±4.5	>0.05
31	7.7	±3.5	±3.3	>0.05
32	5.5	±3.3	±3.0	>0.05
33	24.1	±5.3	±5.5	>0.05
34	12.5	±4.9	±4.2	>0.05
35	22.3	±5.5	±5.1	>0.05
36	19.0	±5.9	±5.4	>0.05
37	9.5	±4.3	±4.4	>0.05
38	4.7	±2.6	±2.3	>0.05
39	3.0	±2.4	±2.3	>0.05

Only $P = 0.000215$ for node 28 is significant at the 5% level when multiple testing is taken into account (see text for details). The MPL and bootstrap values may be transformed to ages in Myr by dividing by 0.79 (see text for details).

4. Discussion

Phylogenetic dating involves many sources of error. Perhaps the most widely discussed source of error is the deviation from a molecular clock in dating based on a clock assumption. In the MPL method, deviations from the clock assumption are treated as stochastic departures. When the number of terminals descending from a node is large (i.e., sample size is large), the MPL estimates can be powerful, even though the molecular clock assumption is not perfectly valid. This feature is thus especially appealing when dating very speciose groups, e.g., flowering plants. As mentioned previously the MPL method is not suitable for a phylogenetic tree with very different molecular clocks in different parts of the tree. However, when age estimates of nodes close to the root of a tree are of main concern and deviations from the clock appear for small subgroups higher up in the tree, as is often the case, then MPL estimates and their confidence intervals are still applicable. This is true because the deviations from the clock assumption have little or no effect on the MPL estimates in such situations. For example, in the tree of Fig. 1, the two taxa descending from node 5, *Drymophila* and *Luzuriaga*, seem to have different clocks even though it was not significant when adjusting for multiple testing. However, when estimating the age of, say, node 2 this difference, in case the clocks really are different, has very little effect. The MPL method also provides the possibility of identifying at which nodes deviations from the clock occur, such that the tree can be subdivided easily into different parts where different reference fossils and different change rates may be applied.

Another source of error, perhaps even the greatest source of both bias and stochastic error, however, only rarely considered, is the uncertainty and incompleteness of the reference fossils used to calibrate the molecular clock. We do not discuss this issue here but see Bremer (2000) for a case study alleviating this problem by the use of several reference fossils for one tree. In situations where the calibration points (the age and systematic position of fossils) themselves are likely to contain the greatest amount of both stochastic error and systematic bias, it appears desirable to use simple approaches, such as the MPL method.

Another obvious source of error is erroneous tree topology. This problem belongs to phylogenetic reconstruction per se; see Baldwin and Sanderson (1998) for a case study assessing the error in age estimates due to mistaken tree topology. Herein we assume the topology of the phylogenetic tree to be correct. Yet another source of error is due to limited taxon sampling, which affects genetic distances, edge lengths and hence age estimates based on such measures. This suggests increased taxon sampling resulting in large trees with numerous terminals, and the dating of large trees is easily done with the MPL method.

Different tree-building methods, and details of them, e.g., ACCTRAN or DELTRAN of parsimony trees (Swofford, 1998), may introduce biases and errors. These differences are usually small and do not significantly affect age estimates. In the MPL method they become, on average, smaller when the number of descending edges are large.

Although the MPL method is applicable to any tree with edge lengths, it is best suited to trees that are corrected for undetected substitutions. Otherwise unde-

tected substitutions will induce underestimated substitution rates. If this bias is evenly distributed over the tree, it is unproblematic. However, the proportion of undetected substitutions usually increases with edge length. As a consequence, long edges in the tree will be up-scaled more when corrected for undetected substitutions. In practice this will often make the tree more clock-like since uncorrected trees often contain terminals with short path lengths, to the root (i.e., few substitutions) but having long terminal edges. Correcting for undetected substitutions will then increase the path lengths of such terminals more than other terminals having longer path lengths consisting of many short edges.

Most phylogenetic dating methods, including the MPL method presented here, assume a molecular clock. The MPL method relies on the assumption that substitution rates are approximately equal over (part of) the tree. A problem is how to decide whether this is a reliable assumption. Most methods, including the test procedure described in the present paper, use relative rate tests to find problematic terminals or sections of the tree. In "linearized" methods (reviewed by Sanderson, 1997), taxa are pruned if they do not pass relative rate tests. Other methods let the user assume different models of evolution along different edges of the tree, but this approach leaves the user with an enormous amount of more or less arbitrary alternatives to choose from. Cooper and Penny (1997) devised a clock-free method for estimating divergence dates of two groups when several fossils are known from each. All possible quartets are used for estimating minimum divergence times. Rambaut and Bromham (1998) extended the method in a maximum-likelihood framework, permitting one, two, or five rates per quartet tree.

