PROBABILISTIC PHYLOGENETIC INFERENCE IN THE FOSSIL RECORD: CURRENT AND FUTURE APPLICATIONS

PETER J. WAGNER

Department of Paleobiology, National Museum of Natural History,
Smithsonian Institution [NHB, MRC 121], P.O. Box 37012, Washington DC 20013-7012

and

JONATHAN D. MARCOT

Department of Animal Biology, University of Illinois, 515 Morrill Hall,
505 S. Goodwin Ave., Urbana, IL, 61801

ABSTRACT.—Quantitative phylogenetic inference estimates the probability of observed character distributions given trees and rates. Most available programs for doing this assume (tacitly or explicitly) that the sampled taxa are contemporaneous. However, paleontologists usually sample taxa over a clade's history. Thus, we must estimate the probability of observed character-state distributions over time given trees and rates. When we include information about sampling intensity, then we really are estimating the probability of the observed record given trees and rates. Some additional problems that *should* be issues for neontologists, but which are much more obvious concerns for paleontologists include: 1) ancestor-descendant relationships; 2) punctuated versus continuous morphological change; and, 3) the effects of extinction and speciation rates on prior probabilities of trees. Future goals of paleosystematists include incorporating these and other "nuisance" parameters so that, ultimately, our tests of phylogeny are really tests of evolutionary histories.

INTRODUCTION

THE ABILITY to accurately reconstruct phylogenies can greatly expand paleobiological research programs. However, although paleontological data presents unique challenges, only a few workers have developed methods accommodating these (e.g., Gingerich, 1976; Fisher, 2008). We have particularly lagged in the development of probabilistic methods of phylogenetic inference. Such methods date back to the 1960's (e.g., Edwards and Cavalli-Sforza, 1964) and methods based on statistical probability (i.e., maximum likelihood, and more recently Bayesian posterior probabilities) have become the most widely used among systematists. The reasons for this popularity include the ability to explicitly model evolutionary processes specific to the data being analyzed, some favorable statistical properties (e.g., Felsenstein, 1988; Hillis, 1995; Huelsenbeck, 1995) and suitability for statistical hypothesis testing (e.g., Huelsenbeck and Rannala, 1997b). All of these should appeal to paleobiologists as well.

Although early presentations of probabilistic methods do not assume characters of any particular data type (e.g., Felsenstein, 1973), the use of probabilistic methods has been restricted predominantly to molecular data. This simply reflects the ease of applying such methods to molecular sequences: the data share not just the same number of possible states, but also the same possible states (e.g., A, G, C and T for DNA sequences). This uniformity allows single models of character evolution to be applied to entire sequences or partitions thereof (e.g., first, second or third codon positions). However, Lewis (2001) presents simple modifications of this approach for the analysis of characters with any number of states, which requires no assumptions about shared states among characters. As a result, some recent analyses of morphologic data use likelihood and Bayesian methods (e.g., Ayache and Near, 2009; Glenner et al., 2004; Wiens et al., 2005), including a handful of studies using fossil taxa (e.g., Snively et al., 2004; Pollitt et al., 2005; Müller and Reisz, 2006; Beck et al., 2008; Clarke and Middleton,

In Quantitative Methods in Paleobiology, pp. 189-211, Paleontological Society Short Course, October 30th, 2010. The Paleontological Society Papers, Volume 16, John Alroy and Gene Hunt (eds.).

2008; Fröbisch and Schoch, 2009).

Numerous works summarize the general methodology and equations for calculating tree likelihoods (e.g., Swofford et al., 1996; Huelsenbeck and Rannala, 1997a; Felsenstein, 2004; Galtier et al., 2005). However, because probabilistic methods of phylogenetic inference are not yet commonly applied to fossil taxa, our first goal in this paper is to familiarize the reader with the general methods and basic calculations. We will show that, while these methods are currently applicable to fossils, they do not yet make full use of information provided in the fossil record. Thus, our second goal is to identify potential modifications of existing methods of probabilistic inference to leverage the wealth of information available in the fossil record for more robust phylogenetic inference of fossil taxa.

PHYLOGENIES AND FOSSILS: SOME NECESSARY COMPONENTS

Important Terms and Concepts.—Probabilistic inference of phylogeny requires two components: a phylogeny and a model of character evolution. These are then used to estimate the probability of the observed character-states among the analyzed taxa (i.e., the character matrix). The maximum likelihood (ML) inference methods choose trees maximizing this probability, whereas Bayesian methods effectively "weight" tree likelihoods by the prior probabilities of those trees (see Wang, this volume).

We require fairly exact phylogenetic hypotheses to determine the probability of observed data. The distinction between "cladogram" and "phylogeny" arises here. Cladograms are very general statements about relationships. Phylogenies are very specific statements about relationships that, for any one cladogram, might differ on hypothesized amounts of change and/or divergence times (see, e.g., Fig. 1). We require the specific information unique to each phylogeny within a cladogram (e.g., amounts of evolution, ancestor-descendant vs. sister-species relationships) to estimate the probability of our observations. The cladogram itself has no likelihood unless we integrate over all possible phylogenies within that cladogram.

Given any phylogeny, calculating probabilities of character-state distributions requires expectations about character change. These expectations are derived from models of character evolution. As we will emphasize, these models usually are flexible rather than invariant assumptions. They even can be estimated as part of the hypothesis in analysis. Possible models that concern paleontologists include not just different rates among different characters, but different rates over phylogeny and/or time, and whether change is concentrated in specific events or distributed through time.

Describing Phylogenies with Fossil Taxa.—Phylogeny has two components: a general topology (Ψ) describing which species are closest relatives, and ages (d) for each speciation event (cladogenetic or anagenetic) on that topology. What distinguishes paleontological phylogenies from neontological phylogenies is that first appearance (FA) and last appearance (LA) data combined with $\mathbf{d}$ (i.e., the set d for all nodes of Ψ) permit three fundamentally different exact relationships given any one Ψ (Fig. 1.1). Having d for node precede the FA the node's oldest descendant (Fig. 1.2) is consistent with "bifurcating" cladogenesis (*sensu* Foote, 1996b). A d that occurs within the stratigraphic range of the node's oldest descendant (Fig. 1.3) necessarily implies "budding" cladogenesis event in which an ancestor persists along with its descendant. Finally, a d that postdates the LA of the node's oldest taxon (Figure 1.4) is consistent with the older and younger taxa being part of an anagenetically evolving lineage (e.g., anagenetic speciation). (Note that d cannot be younger than the FA of the latest-appearing descendant of a node.)

Each branch i of Ψ (Figure 1.1) will have two quasi-independent parameters: its sampled chronostratigraphic range (s_i) and unsampled duration implied by the phylogeny (t_i, which equals a range extension sensu Smith 1988 or a ghost-lineage sensu Norell 1993) (see Fig. 1.2-1.4). Most models of character evolution require one or both of these parameters to calculate a phylogeny's likelihood (see below).

Note that t_i is the unsampled duration of a taxon that necessarily is implied by the phylogeny. As a rule, t_i always precedes s_i (Signor and Lipps 1982). Although the true dates extinction of most taxa also are after their LA, the phylogeny, provides no information about such extensions. The nearest exception to this is the anagenetic lineage depicted in Fig. 1.4: there, the unsampled duration separating Taxa C and D (t_2) might belong to C, D or partially to both. However, this does not affect the calculated likelihoods: as we will show below, the calculated probabilities of character change over t_2 encompass all possibilities. The constraint that

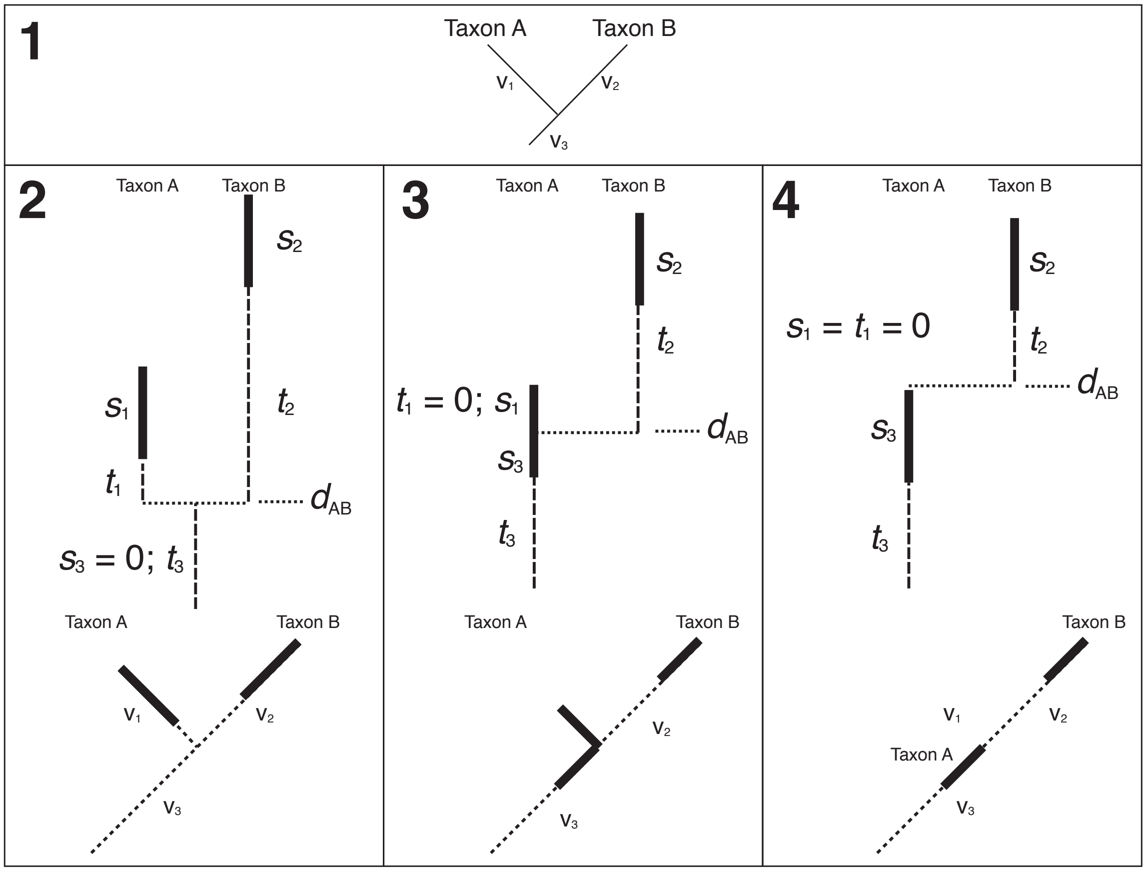

FIGURE *1*.—The use of a topology Ψ, node dates and stratigraphic range data to specify phylogenies of fossil taxa. 1. A topology (e.g., cladogram; Ψ) depicting two sister taxa. Branches of the cladogram are labeled as $\nu_{1\text{-}3}$. 2-4. Three different phylogenies consistent with Ψ. *s* is the portion of each branch that is observed and *t* is the unobserved portion implicit to a particular phylogenetic hypothesis. In 2, the divergence date of the node joining Taxon A and Taxon B (d_{AB}) predates the first appearance (FA) of either Taxon A or B. This relationship between d_{AB} and the FA of all descendants of the node is consistent with bifurcating cladogenesis. In 3, d_{AB} between the FA and last appearance (LA) of Taxon A necessitates budding cladogenesis. In 4, d_{AB} between the LA of A and the FA of B is consistent with anagenesis. Note that we consider "speciation" to encompass both models.

t_i must precede s_i does bias tree-based analyses such as phylogenetic reconstructions of taxonomic richness (e.g., Foote, 1996c;Wagner, 2000c; Lane et al., 2005). However, this constraint does not bias the calculation of tree likelihoods, as there are no character data that bear on extensions subsequent to the LA.

