Constraints on clade ages from fossil outgroups

Matthew M. Hedman

Abstract.—This paper presents a method for constraining the age of a clade with the ages of the earliest fossil specimens in that clade's outgroups. Given a sufficiently deep, robust, well-resolved, and stratigraphically consistent cladogram, this method can yield useful age constraints even in the absence of specific information about the fossil preservation and recovery rates of individual taxa. The algorithm is applied to simulated data sets to demonstrate that this method can yield robust constraints of clade ages if there are sufficient fossil outgroups available and if there is a finite chance that additional outgroups may be discovered in the future. Finally, the technique is applied to actual fossil data to explore the origin of modern placental mammals. Using data from recently published cladograms, this method indicates that if all Mesozoic eutherians are regarded as outgroups of Placentalia, then the last common ancestor of modern placental mammals and their Cenozoic allies lived between 65 and 88–98 million years ago, depending on the assumed cladogram and the number of outgroups included in the analysis.

Matthew Hedman. Department of Astronomy, Space Sciences Building, Cornell University, Ithaca, New York 14853. E-mail: mmhedman@astro.cornell.edu

Introduction

The fossil record is an important tool for reconstructing the history of life on Earth, providing critical constraints on when various groups of organisms arose or went extinct. Even analyses of DNA sequence data use fossil evidence to translate measures of genetic similarity into age estimates for cladogenic events. However, fossils represent only a sample of all the organisms that belonged to a given taxon, so it is not always straightforward to translate the ages of recovered fossils into information about the durations of specific clades. In particular, although a clade must be at least as old as its earliest fossil representative, it is not so obvious how to determine the maximum possible age of any given taxon based on fossil data.

Fossil preservation and recovery rates determine how long taxa could have existed prior to their oldest known fossil representatives, so many efforts to constrain clades' true durations first estimate the recovery rates of the relevant taxa in various ways (e.g., Strauss and Sadler 1989; Marshall 1997; Foote et al. 1999). However, when estimating the age of a given clade, one must consider not only that clade's own stratigraphic record but also its phylogenetic relationships with other taxa.

For example, a large stratigraphic gap between the earliest known examples of two closely related taxa could imply the existence of an extended ''ghost lineage'' without fossil representatives (Norell 1993). Although there is still considerable debate on how best to combine stratigraphic and phylogenetic data (especially when there appear to be incongruities between the reconstructed phylogeny and the fossil record; see, e.g., Fisher 2008 and references therein), combinations of fossil age estimates and phylogenetic analyses have already proven useful for estimating rates of character change (Ruta et al. 2006) and placing older limits on clade ages (Marshall 2008). It therefore seems likely that the best constraints on the age of a given taxon will be those that utilize both phylogenetic and stratigraphic information.

This work presents a method for constraining a particular clade's age with the ages of the oldest fossil representatives of that clade's outgroups. Perhaps surprisingly, it turns out that given a sufficiently deep, robust, well-resolved, and stratigraphically consistent cladogram, these fossil dates by themselves can provide interesting constraints on clade ages. The method presented here therefore differs from previously published methods

that use estimates of fossil recovery rates derived from the fossil record (Hueslenbeck and Rannala 1997; Wagner 2000) or relative lineage durations derived from molecular data (Marshall 2008) to help constrain the ages of the relevant taxa. This method could therefore provide a useful way to quantitatively evaluate hypotheses regarding the timing of cladogenic events in cases where the fossil recovery rates are difficult to constrain or molecular data are not available. On the other hand, the method presented in this paper relies on a relatively secure phylogeny, so its utility may be limited in situations where the phylogenetic reconstruction has significant uncertainties. Nevertheless, the analyses presented here could still help clarify how much information fossil outgroups can provide on clade ages, and so may be useful in efforts to develop a more generally applicable approach to dating the origins of taxa.

For clarity, the basic algorithm for constraining a clade's age is first illustrated by using the relatively simple case of a well-resolved, stratigraphically consistent cladogram with Hennig-comb topology. This is followed by a briefer discussion of how this method can be extended to handle cladograms with more complex or uncertain topologies, unresolved nodes, stratigraphic gaps, etc. Simulated data sets are used to explore the properties of the constraints derived by using this method and the sensitivity of the results to assumed parameters and sampling rates. These preliminary tests demonstrate that in certain limits this method can indeed produce sensible constraints on clade ages.

In order to illustrate how this technique might be applied to actual fossil data, I investigate the origin of modern placental mammals. The earliest clear fossil examples of Placentalia (the crown group that includes the last common ancestor of modern placental mammals and all of that animal's descendants) date from shortly after the K/T boundary 65 Myr ago (Rose and Archibald 2005). However, recent DNA analyses indicate that the last common ancestor of modern placental mammals lived around 105 Ma (e.g., Springer et al. 2003). This 40-Myr gap between the molecular estimate and the earliest fossil members of Placentalia has been discussed at length in the literature (e.g., Rose and Archibald 2005), but despite some important work by Foote et al. (1999) and Tavare et al. (2002), it has been difficult to quantify this seeming inconsistency between the molecular and fossil evidence. The algorithm developed here provides a new way to evaluate the magnitude and the significance of the difference between the molecular analyses and the fossil data.

Methods

The Dating Algorithm

Nomenclature.—The basic algorithm developed here can best be explained by first considering a simple case involving a cladogram like that shown in Figure 1. (More complex situations will be discussed below.) The nodes on this cladogram are labeled by the numbers $j = 1, 2, 3, \ldots, n$ and the times for each node are denoted by $\tau_1, \tau_2, \ldots, \tau_n$. Note that the topology of the cladogram requires that $\tau_1 < \tau_2 < \ldots < \tau_{n-1} < \tau_n$. The taxa derived from each node are designated by the numbers $J = I, II, III, \ldots, N$. The earliest fossil representatives of these taxa occur at times t_I, $t_{II}, t_{III}, \ldots, t_N$. For simplicity, assume for now that the fossil record is stratigraphically consistent with this cladogram, so $t_I < t_{II} < \ldots < t_{N-1} < t_N$. For concreteness, the τ_j and t_J will be described as times throughout this paper, but in practice these parameters could just as well be regarded as depths in a stratigraphic column (Strauss and Sadler 1989) or positions in a sequence of fossiliferous localities (Wagner 1995). In any case, the goal is to obtain constraints on the possible values of τ_j based on the (known) values of t_J.

Statistical Framework.—This analysis will be performed using a Bayesian statistical framework instead of a classical (frequentist) statistical approach. In classical statistics, unknown parameters like τ_j have fixed values and observed data are used to obtain estimates and confidence intervals on these parameters. By contrast, a Bayesian analysis yields the probability that the unknown parameters have particular values based on their consistency with the observed data. For
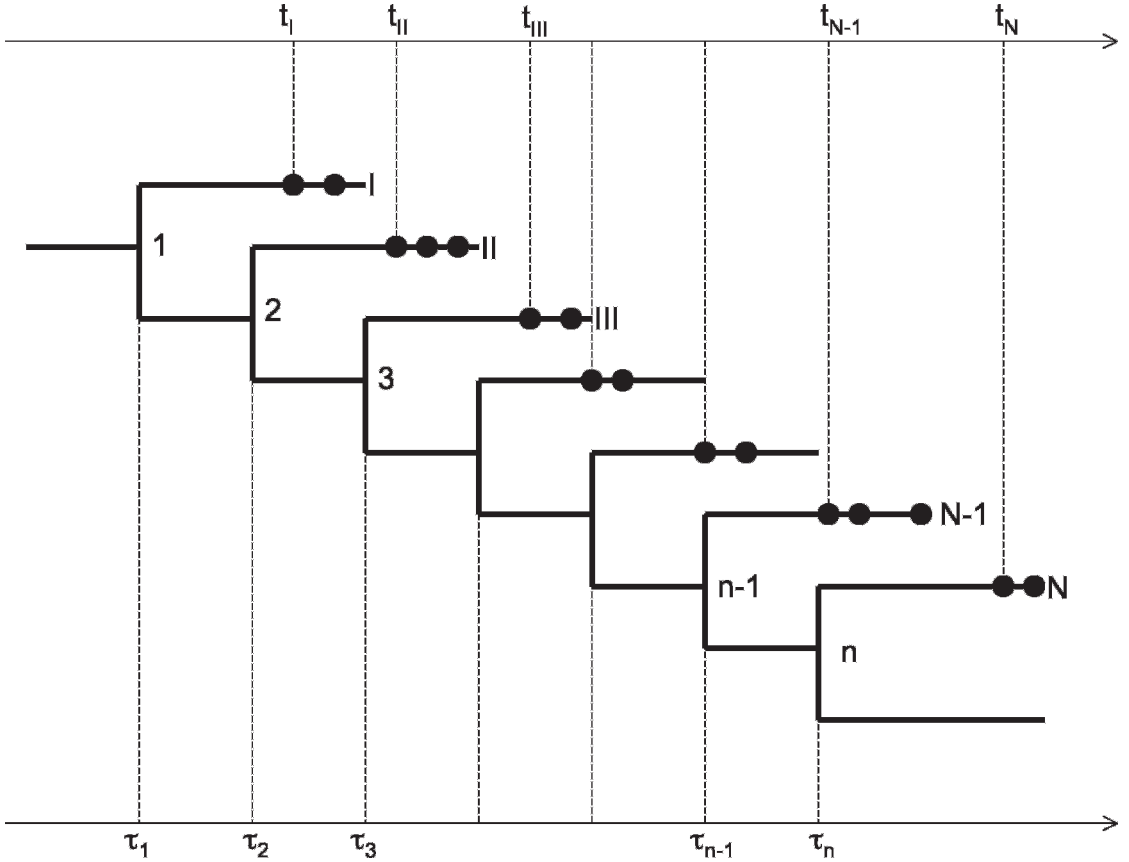
FIGURE 1.   A diagram of a simple cladogram that illustrates the notation system used in this paper. Sequential nodes in this simple cladogram are denoted by the numbers $j = 1, 2, 3... n - 1, n$. The fossil taxa derived from these nodes are named $J = $ I, II, III, … $N - 1, N$, and the dots indicate the age of the known specimens of each taxon. The (known) ages of the earliest fossil specimen of each species are given by t_I, t_{II}, etc. The (unknown) ages of the nodes are given by the variables τ_1, τ_2, etc.

