*Syst. Biol.* 56(5):741–752, 2007

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150701613783

# Estimating Divergence Times in Large Phylogenetic Trees

TOM BRITTON,[1] CAJSA LISA ANDERSON,[2] DAVID JACQUET,[1] SAMUEL LUNDQVIST,[1] AND KÅRE BREMER[3]

[1] *Department of Mathematics, Stockholm University, SE-106 91 Stockholm, Sweden; E-mail: tom.britton@math.su.se (T.B.)*
[2] *Department of Systematic Botany, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, SE-752 36 Uppsala, Sweden*
[3] *Stockholm University, SE-106 91 Stockholm, Sweden*

*Abstract.*— A new method, PATHd8, for estimating ultrametric trees from trees with edge (branch) lengths proportional to the number of substitutions is proposed. The method allows for an arbitrary number of reference nodes for time calibration, each defined either as absolute age, minimum age, or maximum age, and the tree need not be fully resolved. The method is based on estimating node ages by mean path lengths from the node to the leaves but correcting for deviations from a molecular clock suggested by reference nodes. As opposed to most existing methods allowing substitution rate variation, the new method smoothes substitution rates locally, rather than simultaneously over the whole tree, thus allowing for analysis of very large trees. The performance of PATHd8 is compared with other frequently used methods for estimating divergence times. In analyses of three separate data sets, PATHd8 gives similar divergence times to other methods, the largest difference being between crown group ages, where unconstrained nodes get younger ages when analyzed with PATHd8. Overall, chronograms obtained from other methods appear smoother, whereas PATHd8 preserves more of the heterogeneity seen in the original edge lengths. Divergence times are most evenly spread over the chronograms obtained from the Bayesian implementation and the clock-based Langley-Fitch method, and these two methods produce very similar ages for most nodes. Evaluations of PATHd8 using simulated data suggest that PATHd8 is slightly less precise compared with penalized likelihood, but it gives more sensible answers for extreme data sets. A clear advantage with PATHd8 is that it is more or less instantaneous even with trees having several thousand leaves, whereas other programs often run into problems when analyzing trees with hundreds of leaves. PATHd8 is implemented in freely available software. [Divergence times; estimation; molecular clock; phylogenetic trees; substitution rates.]

Estimation of divergence times in phylogenetic trees using sequence data has been increasingly popular during the last decade (Sanderson et al., 2004), and there are now many available methods (e.g., Aris-Brousou and Yang, 2003; Cutler, 2000; Drummond et al., 2006; Huelsenbeck et al., 2000; Kishino et al., 2001; Rambaut and Bromham, 1998; Sanderson, 1997, 2002; Yang and Rannala, 2006; Yoder and Yang, 2000). When estimating divergence times, a key problem is to disentangle duration times from evolutionary substitution rates (Takezaki et al., 1995; Thorne et al., 1998).

In the present paper, we present a new method, PATHd8, for estimating divergence times, that can be used to analyze large phylogenetic trees containing several thousand taxa. The method is a generalization of the mean path length (MPL) method (Bremer and Gustafsson, 1997; Britton et al., 2002). The principal idea in PATHd8 is that the relative age of a specific node in a phylogenetic tree is estimated by the average distance from that node to all its leaves (terminals, taxa), compared to the average distance from the root of the tree to all the leaves. This relative age may be calibrated to absolute age using a reference node from fossil data, but the MPL method did not include such calibration for more than one reference node. The MPL method can be viewed as a clock-based method for a parametric evolutionary model (Britton et al, 2002), but also as a rate-smoothing method without any clock assumption (M. J. Sanderson, personal communication). Sanderson also pointed out that MPL, and hence also PATHd8, as well as his NPRS method (Sanderson, 1997) may be viewed as members of a large class of rate-smoothing methods.

In our new method, the mean path length heuristic is generalized and extended in several ways. The method now allows the use of numerous reference nodes for time calibration and these reference nodes may specify fixed age, minimum age, or maximum age. Because reference nodes may indicate clear deviations from the molecular clock, it follows that PATHd8 is not clock-based but should be viewed as a rate-smoothing method. Further, the input phylogenetic tree need not be fully resolved; polytomies are preserved in the analysis. At least one reference node with an absolute age (a fixed age node) is necessary for time calibration but this reference node does not have to be the root node of the tree. The method and its algorithm are simple to explain, and the number of operations is linear in the number of leaves and linear in the number of constraints, thus enabling effectively instantaneous analysis of trees having thousands of leaves and reference nodes.

## EXISTING METHODS

Estimating divergence times in a phylogenetic tree with known topology and a molecular clock is relatively easy. However, empirical evidence shows that a molecular clock seems to be the exception rather than the rule (e.g., Bell et al., 2005; Elango et al., 2006; Muse, 2000; Near and Sanderson, 2004; Padovan et al., 2005; Porter et al., 2005; Renner, 2005; Sanderson and Doyle, 2001; Schneider et al., 2004; Soltis et al., 2002; Steppan et al., 2004; Wall, 2005; Wilson et al., 1990). When leaving out the molecular clock hypothesis, some other assumption, implicit or explicit, about how the rate of evolution varies over the tree has to be made—otherwise it might not be possible to estimate node ages consistently (Britton, 2005). The most common assumption of rate evolution in the tree is to assume that the substitution rates of edges close to each other in the tree are highly positively correlated,

whereas the farther apart the edges are, the less correlation between the rates of the edges (e.g., Sanderson, 1997, 2002; Thorne et al., 1998). An exception is Drummond et al. (2006), who assume that all substitution rates are uncorrelated.

The most commonly used methods for estimating divergence times include nonparametric rate-smoothing (NPRS) and penalized likelihood (PL) developed by Sanderson (1997, 2002) and implemented in the program r8s (Sanderson 2003), and Bayesian autocorrelated methods by Thorne et al. (1998), Kishino et al. (2001), and Thorne and Kishino (2002), and implemented in the multidivtime package by J. Thorne (Thorne and Kishino, 2002), combined with the PAML package by Z. Yang (1997).

Sanderson (1997) adopted a nonparametric approach in which the rate differences of adjacent edges are minimized in a quadratic function. In Sanderson (2002), a semiparametric approach is taken in which a likelihood model is penalized according to how much substitution rates of mother-daughter edges are changed. Both methods assume that rates change between mother edges and daughter edges, but large rate changes are penalized. In both methods the rates are smoothed simultaneously over the whole tree, but optimizing this rate-smoothing and parameter estimation over the whole sample space is nontrivial and will cause computational problems in large trees (Sanderson, personal communication).

Thorne et al. (1998), further developed by Kishino et al. (2001) and Thorne and Kishino (2002), produced a model for the rate variation, which is analyzed using Bayesian methods. In the Bayesian framework, not only substitutions along edges, but also edge lengths and parameters, are treated as being outcomes of random events. Estimation is performed from the posterior distribution of quantities of interest using the powerful Markov chain Monte Carlo (MCMC) method (Gilks et al., 1996). In large trees the sample space that the MCMC simulation should explore is enormous, and it can always be questioned if the Markov chain has been run long enough for both finding all regions with high posterior probabilities and having done this with high precision. Another serious concern with Bayesian methods is what effect the choice of prior distribution (prior knowledge) has on the posterior distribution (of the tree). For instance, Yang and Rannala (2005) report that the choice of prior for internal edge length affects the posterior distribution substantially in phylogenetic reconstruction, and it is likely that the same phenomenon appears also in divergence time estimation. Yang and Rannala (2006) study the influence of choice of prior on the posterior probabilities in dating (where dating is specified using soft bounds). Their analysis suggests that priors only have a moderate effect on the posterior, but they also state that age estimates can converge to the prior bounds.

It has been concluded in other studies that more taxa (Linder et al., 2005) and more fossils (Bremer et al., 2004) are needed for "better," or at least more stable, age estimates. For this reason, a dating method that can handle very large data sets with multiple constraints is advantageous. The issues raised above have motivated the development of our new method.

