THE ROYAL SOCIETY

Published online 7 October 2004

Assessing the quality of molecular divergence time estimates by fossil calibrations and fossil-based model selection

Thomas J. Near[1*] **and Michael J. Sanderson**[2]

[1]*Department of Ecology and Evolutionary Biology, 569 Dabney Hall, University of Tennessee, Knoxville, TN 37669-1610, USA* (*tnear@utk.edu*)

[2]*Section of Evolution and Ecology, One Shields Avenue, University of California, Davis, CA 95616, USA* (*mjsanderson@ucdavis.edu*)

Estimates of species divergence times using DNA sequence data are playing an increasingly important role in studies of evolution, ecology and biogeography. Most work has centred on obtaining appropriate kinds of data and developing optimal estimation procedures, whereas somewhat less attention has focused on the calibration of divergences using fossils. Case studies with multiple fossil calibration points provide important opportunities to examine the divergence time estimation problem in new ways. We discuss two cross-validation procedures that address different aspects of inference in divergence time estimation. 'Fossil cross-validation' is a procedure used to identify the impact of different individual calibrations on overall estimation. This can identify fossils that have an exceptionally large error effect and may warrant further scrutiny. 'Fossil-based model cross-validation' is an entirely different procedure that uses fossils to identify the optimal model of molecular evolution in the context of rate smoothing or other inference methods. Both procedures were applied to two recent studies: an analysis of monocot angiosperms with eight fossil calibrations and an analysis of placental mammals with nine fossil calibrations. In each case, fossil calibrations could be ranked from most to least influential, and in one of the two studies, the fossils provided decisive evidence about the optimal molecular evolutionary model.

Keywords: molecular clock; penalized likelihood; fossil calibration; divergence time estimation

1. INTRODUCTION

Since the publication of '*On the origin of species*' by Charles Darwin (1859), evolutionary biologists have been fascinated by both the genealogical relationships of organisms and the timing of divergences between lineages. This point is apparent in the one figure in Darwin's (1859) seminal volume, which depicts the diversification of lineages both in terms of ancestor–descendant relationships and in time. In modern evolutionary studies there is a strong and persistent desire to obtain accurate estimates of divergence dates among organisms. Since the early 1960s, methods have been developed that use the degree of genetic divergence between organisms to estimate the age of common ancestry (Zuckerkandl & Pauling 1962, 1965), offering great promise of estimating evolutionary divergence times when fossil information is meagre.

Despite the enthusiasm surrounding divergence time estimation using molecular data, several challenges remain. For instance, the recognition of nearly ubiquitous heterogeneity of nucleotide substitution rates among lineages prevents the straightforward reliance on a 'molecular clock': the simple conversion of observed genetic distances into evolutionary rates, and subsequently into evolutionary divergence times (Britten 1986). Another critical issue for molecular dating methodologies is the additional step of fossil calibration that is required for the transformation of estimated branch lengths in phylogenies into absolute divergence times. There has been increased attention to the role of fossil evidence in molecular dating studies (Lee 1999; Smith & Peterson 2002); however, there have been relatively few critical investigations aimed at assessing consistency between independent fossil calibrations and the use of fossil evidence to model changes in substitution rates among lineages (Springer 1997; Shaul & Graur 2002; Soltis *et al*. 2002; Van Tuinen & Dyke 2003).

The contribution of fossils to divergence time studies is substantial, but it is important to consider the several potential sources of error when using fossils to date lineages in molecular divergence time estimates. Such error in fossil dates can result in substantial disagreement between age estimates derived from fossils and molecular dating methods (Benton & Ayala 2003; Bromham & Penny 2003). Perhaps the most common source of error in fossil age estimates stems from the incompleteness of the fossil record, which necessarily leads to a consistent underestimation of any given lineage's age (Marshall 1990). Other sources of error for fossil dates include issues

* Author for correspondence (tnear@utk.edu).

One contribution of 16 to a Discussion Meeting Issue 'Plant phylogeny and the origin of major biomes'.

Table 1. Fossil calibrations used in fossil cross-validation and fossil-based model cross-validation analyses.

fossil calibration dataset	dated node	age (Myr ago)
(*a*) Bremer (2000).		
1	Tofieldiaceae/*Dicolpopollis*	69.5
2	Araceae/*Pistia*	69.5
3	Cymodoceaceae/*Cymodocea*	69.5
4	Arecaceae/*Spinizonocolpites*	89.5
5	Zingiberales/*Spirematospermum*	83.0
6	Typhaceae/*Typha*	69.5
7	Poaceae/*Monoporites*	69.5
8	Flagellariaceae/Joinvilleaceae/Restionaceae *Milfordia*	69.5
(*b*) Springer *et al.* (2003).		
1	Armadillo/sloth-anteater	60.0
2	Feliform and/caniform carnivores	50.0
3	Hippomorph and/ceratomorph perrisodactyls	54.0
4	Hippo/cetacean	52.0
5	Crown node of Cetartiodactyla	65.0
6	Crown node of Paenungulata	54.0
7	*Mus*/*Rattus*	12.0
8	Flying fox and rousette fruit bat/false vampire bat	43.0
9	Shrew/hedgehog	63.0

involved with taxonomic misidentification and their erroneous placement on a phylogenetic tree (Lee 1999), or the phylogeny does not accurately represent evolutionary relationships. Error can also arise when the age estimates of the fossil-bearing rocks are wrong (Conroy & Van Tuinen 2003). In addition, error is frequently encountered in calibrating molecular phylogenies with fossil information when dates are misapplied to a crown group that a fossil subtends, rather than the appropriate stem group in a phylogenetic tree (Doyle & Donoghue 1993; Magallon & Sanderson 2001).