example, consider a probabilistic process that yields a set of observable parameters O. The probability of observing particular values for O depends on a set of input parameters X as described by the function $P(O \mid X)$. The probability that the input parameters have the values X given the observed values of O can then be expressed as follows:

$$p(X \mid O) = C\, \pi(X)\, P(O \mid X), \qquad (1)$$

where C is a normalization constant and $\pi(X)$ is the so-called prior probability, a function that specifies the probability that the parameters X have a set of given values in the absence of any information about O. The function p, which specifies the probability of X given the observed values of O, is known as the posterior probability distribution function. Further background on Bayesian meth-

ods can be found in statistics textbooks like Lupton (1993), and see Strauss and Sadler (1989) for a demonstration of how this technique can be applied to fossil data.

For this study, the observable parameters are the ages of each taxon's earliest fossil representatives t_J, and the unknown parameters we wish to constrain are the node times τ_j. However, the probability that the first fossil specimen of a given taxon has a certain age t_J depends not only on the corresponding node time τ_j, but also on $r_J(t)$, the (possibly time-dependent) fossil recovery rate of taxon J. In general, the probability of obtaining a given value of t_J can be written as follows:

$$P(t_J \mid \tau_j, r_J(t)) = \theta(t_J - \tau_j) f(\tau_j, t_J, r_J(t)). \quad (2)$$

Here $\theta(x)$ is the Heaviside step function ($\theta(x) = 1$ for $x > 0$ and $\theta(x) = 0$ for $x < 0$), which

imposes the constraint that $t_J > \tau_j$. The function f specifies the probability that the first fossil representative of the clade occurs at a given time t_J after τ_j. If $r_J(t)$ is allowed to vary with time, then f can become almost arbitrarily complex. However, f can always be expressed as the following product:

$$f(\tau_j, t_J, r_J(t)) = r_J(t_J) \prod_{k=0}^{K} [1 - r_J(\tau_j + k\delta t)\delta t], \quad (3)$$

where δt is an infinitely small time step and $K = (t_J - \tau_j)/\delta t$. This product is explicitly the probability that a fossil is recovered at time t_J times the probability that no fossil is recovered in any previous time step. If $r_J(t)$ is assumed to be constant in time, then this product reduces to the simple exponential form given in Strauss and Sadler (1989) and other texts.

Given the above form for $P(t_J | \tau_j, r_J(t))$, equation (1) provides an expression for the posterior probability of τ_j and $r_J(t)$ given an observed value of t_J:

$$p(\tau_j, r_J(t) | t_J) = C \, \pi(\tau_j, r_J(t)) \, P(t_J | \tau_j, r_J(t)). \quad (4)$$

Because the prior probability that τ_j has a given value is not expected to be correlated with the fossil recovery rate $r_J(t)$, the prior probability can be written as the product of two terms, one that gives the prior probability of τ_j and another that gives the prior probability of $r_J(t)$:

$$\pi(\tau_j, r_J(t)) = \pi_1(\tau_j)\pi_2(r_J(t)). \quad (5)$$

Combining equations (3)–(5) then yields

$$p(\tau_j, r_J(t) | t_J) = C \, \pi_1(\tau_j) \, \theta(t_J - \tau_j) \, \pi_2(r_J(t))$$
$$\times f(\tau_j, t_J, r_J(t)). \quad (6)$$

This is the joint posterior probability distribution of τ_j and $r_J(t)$. To obtain the probability distribution of τ_j alone, one must integrate the above expression over all possible values of r_J at every time step:

$$p(\tau_j | t_J) = \int \ldots \int \int p(\tau_j, r_J(t) | t_J)$$
$$dr_J(\tau_j)dr_J(\tau_j + \delta t)\ldots dr_J(t_J). \quad (7)$$

Because only the functions π_2 and f depend on $r_J(t)$, this expression can be written as follows:

$$p(\tau_j | t_J) = C \, \pi_1(\tau_j) \, \theta(t_J - \tau_j) \, F(\tau_j, t_J), \quad (8)$$

where the function F is

$$F(\tau_j, t_J) = \int \ldots \int \int \pi_2(r_J(t))f(\tau_j, t_J r_J(t))$$
$$dr_J(\tau_j)dr_J(\tau_j + \delta t)\ldots dr_J(t_J). \quad (9)$$

Further progress requires specifying the priors. In the absence of any other information, it is reasonable to assume that τ_j is equally likely to occur at any time after the formation of the previous node at τ_{j-1}, so one can assume

$$\pi_1(\tau_j) = C' \, \theta(\tau_j - \tau_{j-1}), \quad (10)$$

where C' is a constant.

The optimal choice of $\pi_2(r_J(t))$ is less obvious, but it turns out to be useful to first rewrite the time-dependent recovery rate as the product of two terms:

$$r_J(t) = R_J \, s(t), \quad (11)$$

where R_J is the mean recovery rate and $s(t)$ represents the time variations in $r_J(t)$. The prior on r_J can then be expressed as

$$\pi_2(r_J(t)) = \pi_2'(R_J) \, \pi_2''(s(t)). \quad (12)$$

Together with equations (3) and (9), this yields the following expression for F:

$$F(\tau_j, t_J)$$
$$= \int \pi_2'(R_J) \times \left\{ \int \pi_2''(s(t_J))R_J s(t_J)ds(t_J) \right.$$
$$\times \prod_{k=1}^{K} \int \pi_2''(s(t_J + k\delta t))$$
$$\times [1 - R_J s(t_J + k\delta t)\delta t]$$
$$\left. ds(t_J + k\delta t) \right\} dR_J$$
$$(13)$$

Given no other information about the variations in the fossil recovery rate, it is reasonable to assume that at any time $s(t)$ is equally likely to have any value between 0 and some arbitrary maximum value S. In this case, the above expression reduces to

$$F(\tau_j, t_J) = F_o \int \pi_2'(R_J)R_J$$
$$\prod_{k=1}^{K} \left(1 - \frac{R_J S}{2}\delta t\right) dR_J, \quad (14)$$

where F_o is a constant. This expression can be rewritten as

$$F(\tau_j, t_J) = F_o \int \pi_2'(R_J) R_J e^{-R_J S(t_J - \tau_j)/2} dR_J. \quad (15)$$

Now to specify the prior on the mean recovery rate. It might at first seem reasonable to assume that R_J is equally likely to have any value between zero and some maximal value. However, with this prior the function F is sensitive to the arbitrarily chosen maximal value. In order to make the posterior probability less explicitly dependent on the possible range of fossil recovery rates, consider an alternative prior with the following form:

$$\pi_2'(R_J) \propto \frac{1}{R_J^2} \left[1 - e^{-R_J/R_C} \right]^2. \quad (16)$$