## PATHd8

The input data for our proposed method is a phylogenetic tree with a specified topology and with specified edge lengths. These edge lengths could be any measurement of amount of evolution but we recommend using the number of substitutions along the edge or estimates thereof. Preferably, multiple substitutions are also included in the edge lengths, as is usually done in edge length calculation by standard model-based methods (e.g., in PAUP*; Swofford, 2003). The input tree need not be fully resolved: polytomies reflecting uncertainty of the topology are allowed and will be preserved. An arbitrary number of reference nodes for time calibration are allowed as input data; each reference node is either specified as "fixed age" (i.e., has an exact age of $x$), "minimum age" (i.e., is at least as old as $x$), or "maximum age" (i.e., is at most as old as $x$). It is also possible to specify a node having both a mimimum age and (an older) maximum age, thereby giving constraints with "soft bounds." At least one fixed age node of absolute age is required for the output tree to be defined in absolute age, but this need not be the root as is sometimes the case in other methods. If no fixed age node is available, the root node is set to have age 1 and relative ages are obtained.

The path length from a node to a leaf (terminal) descended from that node is the sum of the edge lengths from the node to the leaf. The mean path length (MPL) of a node is defined as the average of all path lengths from the node to the leaves descended from that node. The number of paths from a node equals the number of leaves descended from that node.

The basic idea in PATHd8 is that node ages are estimated by their corresponding MPL. This implies that substitution rate variation over the tree is compensated for (or smoothed) by taking averages, under the implicit assumption that averages between sister groups do not vary too much. However, fixed age (or minimum age or maximum age) nodes may indicate a deviation from this assumption, and this is taken into account by recalculating age estimates as weighted averages of MPL estimates from subtrees scaled by the fixed age nodes (precise definitions are given in the algorithm). As an effect, substitution rates are implicitly smoothed over sister groups from the MPL calculations but separated by fixed age nodes indicating different substitution rates. The complete algorithm is found in Appendix 1.

The biological motivation for MPL is that when the substitution rate does not vary too much, the mean path length is a sensible estimate of the node age. However, if there is no molecular clock, the substitution rate should vary "smoothly" in the tree, implying that substitution rates between two fixed age nodes should resemble each other more than rates far away from each other. It is exactly this feature that PATHd8 tries to develop: averages are taken within "regions" separated by fixed age nodes. The underlying biological motivation for PATHd8

is hence the same as for most methods, namely that rate variation changes smoothly.

In our new method the smoothing is done between sister paths (groups of edges). This is different from NPRS (Sanderson, 1997), penalized likelihood (PL) (Sanderson, 2002), and Bayesian (e.g., Kishino et al., 2001) methods where smoothing is done between mother and daughter edges. An effect of the rate smoothing being over sister paths (groups of edges), rather than being between pairs of mother/daughter edges, could be that rate differences between edges close to each other may vary more (future empirical studies will investigate this hypothesis further). In the new method smoothing is done stepwise for each node separately, in contrast to, for example, NPRS, PL, and Bayesian methods, where smoothing is done simultaneously all over the tree. An advantage with the present method is therefore that it is computationally very efficient compared to these methods.

An alternative use of PATHd8 is to use it only for time calibration of ultrametric trees obtained with other methods, in particular those that do not allow for multiple time constraints or "soft bounds" on constraints. The input data is then (1) an ultrametric tree, estimated using some other method but without reference nodes; (2) a list of reference nodes specifying fixed age, minimum age, and maximum age nodes. This means that PATHd8 can use output trees from NPRS, PL, or Bayesian methods as input trees. The PATHd8 output tree is still ultrametric, but edge lengths have been adjusted due to the time calibrations as described above.

## Comparison of Methods on Simulated Data

A disadvantage with empirical data is that the true divergence times are rarely known, thus making it hard to judge which method is better when age estimates differ. For this reason we have simulated data using RateEvolver (Ho et al., 2005). The true underlying tree is a completely symmetric binary tree with eight taxa, all edges having length 5 (the length units are just arbitrary time units), together with an outgroup of length 20 (Fig. 1). From this true tree, rates were simulated using a nonclock (autocorrelated) model in RateEvolver (default parameter values were used in the analysis). This was done 1000 times, giving 1000 phylograms. Each phylogram was then taken as input data and analyzed with PATHd8 and PL (using r8s) to estimate node ages. In the analyses it was assumed that two nodes were fixed age nodes with known age. One fixed age node was the most recent common ancestor of two terminals of age 5 (node 8 in Fig. 1), and the other was the most recent common ancestor of the four taxa in the other subtree, of age 10 (node 3 in Fig. 1). The resulting node age estimates of the two methods are summarized in Table 1. For each node, the mean and standard deviation of the 1000 age estimates for both methods are given, together with the true age. PL gives slightly more accurate mean estimates than PATHd8 for all six nonfixed age nodes. On the other hand, the standard deviations of the estimates are often larger for PL than for PATHd8.

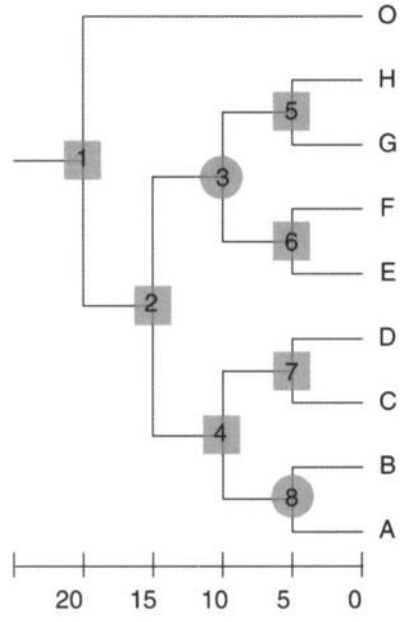


FIGURE 1. Tree from which data are simulated in the simulation study. Nodes 3 and 8 are treated as fixed age nodes in the analyses.

In order to get one summary measure for the deviation between an estimated tree and the true tree, we calculated the square root of the mean sum of squared differences between the estimated node age and the true node age for all nodes. This quantity, the total deviation, hence measures the accumulated node age deviations of the estimated tree as compared to the true tree. The mean of the total deviations from the 1000 trees was 10.33 when using PL, as compared to 13.00 when using PATHd8, thus indicating that PL is somewhat more accurate. On the other hand, the standard deviation of the 1000 total deviations was larger when using PL (31.19 versus 18.80 for PATHd8). In Figure 2, histograms of the 1000 total deviation measures are given for the two methods. Most total deviations of PL (Fig. 2b) are small as compared with PATHd8 total deviations (Fig. 2a), but a few are really large. PATHd8 has fewer and less extreme outliers. The extreme deviations of PL occur when the input tree has large rate variations. In particular, if the

TABLE 1. Total age deviations for the nine-taxon tree in Figure 1 using PL (r8s) and PATHd8. Listed are the means and standard devations from 1000 input trees.

| | | Mean age (time units) | | Standard deviation (time units) | |
|---|---|---|---|---|---|
| Node | True age | PL (r8s) | PATHd8 | PL (r8s) | PATHd8 |
| 1 | 20 | 25.22 | 25.55 | 24.11 | 10.62 |
| 2 | 15 | 18 | 18.12 | 13.89 | 7.08 |
| 3 | 10 | 10 | 10 | 0 | 0 |
| 4 | 10 | 12.33 | 13.14 | 10.58 | 7.91 |
| 5 | 5 | 5.15 | 5.28 | 1.01 | 2.11 |
| 6 | 5 | 5.17 | 5.29 | 1.02 | 2.16 |
| 7 | 5 | 6.46 | 7.95 | 6.37 | 7.7 |
| 8 | 5 | 5 | 5 | 0 | 0 |