If taxa have stratigraphic ranges, then the relationship between *d* and the stratigraphic ranges of descendant taxa determines *s* and *t* along those descendant branches (e.g., ν_1 and ν_2 in Fig. 1.1) *and* the branch immediately leading from the node to its ancestor (e.g., ν_3, in Fig. 1.1). If all taxa descended from a node are sampled from only a single time slice and therefore have no stratigraphic ranges, then s_i=0, and t_i will correspond to the entire temporal duration of branch *i*. If stratigraphic data are available, then s_i and t_i values for all branches of the tree can be determined in a postorder traversal of the tree (i.e., moving from the tips to the root) using the following rules:

1) If the *d* precedes FA for all descendant, then t_i is the difference between *x* and *d*, where *x* is one of two values. If the descendant is a sampled taxon, then *x* is the FA of that taxon. If the descendant is an internal node, then *x* is *d* of

that descendant node. If the branch leads to a terminal taxon, then s equals the stratigraphic range of that taxon (e.g., s_1 in Fig. 1.2). If the descendant branch leads to an internal node, then the next higher node determines its s. For the branch leading from this node to its own ancestor (e.g., v_3 in Fig. 1.1), s is zero, and t is the difference between the d values of this node and its ancestral node (e.g., t_3 in Fig. 1.2).

2) If d is later than the FA of the oldest descendant branch (Fig. 1.3 and 1.4), t_i=0 for the branch leading to the oldest descendant. If the branch leads to a terminal taxon, then s is the difference between that taxon's LA and d. Otherwise, s=d of the next highest node if this descendant branch is entirely sampled. In the latter case, s of the branch leading to this node from its ancestor becomes the segment of the stratigraphic range preceding d (e.g., s_3 in Fig. 1.3). The date of any node must be older than at least one of its descendants, so d must predate the FA of the younger descendant. The values of s and t for this branch are determined using the rules above.

3) If d is later than the LA of a terminal taxon (e.g., Fig. 1.4), then s=t=0 for the branch leading to this oldest descendant. Now s for the branch leading to this node from its descendant will be the entire stratigraphic range of this terminal taxon (e.g., s_3 of Fig. 1.4). s and t for other descendants will be determined as above.

At this point, it is appropriate to make a brief note concerning taxa and coding practices. At the species-level, our discussion equates taxa with unique morphologies. At the species-level, this assumes a "morphospecies" concept, where a species is a unique combination of character states. This concept is appropriate for our purposes for two reasons. First, it follows paleontological practice: fossil species almost inevitably are morphospecies as this really is the only criterion available to us. Second, the morphospecies concept is neutral with regard to any pattern that paleontologists might call "speciation." That is, it allows for bifurcation (Fig. 1.2), budding speciation (Fig. 1.3) and anagenetic speciation (Fig. 1.4). Workers should code and analyze taxa in a manner that allows for any of these three possibilities, so they might be evaluated as part of the phylogenetic analysis. For example, even if one suspects that morphospecies A and B are part of the same anagenetically changing lineage, then one still should code them separately, in order to compare the likelihood of this presumed evolutionary relationship to other possible hypotheses during the phylogenetic analysis (Wagner and Erwin, 1995). These distinctions are important if we are worried about reconstructing evolutionary history rather than general phylogenetic relationships. We will return to how likelihood can distinguish these possibilities in more detail below.

PHYLOGENY LIKELIHOOD: A SIMPLE EXAMPLE

Consider a four-taxon topology, Ψ=((A,B),(CD)). We will begin with phylogeny ψ_1, for which all branch durations t=1.0 and all s=0.0 (Fig. 2). (Note that branch *duration* is not the same as branch *length* common in molecular phylogenetics; see below.) We will examine a single two-state (k=2) character, for which Taxon A shows state 0, while Taxa B, C and D all show state 1. We will assume a rate of character evolution (δ) that is constant on all branches of the tree, with δ=0.05 giving the expected changes per unit time. If we assume that change can happen at any time along a branch, then the probabilities of transition, **P**, along any branch are:

$$\mathbf{P}= \begin{array}{ccc} \rightarrow & 0 & 1 \\ 0 & \frac{1}{k}+\frac{k-1}{k}e^{-k\delta t} & \frac{1}{k}-\frac{1}{k}e^{-k\delta t} \\ 1 & \frac{1}{k}-\frac{1}{k}e^{-k\delta t} & \frac{1}{k}+\frac{k-1}{k}e^{-k\delta t} \end{array} \quad (1)$$

where p_{ij} gives the probability that a branch beginning with state i will end in state j (where j might equal i; Lewis, 2001). Note that these give the probability of *net* stasis and *net* change along a branch and not the probabilities of stasis or a single change along that branch (Felsenstein, 1981). Thus, for our binary character, the diagonal is the summation of the Poisson probabilities of 0, 2, 4, etc., changes, all of which end up in staying in, or reversing to the original state. The off–diagonal is the summation of the Poisson probabilities of 1, 3, 5, etc., changes, all of which end up in a net change to the other state.

Substituting the given parameter values into Equation 1, we get:

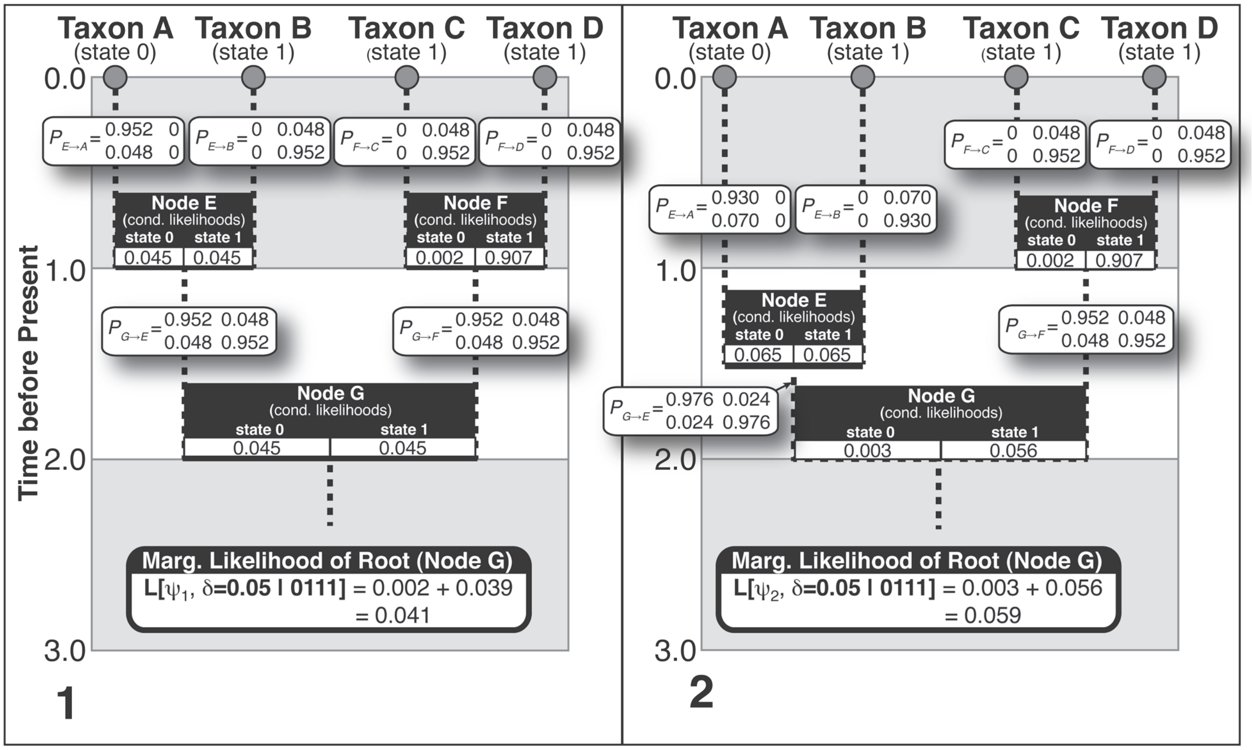

FIGURE *2*.—The likelihoods of two four-taxon phylogenies for contemporaneous taxa given rate of change δ=0.05 and branch durations as figured. $P_{\text{A-D}}$ gives the probability of the observed states for each taxon given either possible ancestral hypothesis. $P_{\text{E-F}}$ give the probabilities of E (= node (AB)) and F (= node (CD) having either state 0 or 1 given either possible condition for the basal node (G). The vectors above the nodes give the marginal likelihoods of the alternative ancestral reconstructions. The likelihood of the entire tree is the sum of the marginal likelihoods of the root node (Felsenstein, 1981). See text for exact calculations.

$$\mathbf{P}=\begin{bmatrix}0.952 & 0.048\\ 0.048 & 0.952\end{bmatrix} \qquad \text{(1a).}$$

Because this particular example has the same t for each branch, we can use the same **P** to determine the likelihoods conditioned on either state 0 or state 1 being the ancestral condition at each node on the tree. In all cases, this is simply the joint probability of going from that state to the two observed states. The likelihood that the ancestral state at the node linking A to B (hereafter, E) is 0 is the joint likelihood (i.e., the product of the probabilities) of net stasis along the branch to A (i.e., p_{00}) and a net change from state 0 at E to state 1 in B (i.e., p_{01}):

$$\begin{aligned}\text{L[E=0|A=0,B=1]} &= p_{00}\times p_{01} \qquad \text{(2a)}\\ &= 0.952\times 0.048\\ &= 0.045.\end{aligned}$$

(Note that here and elsewhere, the final numbers reflect rounding error.) Alternatively, the likelihood that the ancestral state at Node E is 1 is the joint likelihood of a net change in state along the branch to A and net stasis along the branch to B:

$$\begin{aligned}\text{L[E=1|A=0,B=1]} &= p_{10}\times p_{11}\\ &= 0.048\times 0.952\\ &= 0.045.\end{aligned}$$

The ancestral state at Node E is not known, so we want to know the likelihood given either possibility. Thus, the marginal likelihood of Node E is the sum of the conditional likelihoods calculated for all possible ancestral states at the node.

$$\begin{aligned} L[E=(0,1)|\,A=0,B=1] &= L[E=0|\,A=0,B=1] + \\ &\quad L[E=1|\,A=0,B=1] \qquad (2b) \\ &= 0.045 + 0.045 \\ &= 0.090. \end{aligned}$$

Similarly, on the node linking C to D (hereafter, F):

$$\begin{aligned} L[F=0|C=1,D=1] &= p_{01}\times p_{01} \\ &= 0.048 \times 0.048 \\ &= 2.3\times10^{-3} \end{aligned}$$

and

$$\begin{aligned} L[F=1|\,C=1,D=1] &= p_{11}\times p_{11} \\ &= 0.952\times0.952 \\ &= 0.907 \end{aligned}$$

so

$$\begin{aligned} L[F=(0,1)|\,C=1,D=1] &= L[F=0|\,C=1,D=1] + \\ &\quad L[F=1|\,C=1,D=1] \\ &= 2.3\times10^{-3} + 0.907 \\ &= 0.909 \end{aligned}$$