Note that the term in the square brackets prevents this function from becoming infinite as R_J approaches zero. In the limit where the parameter R_C approaches zero, the final integral in equation (15) becomes a constant, which removes any explicitly preservation-dependent terms in the expression for p. This choice of priors is therefore equivalent to assuming that $t_J - \tau_j$ is equally likely to have any positive value. In the absence of other information, this is not an unreasonable assumption to make, and furthermore it has practical advantages for this analysis. In particular, equation (8) reduces to a very simple form (note we also make the dependence on the previous node time explicit):

$$p(\tau_j \mid \tau_{j-1}, t_J) = c\, \theta(\tau_j - \tau_{j-1})\, \theta(t_J - \tau_j), \quad (17)$$

where c is just a constant determined by the normalization criterion:

$$\int_0^\infty p(\tau_j \mid \tau_{j-1}, t_J) d\tau_j = 1, \quad (18)$$

which requires that $c = 1/(t_J - \tau_{j-1})$, so the posterior probability for τ_j as a function of t_J and τ_{j-1} can finally be written as

$$p(\tau_j \mid \tau_{j-1}, t_J) = \frac{\theta(\tau_j - \tau_{j-1})\theta(t_J - \tau_j)}{t_J - \tau_{j-1}}. \quad (19)$$

Equation (19) demonstrates that with the above choice of priors, node j is equally likely to occur any time between the age of the earliest fossil representative of a taxon derived from that node and the appearance of the previous node in the cladogram. These are the weakest possible constraints that can be placed on τ_j, so this choice of priors should produce conservative constraints on the node times. Other choices of the prior π_2' based upon estimates of the fossil recovery rate could potentially place tighter constraints on the clade ages and should be explored in future work. However, as will be shown below, even this very simplistic formulation can provide useful information about the age of certain nodes in the cladogram.

Of course, equation (19) alone does not provide any practical information about the age of node j because it depends explicitly on the age of the earlier node τ_{j-1}, which is a priori an unknown parameter just like τ_j. To obtain useful constraints on node times, it is necessary to extend this analysis to the entire cladogram. The probability that the earliest fossil representatives of the various clades are $t_I, t_{II}, t_{III}, \ldots t_N$, given the node time $\tau_1, \tau_2, \tau_3, \ldots \tau_n$ and the recovery rates $r_I(t), r_{II}(t), r_{III}(t), \ldots r_N(t)$ is simply a product of terms:

$$P(t_I, t_{II}, \ldots t_N \mid \tau_1, \tau_2, \ldots \tau_n, r_I(t), r_{II}(t), \ldots r_N(t))$$
$$= P(t_I \mid \tau_1, r_I(t)) * P(t_{II} \mid \tau_2, r_{II}(t)) * \ldots *$$
$$P(t_N \mid \tau_n, r_N(t)), \quad (20)$$

where each term $P(t_J \mid \tau_j, r_J(t))$ has the form given by equation (2). The desired probability distribution function for the τ_j is this expression multiplied by the appropriate priors and integrated over all possible values for the $r_J(t)$. Again, we can choose the priors on the $r_J(t)$ such that the preservation-dependent terms integrate to constants. Finally, the topology of the cladogram suggests that we assume a prior for the τ_j in which all possible combinations of τ_j are equally likely subject to the constraint that $\tau_1 < \tau_2 < \tau_3 < \ldots < \tau_{n-1} < \tau_n$. For practical reasons that will become evident later, it is also useful to specify that the first node in the cladogram cannot occur earlier than a time τ_0 prior to t_I.

With these assumptions, the joint posterior probability of all node times is given by the product

$$p(\tau_1, \tau_2, \ldots \tau_n \mid \tau_0, t_I, t_{II}, \ldots t_N)$$
$$= p(\tau_1 \mid \tau_0, t_I) * p(\tau_2 \mid \tau_1, t_{II}) * \ldots *$$
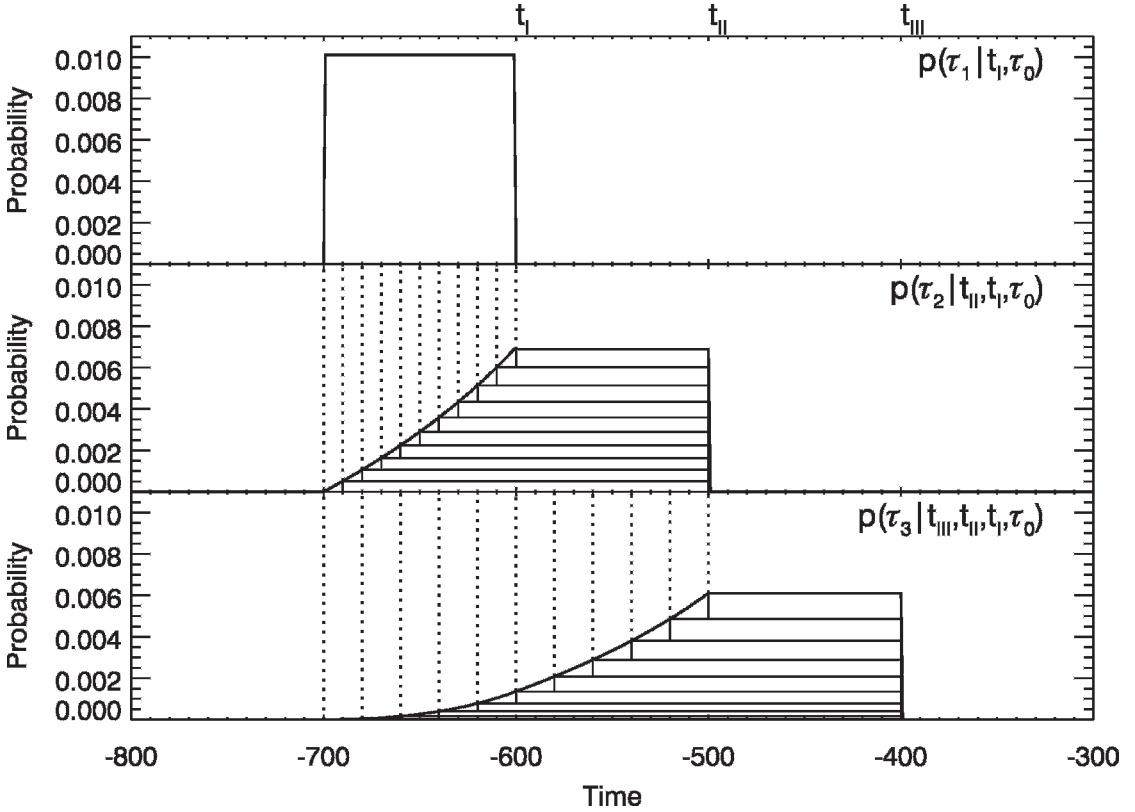$$p(\tau_n \mid \tau_{n-1}, t_N), \quad (21)$$

FIGURE 2. A graphical illustration of the proposed iterative method for computing the probability that a given node occurs at a given time. Each of the three panels above shows the posterior probability distribution function (p.d.f.) of one of the first three nodes in a cladogram like the one illustrated in Figure 1. These functions indicate the probability that a given node occurred as a function of time, which in this case has arbitrary units. The ages of the earliest fossil representatives of taxa I, II, and III are indicated at the top of the figure as t_I, t_{II}, and t_{III}, respectively. The top panel shows the p.d.f. of the age of node 1 (τ_1), which by construction is taken to be equally likely to occur at any time between t_I and an arbitrarily selected time τ_0 (here taken to be -700). The middle panel shows the p.d.f. of the age of node 2 (τ_2) derived using the above p.d.f. for τ_1. The dotted vertical lines indicate different possible values of τ_1. For each possible τ_1 value, τ_2 is equally likely to occur at any time between τ_1 and t_{II}, which corresponds to a simple bar-like p.d.f. (shown as horizontal bars in the plot). By summing together all the p.d.f.s for all different possible values of τ_1, one can obtain the full p.d.f. for τ_2. Similarly, the p.d.f. of τ_3 shown in the bottom panel is constructed from bar-like p.d.f.s for different values of τ_2. Note that in general different values of the prior node times will have different probabilities, so the different p.d.f.s must be weighted appropriately prior to the summation. Graphically, this means that each bar-like p.d.f in τ_3 covers a total area that is proportional to the probability of τ_2.