In view of the potential multiple sources of error when using fossil dates as calibration points in molecular dating, several studies recommend use of multiple fossil calibrations (Smith & Peterson 2002; Soltis *et al.* 2002; Conroy & Van Tuinen 2003; Graur & Martin 2004). This may provide a set of fossil age estimates that contain both accurate and inaccurate fossil dates; however, few methodologies have been developed to identify sets of calibration points that are consistent with one another and presumably represent accurate age estimates, versus fossil calibration points that are erroneous (Shaul & Graur 2002; Soltis *et al.* 2002).

Not only are few strategies available to assess consistency between fossil and molecular age estimates when multiple fossil calibration points are available, but the potential utility of fossil information for determining optimal models of molecular evolution in molecular dating studies has also not been explored. Modern molecular dating methods include several options to account for molecular evolutionary rate heterogeneity. These include pruning genes and lineages that exhibit differing rates of change (Takezaki *et al.* 1995; Hedges *et al.* 1996), the selection of different molecular evolutionary models with different rates on different branches (Hasegawa & Kishino 1989; Yoder & Yang 2000), explicit modelling of the evolution of the rate of evolution using Bayesian methods (Thorne *et al.* 1998; Huelsenbeck *et al.* 2000) and the use of non-parametric or semi-parametric models of rate evolution (Sanderson 1997, 2002, 2003).

Semi-parametric rate smoothing using penalized likelihood (Sanderson 2002) has been used in several studies examining patterns and mechanisms of lineage diversification using absolute age estimates (Conti *et al.* 2002; Des Marais *et al.* 2003; Gray & Atkinson 2003; Near 2004). Penalized likelihood combines the likelihood term of a substitution model, which allows a different rate of evolution on every branch, with a penalty term that prevents rates from varying too much across the tree. The relative contributions of the likelihood term and penalty function are controlled by a smoothing parameter (Sanderson 2002). A cross-validation procedure can be used to provide an objective method for choosing the optimal smoothing parameter value, and thus the optimal 'model'.

In this study, we examine two studies that have used multiple fossil calibration points for molecular divergence time estimation. We use these datasets to examine the consistency of independent fossil calibration points and to test whether fossil calibrations provide an alternative criterion for model selection under penalized likelihood.

2. MATERIAL AND METHODS

(a) *Datasets*

We analysed two published studies that used molecular sequence data and multiple fossil calibrations. Bremer (2000) estimated divergence times in monocot angiosperms using plastid *rbcL* data from 91 monocots and calibrations based on eight reference fossils (table 1*a*). We obtained all sequences from GenBank, aligned them using CLUSTALX (Thompson *et al.* 1997), constructed a tree corresponding to fig. 1 in his paper and estimated branch lengths using maximum likelihood as implemented in PAUP* (Swofford 2000) with a GTR + I + Γ model (Swofford *et al.* 1996). Two of the eight minimum age calibrations are redundant with other calibrations and were deleted in the fossil-based model cross-validation described below (see § 2b). For example, Bremer's node C is a descendant of Node A, but both are assigned minimum ages of 69.5 Myr ago (Bremer 2000). Node A's constraint is

redundant because it must have the minimum age of any descendant node that has an assigned minimum age. In analyses of fossil-based model cross-validation, we used two alternative dates as fixed calibrations for the root node of monocots: at one extreme is 140 Myr ago, which was the result obtained by estimation from the sequence data in both Bremer's (2000) original study and in a more recent three-gene analysis (Wikstrom *et al.* 2003); at the other extreme is 105 Myr ago, which are the earliest fossils from the monocot crown group (Magallon & Sanderson 2001). The first fossil pollen evidence for any angiosperms at all occurs at 132–141 Myr ago (Wikstrom *et al.* 2003).

The second study was based on the Murphy *et al.*'s (2001) multi-gene dataset on 42 placental mammals as used in a later Springer *et al.* (2003) paper, estimating divergence times using the fossil record of mammals. Springer *et al.* (2003) used nine minimum age constraints within placental mammals (table 1*b*), and used 105 Myr ago as a calibration date for the crown group node of the placental mammals, selected to split the difference between rather large extremes reported in the literature (Springer *et al.* 2003). For several of these calibration points, both minimal and maximal ages were reported by Springer *et al.* (2003), and one fossil calibration was reported only as a maximal age estimate. In these cases we used the only the minimal age estimates and treated the single maximal age as a minimal age estimate (table 1). The 105 Myr ago calibration date was treated as the mean of a prior probability distribution, which was then used in a Bayesian analysis of divergence times (Thorne *et al.* 1998). We obtained the complete alignment for Springer *et al.*'s (2003) 'dataset 1' from the *Proceedings of the National Academy of Science USA*'s Web site (www.pnas.org), estimated branch lengths as described above and incorporated all nine fossil age constraints in our analyses. When needed for examining fossil-based model cross-validation, we used 105 Myr ago as a fixed calibration point.