The fact that nodal conditions are unknown is very important as we descend the tree and estimate the likelihood of nodes linking other nodes (e.g., Node G). Now the conditional likelihood of Node G having state 0 is the sum of joint likelihoods of all four possible combinations of states at Nodes E and F. In other words, we estimate the probability of state 0 leading (ultimately) to the observed data through all possible paths. Thus,

$$\begin{aligned} L[G=0] = \quad & [(p_{00} \times L[E=0]) \times (p_{00} \times L[F=0])] + \\ & [(p_{01} \times L[E=1]) \times (p_{00} \times L[F=0])] + \\ & [(p_{00} \times L[E=0]) \times (p_{01} \times L[F=1])] + \\ & [(p_{01} \times L[E=1]) \times (p_{01} \times L[F=1])] \qquad (2c) \end{aligned}$$

These are all possible combinations of states for E and F, given that G has state 0, and therefore all of the possible evolutionary pathways leading to the observed character-states in taxa A, B, C and D. Substituting the results of equations given above and rearranging slightly, the conditional likelihood of G=0 is:

$$\begin{aligned} L[G=0|\text{data}] &= ([0.952\times0.045]+[0.048\times0.045])\times \\ &\quad ([0.952\times2.3\times10^{-3}]\times[0.048\times0.907]) \\ &= 0.002. \end{aligned}$$

Repeating this for G=1, we get:

$$\begin{aligned} L[G=1|\text{data}] &= (p_{10}\times L[E=0]+p_{11}\times L[E=1])\times \\ &\quad (p_{10}\times L[F=0]+p_{11}\times L[F=1]) \\ &= ([0.048\times0.045]+[0.952\times0.045])\times \\ &\quad ([0.048\times2.3\times10^{-3}]\times[0.952\times0.907]) \\ &= 0.039. \end{aligned}$$

Summing the conditional likelihoods, $L[\psi_1,\delta\,|\text{data}]=0.002+0.039=0.041$, which is the marginal likelihood of both Node G, and the entire phylogeny.

THE EFFECT OF BRANCH DURATIONS ON PHYLOGENY LIKELIHOODS

The Effect of Unequal Branch Durations (t).—The phylogeny ψ_1 represents a particular case of cladistic topology Ψ where t=1.0 for all branches. Consider instead ψ_2 (Fig. 2.2) with the same topology, but for which the branches leading to A and B both have durations t=1.5 and for which the branch leading to their ancestor (Node E) has t=0.5. Now, $\mathbf{P}_{E\to A} = \begin{array}{cc} \to & 0 \\ 0 & 0.930 \\ 1 & 0.070 \end{array}$ and $\mathbf{P}_{E\to B} = \begin{array}{cc} \to & 1 \\ 0 & 0.070 \\ 1 & 0.930 \end{array}$. Correspondingly, the conditional likelihoods of Node E given δ=0.05 now increase to L[E=0]=L[E=1]=0.065. In this case, the likelihood increases because we give the branches more time for the change(s) that must have occurred on at least one of them. Because we did not alter the durations of branches descending from Node F, the marginal likelihood of node F is unaffected. The likelihood of the whole tree increases to (L[G=0]=0.003) + (L[G=1]=0.056) = 0.059. Within any general (cladistic) topology, there will be one or a few *phylogenies* that maximize the probability of the character data at any given δ.

The Effect of Sampling Taxa over Time.—With some exceptions (e.g., Drummond et al., 2006; Kim and Sanderson, 2008), neontological studies do not vary δ or t independently as we describe above. Instead, they vary "branch lengths," i.e., the expected numbers of changes and thus a product of rate and time (i.e., $\delta \times t$). Whether branch lengths differ because of time or rate or whether the same branch lengths represent different combinations of time and rate generally is not important for relationships themselves: the distinction becomes important only if we are interested in divergence time. The vast majority of paleontological studies face a dif-

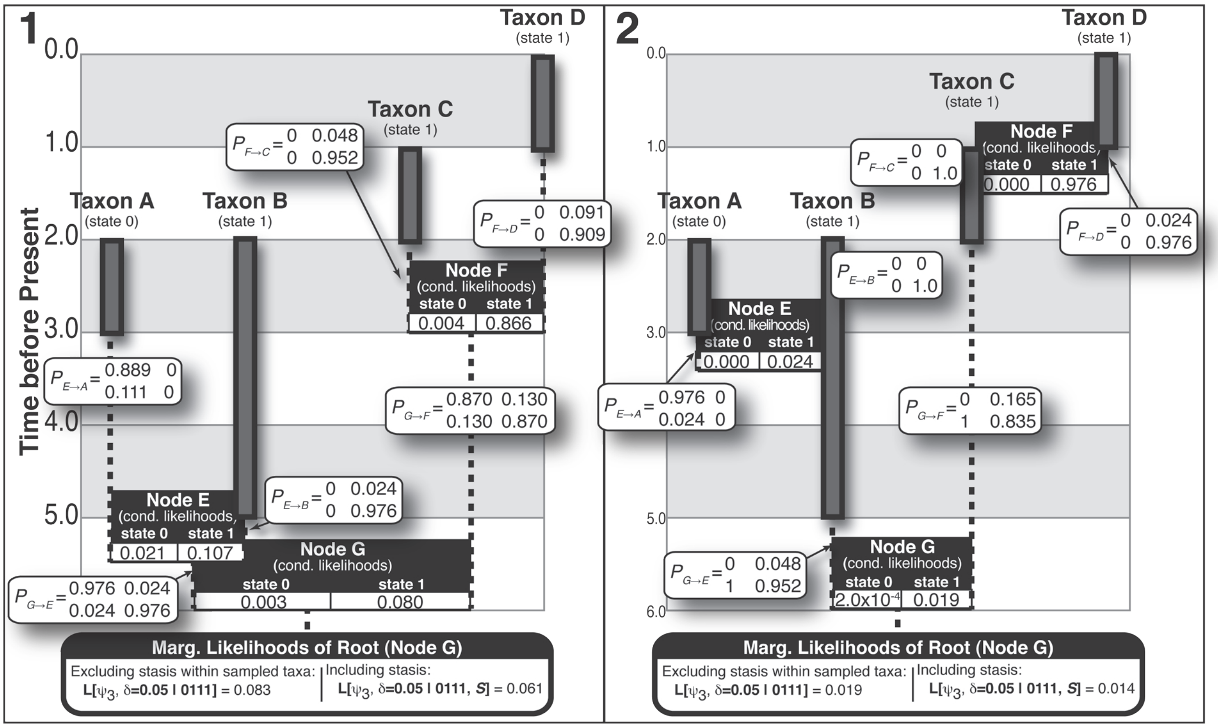

FIGURE *3*.—Calculating the likelihood of two four-taxon phylogenies with stratigraphic ranges. Probabilities of change and marginal likelihoods are given as in figure 1. 1. ψ_3, a phylogeny in which no ancestors are sampled. 2. ψ_4, We assume that each taxon is known from throughout each chronostratigraphic unit rather than being known from only a portion of it. Two tree likelihoods are given. The first (upper) likelihood considers the probability of the data only given the unsampled regions (i.e., the dashed lines). The second (lower) likelihood multiplies this by the probability of stasis between the first and last appearances of the sampled taxa. Positioning B and C as ancestors in ψ_4 means that they must retain the ancestral condition. It also means that we ignore the probability of transition to state 0 for nodes E and F, as the hypothesis specifies that E=B and F=C.

ferent sort of variation in branch duration: taxa being sampled over time rather than from a single time-slice. We note above that this permits three basic types of inference. Let us consider first another phylogeny, ψ_3, consistent with Ψ (Fig. 3.1), in which the taxa have different stratigraphic ranges, but none are ancestral to each other (e.g., Fig. 1.2).

A simplifying aspect of the examples in Figure 2 is no longer true: the branch durations separating observed sister taxa from their last common ancestors are no longer equal. Thus, $\mathbf{P}_{E\rightarrow A}\neq\mathbf{P}_{E\rightarrow B}$ and $\mathbf{P}_{F\rightarrow C}\neq\mathbf{P}_{F\rightarrow D}$. An obvious effect of sister taxa having different branch durations is that they have different influences on the conditional likelihoods of ancestral states. On ψ_3 we see both increases and decreases in most likely ancestral character-states relative to what we had for ψ_1. For node E, states 0 and 1 are equally likely on ψ_1 or ψ_2. However, state 1 is much more likely than state 0 for E on ψ_3. This is because we now posit much more time for change along the branch leading to A than for the branch leading to B. Those taxa closer in time to the most recent common ancestor are more likely to reflect the ancestral states (e.g., Donoghue et al., 1989). For node F, state 1 is the most likely ancestral reconstruction on both ψ_1 and ψ_2, but we see a slight decrease in the likelihood of state 1 and a slight increase in the likelihood of state 0 for this node on ψ_3. This reflects the increased evolutionary time separating the node from both descendants, which makes it slightly less probable that a state would show net stasis and slightly more probable that there would be two net changes.

Another basic assumption in our calculations might seem to be paradoxical. On one hand, we assume that character change can occur at any time on branches separating sampled taxa. On the other hand, we do not take into account the fact that all of the taxa in Figure 3 have some large s over which no change accrued. That is, our calculations do not assess the probability of stasis given s and δ. (We shall address the possibility that some of these taxa are simply morphospecies within an anagenetically evolving lineage below.) Accounting for stasis over all s in the calculations affects the likelihood of any hypothesized δ and thus the likelihood of the overall hypothesized evolutionary history. We do not use p_{ii} here because it corresponds to the probability of *net* stasis, which allows for reversals over some interval of time. Instead, we use the Poisson probability of zero observed changes given an expectation of $(k\text{-}1)\,\delta t$ changes:

$$\mathrm{P}[\text{stasis} \mid t] = e^{-(k-1)\delta t} \tag{3}$$

If there are gaps within a species range, however, then one could use p_{ii} from each gap to the next. We present this likelihood as well as the likelihood given only unsampled lineages in Figure 3.1 and 3.2.

The Effect of Ancestor-Descendant Relationships.—Another possibility introduced by sampling over time is the sampling of ancestor-descendant pairs. At the species-level, we can sample ancestor-descendant pairs among contemporaneous taxa when evolution occurs under budding cladogenetic models (e.g., Taxa A and B in Fig. 3) and the ancestral species persists after the divergence of a daughter taxon. Under this model, we might sample ancestral morphospecies among extant taxa, and many molecular studies suggest that this might be common (e.g., Funk and Omland, 2003). Fossil data, however, offer an additional possibility. Because we can sample over time, we can sample ancestor-descendant morphospecies pairs from a single lineage (e.g., Fig. 1.4). Thus, paleontologists might find ancestor-descendant pairs under any evolutionary model, not simply the "budding" model. We shall return to this topic in more detail below when we contrast the expectations of punctuated character change with those of continuous character change.

Consider phylogeny ψ_4 with the same cladistic topology and stratigraphic ranges as ψ_3 (Fig. 3.1), but with d_E and d_F that require ancestor-descendant hypotheses (Fig. 3.2). There are several points to stress. First, ψ_4 specifies $\mathbf{P}_{E\to B} = \mathbf{P}_{F\to C} = \begin{array}{cc} \to & 1 \\ 0 & 0.000 \\ 1 & 1.000 \end{array}$.

The marginal likelihoods now might seem confusing: even though the ancestral state *must* be 1, the marginal likelihoods of the ancestral states being 1 are less than 1.0. This is, of course, due to the fact that the likelihood reflects the probability of going from 1→0 (B→A) or from 1→1 (C→D), not the probability that the ancestral reconstructions are correct. Moreover, the likelihood of the only alternative (state 0) is zero in both cases, which means that particular hypothesis must be false if the more general hypothesis is true.