where the individual terms $p(\tau_j \mid \tau_{j-1}, t_J)$ are given by equation (19). One can obtain an explicit probability distribution function for the age of the most recent node τ_n from this expression by integrating over all possible values of the earlier node times:

$$p(\tau_n \mid \tau_0, t_I, t_{II}, ...t_N)$$

$$= \int_0^\infty ... \int_0^\infty \int_0^\infty p(\tau_1, \tau_2, ...\tau_n \mid \tau_0, t_I, t_{II}, ...t_N)$$

$$d\tau_1 d\tau_2 ... d\tau_{n-1}. \qquad (22)$$

Calculation of the Probability Distribution

Functions.—The most straightforward procedure for evaluating equation (22) to obtain probability distribution functions for the individual node times is an iterative method illustrated in Figure 2. This approach starts with an explicit probability distribution function for the age of the earliest node in the cladogram τ_1, which is equally likely to occur anytime between the known age of the earliest fossil representative of taxon I (t_I) and the assumed maximum possible age $\tau_0 < t_I$. The posterior probability distribution for τ_1, given τ_0 and t_I, is simply:

$$p(\tau_1 \mid \tau_0, t_\mathrm{I}) = \frac{\theta(\tau_1 - \tau_0)\theta(t_\mathrm{I} - \tau_1)}{t_\mathrm{I} - \tau_0} \qquad (23)$$

For a given value of τ_1, equation (19) tells us that

$$p(\tau_{21} \mid \tau_1, t_\mathrm{II}) = \frac{\theta(\tau_2 - \tau_1)\theta(t_\mathrm{II} - \tau_2)}{t_\mathrm{II} - \tau_1}. \qquad (24)$$

Integrating the overall possible values of τ_1 yields the total probability of τ_2 occurring at a given time, given τ_0, t_I, and t_II:

$$p(\tau_2 \mid \tau_0, t_\mathrm{I}, t_\mathrm{II})$$

$$= \int_0^\infty p(\tau_2 \mid \tau_1, t_\mathrm{II}) p(\tau_1 \mid \tau_0, t_\mathrm{I}) d\tau_1. \qquad (25)$$

This process can then be repeated for each subsequent node in the cladogram to obtain a probability distribution function for each τ_n that depends only on the known fossil ages and the arbitrary parameter τ_0:

$$p(\tau_n \mid \tau_0, t_\mathrm{I}, t_\mathrm{II}, \dots t_N) =$$

$$\int_0^\infty p(\tau_n \mid \tau_{n-1}, t_N) p(\tau_{n-1} \mid \tau_0, t_\mathrm{I}, t_\mathrm{II}, \dots t_{N-1}) d\tau_{n-1},$$
$$(26)$$

where $p(\tau_n \mid \tau_{n-1}, t_N)$ is given by equation (19) and $p(\tau_{n-1} \mid \tau_0, t_\mathrm{I}, t_\mathrm{II}, \dots t_{N-1})$ is the probability distribution function for the prior node's age derived from the previous step in the analysis. In principle, the relevant integrals could be evaluated analytically, but in practice it is easy enough to perform the calculations numerically. A sample computer program that evaluates the probability distribution functions for a given set of t_J and τ_0 is provided in the appendix.

Figures 2 and 3 provide examples of the posterior probability distributions derived with this method. The function for any given node j abruptly goes to zero after t_J, the age of the earliest fossil representative of the taxa derived from this node. This edge reflects the hard constraint that the first fossil representative of taxon J must be more recent than node j. The probability of τ_j is constant between times t_J and t_{J-1}, because this analysis effectively assumes that any node is equally likely to occur at any time between τ_{j-1} (which must occur prior to t_{J-1}) and t_J. Prior to t_{J-1}, the probability of τ_j declines monotonically toward zero. Heuristically, this is because prior to t_{J-1}, there is an increasing

likelihood that node $j - 1$ has not yet split, and node j cannot occur until after τ_{j-1}. Similarly, prior to t_{J-2} there is an increasing chance that neither node $j - 2$ nor node $j - 1$ have split, so the probability of τ_j is even more strongly attenuated. Thus for sufficiently large j, we should expect that the probability τ_j occurs as early as τ_0 will be extremely small and the probability distribution should be very insensitive to this parameter. Quantitative estimates of the influence of τ_0 on this analysis are discussed in the next section.

After computing the probability distribution functions, one can compute various parameters to estimate and constrain the age of each node by using whatever criteria are appropriate to the situation. However, because the functions derived with this technique are asymmetric, I must caution that different methods of computing these parameters will yield different results, so the procedures used to obtain these limits should always be made explicit. In this paper I define the age estimate as the point that bisects the probability distribution function into two equal areas, and define the older (younger) age limit as the time where the probability of obtaining an age older (younger) than the given limit is 2.5%, so there is a 95% probability the age of the node is contained between the two limits.

Extensions of the Basic Algorithm.—Although the above example considered a simple case of a perfectly resolved Hennig comb cladogram where $t_\mathrm{I} < t_\mathrm{II} < t_\mathrm{III} < \dots < t_N$ (implying perfect congruence between the cladogram and the stratigraphic record), this method can be extended to provide age estimates for nodes of cladograms with more complex topologies, unresolved nodes, and even incongruities between the cladogram and the stratigraphic sequence.

Applying the above algorithm to cladograms with more complex topologies is relatively straightforward. For any given node in a cladogram with arbitrary (well-resolved) topology, an age estimate can be computed by simply identifying the times t_J with the ages of the earliest members of the various outgroups of the taxon under consideration.

If there are only a few unresolved nodes in the cladogram, these can also be dealt with
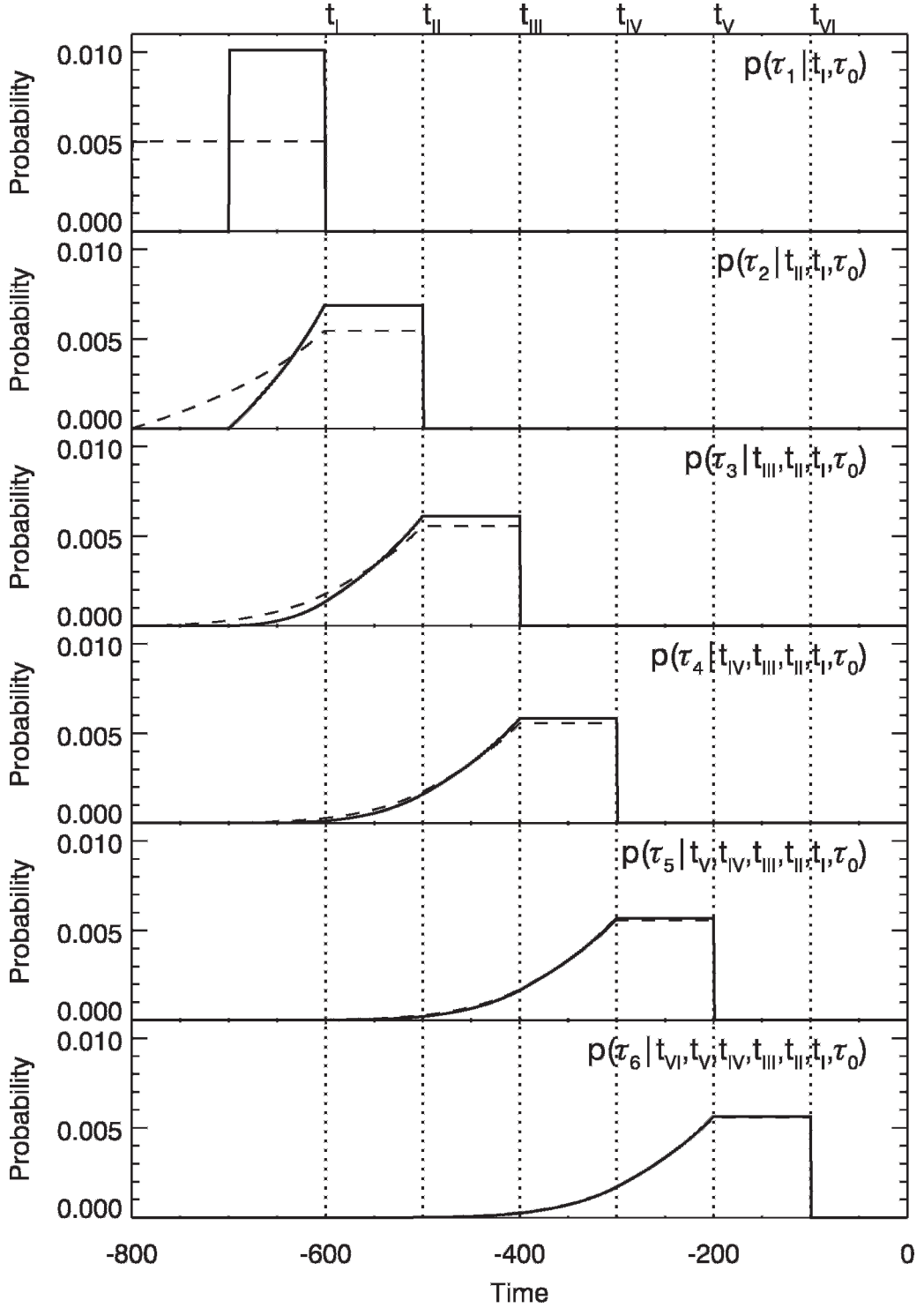
FIGURE 3.   Examples of the probability distribution functions discussed in the text. These distributions are computed assuming the earliest fossil examples of taxa I, II, III, IV, etc. are equally spaced in time (dotted lines). Each panel shows the probability distribution functions derived for a different node in the cladogram, and the solid and dashed curves are the functions derived by using two values of τ_0 that differ by 100 units of time. Note that as the node number n increases, the differences between the two curves get smaller, and the two curves become almost indistinguishable for n greater than 5. This implies that 5 fossil outgroups are sufficient to provide a robust age estimate for a given node.