Sanderson (1997) presented a method (nonparametric rate smoothing; NPRS) in which rates are allowed to vary. Instead changes in rates are minimized between adjacent edges. In the presence of abundant fossil data, several calibration points could in principle be used with Sanderson's method but would lead to serious numerical problems as the number of constraints grow (Sanderson, 1999). In the MPL method, incorporation of multiple calibration points imposes no serious computational problems (see Bremer, 2000 for an example). A different method allowing rate changes is given by Thorne et al. (1998) who study a Bayesian model in which the substitution rate of an edge is assumed to have a log-normal distribution with mean equal to the substitution rate of the ancestor edge. This model also has the effect of making rate changes small for adjacent edges in the tree.

Using trees with edge lengths as input compared to using the data matrix directly results in information loss. Sanderson discussed site-based methods, ignoring the intermediate calculation of edge lengths, and problems with their implementation in his NPRS method. Steel et al. (1996) presented a site-based method for calculating confidence intervals for divergence times of two groups only assuming the two groups are defined correctly and that sites evolve independently. Takezaki et al. (1995) proposed another site-based method in the same spirit but assuming a molecular clock. First sequences evolving at different rates are detected and eliminated and then node ages in the remaining tree are estimated by the average distances between all pairs of sequences in the two subtrees having the node in question as the root. However, while elegant for calculating confidence intervals for particular groups, these methods appear cumbersome for dating all nodes in a large tree (also giving confidence bounds). The purpose of the MPL method is to aid simple age estimates feasible also in large trees. This is done by assuming that the tree topology is correct and that edge lengths denote the number of substitutions along the edge. Then the tree with edge lengths is a sufficient statistic for estimating relative ages under the assumption of an approximately constant substitution rate.

The simple variance calculations of the MPL method rely on the observation that substitutions follow the Poisson distribution, a consequence of the model assumptions. Several studies (reviewed by Cutler, 2000) have demonstrated deviations from these assumptions. An alternative nonparametric approach is to use a bootstrap procedure, as done by Vinnersten and Bremer (2001) using MPL, and suggested by Sanderson (1997) as a means for obtaining confidence intervals for his NPRS method. Bootstrap analysis is generally assumed to be a reliable procedure for obtaining confidence intervals, but it is more tedious than the simple MPL intervals. Implementation of the bootstrap actually provides a means for testing the Poisson assumption. If the confidence intervals obtained by bootstrap analysis are similar to those obtained by the MPL method, this indicates that the underlying Poisson assumption is satisfactory. Remarkably, there is a great correspondence in the confidence intervals obtained by the MPL method and those calculated by bootstrap analysis for the Liliales example (Table 1). The correlation between the two types of confidence intervals is high ($r = 87.4\%$). On average, the confidence intervals from our MPL method and from bootstrap analysis are approximately equally long, ± 5.3 vs ± 5.2. The difference is nonsignificant ($P = 0.46$).

Acknowledgment

The authors gratefully acknowledge support from the Swedish Research Council.

References

Baldwin, B.G., Sanderson, M.J., 1998. Age and rate of diversification of the Hawaiian silversword alliance (Compositae). Proc. Natl. Acad. Sci. USA 95, 9402–9406.

Bremer, K., 2000. Early Cretaceous lineages of monocot flowering plants. Proc. Natl. Acad. Sci. USA 97, 4707–4711.

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

Cooper, A., Penny, D., 1997. Mass survival of birds across the Cretaceous–Tertiary boundary: molecular evidence. Science 275, 1109–1113.

Cutler, D.J., 2000. Understanding the overdispersed molecular clock. Genetics 154, 1403–1417.

Hillis, D.M., Moritz, C., Mabel, B.K., 1996. Molecular Systematics, second ed. Sinauer, Sunderland, MA.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Statist. 6, 65–70.

Maddison, D.R., Swofford, D.L., Maddison, W.P., 1997. Nexus: an extensible file format for systematic information. Syst. Biol. 46, 590–621.

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

Ross, S.M., 1997. Introduction to Probability Models, sixth ed. Academic Press, San Diego.

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., 1998. Estimating rate and time in molecular phylogenies: beyond the molecular clock. In: Soltis, D.E., Soltis, P.S., Doyle, J.J. (Eds.), Molecular Systematics of Plants, vol. II, DNA Sequencing. Kluwer Academic Publishers, Boston, pp. 242–264.

Sanderson, M.J., 1999. r8s. Computer program and documentation. Available from <http://phylo.ucdavis.edu/r8s/r8s.html>.

Steel, M., Cooper, A., Penny, D., 1996. Confidence intervals for the divergence time of two clades. Syst. Biol. 45, 127–134.

Swofford, D.L., 1998. PAUP*. Phylogenetic Analysis Using Parsimony, 4th ver. Sinauer, Sunderland, MA.

Takezaki, N., Rzhetsky, A., Nei, M., 1995. Phylogenetic test of the molecular clock and linearized trees. Mol. Biol. Evol. 12, 823–833.

Thorne, J.L., Kishino, H., Painter, I.S., 1998. Estimating the rate of evolution of the rate of evolution. Mol. Biol. Evol. 15, 1647–1657.

Vinnersten, A., Bremer, K., 2001. Age and biogeography of major clades in Liliales. Am. J. Bot. 88, 1695–1703.