Positing observed ancestral conditions means that $\mathbf{P}_{G\to E}$ and $\mathbf{P}_{G\to F}$ now are vectors rather than matrices. Note, however, that we do not calculate the likelihood of the root node as if it links only two observed taxa. Instead, it links two segments of phylogeny: thus we must multiply $\mathbf{P}_{G\to E}$ and $\mathbf{P}_{G\to F}$ by the marginal likelihoods of those nodes to get the likelihoods of states 0 and 1 for the root node. Correspondingly, the likelihood of state 0 is very low.

Finally, note that the likelihood of ψ_4 is lower than that of ψ_3. This is largely an effect of sampled versus unsampled ancestors. Phylogenies with unsampled ancestors actually are a collection of phylogenies: for example, $\psi_{1\text{-}3}$ each represent $2^3=8$ distinct evolutionary histories based on the alternative ancestral reconstructions for nodes E, F & G (see Appendix). Each of these alternatives contributes to the overall likelihood of the phylogeny when we sum the marginal likelihoods. Positing sampled ancestors reduces the summed histories to only two (based on the basal node) for ψ_4. Thus, adding hypothetical ancestors is akin to adding a parameter to a model (see Wang this volume; Hunt this volume): unsampled ancestors provide the hypotheses with greater flexibility to predict the data. Likelihood theory, information theory and graph theory all demand that the improvement in likelihood be substantial to justify this flexibility (Edwards, 1992; Alroy, 1995; see also Fisher, 1994; Smith, 1994). However, this also is not a case where the simpler hypothesis is a special case of the more complicated hypothesis. Thus, we cannot evaluate the significance of likelihood improvement using log-likelihood ratio tests (see, e.g., Goldman, 1990; Wang, this volume). It is possible that information theory criteria (e.g., AIC; Burnham and Anderson, 2002; Hunt this volume) could be used, but this requires further exploration, as the amount of simplification is dependent on numerous factors (see Appendix). Parametric bootstrapping, where the null hypothesis is used to parameterize simulations and determine the expected improvement from incorrect test hypotheses (Huelsenbeck et al., 1996) might be required here.

This discussion above assumes that one is analyzing species, but the basic principles apply to supraspecific analyses. Indeed, because the sampling of higher taxa typically is better than that of species, we expect to sample ancestral morphotypes at the supraspecific level more frequently than we sample those at the species level (see, e.g., Foote, 1996b). The major caveat here is that the stratigraphic range of, say, a genus often might not be not the stratigraphic range of the coded morphotype representing that genus. For example, early representatives of genera often lack some of the diagnostic traits of the whole genus (see, e.g., Brusatte, 2010). Conversely, many genera include species that have one or more synapomorphies of a "daughter genus." However, because they retain the primary diagnostic characters of the "ancestral genus," they still are classified in that genus. As a result, it often might be more difficult to ancestor-descendant hypotheses among genera when using only genera rather than species within genera.

EFFECTS OF CHARACTER EVOLUTION MODELS

Statistical models versus biological models.—As noted above, we need character evolution models to calculate the probability of data. The diversity of currently available models for probabilistic inference and their use has been reviewed elsewhere (e.g., Galtier et al., 2005; Kelchner and Thomas, 2007), as have cautions on their application (Gatesy, 2007). The model we have used so far is a Markov process in which the numbers of state changes follow a Poisson distribution (e.g., Galtier et al., 2005). The Poisson model is only a convention: we can use *any* model of change that gives the probability of the observed distribution of character-states among taxa (see below). We also must emphasize that these statistical models used to describe patterns of evolutionary change typically are agnostic with respect to biological process models: that is, large numbers of biological processes will yield virtually identical distributions such as Poisson distributions. Statistical models used in probabilistic inference rarely preclude or specify particular biological processes.

Nevertheless, our examples above assume that rates of change among states are equal (i.e., symmetrical) and that δ is the same on all branches. Empirical studies demonstrate that rates of change among states of a single character can be unequal (e.g., Sanderson, 1993), and that rates can vary in different portions of clades (e.g., Wagner, 1997; Ruta et al., 2006). In addition, rates might vary among characters (e.g., site-to-site rate variation Yang, 1996). In the subsequent sections, we shall outline how to accommodate these possibilities. As these problems are not unique to paleontology, and as paleontologists themselves have contributed to general solutions, we will not go into these issues in the level of detail used above.

Variation in Rates Among Characters.—In the examples above, we arbitrarily set δ=0.05. Because the calculated likelihoods give L[ψ_i,δ|data], we will get a different likelihood when δ=0.05 than, say, when δ=0.06. On such a small phylogeny, likelihoods usually increase continuously as rates do, asymptotically approaching 0.125: the value where all eight possible vectors for binary characters over four taxa are equally probable. Of course, most studies examine many more than 4-taxa. Consider a 16-taxon phylogeny that is symmetrical: i.e., two clades of 8 taxa that both include two clades of four taxa identical to ψ_1. When we look at numerous characters with different numbers of minimum changes (Fig. 4), then several important features emerge:

1) characters with low minimum numbers of changes show peak likelihoods at low δ;
2) characters with low minimum numbers of changes show higher peak likelihoods than do those characters with high minimum numbers of changes;
3) as rates increase, we retain an asymptotic approach to likelihoods equal to the probability of a randomly generated binary vector, but with this likelihood now very much lower than the peak likelihoods.

The log-likelihood of the 16-taxon tree for any single δ is the sum of the heights of these curves at that δ. However, if we use separate δ for different characters (or sets of characters), we can substantially increase overall likelihood (e.g., Wagner, 2001). For example, the combination of δ_1=0.04, δ_2=0.12 and δ_3=0.17 would increase the log-likelihood given these 22 characters by a factor greater than 15. Given log-likelihood ratio tests, this is well worth adding two parameters. (Note that log-likelihood ratio tests become inappropriate if we modify ψ to accommodate these new rates; Gold-

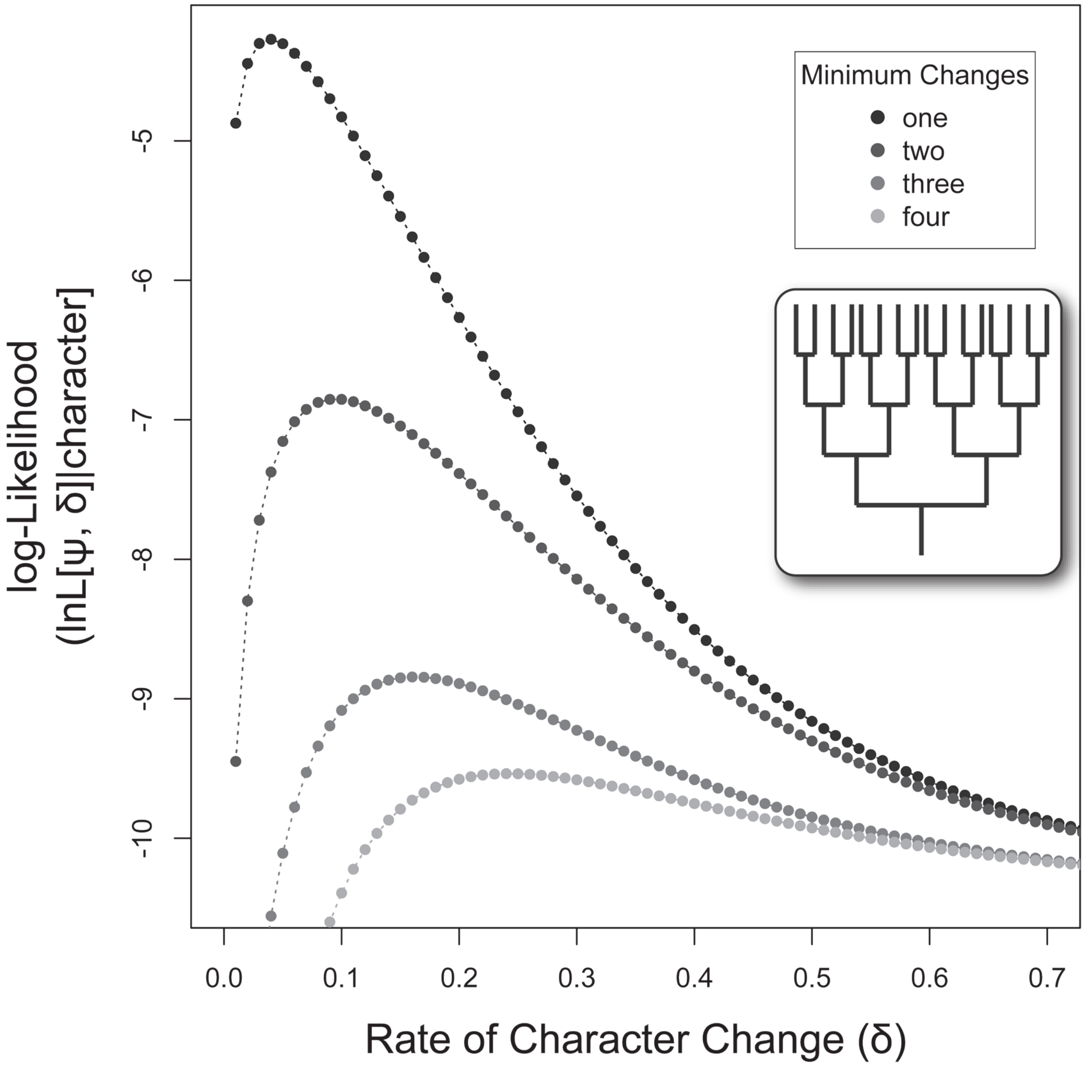

FIGURE *4*.—1. Log-likelihoods of δ given numerous character vectors for a symmetrical 16-taxon phylogeny with t=1.0 for all internal branches. (We ignore the effects of stasis within sampled taxa, but this will have a uniform effect on each character at any given δ). "Changes" give the minimum number of changes possible for each character on this hypothetical tree.

man, 1990).

Unfortunately, optimizing a separate value of δ for each character slows down computer analyses greatly and can easily lead to over-parameterization supporting unnecessary evolutionary hypotheses. An alternative is to calculate the probability of the data using a range of rates, with the subsequent likelihood an average of these probabilities weighted by the probability of the rates themselves.. Systematists commonly use the gamma (Γ) distribution for this purpose (e.g., Yang, 1993, 1994, 1996). The Γ distribution has two parameters (shape parameter α and scale parameter β), but in practice β is constrained to equal α, which both reduces the distribution to a single parameter and sets the mean of the distribution to 1.0.

For computational simplicity, workers typically use a discrete approximation of the Γ distribution (Yang, 1994) that divides the area beneath the distribution into a fixed number of equal partitions, each of which corresponds to a "rate class". Thus, we can use the mid-point of each rate class (i.e., the points separating 12.5%, 37.5%, 62.5% and 87.5% of the total area from the end of the curve) as a scalar multiplier of δ, the "overall" rate. However, we do not assign individual characters to a particular rate class. Instead, we calculate marginal likelihoods of the tree over all possible rate class assignments to all characters.