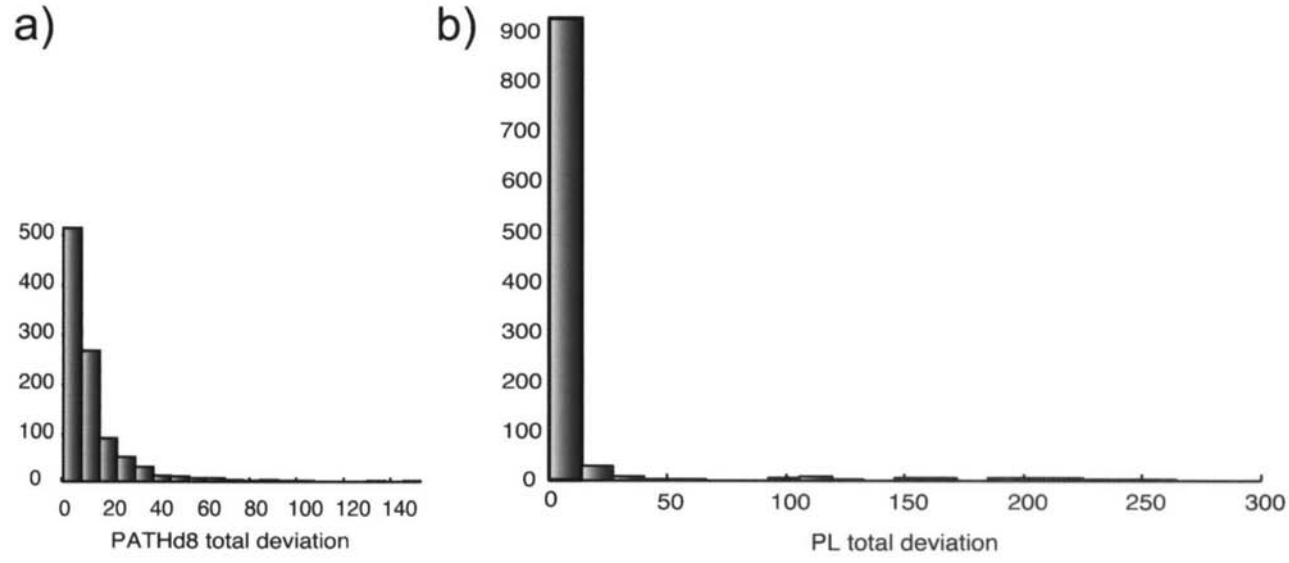


FIGURE 2. Histograms of total deviations from 1000 simulated input trees originating from the tree in Figure 1 (see text for further explanation of "total deviation"). PATHd8 values to the left (a) and PL values to the right (b).

outgroup happens to have a high substitution rate, then the estimated ages of nodes close to the root using PL are much older than their true ages and also older than the corresponding PATHd8 estimates. Similar analyses were performed for other pairs of fixed age nodes. The general trend was that PL (using r8s) is relatively better than PATHd8 the closer to the root where there is a fixed age node and, conversely, that PATHd8 is superior to PL when all fixed age nodes are far away from the root. This has not been studied thoroughly, but a possible explanation could be that the complicated mathematical smoothing in PL is generally much harder when smoothing towards a root without maximum age than when smoothing to a leaf with fixed age (namely 0!).

## COMPARISON OF METHODS ON EMPIRICAL DATA

PATHd8 can be used for analyzing large data sets containing several thousand taxa. Due to computational problems when using other methods than PATHd8 for dating of large data sets, it is not possible to present comparisons with other methods for data sets with thousands of taxa. We have therefore analyzed three empirical data sets of different sizes and compared age estimates from PATHd8 with estimates from methods that can handle the respective data sets. The data sets were the 200+ eudicot data set from Anderson et al. (2005), a reduced version of that data set (having 35 basal eudicot, 2 outgroup, and 10 placehoder taxa), and the 800+ monocot taxa data set from Janssen and Bremer (2004). The phylograms of all three empirical data sets, and the sequence data set for the reduced eudicot data, are available at http://www.systematicbiology.org.

One of the fossil constraints has to be fixed in all methods. None of the methods require, in theory, the root node to be fixed, but because PL and Langley-Fitch can give unreasonable results without the root being fixed (Sanderson, 2004), we choose to fix a node close to the root in all analyses. Fixing the root also facilitates the comparison between the methods. Because only minimum ages can be determined with some confidence, we have chosen to use all other fossils as minimum age constraints.

The 200+ eudicot data set (Fig. 3a) was analyzed with PATHd8 and compared with the Langley-Fitch (LF) (Langley and Fitch 1974) clock method (in which a strict molecular clock is assumed and divergence times are reconstructed by maximum likelihood of the edge lengths) as implemented in r8s, and PL dating taken from Anderson et al. (2005). (Multidivtime failed to come up with a result in a reasonable amount of time, therefore no dating could be performed using the Bayesian approach.) The same 14 fossil constraints (1 fixed age and 13 minimum ages) used in the original study were used in the PATHd8 analysis. The results from this comparison are summarized in Table 2, and chronograms are found in Figure 4. LF and PL give very similar results for all nodes. PATHd8 gives comparable estimates for most stem groups, but younger ages for some crown groups, especially the ones that are less constrained. Papaveraceae and Sabiales are two examples of this. The large age differences for these two nodes are due to the large distance to fossil constraints at these nodes, in combination with the different smoothing directions of PATHd8 versus the other methods. The whole core eudicot crown group is also younger when analyzed with PATHd8.

The reduced eudicot data set (Fig. 3b) was analyzed using four methods: PATHd8, PL, LF, and the Bayesian autocorrelation method implemented in the combination of the PAML package by Yang (1997) and the multidivtime package described by Thorne et al. (2002). A selection of 6 fossil constraints, applicable to the reduced data set (1 fixed age and 5 minimum ages), from the original 200+ data set was used.