rather easily. An unresolved node provides insufficient topological information to determine which of several clades is the relevant outgroup for any given taxon, so the most logical way to handle these situations is to treat this part of the cladogram as a single node, with the earliest representative of any of the multiple clades providing the relevant outgroup age.

On the other hand, if many nodes in the tree are unresolved, so that multiple cladograms with very different topologies are viable, then the above approach for constraining node ages will require more substantial modifications. If one can assign relative likelihoods to the different possible cladograms (perhaps using Bayesian methods as discussed in Huelsenbeck et al. 2001), then it should be possible to combine the probability distribution functions computed from the different cladograms into a single age estimate for certain clades. This sort of analysis is beyond the scope of this paper, so at this point I will simply caution that the reliability of the relevant cladogram should be carefully evaluated prior to using the algorithms described here.

Coping with incongruities between the cladogram topology and the stratigraphic record also requires careful consideration. Say … $t_{J-2} < t_J < t_{J-1} < t_{J+1}$ … so the earliest representative of taxon J occurs earlier than the earliest fossil of taxon $J - 1$. There are two different ways one might handle such a stratigraphic gap, which I will refer to as the ''consistent'' and ''conservative'' approaches. With the ''consistent approach'' we recognize that because the assumed cladogram requires that the node leading to taxon $J - 1$ must be earlier than that leading to J, some member of $J - 1$'s clade existed as least as early as t_J, and we can extend a ''ghost lineage'' (Norell 1993) back to this time. Therefore, we replace t_{J-1} with t_J and proceed as usual. With the ''conservative approach'' we simply ignore taxon $J - 1$ and include only those fossil outgroups that follow the correct stratigraphic sequence. If the cladogram is well resolved and well supported, then the consistent method should obviously be preferred, as it uses all the relevant data. However, if there are uncertainties in the cladogram or many inconsistencies with the stratigraphic record,

then it is less clear which approach is most appropriate. If nothing else, the conservative approach provides a way to gauge how strongly the stratigraphically incongruent taxa influence the age estimates. In lieu of a full analysis of the relative merits of these different approaches (which would require dealing with uncertainties in the cladogram's topology), the following section includes a preliminary study of certain aspects of both methods' performance.

Finally, although the basic algorithm presented here does not require any specific information about fossil recovery rates, such information could be incorporated into the age estimations as different priors on the $r_J(t)$ parameters. For example, if there were significant changes in the number of the relevant fossiliferous localities over time, it might be more appropriate to assume priors that result in node j being equally likely to be found in any of the localities dating between t_J and τ_{j-1}, rather than at any time between those two events (see, e.g., Wagner 1995). In this case, one could use the same basic algorithms derived above, but interpret the parameters t_J and τ_j as positions in the sequence of localities (or depths in the stratigraphic column) rather than as absolute times (as mentioned above). Beyond this, there is an extensive literature on estimating fossil recovery rates and how they vary over time and across taxa (e.g., Solow and Smith 1997; Foote 1997, 2001; Wagner 2000). However, incorporating such estimates into this sort of analysis and exploring how variable preservation rates might influence the age constraints is a complex task and therefore must be the subject of a future work.

Tests of the Algorithm and Sensitivity to Assumed Parameters

To explore how the derived constraints on the node ages depend on assumptions (such as the parameter τ_0) and the sampling of the fossil record, it is useful to consider some concrete examples using simulated data sets. These preliminary simulations do not test the sensitivity of this method to uncertainties in the clade topology or variations in recovery rates over time and among taxa, so they should not be considered exhaustive tests of

the reliability and range of applicability of the above algorithms. Nevertheless, these studies demonstrate the utility of these algorithms and suggest tests that could be done to gauge the results' sensitivity to unknown parameters.

Sensitivity to τ_0.—The parameter τ_0 specifies the assumed maximum possible age of the node closest to the cladogram's root. It initializes the algorithm and enables explicit probability distribution functions to be derived for the ages of all the other nodes in the cladogram. All the age constraints derived with this method therefore depend to some extent on the arbitrarily chosen value of τ_0. However, different nodes in the cladogram do not have the same sensitivity to τ_0. For example, the solid and dashed lines in Figure 3 show the probability distribution functions derived from a simple cladogram (with equally spaced t_J) assuming two different values of τ_0. Clearly the two probability distribution functions of τ_1 derived by this method are very different. However, the probability distribution functions for the other nodes are more similar to each other, and for nodes 5 and 6, the two functions are almost indistinguishable. The latter probability distribution functions are already well below their peak values at τ_0, so it is relatively unlikely that these nodes will occur as early as τ_0. Therefore, it should not be surprising that the exact value of this parameter has limited effects on the shape of these functions.

To better quantify the sensitivity of the probability distribution functions to τ_0, I computed the older limit of the confidence interval for each node, assuming two values of τ_0 that differ by $\Delta\tau_0$. I then determined the difference in the older limit between these two cases and divided that number by $\Delta\tau_0$ (I do not consider the younger limit here because it is much less sensitive to τ_0). Figure 4 shows a plot of the resulting ratios as a function of node number. For the first node this ratio is essentially 1, because the older limit on τ_1 is τ_0. However, for the other nodes this ratio is less than unity, and indeed the ratio monotonically decreases with increasing n. For $n = 6$, this ratio is about 0.1, meaning that an uncertainty in τ_0 of 100 Myr
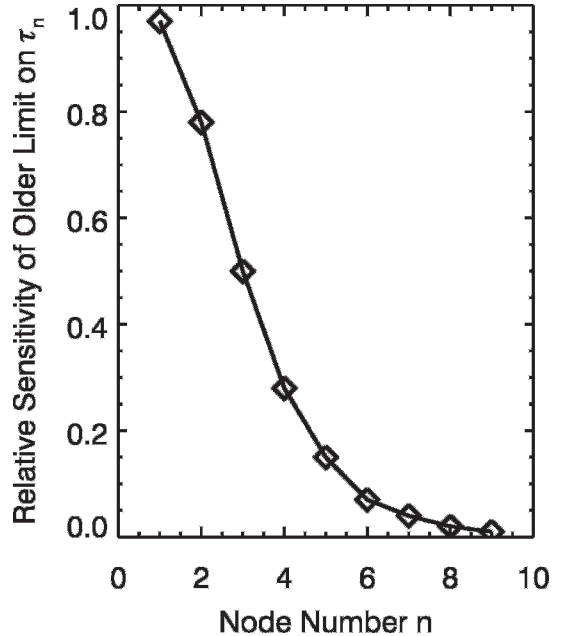

FIGURE 4. The sensitivity of the older limit on τ_n to the assumed value of τ_0. Each point on this plot shows the ratio of the change in the older limit on a given τ_n to the change in the assumed τ_0. The monotonic decline in this ratio with node number confirms that the older limits on the ages of nodes with sufficient numbers of outgroups are largely independent of the assumed τ_0. This graph was constructed with the assumption that the earliest fossil examples of the taxa are equally spaced in time. Assuming different values for fossil ages changes the exact values of the ratios, but does not change the overall trend.

would correspond to an uncertainty in the older limit of τ_6 of only 10 Myr. Note that although the data shown in Figure 4 assume that the ages t_J are regularly spaced in time, trials using t_J with increasingly large or small spacings show the same qualitative trend.