Now return to our initial example with ψ_1, where we arbitrarily used δ=0.05. The optimal parameter for the gamma distribution is typically estimated using all characters in the data matrix. In our example, we calculated the likelihood of the tree for only a single character, so, for simplicity, let us assume that other characters suggest a Γ distribution with $\alpha=\beta=1.0$. The midpoints of our rate classes of the Γ distribution are 0.134, 0.470, 0.981 and 2.079. We now calculate the likelihood of ψ_1 four times, with $\delta_1=0.05\times0.134=0.0067$, $\delta_2=0.05\times0.470=0.0235$, $\delta_3=0.05\times0.981=0.0491$ and $\delta_4=0.05\times2.079=0.1040$. Inserting each δ_i into Equation 2 above gives us likelihoods of 0.007, 0.0214, 0.041 and 0.071 respectively. Now, the likelihood of the tree,

$$L[\psi_1,\delta=0.05,\Gamma_{\alpha=\beta=1.0} \mid \text{data}] = \frac{(0.007 + 0.021 + 0.041 + 0.071)}{4}.$$

This example demonstrates the application of the Γ distribution to a single character, and these calculations would be repeated then for all other characters in the data matrix.

Because we set $\beta=\alpha$, the Γ distribution has only one freely varying parameter. This parameter essentially describes the how characters are partitioned among the rate classes: as α increases, the difference between δ_1 and δ_4 decreases and thus the modeled variation in rates decreases. The value of α estimated for a set of characters is not dependent on the topology: it can be reliably estimated from an initial parsimony tree or any other starter tree (Yang, 1993). Among-character rate variation also can be modeled with alternative statistical distributions such as the Dirichlet distribution (Huelsenbeck and Suchard, 2007) which can provide greater flexibility in the rate classes while using two parameters. Mixture modeling (e.g., Pagel, 1997, 1999) also can be used. The point that bears emphasizing is that we can model considerable rate variation among characters with few parameters.

Variation in Transition Rate among States.—For some characters, some states might evolve more often than others. This posits an asymmetrical transition matrix for equation 1, where $p_{01}\neq p_{10}$ and thus $p_{00}\neq p_{11}$. "Dollo's Law" (e.g., Goldberg and Igić, 2008) is an extreme version of this model where change is irreversible. Driven trends (*sensu* McShea, 1994), where we see biased gains/loses or "increase/decrease" along multistate characters, represent a more intermediate example. Because of the macroevolutionary implications of this model, it has been well researched by both paleontologists and neontologists. Numerous methods exist that basically employ asymmetrical transition matrices (e.g., Sanderson, 1993; Pagel, 1997), which have been applied to paleontological data (e.g., Wagner, 2001). Obviously, we cannot consider this for all characters easily. However, several methods for detecting driven trends without phylogenetic models exist (e.g., McShea, 1994; Alroy, 2000; Wang, 2001). Workers should use these prior to phylogenetic analysis to identify likely candidates for biased change.

Variation in Rates over Phylogeny.—The examples above assume a constant rate of change throughout the tree. Numerous paleontological and neontological studies indicate that rates vary across phylogeny. This might represent secular shifts over time due to any number of factors (e.g., ecologic turnover, sudden climate change, etc.) Alternatively, subgroups within a larger clade might show different characteristic rates due to differences in intrinsic constraints or ecologic restrictions (e.g., Wagner, 1997). Thus, even if there are general rates for characters, we do not wish to assume these rates will be constant over time or across all branches of the phylogeny.

Interest in dating nodes accurately has led molecular systematists to develop "relaxed clock" methods that allow for shifts in rates. "Shift" methods assume that δ itself evolves, according to Brownian motion process (Thorne et al., 1998), Poisson process (Huelsenbeck et al., 2000) or other statistical processes (Sanderson, 2002). The assumption of rate heritability reduces the number of free parameters in the overall model and also lends itself conceptually to paleontological studies testing for shifts in rates over time. Paleontologists have two advantages over neontologists here. First, whereas neontologists must parse δ and t from branch lengths, we have much more exact data concerning t. This comes not simply from first appearances, but frequently also from dense sampling, which can tightly constrain plausible divergence times (see below; Marshall this volume.) Second, paleontologists can employ disparity analyses that can point to the need for shifting rates in advance of phylogenetic analyses (Foote, 1996a). Thus, we often have other tests that can indicate such shifts prior to phylogenetic analysis.

Even if there are no secular changes in character evolution rates, then we still can imagine numerous processes that might elevate or deflate rates of character change from one branch to the next (i.e., “local rates”). Drummond et al. (2006) use log-normal distributions around “average” rates to model local rate variation. As with the Γ distribution for character rates, one can divide the lognormal distribution into equal area units, with the midpoint of each area used as a multiplier for the “base” δ. The lognormal requires two parameters, the mean and a standard deviation term describing how many times the rate increases/decreases. The mean for any rate is simply δ, which means that we really are dealing with only one parameter. If we unite this with the Γ distribution described above, then we have four means based on one parameter (α from the Γ distribution). Thus, we now need only two parameters, α and the “variance” from the lognormal distribution, to get 4 lognormal rate distributions that model variation among characters and among branches.

PUNCTUATED VERSUS CONTINUOUS CHANGE

The Effects of Punctuation on Expectations.—A very different type of rate variation over phylogeny is the possibility that character change is punctuated. Punctuated change is concentrated into particular events that might occur within lineages (Wright, 1931; Malmgren et al., 1983), during cladogenesis (Eldredge and Gould, 1972) or a combination of cladogenesis and anagenesis (e.g., Futuyma, 1987). Punctuated change as posited by ψ_5 (Fig. 5) presents a challenge for Equation 1: if d_E=FA$_A$, then t=0 for the branch leading to A. Now, p_{01}=0.0. In other words, the instantaneous derivation of A from B is considered to be impossible!

Of course, the Poisson process that we have assumed in the prior examples really is a proxy for a discrete time process: the unit of evolutionary time in continuous models of evolution is the generation, not absolute time (e.g., Fisher, 1930; Haldane, 1949). However, generations are so short on geologic/phylogenetic time scales that the difference between mathematical assumption and biological reality is trivial. Although the difference is not trivial for the punctuated model, continuous statistical processes actually do a fair job of modeling discrete change. For a single event, a binomial distribution is appropriate rather than a Poisson distribution. However, when δ is low the probability of stasis for one branch (t=1) is nearly identical given either Poisson or binomial distributions: $e^{-\delta} \cong (1\text{-}\delta)$. A compound Poisson model, with a Poisson distribution of change over speciation events that themselves had a Poisson distribution over time, increases the differences only slightly: at δ=0.05, $P_{E\rightarrow A}= \begin{array}{cc} 0 & 0.000 \\ 1 & 0.048 \end{array}$ and $\mathbf{P}_{F\rightarrow D}= \begin{array}{cc} 0 & 0.000 \\ 1 & 0.050 \end{array}$ under the compound Poisson, whereas $\mathbf{P}_{E\rightarrow A}= \begin{array}{cc} 0 & 0.000 \\ 1 & 0.952 \end{array}$ and $\mathbf{P}_{F\rightarrow D}= \begin{array}{cc} 0 & 0.000 \\ 1 & 0.950 \end{array}$ under the binomial. Thus, the effects of different δ shown in Fig. 4 will quickly exceed the effects of the two different models.

If change is restricted to speciation events, then the total opportunity for character change will then be proportional to the number of speciation events. Unfortunately, this means we cannot know how many character-change opportunities actually happened along any branch: not all species are sampled from the fossil record, and it is common in phylogenetic analysis to include only a subset of those that are sampled. On ψ_5, this is most obvious for Taxon C, where three time units separate its first occurrence in the fossil record from its common ancestor with B. However, we also should not assume just a single speciation event along the branches separating A from B or C from D: numerous speciation events might happen (geologically) instantaneously. In either case, we expect a Poisson distribution for the speciation events themselves (see, e.g., Bokma, 2008).

Now character evolution is a compound model of binomial character change with Poisson distributions over either time or turnover events. The probability distribution functions for net change under punctuated evolution now are:

$$\begin{aligned}
&P[\text{net change} \mid \delta,\lambda,t] = \\
&(1-\delta)\sum_{m=0}^{\infty} P[\text{net change} \mid \delta,\, m \text{ events}] \times P[m \text{ events} \mid \lambda, t] \\
&+\ \delta\sum_{m=0}^{\infty} P[\text{net stasis} \mid \delta,\, m \text{ events}] \times P[m \text{ events} \mid \lambda, t] \qquad (4).
\end{aligned}$$

where the first term gives the probability of net change

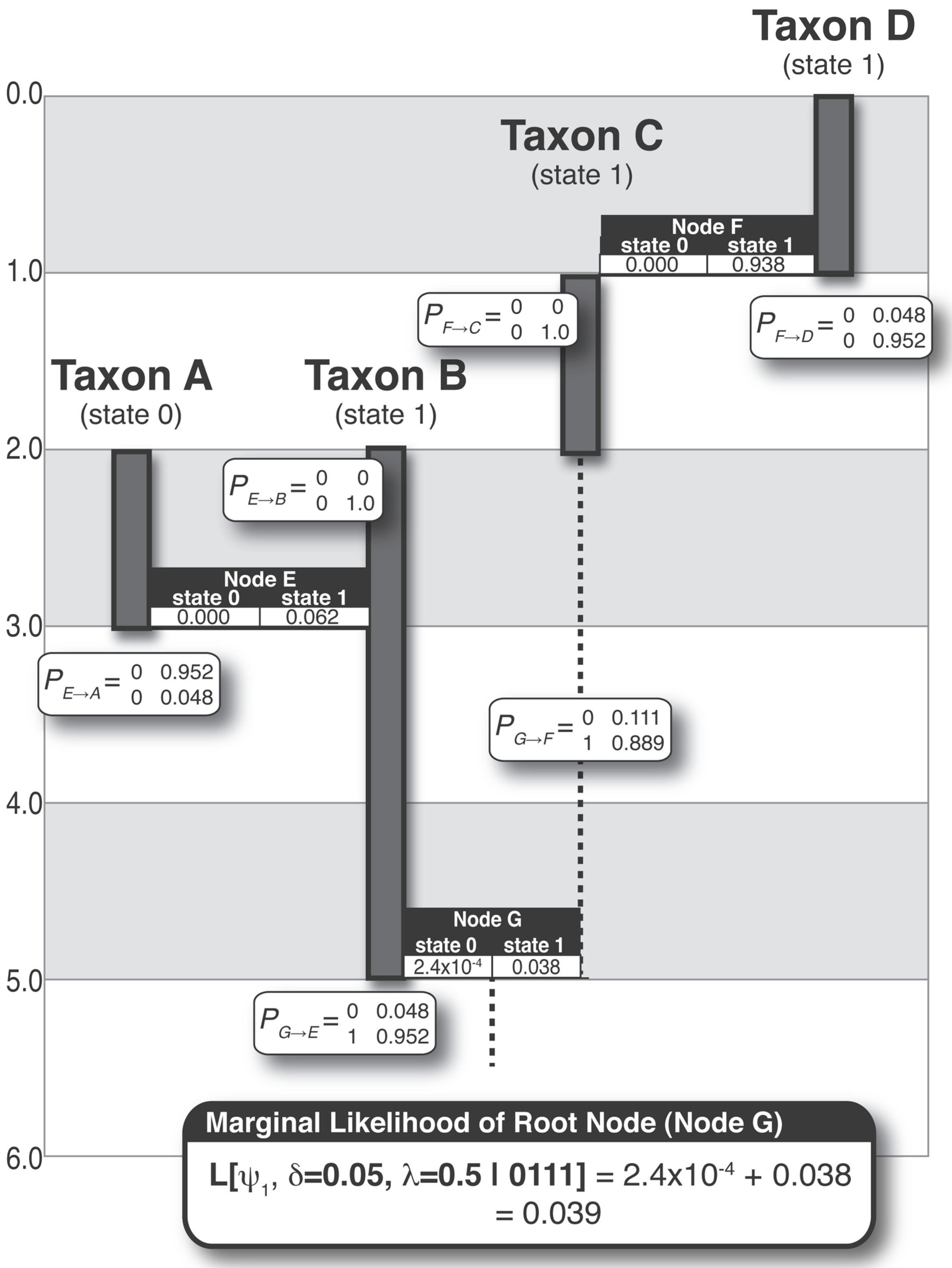

FIGURE 5.—A four-taxon tree, ψ_5, for taxa sampled over time with two ancestor-descendant pairs and punctuated morphological change. We assume that speciation events (either budding cladogenetic as in B→A or anagenetic as in C→D) have a Poisson distribution with expectation λ per time unit.

after no initial change given subsequent speciation events and the second term gives the probability of net stasis after initially changing to any one state. Here, *m* is the number of speciation events, λ is the speciation rate, $P[m \geq 1|\lambda, t]$ is the conditional probability of at least one such event, and δ and *t* are instantaneous rate and evolutionary time (continuous time or number of speciation events) as before. For a binary character, "net change" encompasses all odd number of changes, as even numbers of changes necessarily lead back to the original state. Although we cannot sum to infinity, it is unnecessary to do so: Equation 4 quickly converges to $p_{01} \approx \frac{1}{k} - \frac{1}{k} e^{-k\delta\lambda(1+t)}$. Similarly, $p_{00} \approx \frac{1}{k} + \frac{k-1}{k} e^{-k\delta(1+\lambda t)}$.