(b) *Fossil cross-validation*

We used a method developed by Thomas Near and H. Bradley Shaffer (Near *et al.* 2004) to measure the agreement between molecular age estimates derived using any one single fossil calibration point and all other available fossil calibration points. This method attempts to identify fossil calibrations that generate inconsistent, and potentially erroneous, molecular age estimates.

Given a phylogenetic tree with multiple nodes dated with fossil information, Near *et al.* (2004) fixed the age of a single fossil-dated node and calculated the difference between the molecular and fossil estimates for all other fossil-dated nodes in the phylogeny. The difference between the fossil and molecular ages using a single fossil-dating node, χ, was defined as $D_i = (\mathrm{MA}_i - \mathrm{FA}_i)$, where FA_i is the fossil age estimate and MA_i the molecular age estimate for node i. When fossil age estimates are available for n nodes in the phylogeny, Near *et al.* (2004) defined $\bar{D}_\chi$ as

$$\bar{D}_\chi = \frac{\sum_{i \neq \chi} D_i}{n-1}, \tag{2.1}$$

the mean $\bar{D}_\chi$ for all available nodes (other than node χ) based on the fossil calibration at node χ. In a given step of the cross-validation analysis, the fossil age for a single node (χ) was used as the calibration point in penalized likelihood analysis, and the $\bar{D}_\chi$ and its standard error were calculated from the remaining available fossil-dated calibration point nodes. A plot of $\bar{D}_\chi$ (figure 1*a*) provides a visual assessment of the performance of each fossil, although the interdependence of each $\bar{D}_\chi$ with all other values (because each fossil and its associated error contributes to all other values of D) limits any statistical analyses of these values.

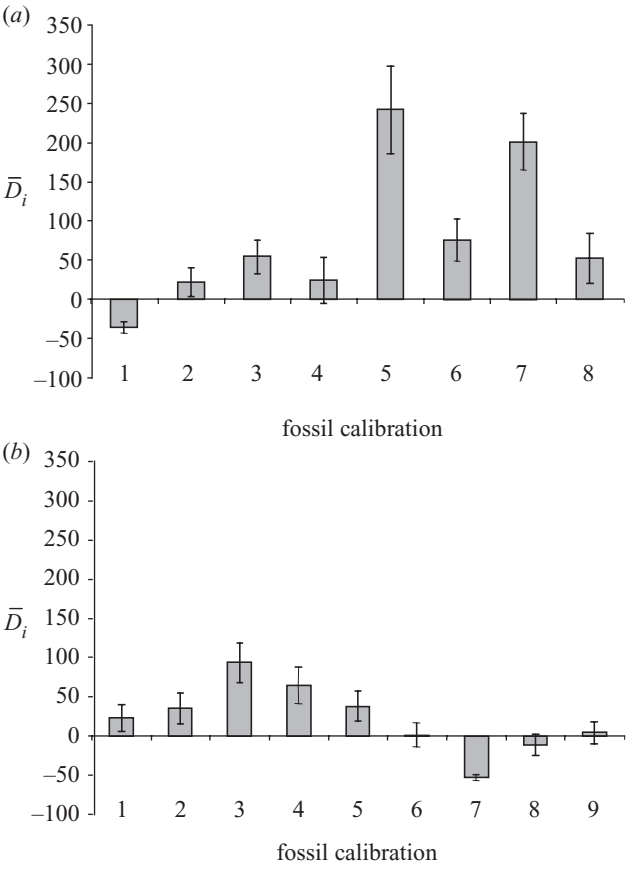

Figure 1. Histogram of mean percentage deviation ($\bar{D}_\chi$) between molecular and fossil age estimates for all nodes using a single fossil calibration point. (*a*) The result from the monocot dataset, and (*b*) the mammal dataset (Springer *et al.* 2003). Error bars, ±1 s.e.m.

A two-step procedure was developed by Near *et al.* (2004) to identify and remove inconsistent fossils from the analysis. First, for each fossil calibration the sum of the squared differences between the molecular and fossil age estimates, SS, was calculated:

$$\mathrm{SS}_\chi = \sum_{i \neq \chi} D_i^2. \tag{2.2}$$

Each calibration point was then ranked based on the magnitude of SS and the fossil with the greatest SS value was identified as the most inconsistent with respect to all other fossils in the analysis. Second, the average squared deviation, s, for all fossils in the analysis was calculated:

$$s = \frac{\sum_{\chi=1}^{n} \sum_{i \neq \chi} D_i^2}{n(n-1)}. \tag{2.3}$$

Following the method of Near *et al.* (2004) to determine the impact of removing fossil calibration from the analysis, we removed the fossil with the greatest SS and recalculated s with the remaining fossil calibration points. This process was continued until only the two fossil calibration points with the lowest and second lowest magnitudes of SS remained. If all calibration points are approximately equally accurate, Near *et al.* (2004) propose that the magnitude of s should decrease by only a small fraction as fossils are removed. However, the removal from the analysis of

extreme outliers that provide very inaccurate calibrations with respect to other fossils should cause an appreciable drop in *s*.