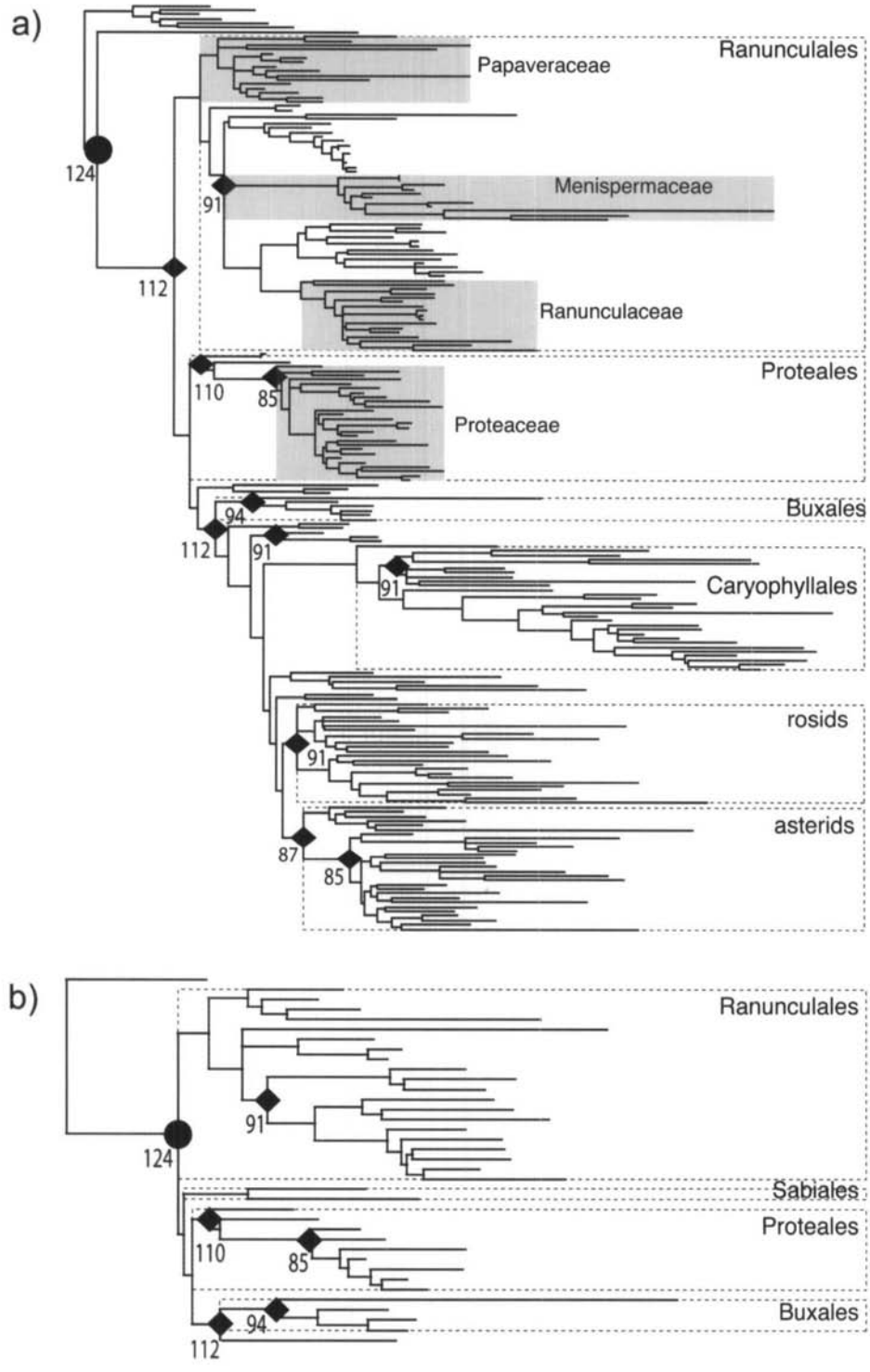


FIGURE 3. Phylograms of the 200+ eudicot (a) and the reduced eudicot (b) data sets. Fossil constrained nodes are marked with absolute age in million years. The node closest to the root is fixed at 124 Ma; other constraints are used as minimum ages and marked with diamonds in the figure. For information on fossils used, see Anderson et al. (2005).

For the penalized likelihood analysis, a smoothing value of 0.000016 was chosen based upon the lowest fraction error obtained from the fossil-based cross-validation procedure implemented in r8s.

Prior settings for the multidivtime analysis followed the recommendations from the program's manual. Brownmean (the rate variation parameter) and standard deviation were set to 0.012 (calculated, as suggested, by dividing 1.5 by the expected number of time units between root and leaves). Rtrate (rate at the root node) was set to 0.0005 (calculated by taking the mean path length and dividing it by the expected number of time units

TABLE 2. Age estimates in My of some internal nodes (see Figs. 3, 4) in the 200+ eudicot tree obtained by Langley-Fitch clock (r8s), PL (r8s), and PATHd8.

| | LF clock (r8s) | PL (r8s) | PATHd8 |
|---|---|---|---|
| crown Ranunculales | 109 | 114 | 91 |
| crown Papaveraceae | 96 | 106 | 42 |
| crown Ranunculaceae | 74 | 73 | 62 |
| crown Menispermaceae | 71 | 70 | 67 |
| stem Proteales | 118 | 119 | 112 |
| crown Proteales | 114 | 115 | 110 |
| crown Proteaceae | 85 | 85 | 85 |
| stem Buxales | 115 | 117 | 112 |
| crown Buxales | 96 | 91 | 94 |
| crown Sabiales | 83 | 99 | 40 |

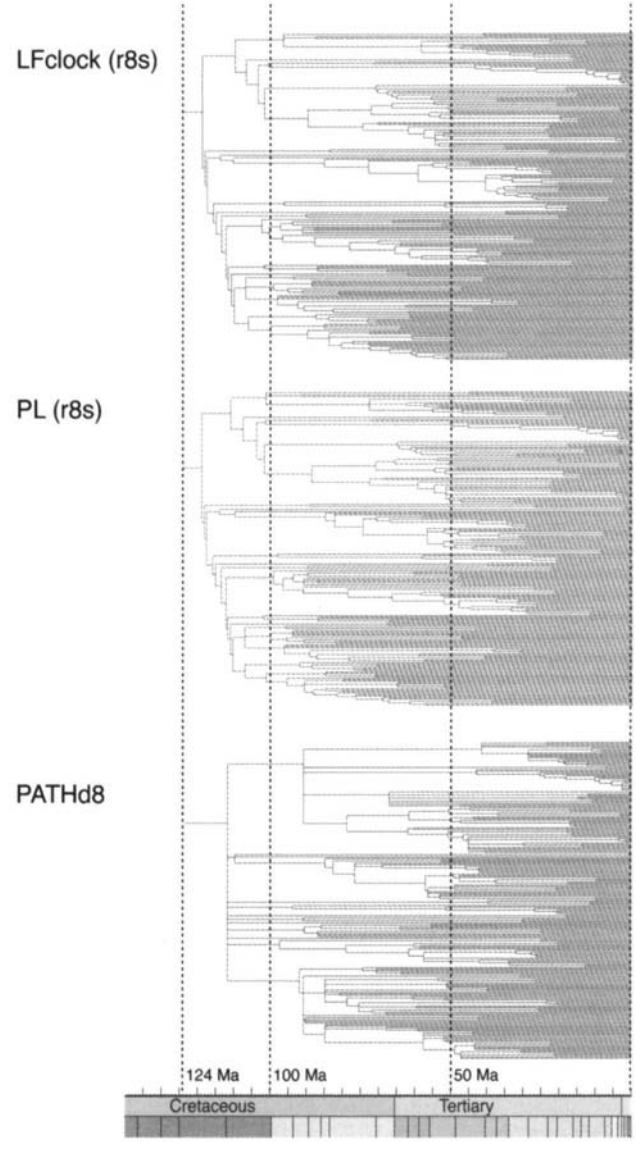


FIGURE 4. Chronograms obtained from the dating analyses of the 200+ eudicot data set. Outgroup taxa are pruned. The analyses were performed using the Langley-Fitch clock method (r8s), penalized likelihood (r8s), and PATHd8. Absolute ages (for the named groups in Fig. 3) are given in Table 2.

between root and leaves), with the same value as standard deviation. Minab (the prior specifying whether the internal nodes should repel or attract each other) and the MCMC parameters were the same as the example infile distributed with the program. Following a burn-in of 100,000 cycles, 10,000 samples were taken from the MCMC chain with a sample frequency of 100 cycles. Three separate Markov chains were simulated in order to see if the chains gave approximately the same age estimates.

The divergence times from the analysis of the reduced eudicot data set should not be taken as reliable estimates, because the data set is too small for obtaining stable age estimates (Linder et al., 2005). Furthermore, in the reduced topology, the eudicot crown group is constrained to 124 million years (Myr), as opposed to the stem group in the 200+ data set, which complicates further comparisons. The analyses are primarily performed to investigate differences between the methods. The large number of fossil constraints in relation to the low number of taxa should mimimize the differences in age estimates between the methods, because all nodes in the tree will be close to a constrained node. Chronograms from the four analyses are found in Figure 5. The estimated ages are presented in Table 3. All methods give similar results, but there are a few rapid rate changes in PATHd8 (e.g., within Ranunculales) that we do not see in the LF, PL, or multidivtime results. Divergence times are most evenly spread over the chronograms obtained from the Bayesian implementation and clock-based method. Ages for all named nodes are also very similar with these two methods, differing the most by 11 Myr. The largest differences between methods are found in less constrained crown groups, where age estimates in general become older using PL and younger using PATHd8. The most obvious difference is the age estimates of crown group Papaveraceae, which differ by 62 Myr between PATHd8 and PL, and have no overlap in confidence intervals (but see further discussion below).

In this comparison, all methods give results that are concordant with the fossil record, and it is therefore not possible to say which method gives the results closest to the true ages.