In other words, it is just the probability of net change/stasis with evolutionary time now the expected number of events (with a minimum of one event). Thus, the very different assumptions about biological patterns underlying phylogeny have little effect on phylogeny likelihood, at least insofar as the.

Model choice.—Numerous methods exist for estimating turnover rates without a model phylogeny, both within clades and among different intervals (e.g., Foote, 2001; Alroy, 2008; Liow and Nichols this volume). Indeed, our ability to independently estimate shifts in λ actually should allow paleontologists to recognize that local shifts in rates of character change actually reflect elevated speciation rates rather than (or in addition to) local elevations in δ. There are taxonomic groups for which the fossil record is not sufficient to estimate turnover rates (see Foote and Raup, 1996). However, with poor fossil records, most (if not all) branches will be like the Taxon C's branches on ψ_5, and the expected change will reflect duration of the branch under either model.

We should not assume either continuous or punctuated change. Instead, the likelihoods under the different models allow them to be evaluated during the course of phylogenetic inference. If we recalculate the likelihoods of our three example trees (Table 1) under these different models, then we can see that the punctuated model not only makes ψ_5 possible, but it also triples the likelihood of ψ_4 while leaving the likelihood of ψ_3 almost unchanged. If this pattern were repeated with additional characters and over more branches, then we very rapidly would reach the point where the punctuated model would be significantly more likely given log-likelihood ratio tests for numerous particular phylogenies in this example. Yet again, probabilistic phylogenetic analysis informs us not only about relationships, but also about the evolutionary history of a clade.

Supraspecific taxa.—Our associations between evolutionary models and phylogenetic patterns do not apply easily to supraspecific taxa. Even there is continuous anagenetic change among species, anagenesis is an inappropriate model for change within a genus with 2+ contemporaneous lineages. Moreover, the fact that multiple species share the suite of characters also indicates that there is some degree of stasis. However, it does not follow from this that change between genera is "punctuated": instead, it might happen over multiple species, unsampled and/or sampled and assigned to either genus. Nevertheless, it is essentially assumed that change is between taxa rather than among taxa in this case. Until one can conduct species-level analyses, it probably is best to simply examine change over branches linking sampled taxa.

TABLE *1*. Likelihoods of the three example phylogenies under different likelihood models. In all cases, δ=0.05. For the punctuated models, λ=0.50 and it is assumed that there is at least one opportunity for speciation on each branch. Note that likelihoods for ψ_3 are necessarily higher than for ψ_4 or ψ_5 because of the increased alternative pathways offered by unsampled ancestors in ψ_3.

Phylogeny	Continuous, unsampled only	Continuous, sampled + unsampled	Punctuated, sampled + unsampled
ψ_3	0.0826	0.0612	0.0725
ψ_4	0.0191	0.0142	0.0453
ψ_5	0.0000	0.0000	0.0386

CHARACTER CODING: A BRIEF COMMENT

There are no universal protocols for defining characters and character-states. However, some common practices that are currently suitable for parsimony-based phylogenetic inference are not optimal for probabilistic inference. Specifically, highly homoplastic characters (those known or suspected to change frequently among included taxa) and autapomorphies are useless or misleading for parsimony-based inference, and are therefore regularly excluded from phylogenetic data matrices. Yet, in the context of probabilistic inference, such characters improve parameter estimation for the previously described models of character evolution. While they might not unequivocally indicate shared ancestry, such characters are useful for rate estimation. Indeed, most parameterizations of the Γ distribution predict that there will be a substantial number of characters that change at a high rate. Autapomorphies might accumulate along lineages with long *t*, but failure to include them could lead to artifactually low rate estimates along such branches. Autapomorphies can also make some trees less likely than they do others (Lewis, 2001). This is particularly true for paleontological studies as autapomorphies can decrease the plausibility of particular ancestor-descendant hypotheses (Fisher, 1994; Wagner, 1995). In order to test evolutionary histories properly, the inclusion autapomorphies (when they exist) and homoplastic character needs to become standard practice.

PHYLOGENY LIKELIHOOD GIVEN STRATIGRAPHIC DATA

Phylogenetic Predictions about the Fossil Record.—Phylogenetic hypotheses do not make predictions only about distributions of character-states: they also make predictions about the stratigraphic distributions of taxa in the fossil record. When sampling is dense, then large gaps in a taxon's stratigraphic range are improbable and hypotheses requiring such gaps have low likelihoods. The ability to constrain plausible *t* and thus *d* in this way potentially adds great power to phylogenetic studies of fossil taxa. One of the most consequential advantages of a probabilistic approach to phylogenetic inference is that different forms of data (e.g., character-states and stratigraphic ranges) can be integrated using the common currency of probability, so long as the same phylogeny makes quantifiable predictions about them.

Numerous papers present methods exist for assessing duration likelihoods given stratigraphic data (Huelsenbeck and Rannala, 1997a; Solow and Smith, 1997; Wagner, 2000a; Wang and Everson, 2007). The methods are basically identical to each other in that L[duration | stratigraphic data] relates to P[0 finds| gap required by duration, θ] where θ is preservation rate. For example, if we use individual localities, then is $(1-\theta)^g$ where g is the number of possible occurrences over a stratigraphic gap. (See Marshall, this volume, for fuller discussion of such methods)

Rates of preservation for testing gaps.—Even for groups with good fossil records, preservation can vary considerably among members of a clade. Wagner (2000a) recommends treating θ as an evolving parameter the way that some analyses mentioned above treat δ as an evolving parameter (e.g., Thorne et al., 1998). However, an analog to the local rates solution probably is more appropriate than this. Consider a distribution of θ's (Fig. 6, based on Ordovician-Silurian bellerophonts). This example shows a strong lognormal distribution, where the variation likely represents differences in abundances and geographic distributions (see, e.g., Buzas et al., 1982). Given the considerable variation around the mean, we clearly cannot assume a single value of θ for all bellerophonts. Instead, we can use this lognormal distribution to put probabilities on different *θ's* the way that character likelihood uses a Γ distribution to weight rates of character change by their probabilities. Now the likelihood of a hypothesized gap is:

$$\text{L[duration | stratigraphic ranges]} = \int_0^1 \left((1-\theta)^g \times P[\theta \,|\, \text{distribution}]\right) d\theta \qquad (5).$$

By doing this, we allow the ancestral lineages to have any preservation rate, including rates lower than actually observed in the clade; however, we also weight those θ's by empirically determined prior probabilities (see Wang this volume).

Let us suppose further that our four-taxon clade has the following number of appropriate fossiliferous localities in each time unit: 20 from Unit 2→1,

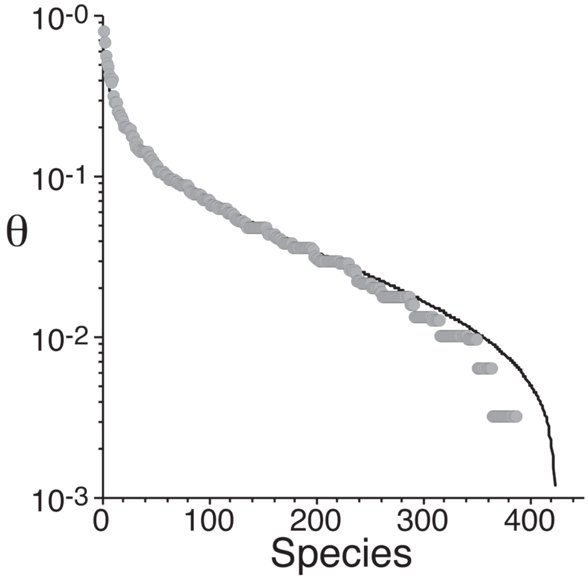

FIGURE *6*.—Preservation rates for Ordovician-Silurian bellerophontoid gastropods. Here, θ=occurrences/possible occurrences where "possible occurrences" are all the localities from formations bearing bellerophonts within a stratigraphic bin and within a biogeographic unit. The best fit lognormal has mean = 0.03 with a threefold increase/decrease in θ every standard deviation.

40 from Unit 3→2, 30 from Unit 3→4, 30 from Unit 4→5 and 20 from Unit 5→6. We will further assume that localities are evenly distributed through the units so that a lineage present in half a unit (e.g., A in ψ_3) might have been found in half of the localities in that unit. Finally, suppose that the bellerophont θ distribution is appropriate for this clade. The stratigraphic data entail large differences in the likelihoods of three alternative phylogenies (Table 2): ψ_4 is over 20 times more likely than ψ_3, and ψ_5 is over 5 times more likely than ψ_4. We also see large differences in the overall likelihoods of the trees given both morphology and stratigraphy: lnL[ψ_3| all data]=-9.52, lnL[ψ_4| all data]=-6.53 and L[ψ_5| all data]=-4.99 (see Table 1 for the best morphologic likelihoods for each phylogeny). Thus, the very different evolutionary histories perform very differently in their ability to predict not just the distribution of character-states, but also the distribution of morphologies in the fossil record.

The example above uses a large database of occurrences. Such data are unavailable for many taxa. However, methods exist for estimating per unit preservation rates based on first and last appearance data (e.g., Foote, 1997). Here again we expect that preservation rates will represent an average. As occurrences typically follow a log-normal distribution (e.g., Alroy et al., 2008), the log-normal distribution documented for bellerophonts probably will be common. We should use databases such as the Paleobiology Database to de-

TABLE *2*.—Stratigraphic gaps implied by each branch of alternative phylogenies and their likelihoods. The tree likelihoods given stratigraphic data alone are the products of these numbers. "•" reflects species hypothesized to be ancestral to another sampled species. Any gaps implied by those species are attributed to the node that they represent.