(c) *Fossil-based model cross-validation*

Sanderson (2002) proposed a cross-validation method to determine the optimal level of rate-smoothing in the context of penalized likelihood for a given dataset. This is essentially a model selection procedure that uses a measure of the predictive ability of the family of rate-smoothing models indexed by different values of their smoothing parameter. As originally proposed, cross-validation uses only the information on the estimated number of substitutions on each branch to determine the optimal level of smoothing. In practice, each terminal branch is removed in turn, all times and model parameters are re-estimated and these are used to predict the expected number of substitutions on that removed branch. An overall score is calculated based on the goodness-of-fit of observed and expected numbers of substitutions summed over the tree (Sanderson 2002).

Here, we develop an alternative cross-validation procedure that uses fossils instead of the molecular data. The procedure is restricted to the set of nodes that have minimum and/or maximum age constraints. Any nodes with fixed ages are not considered (although they obviously help in the estimation of all node times). Each constrained node's constraint is removed in turn, all times and parameters are re-estimated and the node's new age estimated in the absence of its former constraint. A score for that node is determined as follows: if the new age violates the former constraint, the score is equal to the absolute value of the difference in age between the estimate and the constraint; if it does not violate the constraint, the score is zero. Alternatively, the score can be calculated in terms of percentage error (or zero, respectively). Notice that any estimate can only violate either a minimum or a maximum age constraint, but not both. Scores for each constrained node are calculated and summed across the tree to obtain an overall cross-validation score. This score is then calculated in turn for a range of model smoothing parameters, varying from effectively clock-like to highly variable, and the smoothing value that corresponds to the lowest cross-validation score is regarded as optimal for the dataset.

3. RESULTS

(a) *Fossil cross-validation*

Cross-validation analysis of the fossil calibration points revealed appreciable average deviation ($\overline{D}_{\chi}$) between the fossil and molecular age estimates in both the monocot and mammal datasets (figure 1). In the monocot dataset, fossil calibrations for the nodes subtending the Zingiberales and the Gramineae (Poaceae) (nodes 5 and 7; table 1) exhibited the largest average deviation, exceeding 200% (figure 1*a*). Interestingly, one fossil calibration in each of the datasets exhibited appreciable average negative deviation (node 1 in the monocot dataset and node 7 in the mammal dataset), indicating that when these fossils were used as calibration points they consistently resulted in molecular age estimates for other nodes that were much younger than their fossil ages.

The summed squared values of the deviations between molecular and fossil ages, SS, are plotted in figure 2. The magnitude of SS matched the average percentage deviation between fossil and molecular ages, and the ranking of these values determined the order in which fossils were excluded from the analysis. In the monocot dataset, removal of the fossil calibrations for the Zingiberales (calibration 5, table 1*a*) and Gramineae (calibration 7, table 1*a*) resulted in a 10-fold decrease of *s* (figure 3*a*). Removal of the remaining fossil calibrations had no impact on the magnitude of *s* for the monocot dataset (figure 3*a*). In the mammal dataset, the fossil calibrations for the split within the perrisodactyls (fossil calibration 3, table 1*b*) and the divergence between the rodents *Mus* and *Rattus* (fossil calibration 7, table 1*b*) exhibited the largest magnitude of SS (figure 2*b*). Removal of these two fossil calibrations resulted in a fivefold decrease in *s* (figure 3*b*). Removal of the remaining fossil calibrations resulted in a slight and continual decrease up to the point where only two fossil calibrations remained (figure 3).