Bayesian methods, such as multidivtime, have the advantage of giving information about uncertainty in the original (albeit time-consuming) analysis. In order to investigate uncertainty of the estimated node ages of LF, PL, and PATHd8 methods, a simulation study was performed for the reduced eudicot data set. One hundred bootstrap replicates of the original data matrix were produced by Phylip (Felsenstein, 2006) and used to infer 100 phylograms with the same topology using PAUP* (Swofford, 2003). Node ages were estimated using LF, PL, and PATHd8. The resulting age estimates show how the methods "inherit" uncertainty from the input trees. In Table 3 the age estimates of some internal nodes are given with their respective central 95% distribution, together with the credibility intervals from the multidivtime analysis. We emphasize that the credibility intervals cannot be directly compared with the bootstrap intervals of LF,

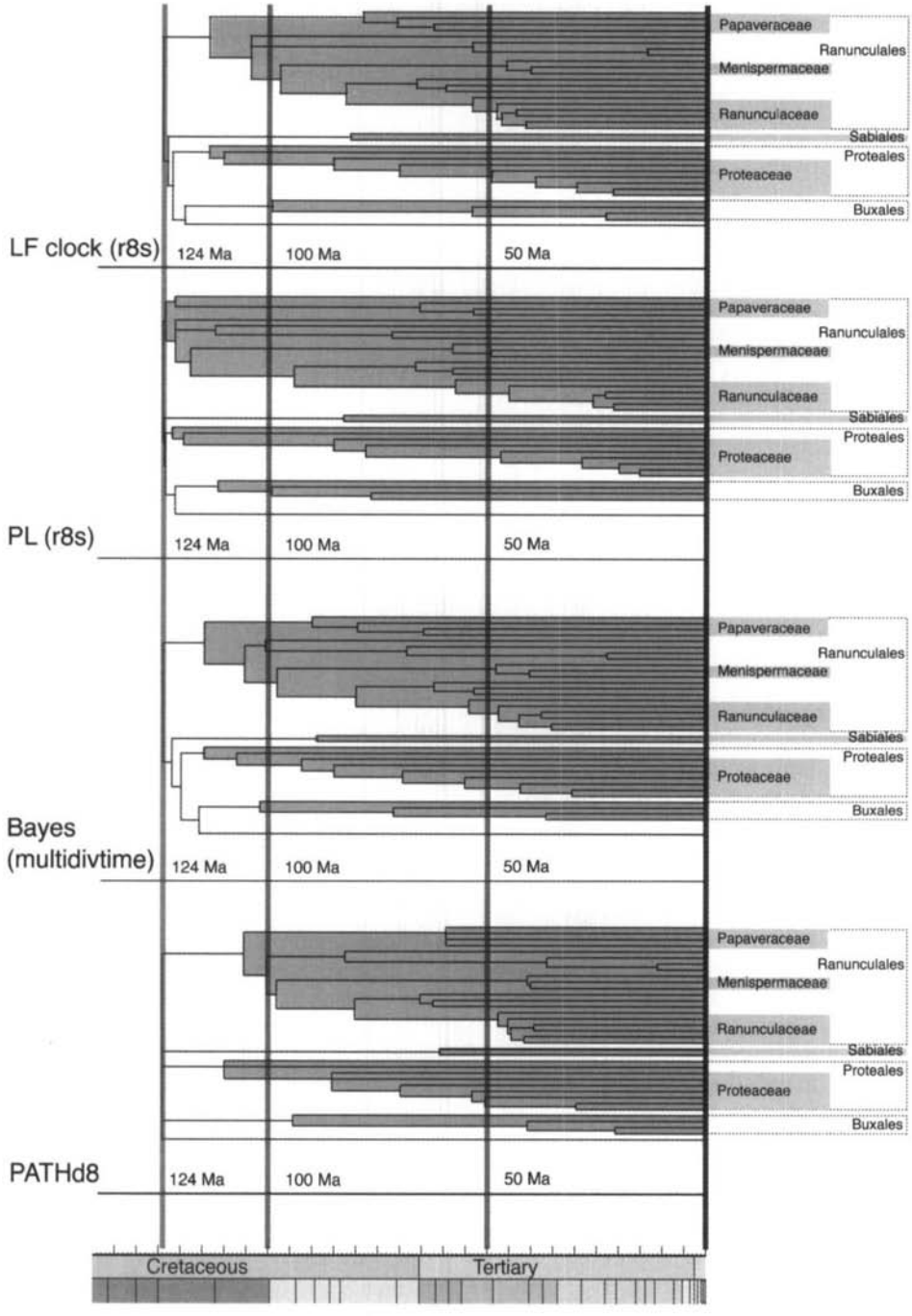


FIGURE 5. Chronograms obtained from the dating analyses of the reduced eudicot data set. Outgroup and placeholder taxa are pruned. The analyses were performed using the Langley-Fitch clock method (r8s), penalized likelihood (r8s), the Bayesian autocorrelation method implemented in multidivtime, and PATHd8. Absolute ages for the named groups are given in Table 3.

PL, and PATHd8. The credibility intervals reflect which age estimates have highest posterior probability assuming a probability model, prior distributions, and original input data, whereas the bootstrap intervals reflect uncertainty in input data and how this affects the nonprobabilistic age estimates of PATHd8, semiprobabilistic age estimates of PL, and probabilistic age estimates of LF.

When comparing the confidence intervals and credibility intervals (Table 3), it is seen that all intervals resulting from the mother-daughter smoothing approaches (PL, LF, and Bayesian) are partly overlapping, which is not surprising. However, as seen from Table 3, r8s/PL produces an abnormal confidence interval for crown Sabiales; it seems to be an artefact resulting from computational problems inherent in r8s/PL when using a

TABLE 3. Age estimates in My of some internal nodes (see Figs. 3, 5) in the reduced eudicot tree obtained by four different dating methods. Numbers within parentheses represent bootstrap confidence intervals for LF, PL, and PATHd8, and credibility intervals from the multidivtime analysis. For three estimates (crown Proteace estimated by LF, and crown Proteales and crown Proteaceae by PATHd8), no confidence interval is given, because all replicates gave the same age as the age constraints applied to these nodes.

| | Langley Fitch clock (r8s) | Penalized likelihood (r8s) | Bayesian autocorrelation (multidivtime) | PATHd8 |
|---|---|---|---|---|
| crown Ranunculales | 113 ( 103–121) | 123 (123–124) | 115 (103–123) | 105 (94–118) |
| crown Papaveraceae | 79 (64–98) | 121 (119–123) | 90 (69–110) | 59 (46–71) |
| crown Ranunculaceae | 53 (44–63) | 57 (35–96) | 54 (35–77) | 48 (46–77) |
| stem Menispermeaceae | 98 (91–106) | 118 (114–122) | 98 (91–111) | 98 (91–111) |
| crown Menispermeaceae | 45 (37–58) | 58 (52–104) | 48 (28–72) | 40 (31–52) |
| stem Proteales | 122 ( 119–124) | 123 (122–124) | 120 (115–123) | 124 (116–124) |
| crown Proteales | 114 (111–123) | 122 (122–124) | 114 (110–121) | 110 (–) |
| crown Proteaceae | 85 (–) | 85 (85–106) | 92 (85–106) | 85 (–) |
| stem Buxales | 119 (116–123) | 121 (118–124) | 116 (112–121) | 124 (112–124) |
| crown Buxales | 99 ( 94–112) | 111 (101–120) | 102 (94–114) | 94 (94–98) |
| crown Sabiales | 81 (64–104) | 83 (117–123) | 89 (62–115) | 60 (44–77) |

very low smoothing value (0.000016); we have double-checked the result and this is the interval delivered by using r8s/PL. It may be that other intervals resulting from the described bootstrap procedure are artefacts. Furthermore, r8s/PL failed to find solutions in 11 bootstrap replicates. These caveats should be remembered in comparisons with the other methods.

Intervals from PATHd8 overlap partly with the ones obtained by multidivtime and LF, but with PL only for five (or possibly four) well-constrained nodes and the stem nodes of the orders. The intervals from all methods are larger for nodes that are not constrained, and largest when nodes are not bracketed by reference nodes above (i.e., towards the leaves). In one case all 100 replicates in PATHd8 and LF result in the same age, the minimum age of crown Proteaceae, indicating that both methods would have estimated this age younger had it not been for the minimum age constraint. The same node as calculated by multidivtime and PL has intervals of 21 Myr. PATHd8 also does not produce an interval for crown Proteales and results in the age of the minimum age of the node below, whereas LF and multidivtime produce intervals of about 10 Myr for this node.