Branch	ψ_3	ψ_4	ψ_5	L[ψ_3|data]	L[ψ_4|data]	L[ψ_5|data]
A	70	15	0	0.2030	0.5691	1.0000
B	10	•	•	0.6609	•	•
(AB)	10	10	0	0.6609	0.6609	1.0000
C	40	•	•	0.3266	•	•
D	60	10	0	0.2345	0.6609	1.0000
(CD)	80	120	100	0.1774	0.1114	0.1389
			L[ψ|stratigraphic data]	0.0012	0.0277	0.1389
			L[ψ|character data]	0.0612	0.0453	0.0386
			L[ψ|character+stratigraphic data]	7.34×10^{-5}	0.0013	0.0054

termine the typical standard deviations accompanying average per-stage preservation rates and then estimate the probability of missing lineages as done above.

Ancestor-Descendant Hypotheses Revisited.—It is important to note that likelihoods given character data and stratigraphic data have the opposite effects on ancestor-descendant hypotheses: whereas sister-species hypotheses increase likelihoods given character data, they decrease likelihood given stratigraphy. This is simply because ancestor-descendant hypotheses almost always reduce stratigraphic gaps implied by the phylogeny. However we have a similar effect concerning hypothesis complexity: whereas ancestor-descendant hypotheses simplify morphologic evolution by reducing unknown ancestral character-states, they complicate duration hypotheses by allowing separate origins for closely related species (see Wagner, 2000b; Wang and Everson, 2007). Of the trees considered here, ψ_3 requires that both A+B and C+D share origination times whereas ψ_4 and ψ_5 allow asynchronous originations for A&B as well as C&D.

In this example, stratigraphic data favors punctuated phylogenies, as these minimize the necessary time over which intermediate forms might have existed. This might appear to represent a bias. However, with dense sampling of the fossil record, we expect to sample individual lineages multiple times (Foote, 1997). If anagenetic speciation occurs within a lineage, then we expect to find 2+ morphospecies within that lineage, and we should frequently reconstruct patterns such as Taxa C-D in Figure 5. Conversely, with poor sampling, we will rarely sample individual lineages more than once. In such cases, most taxa will be have "point occurrences" rather than stratigraphic ranges and adding ancestor-descendant relationships (punctuated or otherwise) would have had little (if any) effect on the number of necessary gaps. Thus, it is more appropriate to state that dense sampling creates the potential to recognize particular speciation patterns rather than biasing results towards suggesting them.

Finally, the resolution of stratigraphic data will play a large role in our ability to distinguish modes of anagenetic change. In our simple example, the stratigraphic range of each taxon is an entire chronostratigraphic unit. Ideally, we would order localities by ordinating the faunas including the clade of interest (see Sadler this volume). (Note that using ordinated horizons also alters the way we must calculate θ; e.g., Wagner, 2000a). If we have a continuous analogues of the C-D pair in ψ_5 (e.g., where there is a stratigraphic gap between LA_C and FA_D as in Fig. 1.4) will (probably) have with short ranges within their respective units. Continuous change now will predict morphologic change approximately as well as punctuated change without adding a "stasis" parameter (see above). As the same stratigraphic gap will apply to both hypothesized modes, we typically will favor the continuous change model in this case.

TREE PROBABILITIES

Basic concepts.—Bayesian inference of phylogenetic trees weights the likelihood of the tree by its prior probability (see Wang this volume). Two aspects of phylogenies have different probabilities given different evolutionary parameters: tree balance and the distribution of unsampled durations. Tree balance describes how evenly numbers of descendants tend to be distributed between sister clades. Although tree balance is the by-product of speciation and extinction rates (λ and μ; e.g., Heard, 1996; Mooers and Heard, 1997), current Bayesian phylogenetic algorithms assume uniform probabilities for tree shapes (e.g., Huelsenbeck et al., 2001). The second aspect is the distribution of branch durations (Rannala and Yang, 1996; Yang and Rannala, 1997). Here, algorithms use an exponential distribution for prior probabilities, which is the expectation of most speciation models (e.g., Raup, 1985; Nee et al., 1994).

At any set of turnover rates, we expect very different distributions of tree shapes and branch distributions for taxa sampled over time than for taxa sampled from a single time slice. In particular, we expect paleontological trees to be more pectinate (i.e., linking a species and a clade instead of two clades) than neontological trees (Harcourt-Brown et al., 2001). This reflects two factors. One, we can sample short-lived, species-poor subclades that go extinct long before the rest of the clade. Second, we can sample ancestors, which necessarily represent a single lineage rather than a clade of any richness. Anagenetic series do not simply limit tree shapes, but actually determine (portions) of shapes. If we sample an anagenetic series A→B→C→D, then the only possible topology for an anagenetic series is purely pectinate. Thus, the probability of a pectinate tree-shape if there is only anagenesis is 1.0. "Balanced" nodes (i.e., those linking at least two clades) can happen

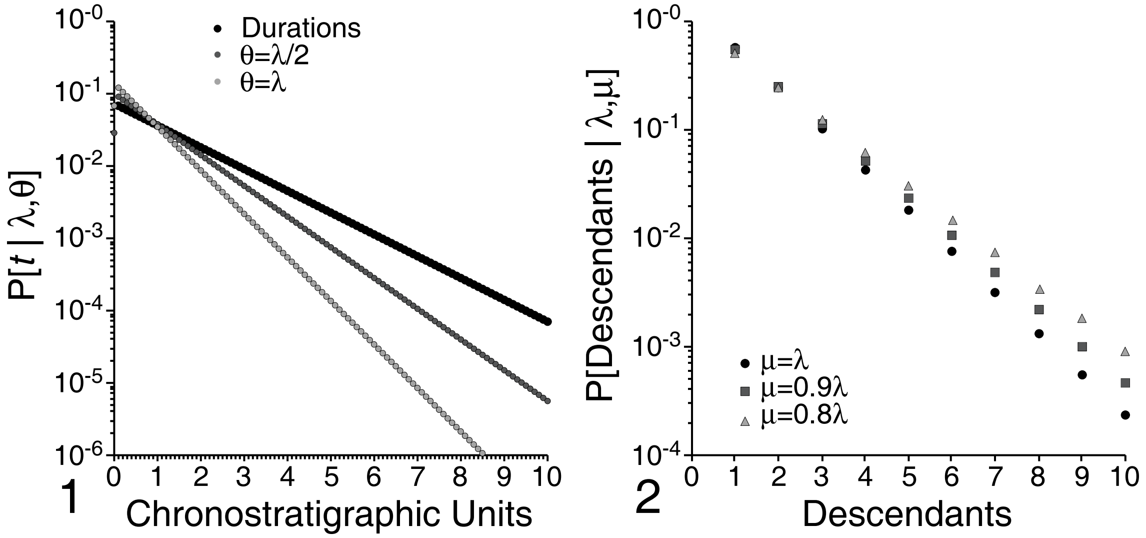

FIGURE 7.—Tree shape probabilities. 1. Expected distributions of the durations of branches linking taxa. Raw durations reflect origination rates (λ) only. When we add preservation rate θ, then the expected distribution of unsampled branch durations decreases as θ/λ increases. 2. Expected numbers of descendants per ancestral species at different relative origination (λ) and extinction (μ) rates, given a budding model. The expected number increases as λ/μ increases. However, note that under other models of speciation, high numbers of descendants are impossible.

only when there is cladogenesis. Thus, the prior probability of tree shapes for fossil clades depends in part on rates of anagenesis relative to rates of cladogenesis.

Sampling also alters the expectations for unsampled branch durations. We still expect durations to have exponential distributions, but we also expect to sample taxa with long distributions. Thus, *unsampled* ancestral lineages (or the unsampled portions of ancestral lineages) will reflect sampling rates as well as speciation rates: the exponential distribution now is the product of $(1\text{-}\lambda)(1\text{-}\theta)$ rather than just $(1\text{-}\lambda)$ (Fig. 7.1).

The final tree shape issue is actually not unique to paleontology, but it has been much neglected by neontologists. Under the budding cladogenesis model, ancestral taxa can give rise to any number of direct descendants. This has two ramifications. First, when extinction rates (μ) are close to speciation rates, then we expect most ancestors to have only one daughter taxon. Here again, pectinate topologies will be favored (Fig 7.2). However, we also expect as many as 25% of ancestors to have 3+ descendants. This, in turn, leads to polytomies, i.e., multiple lineages stemming from a single node (Wagner and Erwin, 1995). The frequency of polytomies of different sizes will be a product of the difference between λ and μ. However, as rates of anagenesis increase, the probability of polytomies will decrease.

PROSPECTUS

Although we have highlighted some potential modifications of existing approaches to calculate likelihoods of phylogenies, numerous other modifications can be made in the future. The models of character evolution are one example. Models of molecular evolution have been continually refined over the past 40 years. We can continue to adapt some of these for morphologic data, especially those that pertain to general rates of change. Furthermore, we expect empirical study of morphologic evolution in the fossil record will allow such models to be explicitly modified to address evolutionary phenomena that are only observable in the fossil record (e.g., punctuation and trends).

In this paper, we exclusively discuss new approaches to calculate the likelihood of phylogenies of fossil taxa. Likelihood is the sole optimality criterion for ML inference, and also contributes greatly to the posterior probability that characterizes Bayesian inference. We anticipate the use of Bayesian inference to increase in the near future, and the methods described herein can be easily adapted for this use. Bayesian methods have some advantages over ML inference, particularly in their ability to "show" uncertainty, and the speed of algorithms to find optimal trees. Although we focus on phylogeny here, Bayesian analyses can use prior probabilities for any of the relevant parameters (e.g., Ψ, d, δ, etc.). We anticipate that continued empirical study of fossil taxa will inform our understanding of appropriate distributions of these prior probabilities, and allow such distributions to be tailored specifically to fossil taxa.

Throughout this paper, we emphasize that phylogenetic inference in a vacuum is impossible, and that what we really must attempt is to infer the evolutionary history of a clade that includes phylogenetic relationships. To properly unravel relationships, we must simultaneously unravel how characters evolved. Data that bear on this include not just the characters themselves, but the stratigraphic distributions of taxa bearing those characters and even consideration of general speciation and extinction rates. This might make it seem that we are making phylogenetics much more assumption laden than desirable. In fact, the opposite is true: by explicitly considering different evolutionary parameters, we reduce the chances of incorrect assumptions distorting any particular aspect of reconstructed evolutionary history.

Goals for paleosystematists over the next decade should include:

1) Adjusting existing phylogenetic methods to accommodate our unique concerns stemming from sampling taxa over time instead of from a single time slice;
2) devising and refining tests of basic character evolution parameters;
3) refining tests of ancestor-descendant vs. sister-species hypotheses;
4) aggressively assessing the frequencies of anagenesis and budding cladogenesis in order to provide future workers with (semi) empirical prior probabilities of expected tree shapes.

A successful execution of this program will not only expand our knowledge of evolution, but also provide our neontological colleagues with valuable insights into general phylogenetic patterns that only we can provide.

ACKNOWLEDGEMENTS

We thank P. Novack-Gottshall, A. Purvis and G. Hunt for critical comments. We also thank G. Hunt and J. Alroy for inviting us to contribute to this Short Course, and their immeasurable patience as they realized what they had done. The bellerophontoid data were collected under NSF-EAR-0207874.

REFERENCES

ALROY, J. 1995. Continuous track analysis: a new phylogenetic and biogeographic method. Systematic Biology 44:152 - 178.

ALROY, J. 2000. Understanding the dynamics of trends within evolving lineages. Paleobiology 26:319 - 329.

ALROY, J. 2008. Dynamics of origination and extinction in the marine fossil record. Proceedings of the National Academy of Sciences 105:11536 – 11542.