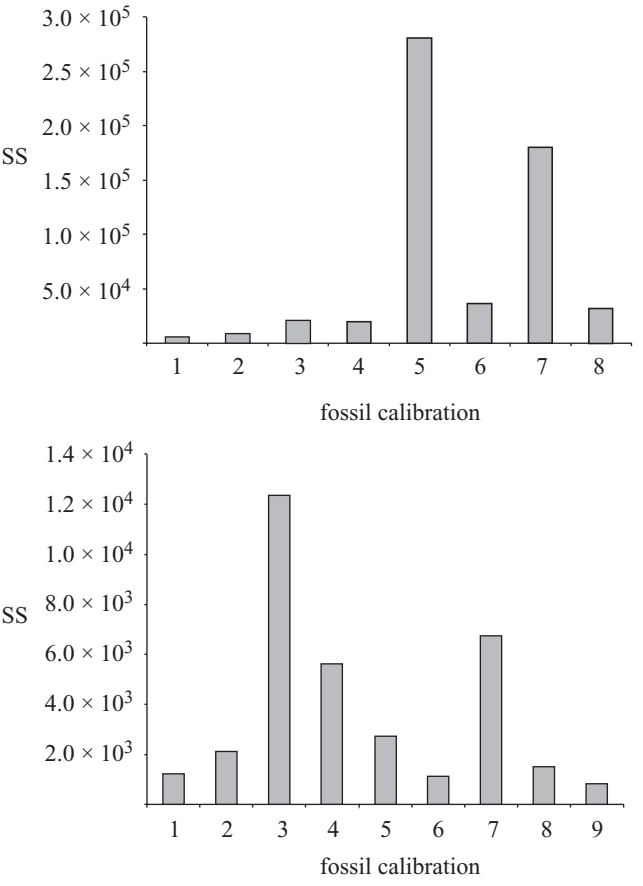

Figure 2. Histogram of the SS values for each fossil calibration. (*a*) The result from the monocot dataset (Bremmer 2000), and (*b*) the mammal dataset (Springer *et al.* 2003).

(b) *Fossil-based model cross-validation*

Results from 'conventional' cross-validation based on the sequence data alone and fossil-based cross-validation are shown for the two datasets in figures 4 and 5. The sequence-based cross-validation score varies smoothly and has an optimal value at an intermediate smoothing level for both datasets. Thus, with respect to the model's ability to predict the distribution of substitutions on branches of the tree, an intermediate level of rate variation performs best: rates that are neither too clock-like, nor too rapidly varying. This agrees with the findings of Sanderson (2002) and many other papers that have used this cross-validation criterion.

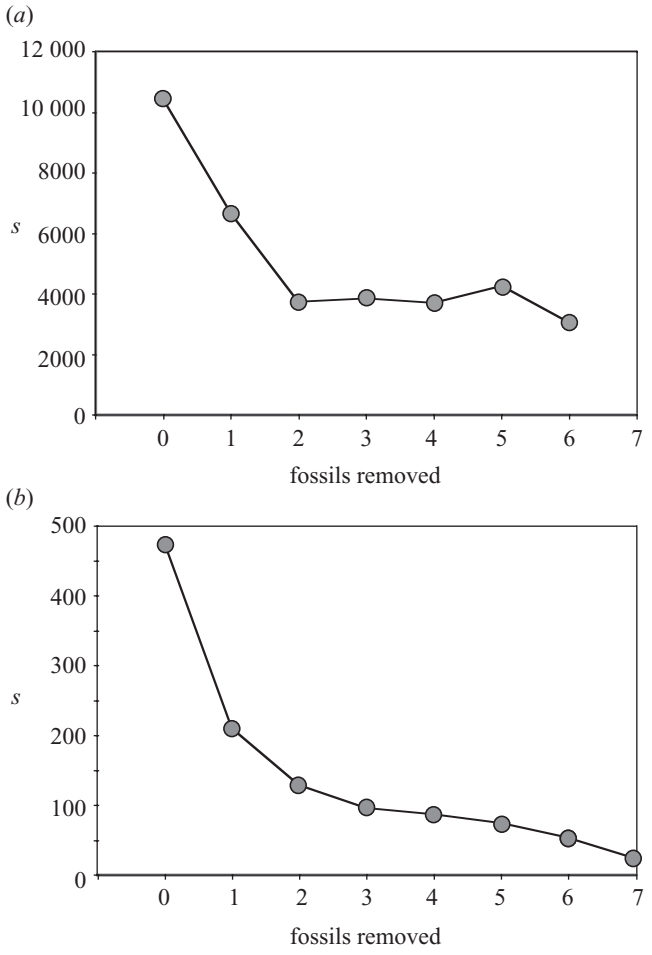

Figure 3. Plots illustrating the effect of removing fossil calibration points on *s*. (*a*) The result from the monocot dataset (Bremer 2000), and (*b*) the mammal dataset (Springer *et al*. 2003).