Of course, the degree of uncertainty one is willing to tolerate in τ_0 or the chosen node age will depend on the specific situation, but as a rule of thumb it appears that any node with 5 or more outgroups in the cladogram should have an age estimate that is reasonably insensitive to the assumed τ_0. Even so, because the calculation of a node age is not computationally intensive, I strongly urge that anyone using this method try several different values of τ_0 to determine precisely how sensitive the result is to this parameter and whether the uncertainty in τ_0 will be an issue in their particular analysis.
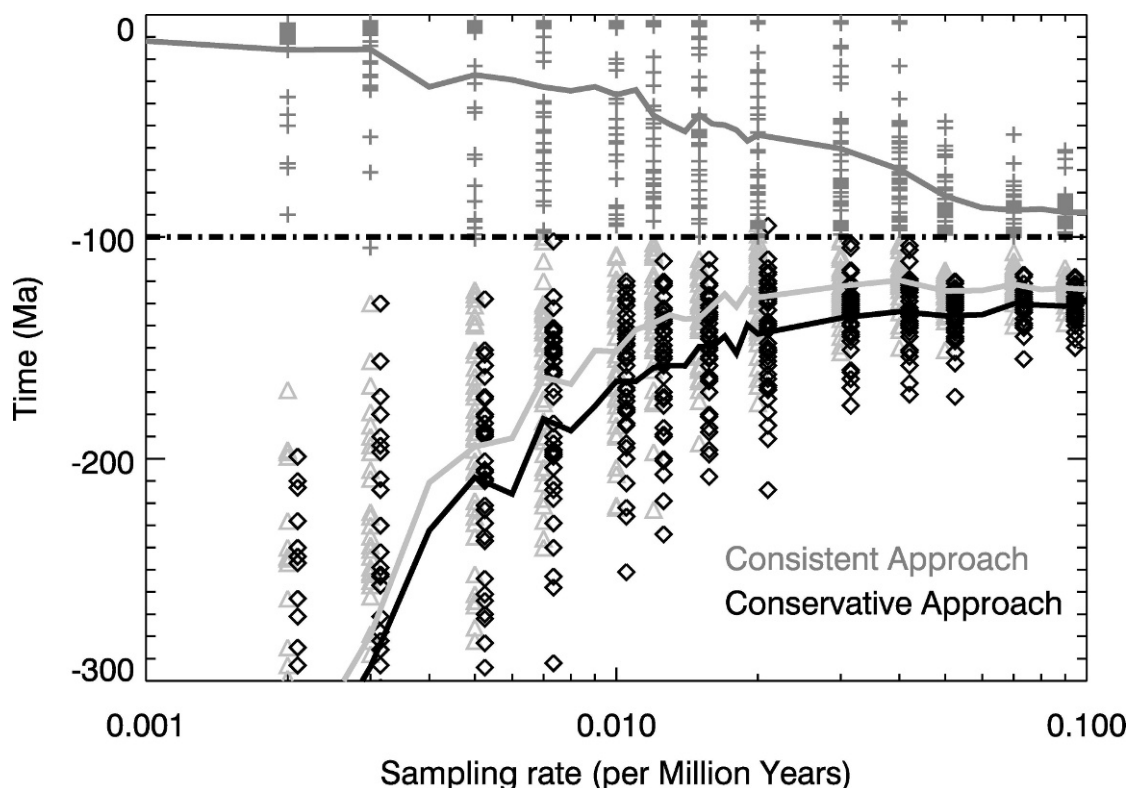
FIGURE 5. This plot shows the age limits derived for one node from a collection of simulated cladograms. Each cladogram consists of a series of nodes separated by 10 Myr, giving rise to a series of taxa that last for a random interval between 0 and 100 Myr. The last node occurs at 100 Ma and is assumed to give rise to taxa that persist to the present. A common sampling rate is assumed for all the taxa. The symbols show some of the younger (pluses) and older (diamonds and triangles) limits on the age of the final node derived from individual simulations using the methods described in the text. Note that the younger limits derived from the consistent and conservative approaches are almost identical and so are not plotted separately on this graph. The lines show the average value of the limits as a function of sampling rate. The mean younger limit on the node age moves further into the past as the sampling rate increases, while the mean older limit becomes moves closer to the present, as desired.

Trends with Sampling Rates.—If this method yields reasonable confidence intervals on the clade ages, we should expect that the limits on the age of a given node will converge toward the true age of the node as the fossil sampling and recovery rates improve. To verify that this is indeed the case, the above analysis was run on a collection of simulated cladograms. Each cladogram has the basic topology shown in Figure 1, with a series of nodes that are assumed to be equally spaced in time (with 10 Myr between each node) up to a final node that formed 100 Myr ago. For simplicity, I assume a common sampling rate for all lineages in this cladogram. For each simulation, I assume the taxa derived from the last node persist until the present day (when they are always recovered), assign a

randomly determined duration between 0 and 100 Myr to each taxon derived from all earlier nodes, choose a number of fossil specimens for each taxa using the appropriate Poisson distribution for a given sampling rate and distribute the fossils uniformly along the duration of the lineage. Finally, using the techniques described above, I compute the probability distribution function and age limits for the final node in the cladogram. Both conservative and consistent approaches are used.

Figure 5 plots the limits derived from these simulations as a function of the assumed sampling rate. The results of individual runs (40 per sampling rate) are shown as symbols and the lines show the average value of the individual runs as a function of the sampling

rate. Note that younger limits from the simulations are almost always more recent that the true age of the node (100 Myr ago), and the older limits from both the conservative and the consistent approaches are almost always earlier than the true age of the node. This suggests that these limits can provide reliable constraints on node ages. We also note that for sampling rates below roughly 0.05/Myr, both the older and younger confidence limits converge toward the true age of the node, as desired.

Because the younger limit on the node's age is set by the age of the oldest fossil representative of the taxa derived from that node, it makes sense that the younger limit on the age steadily moves backward in time as the sampling rate improves. The trends in the older limits, by contrast, result from a more complex interaction between two processes. On the one hand, improving the sampling rate means that the earliest members of a given lineage will typically be found at earlier times, causing the t_f to move back in time and driving the older limit earlier. On the other hand, increasing the sampling rate will also result in the recovery of additional taxa, increasing the number of outgroups within a given time span and thus moving the older limit later in time. In these simulations, the latter effect wins out for sampling rates below about 0.05/Myr, producing the desired trend of increasingly tight limits on the age of the node.

These simulations also indicate that when the sampling rate is above 0.05/Myr, the older limit saturates at a value roughly 20 Myr before the true age of the node, whereas the younger limit continues to move to earlier times. This is partially a limitation of the simulations, which assumed that internal branches in the cladogram are never sampled. Thus, in the limit of perfect sampling, the first fossil representative of the taxa derived from the final node will appear 100 Myr ago and the first fossil representative of its various outgroups will occur at 110 Ma , 120 Ma, and so on. In this limit, the probability distribution function on the node age reaches a limiting form with an older limit around 120 Ma. Although this limiting case is clearly

unrealistic, it does demonstrate an important limitation of the method proposed here. These algorithms are most likely to produce confidence intervals that converge toward the true node age if it is probable that additional outgroups will be discovered. By contrast, methods that explicitly include sampling rates (e.g., Strauss and Sadler 1989) are more appropriate if the fossil recovery rate is sufficiently good that it is unlikely that additional outgroup taxa are going to be recovered (or if direct ancestor-descendent relationships between taxa are expected).

Additional simulations (using 1000 replicates for each sampling rate) provide more quantitative data on the validity of the older limits derived by this analysis. Given the criteria used to define this older limit, one expects that these will be more recent than the true node age 2.5% of the time. In reality, for simulations with sampling rates between 0.006/Myr and 0.03/Myr, the older limit is more recent than 100 Ma in roughly 2% of the simulations when the consistent approach is used and only 0.5% of the simulations when the conservative approach is used. The results for the consistent approach are close to expectations, perhaps a little low because of the conservative priors assumed for the fossil recovery rates. The lower value for the conservative approach confirms that this approach gives more conservative constraints on the node age. At sampling rates above 0.03/Myr or below 0.0006/Myr the percentage of simulations with upper limits more recent than 100 Ma decrease to below 0.1%. For high sampling rates, this reflects the limitation of the simulations discussed above. For low sampling rates it becomes unlikely that any fossil is recovered anytime in the interval between the node's true age and the present, so the only evidence for taxa derived from the node of interest comes from the specimens assumed to be recovered at the present. Such a situation violates the assumptions of the above calculations, so deviations from predictions are not unexpected. These models therefore indicate that the consistent approach provides a reasonable older limit on the node time for the range of recovery rates where the assumptions behind this model are

justified, and provides conservative constraints on the node age outside of this range.