The large 800+ monocot phylogram, shown in Figure 6 (and first analyzed by Janssen and Bremer 2004), was analyzed with PATHd8 and NPRS. (The reason for using NPRS, and not PL, in the 800+ monocot data set was simply because the PL algorithm could not find a solution. At the time of the original study, no other method, except for strict clock methods, could be used.) The chronogram obtained from the PATHd8 analysis is found in Figure 7 (see Janssen and Bremer, 2004, for the corresponding NPRS chronogram). The estimated ages from the two analyses are presented in Table 4. For the 800+ monocot data set, there is an overall tendency for PATHd8 to yield younger ages than NPRS, with the exception of stem and crown Poaceae. In particular, a few crown nodes are much older when estimated with NPRS. This illustrates the different smoothing approaches. In general, groups with long branches get older ages when estimated with PATHd8, and groups with shorter branches get younger ages, because of the sister-group smoothing.

Generally speaking, chronograms from LF, PL, and Bayesian autocorrelation analyses show systematic differences in the way their respective chronograms look. The Bayesian trees look more smoothed in the sense of having more evenly spread edge lengths. Even trees with apparently very heterogeneous rates, judged from the phylograms, have this smooth appearance. PL chronograms are more varied in age differences between clades. Compared to the other methods, PATHd8 gives the results with the least smooth appearance and could be said to preserve more of the heterogeneity seen in the original phylogram. This is an effect of the local smoothing as opposed to the other methods where smoothing is done simultaneously over the whole tree. Local smoothing need not be a disadvantage—it may be that there sometimes is an excess smoothing in other methods.

The most important factor for obtaining reasonable age estimates, regardless of method, is age constraints (Bremer et al., 2004). In the absence of age constraints, the differences between methods become more apparent. When using fossil constraint, the placement of the fossil taxa on the correct node in the phylogeny is crucial. Fossils to be used as minimum ages are implicitly placed on a stem lineage, and hence the age is placed on the node where the stem lineage splits from its sister group. In some cases this means that the resulting crown group age can be much younger than the stem group age. This is true for all methods. Because PATHd8 smoothes between sister groups, as opposed to the other methods that smooth mother-daughter lineages, this can lead to very different age estimates compared to the other available methods. An example of this can be seen in the comparison of NPRS and PATHd8 on the 800+ monocot data set. The palm clade, family Arecaceae, have much shorter branches in the phylogram, compared to the rest of the monocots, indicating a slowdown in evolutionary rate at some point (Janssen and Bremer, 2004; Wilson et al. 1990), and the absolute age estimates for the crown group resulting from PATHd8 and NPRS are differing by almost 100 Myr. If a fossil belonging to the crown group had been used instead of the minimum age on the stem lineage of Arecaceae, we would obtain an older age of the

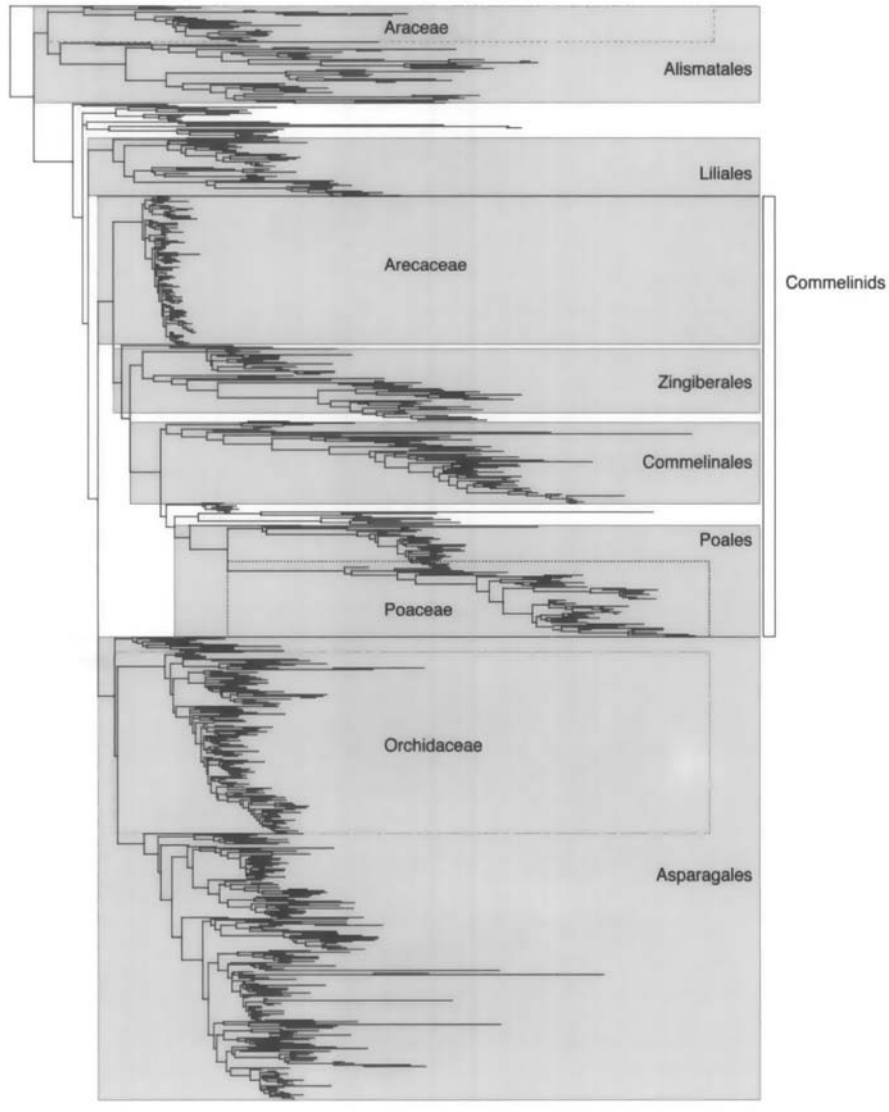


FIGURE 6. Phylogram of the 800+ monocot data set. For information on fossil constraints, see Janssen and Bremer (2004).

crown group using PATHd8, although not as old as the age calculated by NPRS. This is not only due to the fossil constraint. The clade can be said to be overrepresented in number, compared to the total number of taxa in the total data set. Increased sampling is thought to lead to older age estimates in methods smoothing between mother-daughter lineages, and therefore systematically often results in overestimates (Janssen and Bremer, 2004; Sanderson and Doyle, 2001). This phenomenon does not occur in PATHd8. On the other hand, in the absence of constraints, one can suspect PATHd8 of systematically underestimating the age of big clades with short internal branches. Because placement and age of age constraints is such an important issue, it is crucial to choose as many reliable fossil constraints as possible.

In Ericson et al. (2006) PATHd8 and PL were used for estimating divergence times of Neoaves. The analyses included 91 taxa and 23 fossil constraints. The methods give similar results, PL generally giving slightly older age estimates. As the PATHd8 estimates are more in agreement with the fossil record, in the sense that the analysis produces smaller "ghost intervals" between the age of the first fossil found and the molecular age estimate, Ericson et al. considered the PATHd8 estimates to be more reliable than those obtained by PL.

Another difference between PATHd8 and the other methods is that PATHd8 will collapse zero- or near-zero-edge lengths in the analysis, as opposed to the other programs that do not allow zero-edge lengths (polytomies) and hence have to increase such edge lengths to make them positive. Because short edge lengths either reflect uncertainties in the phylogeny, or else a very short time elapsed between the divergences, this motivates the possibility to collapse edges. The collapsing is one reason why PATHd8 chronograms look different from the smooth, and fully dichotomous, appearances of the Bayes chronograms.