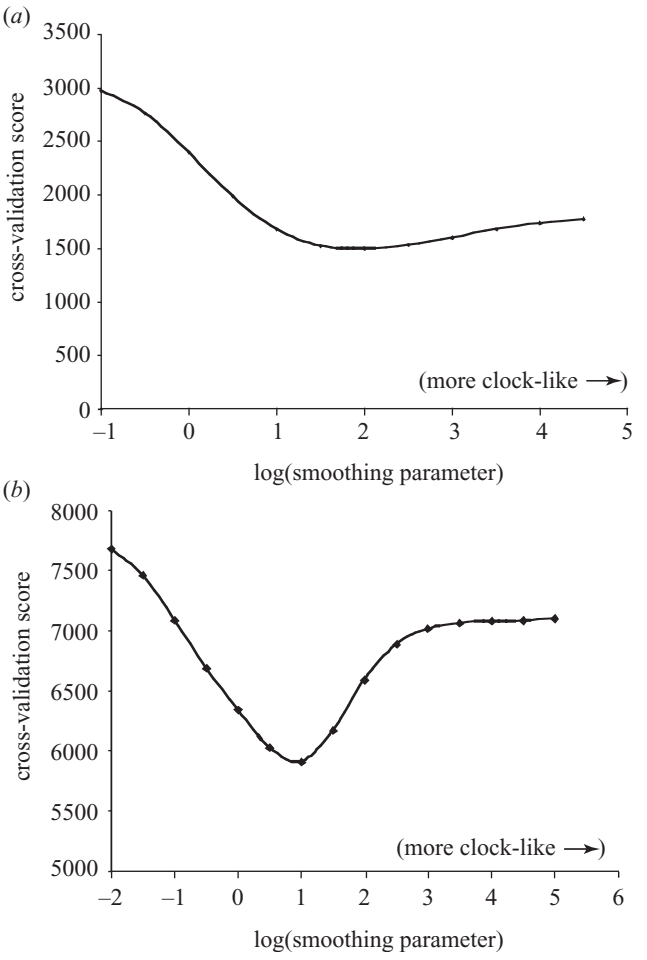

Figure 4. Sequence-based model cross-validation analysis. (*a*) The result from the monocot dataset (Bremer 2000), and (*b*) the mammal dataset (Springer *et al*. 2003).

The results from fossil-based model cross-validation are more variable. A plot of the cross-validation score versus smoothing value (figure 5*a*) for the monocot data shows different results for the two different fixed calibration times of monocots at 105 and 140 Myr ago. The cross-validation error rates are uniformly higher for the 105 Myr ago calibration, peaking at *ca*. 17% at a log smoothing value of *ca*. 2.8. Error rates for the 140 Myr ago calibration are highest (*ca*. 12%) at low smoothing values and lowest (*ca*. 10.8%) at clock-like values *and* at intermediate smoothing values. This probably arises because the 105 Myr ago date, in the context of the sequence rate variation, conflicts the most with the fossil minimum age constraints. Evidence for this is found in a clock-based analysis of the same data, which is consistent with a much older date for crown group monocots of *ca*. 140 Myr ago. The 105 Myr ago date requires much more rapid fluctuation in rates of evolution on the tree, and hence lower smoothing values (and higher error).

Based on *ca*. 10 times the sequence length of the monocot dataset (partial sequences from 22 genes), the mammal dataset exhibits less dramatic shifts in apparent rate than the monocot data. A plot of the cross-validation score versus smoothing value (figure 5*b*) shows that the error rate is highest for a clock model (*ca*. 6.5%) and that the lowest rate occurs at a log smoothing value of *ca*. 0, corresponding to *ca*. 2% average error across constraints. The dataset is a bit unusual because, even though it is non-clock-like (χ^2 test rejects clock: $p \ll 0.001$), the magnitude of rate variation never exceeds about fourfold across the tree regardless of how low a smoothing value is chosen for the model. Cross-validation suggests that it does not matter much what model is selected below a log smoothing value of *ca*. 0.5, but it decisively prefers models that are far from clock-like.

4. DISCUSSION

The great interest in reconstructing divergence times from sequence data and the rapidly accumulating number of case studies is beginning to foster critical investigation of methodologies. Setting them in the context of multiple fossil calibrations can significantly enrich these investigations. We have examined two such approaches: assessing the impact of individual fossil calibrations on divergence time estimates; and using those calibrations to help to select among models of rate variation that vary in how quickly rates change. A third issue can also be addressed based on these two: how well are methods performing? In particular, are so-called relaxed clock methods estimating divergence times more accurately than methods that assume a clock?

Fossil cross-validation analysis of the monocot and mammal datasets both indicated the same thing. Fossil

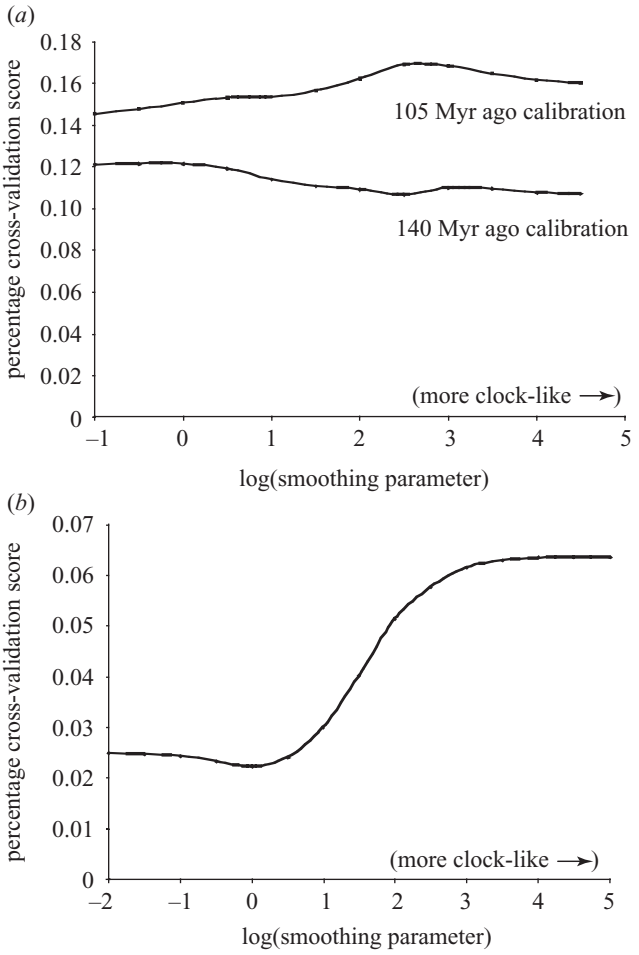

Figure 5. Fossil-based model cross-validation analysis. (*a*) The result from the monocot dataset (Bremer 2000), and (*b*) the mammal dataset (Springer *et al.* 2003).