Although the consistent approach yields limits with the appropriate significance in the above simulations, this result is somewhat dependent on the durations and node spacings of the assumed lineage. If the time interval between adjacent nodes is decreased relative to the lineages' mean duration, then it becomes more probable that the ages of the earliest fossil representatives of the various outgroups will be found out of the expected stratigraphic sequence. This could potentially make the consistent approach more likely to underestimate the older limit on the node age. Simulations where the interval between two nodes in the cladogram were reduced by a factor of two (to 5 Myr) while keeping the lineages durations the same confirm that this in fact occurs, and also show that the conservative approach continues to produce more conservative older limits. These results indicate that the limits derived with this method are most reliable when the fossil record is stratigraphically consistent with the cladogram, and the consistent approach in particular should be used with caution when many taxa are out of the expected stratigraphic sequence.

Application of the Algorithm to the Origin of Placental Mammals

To demonstrate how this method can be applied to actual fossil data, let us consider the origin of placental mammals. The sequence of outgroups used in this analysis derive from two cladograms of stem Mesozoic eutherians from the recent literature, both of which contain at least five stratigraphically consistent outgroups to placentalians (Ji et al. 2002; Wible et al. 2007). The Ji et al. (2002) cladogram only contains a single placentalian (*Erinaceus*), which has the Cenozoic *Protungulatum* as a sister group. The Wible et al. (2007) cladogram shows the Cenozoic animals *Protungulatum*, *Purgatorius*, and *Oxyprimus* forming an outgroup to Placentalia, but the authors cannot rule out a cladogram with these animals located inside Placentalia (see supplemental online material of Wible et al. 2007). To keep things simple, we will treat

Protungulatum, *Purgatorius*, and *Oxyprimus* as ''Cenozoic allies'' of Placentalia and consider the age of this entire clade, whose earliest members date back to approximately 65 Ma.

Table 1 lists the sequence of outgroups to this placental clade in the two cladograms, with ages obtained from Kielan-Jaworowska et al. (2004). Both cladograms have at least one outgroup that is out of stratigraphic sequence, so we tabulate the dates used in the consistent and conservative approaches separately for clarity (note the Ji et al. 2002 cladogram also contains one trichotomy).

Results and Discussion

Figure 6 shows the probability distribution functions derived from the fossil outgroup data, and the corresponding age estimates and limits are included in Table 1. As expected, the conservative approach yields somewhat looser constraints and higher age estimates than the consistent approach.

Given that the two sequences of outgroups are quite different, it is perhaps surprising that the two cladograms give very similar estimates for the age of modern placentalians and their Cenozoic allies, with central values ranging from 74 to 79 Ma, younger limits at 65–66 Ma, and older limits between 88 and 98 Ma. Upon closer inspection, we can see that in both cladograms numerous outgroups date back to ca. 90 Ma, causing all the probability distribution functions to be strongly attenuated before this time. In fact, the primary difference between the two cladograms for the purposes of this analysis is that the Wible et al. cladogram has the first outgroup (*Gypsonictops*) dating back to 76 Ma, whereas the Ji et al. data has its first outgroup (*Eoungulatum*) at 87 Ma. This leads to the narrower probability distribution function for the Wible et al. analysis and the somewhat more recent age estimate.

We may compare these results with the recent molecule-based estimates for the origin of modern placental mammals tabulated in Springer et al. (2003). This source provides multiple age estimates based on different subsets of their molecular data. For the full data set, the estimated age and 95% confidence interval for the base of Placentalia is

TABLE 1. Data used to estimate age of Placentalia.

| Outgroup[a]	Earliest age[b]	Assumed date[c] (Ma)	Consistent approach (Ma)	Conservative approach (Ma)
		Ji et al. 2002 cladogram		
Eoungulatum	Coinacian	87	87	87
Aspanlestes	L. Turonian	90	90	90
Gypsonictops/Cimolestes	Judithian	76	90	—
Daulestes	L. Turonian	90	90	90
Montanalestes[d]	Aptian-Albian	112	112	112
Eomaia	M. Barremian	125	125	125
Estimated age of Placentalia (and Cenozoic allies)[e]			78 (66–93)	79 (66–98)
		Wible et al. 2007 cladogram		
Gypsonictops	Judithian	76	76	76
Zalambdalestidae	L. Turonian	90	90	90
Deccanolestes	Maastrichtian	69	90	—
Asioryctitheria etc.	Coniacian	87	90	—
Eozhelestes	E. Cenomanian	97	97	97
Zhelestidae	L. Turonian	90	97	—
Montanalestes	Aptian-Albian	112	112	112
Bobolestes	L. Albian	100	112	—
Eomaia	M. Barremian	125	125	125
Estimated age of Placentalia (and Cenozoic allies)[e]			74 (65–88)	76 (66–98)

[a] Possible Cenozoic outgroups including *Protungulatum* not included (see text).
[b] Earliest ages of each outgroup are derived from Kielan-Jaworowska et al. (2004).
[c] Assumed date used for calculations presented here, based on the formations recorded in the previous column.
[d] As earliest member of trichotomy involving Zalambdalestidae.
[e] Age estimates assume the earliest Placentalian fossil are from 65 Ma and use $\tau_0 = 150$ Ma; varying τ_0 between 130 Ma and 170 Ma shifts the older limit by less than 3 Myr and the younger limit by less than 1 Myr.
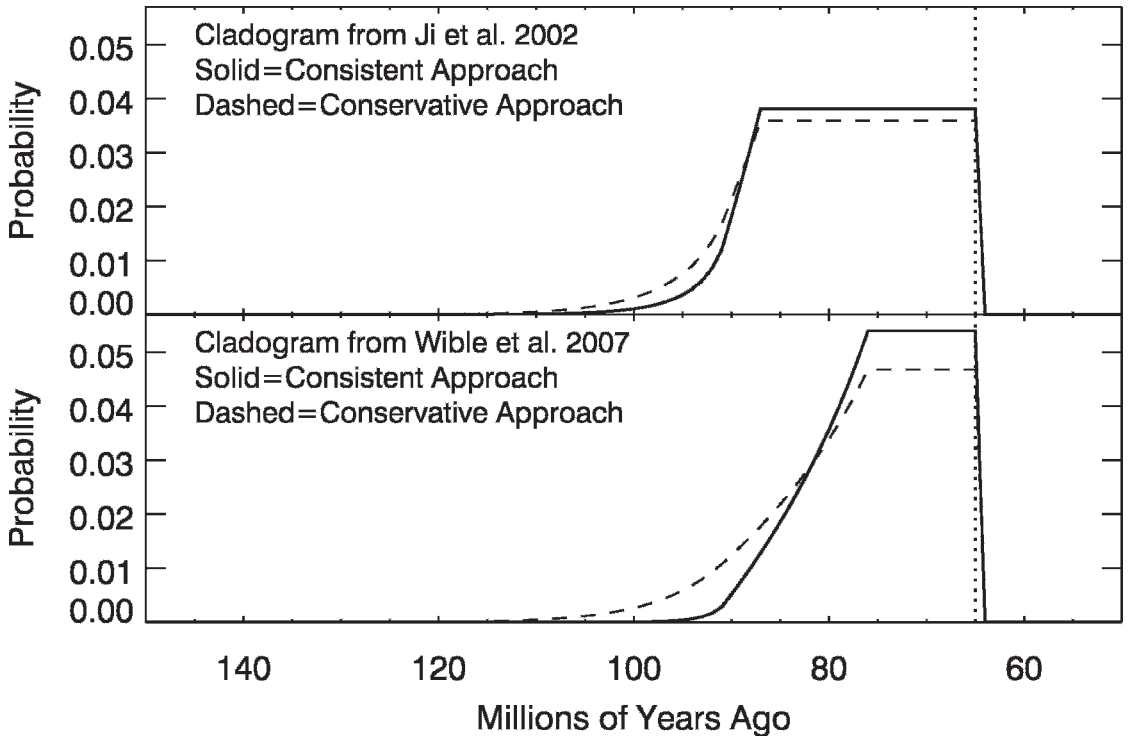
FIGURE 6. Probability distribution functions of the age of the last common ancestor of placental mammals and their Cenozoic allies assuming outgroups from cladograms in Ji et al. (2002) and Wible et al. (2007) and outgroup ages contained in Table 1. Functions calculated with the consistent and conservative approaches are shown as solid and dashed lines, respectively. The units on the y-axes are per million years.