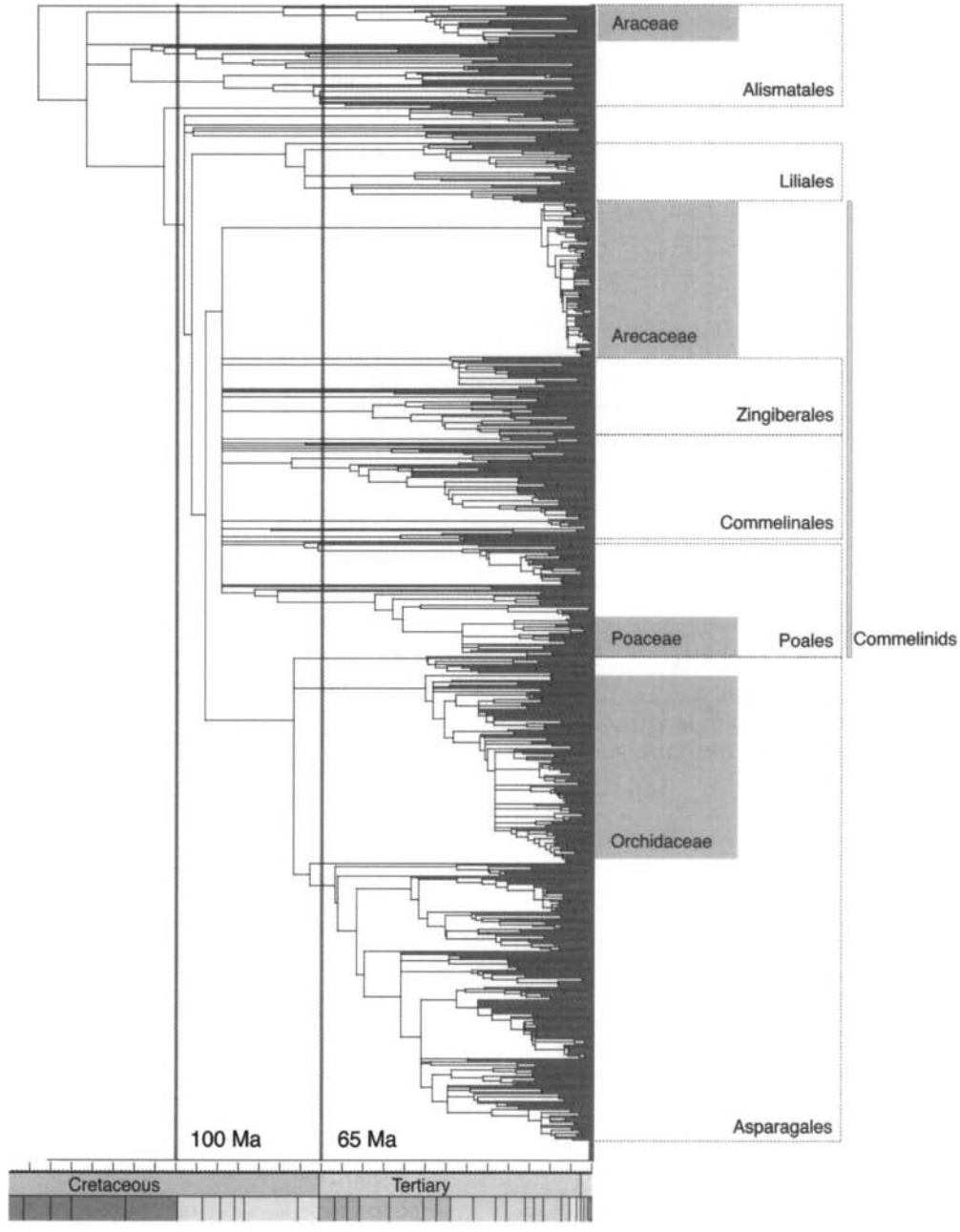


FIGURE 7. Chronogram resulting from the PATHd8 analysis of the 800+ monocot data set. Absolute ages for the named groups are given in Table 3. For chronogram from the NPRS analysis, see Janssen and Bremer (2004).

Worth noticing is the short time that the program needs to analyze even very large data sets. The 800+ data set was dated in approximately 1 second, and even the largest data set so far analyzed by PATHd8 (5100+ taxa) took only a few seconds to complete. This should be compared with the other available methods. Bayesian autocorrelation using multidivtime fail to give results for both the full 200+ eudicot data set and the monocot 800+ data set. Similarly, penalized likelihood (PL) could not be used to analyze the monocot 800+ data set.

## PATHd8 COMPUTER PROGRAM

The new method is implemented in the C-program PATHd8 and is freely available at www.math.su.se/PATHd8/ where compiled UNIX, Linux, Mac OSX, and Windows versions are available. The source code, a description of the algorithm, a manual, and some simple examples are also given there.

The input tree file should be written in Newick format, followed by an arbitrary number of age constraints of nodes, specified either as fixed age, minimum age, or maximum age. The output file contains the PATHd8

TABLE 4. Age estimates in Ma of some internal nodes (see Figs. 6, 7) in the 800+ monocot tree obtained by NPRS (r8s) and PATHd8.

| | PATHd8 | NPRS (r8s) |
|---|---|---|
| crown Alismatales | 122 | 128 |
| crown Araceae | 75 | 117 |
| crown Liliales | 74 | 117 |
| crown Asparagales | 72 | 119 |
| crown Orchidaceae | 40 | 111 |
| crown Commelinids | 100 | 120 |
| crown Arecaceae | 14 | 110 |
| crown Commelinales | 97 | 110 |
| crown Zingiberales | 37 | 86 |
| crown Poales | 100 | 113 |
| crown Poaceae | 89 | 83 |
| stem Alismatales | 123 | 131 |
| stem Araceae | 122 | 128 |
| stem Liliales | 100 | 124 |
| stem Asparagales | 72 | 122 |
| stem Orchidaceae | 72 | 119 |
| stem Commelinids | 97 | 122 |
| stem Arecaceae | 97 | 120 |
| stem Commelinales | 100 | 114 |
| stem Zingiberales | 97 | 114 |
| stem Poales | 97 | 117 |
| stem Poaceae | 97 | 89 |

analyses as a tree in Newick format as well as a list of estimated node ages, their mean path lengths, and their estimated substitution rates. The output also contains results from the MPL analysis (Britton et al., 2002). The MPL method is of interest when there are no fossil datings (so the estimated MPL tree is given in relative time) and also when the clock hypothesis is of interest. In each node a test is performed to determine if the subtrees descending from the node have significantly different substitution rates or not. The original test (Britton et al., 2002) was for a binary tree but has been extended to allow for polytomies when all sister group mean path lengths are compared with the overall average in a chi-square test. The molecular clock hypothesis was rejected on all three empirical data sets analyzed in the previous section.

## REFERENCES


Anderson, C. L., K. Bremer, and E. M. Fries. 2005. Dating phylogenetically basal eudicots using *rbcl* sequences and multiple fossil reference points. Am. J. Bot. 92:1737–1748.

Aris-Brosou, S., and Z. Yang. 2003. Bayesian models of episodic evolution support a late Precambrian explosive diversification of the Metazoa. Mol. Biol. Evol. 20:1947–1954.

Bell, C. D., D. E. Soltis, and P. S. Soltis. 2005. The age of the angiosperms: A molecular timescale without a clock. Evolution 59:1245–1258.

Bremer, K., E. M. Fries, and B. Bremer. 2004. Molecular phylogenetic dating of asterid flowering plants shows early Cretaceous diversification. Syst. Biol. 53:496–505.

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

Britton, T. 2005. Estimating divergence times in phylogenetic trees without a molecular clock. Syst. Biol. 54:500–507.

Britton T., B. Oxelman, A. Vinnersten, and K. Bremer. 2002. Phylogenetic dating with confidence intervals using mean path lengths. Mol. Phyl. Evol. 24:58–65.

Cutler, D. J. 2000. Estimating divergence times in the presence of an overdispersed molecular clock. Mol. Biol. Evol. 17:1647–1660.

Drummond, A. J., S. Y. W. Ho, M. J. Phillips, and A. Rambaut. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Elango, N., J. W. Thomas, and S. V. Yi. 2006. Variable molecular clocks in hominoids. Proc. Nat. Acad. Sci. USA 103:1370–1375.