calibrations are not created equal. Some of them have a much more dramatic impact on estimating the age of uncalibrated nodes in the tree than do others. This effect may depend on the position of the nodes that are dated with error-laden fossils relative to nodes calibrated with more accurate fossil ages (Near *et al.* 2004). Not surprisingly, removal of just one or two of these high-impact calibrations leaves a set of calibrations that are much more consistent with each other. Apparently the effect of calibrations in these datasets is skewed: one or two fossils are 'outliers' and the remaining ones are more or less consistent with each other (figure 3). Interestingly, the plots showing the effect of fossil removal have slightly different patterns between the two datasets. After the removal of two fossils in the monocot dataset (figure 3*a*) there is very little change in the *s* value. However, removal of fossils in the mammal dataset results in a continuous decrease in the *s* value to the point where only two fossil calibration points remain (figure 3*b*). This pattern is reflected in the distribution of SS values in the two datasets: the monocot dataset is more bimodal than the mammal dataset, with individual fossil calibration points having either large or small SS values. Because there are many ways in which a fossil calibration can introduce error into divergence time studies (for example, underestimate of true age as a result of an incomplete fossil record, mistaken assignment to wrong node, error in stratigraphic correlations, etc.), this type of cross-validation may be useful in providing hints about which calibrations might deserve further scrutiny. However, we are not advocating an over-interpretation of the molecular data, as our methods assume both that the phylogeny is accurate and molecular branch lengths are estimated with negligible error.

With respect to rate heterogeneity, once the model of molecular evolution departs from a simple one-rate molecular clock, the divergence time problem enters a realm of model selection in which the number of models is effectively infinite. However, fossil calibrations can add additional criteria for model selection. Based on the mixed results from the two datasets, it seems that the fossil-based model cross-validation has some potential for helping to choose models when there are sufficient data. The monocot *rbcL* data were apparently too noisy for any consistent pattern to emerge from incorporation of several calibrations. The mammal data, however, revealed a clear picture pointing to models performing best when they were sufficiently non-clock-like (figures 4 and 5).

Together, these two cross-validation procedures also shed light on the basic question of whether relaxed clock methods are working. By discarding one or two problematic fossil calibrations that impose large deviations on the remaining calibrated nodes, it is possible for penalized likelihood methods to do a reasonable job of estimating divergence times, at least in the sense of obtaining sets of ages that are largely consistent. Moreover, in the fossil-based model cross-validation it is possible to directly compare the average age error (cross-validation score) under a range of models, from completely clock-like to highly rate variable. For the well-behaved mammal data, relaxing the clock improves the average age error of a node from 6.5% to just over 2%. The effect is not as consistent or dramatic in the monocot data, but even there, for the 105 Myr ago calibration, clock-like models perform worse than very unclock-like models. However, for the 140 Myr ago calibration the opposite is true. We think these procedures will form a useful set of tools to assay other divergence time methods.

Fossil calibrations are simultaneously both informative about divergence time and rates and also subject to significant error (Lee 1999; Smith & Peterson 2002). Addition of a substantial number of fossils to empirical studies promises to reveal the extent of rate variation in sequence evolution and to provide better information on the limits to accurate age estimation. Ultimately, we may well conclude that accurate divergence time estimates require multiple reliable calibrations. That is too bad, if true, but it may be true. Until this largely empirical question is resolved, it may be desirable to evaluate existing and new methodologies in the context of a few carefully chosen systems that offer numerous fossil calibrations. If methods can be fine-tuned to succeed in problems where cross-validation is feasible, then there may be some hope to extend them to more difficult problems with fewer available fossil calibrations. Ironically, the systems in which divergence time estimation from sequence data is needed most critically are the ones with few or no good calibrations (e.g. Darwin's finches, East African cichlids). However, the fossil record is rich in many taxa. Perhaps we should learn to walk in the context of these systems before learning to run in the real world.

We thank the organizers for their invitation of our participation in the 'Plant phylogeny and the origin of major biomes' Discussion Meeting at The Royal Society. The methods of fossil cross-validation presented in this paper were developed with H. Bradley Shaffer, and we appreciate his valuable insight and suggestions.

REFERENCES

Benton, M. J. & Ayala, F. J. 2003 Dating the tree of life. *Science* **300**, 1698–1700.

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

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

Bromham, L. & Penny, D. 2003 The modern molecular clock. *Nature Rev. Genet.* **4**, 216–224.