107 (98–117) Ma, but some subsets (for 3' UTRs and mitochondrial RNA) give ages as low as 97 (86–109) Ma. These lower dates from some of the molecular data subsets do overlap with the older limits from this analysis, but even in these cases the hypothesis that the two dates are consistent is disfavored (the probability of no difference being less than 10%). However, this analysis suggests that the magnitude of the difference between the two age estimates may be smaller than previously considered. The fossil-based age limits derived here indicate the present fossil evidence cannot rule out the possibility that placentalians and their Cenozoic allies originated as much as 20 or even 30 Myr before the K/T boundary. If the last common ancestor of placentalians did live this long ago, then the molecular estimates would only be off by 20 Myr (or about 25%), not 40 Myr, as would be required to move the base of Placentalia all the way to the K/T boundary.

Again, I must remind the reader that the age estimates derived here rest on several assumptions, and these must be carefully evaluated. Obviously, this includes the assumed priors on the fossil recovery rates and node ages. As noted above, the priors assumed for this analysis tend to be conservative, so other priors, which might be based on estimates the preservation and fossil recovery rates, will tend to tighten the limits on the age of Placentalia and increase the discrepancy between the fossil and molecular evidence.

A bigger issue is that a given age estimate is based on a particular cladogram, so the estimate is sensitive to the topology of the resulting tree. In this example, two different cladograms gave similar results, which may give us some confidence in the reliability of these calculations. However, the topologies of these two cladograms were similar in that all of the Mesozoic animals formed a highly nested set of outgroups to Placentalia. If some Mesozoic eutherians are not outgroups to Placentalia but instead part of it (Archibald et al. 2001; Archibald 2003), then the age estimate will shift significantly earlier. (Unfortunately, the cladograms showing this pattern contain insufficient stratigraphically consistent fossils below the base of Placentalia to derive a robust date with the present method.) Also, if the eutherian cladogram had fewer discrete outgroups leading to Placentalia, the age estimate would tend to move earlier. Conversely, if more of the late Cretaceous and early Cenozoic eutherians turn out to be closely related outgroups to Placentalia and not part of Placentalia itself, the age estimate will shift later, perhaps more consistent with ''explosive'' models of placentalian origins (Archibald and Deutschman 2001). An exhaustive analysis of the consistency between the molecular analyses and the fossil analysis would therefore have to account for uncertainties in the topology of the eutherian phylogeny. Even though such an analysis is beyond the scope of this report, the age estimates presented here demonstrate that this technique can be a useful way to investigate the consistency between molecular age estimates and fossil evidence.

Conclusions

The algorithms presented here provide a way to constrain the ages of nodes in a cladogram based on the ages of the earliest fossil specimens of that clade's outgroups. The age estimate for the last common ancestor of modern placentalians and their Cenozoic allies derived by using this method permit us to quantify the significance and magnitude of the difference between molecular-based age estimates and the fossil evidence. Of course, additional work still needs to be done to verify whether the age estimates derived with this method are reliable. In particular, it will be important to evaluate how uncertainties in the cladogram topology and variations in the fossil recovery rate in time and space affect the age estimates derived with this method. Even so, some form of these algorithms could be useful for a wide range of applications, including calibrating DNA-based age estimates and identifying times of rapid cladogenesis. Furthermore, because the calculations presented here are not computationally intensive, anytime researchers create a sufficiently deep, well-resolved, and stratigraphically consistent cladogram of fossil taxa,

they could easily also provide age estimates for all of the nodes sufficiently far from the root.

Acknowledgments

Literature Cited

Archibald, J. D. 2003. Timing and biogeography of the eutherian radiation: fossils and molecules compared. Molecular Phylogenetics and Evolution 28:350–3009.

Archibald, J. D., and D. H. Deutschman. 2001. Quantitative analysis of the timing of the origination and diversification of extant placental orders. Journal of Mammalian Evolution 8:107–124.

Archibald, J. D., A. O. Averianov, and E. G. Ekdale. 2001. Late Cretaceous relatives of rabbits, rodents and other extant eutherian mammals. Nature 414:62–65.

Fisher, D. C, 2008. Stratocladistics: integrating temporal data and character data in phylogenetic inference. Annual Review of Ecology, Evolution, and Systematics 39:365–385.

Foote, M. 1997. Estimating taxonomic durations and preservation probability. Paleobiology 23:278–300.

———. 2001. Inferring temporal patterns of preservation, origination and extinction from taxonomic survivorship analysis. Paleobiology 27:602–630.

Foote, M., J. P. Hunter, C. M. Janis, and J. J. Sepkoski Jr. 1999. Evolutionary and preservational constraints on origins of biotic groups: divergence times of eutherian mammals. Science 283:1310–1314.

Huelsenbeck, J. P., and B. Rannala. 1997. Maximum likelihood estimate of phylogeny using stratigraphic data. Paleobiology 23:174–180.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314.

Ji, Q., Z-X. Luo, C-X. Yuan, J. R. Wible, J-P. Zhang, and J. A. Georgi. 2002. The earliest known eutherian mammal. Nature 416:816–822.

Kielan-Jaworowska, Z., R. J. Cifelli, and Z-X. Luo. 2004. Mammals from the age of dinosaurs. Columbia University Press, New York.

Lupton, R. 1993. Statistics in theory and practice. Princeton University Press, Princeton, N.J.

Marshall, C. A. 2008. A simple method for bracketing absolute divergence times on molecular phylogenies using multiple fossil calibration points. American Naturalist 171:726–742.

———. 1997. Confidence intervals on stratigraphic ranges with nonrandom distributions of fossil horizons. Paleobiology 23:165–173.

Norell, M. A. 1993. Tree-based approaches to understanding history: comments on ranks, rules and the quality of the fossil record. American Journal of Science 293A:407–417.

Rose, K. D., and J. D. Archibald. 2005. The rise of placental mammals. Johns Hopkins University Press, Baltimore.

Ruta, M., P. J. Wagner, and M. I. Coates. 2006. Evolution patterns in early tetrapods. I. Rapid initial diversification followed by decrease in rates of character change. Proceedings of the Royal Society of London B 273:2107–2111.

Solow, A. R., and W. Smith. 1997. On fossil preservation and the stratigraphic ranges of taxa. Paleobiology 23:271–277.

Springer, M. S., W. J. Murphy, E. Eizirik, and S. J. O'Brien. 2003. Placental mammal diversification and the Cretaceous-Tertiary boundary. Proceedings of the National Academy of Sciences USA 100:1056–1061.

Strauss, D., and P. M. Sadler. 1989. Classical confidence intervals and Bayesian probability estimates for ends of local taxon ranges. Mathematical Geology 21:411–427.

Tavare, S., C. R. Marshall, O. Will, C. Soligo, and R. D. Martin. 2002. Using the fossil record to estimate the age of the last common ancestor of extant primates. Nature 416:726–729.

Wagner, P. J. 1995. Stratigraphic tests of cladistic hypotheses. Paleobiology 21:153–178.

———. 2000. Likelihood tests of hypothesized durations: determining and accommodating biasing factors. Paleobiology 26:431–449.

Wible, J. R., G. W. Rougier, M. J. Novacek, and R. J. Asher. 2007. Cretaceous eutherians and Laurasian origin for placental mammals near the K/T boundary. Nature 447:1003–1006.

Appendix

Sample Code for Computing Clade Ages

This is the basic algorithm for calculating the probability distribution function for a given clade, written in the R computing language (thanks to M. Foote for translating the original IDL script into this form).

```
#declare function
nodeage<-function(tsteps,tnodes,t0)
#function returns p.d.f. for node ages
#t0 is an arbitrary oldest age to consider
#tnodes are the ages of oldest known fossil stemming from
each node
#tsteps is a vector of arbitrary time steps on which the p.d.f. is
calculated
    {
    nn<-length(tnodes)
    nt<-length(tsteps)
    pnodes<-matrix(0,nn,nt) #initialize array
    # first get pdf for node 1, at discrete values
    ii<-which(tsteps > t0&tsteps<tnodes[1])
    pnodes[1,ii]<-1.0/length(ii) #assume uniform distribution
    for (i in 2:nn) #cycle through remaining nodes
        {
        for (j in 1:nt) #cycle through series of time steps
            {
            p21<-rep(0,nt) #initialize vector
            ii<-which(tsteps > = tsteps[j]&tsteps<tnodes[i])
            p21[ii]<-1.0/length(ii)*pnodes[(i-1),j]
            #cond. probability of this age, given previous node age,
            #times prob. of previous node age
            pnodes[i,ii]<-pnodes[i,ii]+p21[ii]
            #add to cumulative sum
            }
        }
list(pp = as.vector(pnodes[nn,]))
    #return the p.d.f. vector for the youngest node
}
```