Ericson, P. P. G., C. L. Anderson, T. Britton, A. Elzanowski, U. S. Johansson, M. Källersjö, J. I. Ohlson, T. J. Pahrsons, D. Zuccon, and G. Mayr. 2006. Diversification of Neoaves: Integration of molecular sequence data and fossils. Biol. Lett. 2:543–547.

Felsenstein, J. 2006. PHYLIP (phylogeny inference package). Version 3.66. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (eds). 1996. Markov chain Monte Carlo in practice. Chapman & Hall, New York.

Ho, S. Y. W., M. J. Phillips, A. Cooper, and A. J. Drummond. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Mol. Biol. 22:1561–1568.

Huelsenbeck, J. P., B. Larget, and D. Swofford. 2000. A compound Poisson process for relaxing the molecular clock. Genetics 154:1879–1892.

Janssen, T., and K. Bremer. 2004. The age of major monocot groups inferred from 800+ sequences. Bot. J. Linn. Soc. 146:385–398.

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. Fitch. 1974. An estimation of the constancy of the rate of molecular evolution. J. Mol. Evol. 3:161–177.

Linder, H. P., C. R. Hardy, and F. Rutschman. 2005. Taxon sampling effects in molecular clock dating: An example from the African Restionaceae. Mol. Phylogenet. Evol. 35:569–582.

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

Near, T. J., and M. J. Sanderson. 2004. Assessing the quality of molecular divergence time estimates by fossil calibrations and fossil-based model selection. Phil. Trans. R. Soc. Lond. B Biol. Sci. 359:1477–1483.

Padovan, A. C. B., G. F. O. Sanson, A. Brunstein, and M. R. S. Briones. 2005. Fungi evolution revisited: Application of the penalized likelihood method to a bayesian fungal phylogeny provides a new perspective on phylogenetic relationships and divergence dates of ascomycota groups. J. Mol. Evol. 60:726–735.

Porter, M. L., M. Perez-Losada, and K. A. Crandall. 2005. Model-based multi-locus estimation of decapod phylogeny and divergence times. Mol. Phylogenet. Evol. 37:355–369.

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

Renner, S. S. 2005. Relaxed molecular clocks for dating historical plant dispersal events. Trends Plant Sci. 10:550–558.

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

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

Sanderson, M. J. 2003. r8s: Inferring absolute rates of molecular absence of a molecular clock. Bioinformatics 19:301–302.

Sanderson, M. J., and J. A. Doyle. 2001. Sources of error and confidence intervals in estimating the age of angiosperms from rbcL and 18S rDNA data. Am. J. Bot. 88:1499–1516.

Sanderson, M. J., J. L. Thorne, N. Wikström, and K. Bremer. 2004. Molecular evidence on plant divegence times. Am. J. Bot. 91:1656–1665.

Schneider, H., E. Schuettpelz, K. M. Pryer, R. Cranfill, S. Magallon, and R. Lupia. 2004. Ferns diversified in the shadow of angiosperms. Nature 428:553–557.

Soltis, P. S., D. E. Soltis, V. Savolainen, P. R. Crane, and T. G. Barraclough. 2002. Rate heterogeneity among lineages of tracheophytes: Integration of molecular and fossil data and evidence for molecular living fossils. Proc. Nat. Acad. Sci. Proc. USA 99:4430–4435.

Steppan, S. J., R. M. Adkins, and J. Anderson. 2004. Phylogeny and divergence-date estimates of rapid radiations in muroid rodents based on multiple nuclear genes. Syst. Biol. 53:533–553.

Swofford, D. L. 2003. PAUP*: Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Takezaki, N., A. Rzhetsky, 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 the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15:1647–1657.
Wall, D. P. 2005. Origin and rapid diversification of a tropical moss. Evolution 59:1413–1424.
Wilson, M. A., B. Gaut, and M. T. Clegg. 1990. Chloroplast DNA evolves slowly in the Palm family (Arecaceae). Mol. Biol. Evol. 7:303–314.
Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. CABIOS 13:555–556 (http://abacus.gene.ucl.ac.uk/ software/paml.html).
Yang, Z., and B. Rannala. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. Syst. Biol. 54:455–470.
Yang, Z. H., and B. Rannala. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol. Biol. Evol. 23:212–226.
Yoder, A. D., and Z.Yang, Z. 2000. Estimation of primate speciation dates using local molecular clocks. Mol. Biol. Evol. 17:1081–1090.

*First submitted 31 October 2006; reviews returned 16 January 2007; final acceptance 30 May 2007*
*Associate Editor: Frank Anderson*

## APPENDIX 1

## ALGORITHM FOR PATHd8

1. Check that there are no conflicts in the fixed age, minimum age, or maximum age nodes of the input tree. If present, terminate the program and display an error message.
2. Remove redundant constraints. This means that if for a minimum age node $m_1$ there exists a younger min-node $m_2$ whose minimum age exceeds the minimum age for $m_1$, then $m_1$ is not regarded as a min-node anymore. Also, if for a maximum age node $m_1$ there exists a younger max-node $m_2$ whose maximum age exceeds the maximum age for $m_1$, then $m_2$ is not regarded as a max-node anymore.
3. If the root is not a fixed age node, then the root age $a_r$ is estimated by a weighted average of fixed age node ages multiplied by relative mean path lengths (MPLs) of the root and the fixed age node, where the weights are proportional to the number of leaves of the fixed age nodes. More specifically, suppose there are $k$ fixed age nodes adjacent to the root (where "adjacent" means that there are no fixed age nodes in between), and note that by assumption $k>0$. Let $p_r$ denote the MPL of the root and let $p_1, \ldots, p_k$ denote the MPL of the k fixed age nodes having $n_1, \ldots, n_k$ leaves and fixed ages $a_1, \ldots, a_k$. The estimated root age is then defined by

$$a_r = (n_1 a_1 p_r/p_1 + \ldots + n_k a_k p_r/p_k)/(n_1 + \ldots + n_k).$$

Check that the calculated root age $a_r$ is not in conflict with any of the fixed ages $a_1, \ldots, a_k$. If it is, set $a_r$ equal to the oldest of $a_1, \ldots, a_k$ and set all undated nodes in between these nodes equal to $a_r$. Treat the root as a fixed age node with age $a_r$ from now on.
4. Go through the tree, starting next to the root and moving towards the leaves, and calculate the age estimate of a given (nonfixed age) node $x$ as follows. Let 0 be the fixed age node closer to the root than $x$ and $a_0$ its age, and let $a_1, \ldots, a_k$ be the ages of fixed age nodes between $x$ and the leaves ($k$ can be 0 or positive). Let $n_x$ be the number of paths from $x$ to the leaves without passing through any fixed age node and let $n_i$ denote the number of leaves of fixed age node $i$, $i = 1, \ldots, k$. The age $a_x$ is estimated by a weighted average of the relative age of $x$ and the root, and of the age of each fixed age node $i$ plus the relative additional age of $x$ from the fixed age node up to 0, where the weights are proportional to the number of paths going through the fixed age nodes. More precisely we have

$$a_x = (n_x a_0 p_x/p_0 + s_1 + \cdots + s_k)/(n_x + n_1 + \cdots + n_k),$$

*where*

$$s_i = n_i[a_i + (a_0 - a_i)(p_x - p_i)]/(p_0 - p_i),$$
$$\text{for } i = 1, \ldots, k.$$

Check that $a_x$ is older than $a_1, \ldots, a_k$. If not, set $a_x$ equal to the oldest of them and set all nodes in between equal to the same age. Check also that all estimated ages between $x$ and its leaves are younger than $a_x$. If not, change the older age estimates to equal $a_x$.
5. For all minimum age and maximum age nodes, check that the calculated age estimates are not in conflict with them. If there are conflicts, set all violated node ages equal to the constrained minimum age or maximum age and treat these modified node ages as fixed age nodes. Then repeat steps 3 and 4, treating the original fixed age nodes and the new fixed age nodes as original fixed age nodes and all other node ages as unknown.