Conroy, C. J. & Van Tuinen, M. 2003 Extracting time from phylogenies: positive interplay between fossil and genetic data. *J. Mammal.* **84**, 444–455.

Conti, E., Eriksson, T., Schonenberger, J., Sytsma, K. J. & Baum, D. A. 2002 Early tertiary out-of-India dispersal of Crypteroniaceae: evidence from phylogeny and molecular dating. *Evolution* **56**, 1931–1942.

Darwin, C. 1859 *On the origin of species*. London: John Murray.

Des Marais, D. L., Smith, A. R., Britton, D. M. & Pryer, K. M. 2003 Phylogenetic relationships and evolution of extant horsetails, equisetum, based on chloroplast DNA sequence data (*rbc*L and *trn*L-F). *Int. J. Pl. Sci.* **164**, 737–751.

Doyle, J. A. & Donoghue, M. J. 1993 Phylogenies and angiosperm diversification. *Paleobiology* **19**, 141–167.

Graur, D. & Martin, W. 2004 Reading the entrails of chickens: molecular time-scales of evolution and the illusion of precision. *Trends Genet.* **20**, 80–86.

Gray, R. D. & Atkinson, Q. D. 2003 Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**, 435–439.

Hasegawa, M. & Kishino, H. 1989 Confidence limits on the maximum-likelihood estimate of the homonid tree from mitochondrial-DNA sequences. *Evolution* **43**, 672–677.

Hedges, S. B., Parker, P. H., Sibley, C. G. & Kumar, S. 1996 Continental breakup and the ordinal diversification of birds and mammals. *Nature* **381**, 226–229.

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

Lee, M. S. Y. 1999 Molecular clock calibrations and metazoan divergence dates. *J. Mol. Evol.* **49**, 385–391.

Magallon, S. & Sanderson, M. J. 2001 Absolute diversification rates in angiosperm clades. *Evolution* **55**, 1762–1780.

Marshall, C. R. 1990 The fossil record and estimating divergence times between lineages: maximum divergence times and the importance of reliable phylogenies. *J. Mol. Evol.* **30**, 400–408.

Murphy, W. J. (and 10 others) 2001 Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294**, 2348–2351.

Near, T. J. 2004 Estimating divergence times of notothenioid fishes using a fossil-calibrated molecular clock. *Antarctic Sci.* **16**, 37–44.

Near, T. J., Meylan, P. A. & Shaffer, H. B. 2004 Assessing concordance of fossil calibration points in molecular clock studies: an example using turtles. *Am. Nat.* (In review.)

Sanderson, M. J. 1997 A non-parametric approach to estimating divergence times in the absence 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 evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**, 301–302.

Shaul, S. & Graur, D. 2002 Playing chicken (*Gallus gallus*): methodological inconsistencies of molecular divergence date estimates due to secondary calibration points. *Gene* **300**, 59–61.

Smith, A. B. & Peterson, K. J. 2002 Dating the time of origin of major clades. *A. Rev. Earth Planet. Sci.* **30**, 65–88.

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

Springer, M. S. 1997 Molecular clocks and the timing of the placental and marsupial radiations in relation to the Cretaceous–Tertiary boundary. *J. Mammal. Evol.* **4**, 285–302.

Springer, M. S., Murphy, W. J., Eizirik, E. & O'Brien, S. J. 2003 Placental mammal diversification and the Cretaceous–Tertiary boundary. *Proc. Natl Acad. Sci. USA* **100**, 1056–1061.

Swofford, D. L. 2000 *PAUP*; phylogenetic analysis using parsimony (* and other methods)*. Sunderland, MA: Sinauer.

Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. 1996 Phylogenetic inference. In *Molecular systematics* (ed. D. M. Hillis, C. Moritz & B. K. Mable), pp. 407–514. Sunderland, MA: Sinauer.

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

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. 1997 The CLUSTALX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **22**, 4673–4680.

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

Van Tuinen, M. & Dyke, G. J. 2003 Calibration of galliform molecular clocks using multiple fossils and genetic partitions. *Mol. Phylogenet. Evol.* **30**, 74–86.

Wikstrom, N., Savolainen, V. & Chase, M. W. 2003 Angiosperm divergence times: congruence and incongruence between fossils and sequence divergence estimates. In *Telling the evolutionary time: molecular clocks and the fossil record* (ed. P. C. J. Donoghue & M. P. Smith), pp. 142–165. London: Taylor & Francis.

Yoder, A. D. & Yang, Z. 2000 Estimation of primate speciation dates using local molecular clocks. *Mol. Biol. Evol.* **17**, 1081–1090.

Zuckerkandl, E. & Pauling, L. 1962 Molecular disease, evolution, and genic heterogeneity. In *Horizons in biochemistry* (ed. M. Kasha & B. Pullman), pp. 189–225. New York: Academic.

Zuckerkandl, E. & Pauling, L. 1965 Evolutionary divergence and convergence in proteins. In *Evolving genes and proteins* (ed. V. Bryson & H. J. Vogel), pp. 97–166. New York: Academic.