ALROY, J., M. ABERHAN, D. J. BOTTJER, M. FOOTE, F. T. FÜRSICH, P. J. HARRIES, A. J. W. HENDY, S. M. HOLLAND, L. C. IVANY, W. KIESSLING, M. A. KOSNIK, C. R. MARSHALL, A. J. MCGOWAN, A. I. MILLER, T. D. OLSZEWSKI, M. E. PATZKOWSKY, S. E. PETERS, L. VILLIER, P. J. WAGNER, N. BONUSO, P. S. BORKOW, B. BRENNEIS, M. E. CLAPHAM, L. M. FALL, C. A. FERGUSON, V. L. HANSON, A. Z. KRUG, K. M. LAYOU, E. H. LECKEY, S. NÜRNBERG, C. M. POWERS, J. A. SESSA, C. SIMPSON, A. TOMASOVYCH, AND C. C. VISAGGI. 2008. Phanerozoic trends in the global diversity of marine invertebrates. Science 321:97 - 100.

AYACHE, N. C., AND T. J. NEAR. 2009. The utility of morphological data in resolving phylogenetic relationships of darters as exemplified with *Etheostoma* (Teleostei: Percidae). Bulletin of the Peabody Museum of Natural History, Yale University 50:327-346.

BECK, R. M. D., H. GODTHELP, V. WEISBECKER, M. ARCHER, AND S. J. HAND. 2008. Australia's oldest marsupial fossils and their biogeographical implications. PLoS ONE 3:e1858.

BOKMA, F. 2008. Bayesian estimation of speciation and extinction probabilities from (in)complete phylogenies. Evolution 62:2441-2445.

BRUSATTE, S. L. 2010. Representing supraspecific taxa in higher-level phylogenetic analyses: guidelines for palaeontologists. Palaeontology 53:1-9.

BURNHAM, K. P., AND D. R. ANDERSON. 2002. Model Selection and Inference: a practical Information-Theoretic approach. Springer, New York, 488 p.

BUZAS, M. A., C. F. KOCH, S. J. CULVER, AND N. F. SOHL. 1982. On the distribution of species occurrence. Paleobiology 8:143 - 150.

Clarke, J. A., and K. M. Middleton. 2008. Mosaicism, modules, and the evolution of birds: results from a Bayesian approach to the study of morphological evolution using discrete character data. Systematic Biology 57:185 - 201.

Donoghue, M. J., J. A. Doyle, J. A. Gauthier, A. G. Kluge, and T. Rowe. 1989. The importance of fossils in phylogeny reconstruction. Annual Review of Ecology and Systematics 20:431-460.

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

Edwards, A. W. F. 1992. Likelihood - Expanded edition. Johns Hopkins University Press, Baltimore, 275 p.

Edwards, A. W. F., and L. L. Cavalli-Sforza. 1964. Reconstruction of evolutionary trees, p. 67 - 76. *In* J. H. Heywood and J. McNeil (eds.), Phenetic and phylogenetic classification. Systematic Association, London.

Eldredge, N., and S. J. Gould. 1972. Punctuated Equilibria: an Alternative to Phyletic Gradualism, p. 82 - 115. *In* T. J. M. Schopf (ed.), Models in Paleobiology. Freeman, San Francisco.

Felsenstein, J. 1973. Maximum-likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Systematic Zoology 22:240 - 249.

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

Felsenstein, J. 1988. Phylogenies from molecular sequences: inferences and reliability. Annual Review of Genetics 22:521-565.

Felsenstein, J. 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, Massachusetts, 664 p.

Fisher, D. C. 1994. Stratocladistics: Morphological and Temporal Patterns and Their Relation to Phylogenetic Process, p. 133 - 171. *In* L. Grande and O. Rieppel (eds.), Interpreting the Hierarchy of Nature - from Systematic Patterns to Evolutionary Process Theories. Academic Press, Orlando, Florida.

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

Fisher, R. A. 1930. The Genetical Theory of Natural Selection. Clarendon Press, Oxford, 272.

Foote, M. 1996a. Models of Morphologic Diversification, p. 62 - 86. *In* D. Jablonski, D. H. Erwin, and J. H. Lipps (eds.), Evolutionary Paleobiology: Essays in Honor of James W. Valentine. University of Chicago Press, Chicago.

Foote, M. 1996b. On the probability of ancestors in the fossil record. Paleobiology 22:141 - 151.

Foote, M. 1996c. Perspective: evolutionary patterns in the fossil record. Evolution 50:1 - 11.

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

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

Foote, M., and D. M. Raup. 1996. Fossil preservation and the stratigraphic ranges of taxa. Paleobiology 22:121 - 140.

Fröbisch, N. B., and R. R. Schoch. 2009. Testing the impact of miniaturization on phylogeny: Paleozoic dissorophoid amphibians. Systematic Biology 58:312 - 327.

Funk, D. J., and K. E. Omland. 2003. Species-level paraphyly and polyphyly: frequency, causes and consquences, with insights from animal mitochondrial DNA. Annual Review of Ecology, Evolution, and Systematics 34:397 - 423.

Futuyma, D. J. 1987. On the role of species in anagenesis. The American Naturalist 130:465 - 473.

Galtier, N., O. Gascuel, and A. Jean-Marie. 2005. Markov Models in Molecular Evolution, p. 3-24. *In* R. Nielsen (ed.), Statististical Methods in Molecular Evolution. Springer, New York.

Gatesy, J. E. 2007. A tenth crucial question regarding model use in phylogenetics. Trends in Ecology and Evolution 22:509.

Gingerich, P. D. 1976. Paleontology and phylogeny: patterns of evolution at the species level in early Tertiary mammals. American Journal of Science 276:1-28.

Glenner, H., A. J. Hansen, M. V. Sørensen, F. Ronquist, J. P. Huelsenbeck, and E. Willerslev. 2004. Bayesian inference of metazoan phylogeny: a combined molecular and morphological approach. Current Biology 14:1644-1649.

Goldberg, E. E., and B. Igić. 2008. On phylogenetic tests of irreversible evolution. Evolution 62:2727-2741.

Goldman, N. 1990. Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analysis. Systematic Zoology 39:345 - 361.

Haldane, J. B. S. 1949. Suggestions as to quantitative measurement of rates of evolution. Evolution 3:51 - 56.

Harcourt-Brown, K. G., P. N. Pearson, and M. Wilkinson. 2001. The imbalance of paleontological trees. Paleobiology 27:188 - 204.

Heard, S. B. 1996. Patterns in phylogenetic tree balance with variable and evolving speciation rates. Evolution 50:2141 - 2148.

Hillis, D. M. 1995. Approaches for assessing phylogenetic accuracy. Systematic Biology 44:3-16.

Huelsenbeck, J. P. 1995. Performance of Phylogenetic Methods in Simulation. Systematic Biology 44:17-48.

Huelsenbeck, J. P., D. M. Hillis, and R. Jones. 1996. Parametric bootstrapping in molecular phylogenetics: applications and performance, p. 19 - 45. *In* J. D. Ferraris and S. R. Palumbi (eds.), Molecular Zoology: Advances, Strategies and Protocols. Wiley - Liss, New York.

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

Huelsenbeck, J. P., and B. Rannala. 1997a. Maximum likelihood estimation of topology and node times using stratigraphic data. Paleobiology 23:174 - 180.

Huelsenbeck, J. P., and B. Rannala. 1997b. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science 276:227 - 232.

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.

Huelsenbeck, J. P., and M. A. Suchard. 2007. A nonparametric method for accommodating and testing across-site rate variation. Systematic Biology 56:975 - 987.

Kelchner, S. A., and M. A. Thomas. 2007. Model use in phyogenetics: nine key questions. Trends in Ecology and Evolution 22:88-94.

Kim, J., and M. J. Sanderson. 2008. Penalized Likelihood Phylogenetic Inference: bridging the parsimony-likelihood gap. Systematic Biology 57:665 - 674.

Lane, A., C. M. Janis, and J. J. Sepkoski, Jr. 2005. Estimating paleodiversities: a test of the taxic and phylogenetic methods. Paleobiology 31:21-34.

Lewis, P. O. 2001. Maximum likelihood phylogenetic inference: modeling discrete morphological characters. Systematic Biology 50:913 - 925.

Malmgren, B. A., W. A. Berggren, and G. P. Lohmann. 1983. Evidence for punctuated gradualism in the Late Neogene *Globorotalia tumida* lineage of planktonic foraminifera. Paleobiology 9:319-332.

McShea, D. W. 1994. Mechanisms of large-scale evolutionary trends. Evolution 48:1747 - 1763.

Mooers, A. Ø., and S. B. Heard. 1997. Inferring evolutionary processes from phylogenetic tree shape. Quarterly Review of Biology 72:31 – 54.

Müller, J., and R. R. Reisz. 2006. The phylogeny of early eureptiles: comparing parsimony and Bayesian approaches in the investigation of a basal fossil clade. Systematic Biology 55:503 - 511.

Nee, S., R. M. May, and P. H. Harvey. 1994. The reconstructed evolutionary process. Philosophical Transactions of the Royal Society of London, series B 344:305-311.

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 293:407-417.

Pagel, M. D. 1997. Inferring evolutionary processes from phylogenies. Zoologica Scripta 26:331 - 348.

Pagel, M. D. 1999. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. Systematic Biology 48:612 - 622.

Pollitt, J. R., R. A. Fortey, and M. A. Wills. 2005. Systematics of the trilobite families Lichidae Hawle & Corda, 1847 and Lichakephalidae Tripp, 1957: the application of Bayesian inference to morphological data. Journal of Systematic Palaeontology 3:225 - 241.

Rannala, B., and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. Journal of Molecular Evolution 43:304 - 311.

Raup, D. M. 1985. Mathematical models of cladogenesis. Paleobiology 11:42 - 52.

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

Sanderson, M. J. 1993. Reversibility in evolution: a maximum likelihood approach to character gain/loss bias in phylogenies. Evolution 47:236-252.

Sanderson, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Molecular Biology and Evolution 19:101-109.

Signor, P. W., III and J. H. Lipps. 1982. Sampling bias, gradual extinction patterns and catastrophies in the fossil record. Geological Society of America Special Paper 190:291-296.

Smith, A. B. 1988. Patterns of diversification and extinction in early Palaeozoic echinoderms. Palaeontology 31:799-828.

Smith, A. B. 1994. Systematics and the Fossil Record - Documenting Evolutionary Patterns. Blackwell Scientific Publications, Oxford, 223 p.

Snively, E., A. P. Russell, and G. L. Powell. 2004. Evolutionary morphology of the coelurosaurian arctometatarsus: descriptive, morphometric and phylogenetic approaches. Zoological Journal of the Linnean Society 142:525-553.

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

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference, p. 407-514. *In* D. M. Hillis, C. Moritz, and B. K. Mable (eds.), Molecular Systematics. Sinauer Associates, Sunderland, MA.

THORNE, J. L., H. KISHINO, AND I. S. PAINTER. 1998. Estimating the rate of evolution of the rate of molecular evolution. Molecular Biology and Evolution 15:1647-1657.

WAGNER, P. J. 1995. Stratigraphic tests of cladistic hypotheses. Paleobiology 21:153 - 178.

WAGNER, P. J. 1997. Patterns of morphologic diversification among the Rostroconchia. Paleobiology 23:115 - 150.

WAGNER, P. J. 2000a. Likelihood tests of hypothesized durations: determining and accommodating biasing factors. Paleobiology 26:431 - 449.

WAGNER, P. J. 2000b. Phylogenetic analyses and the fossil record: tests and inferences, hypotheses and models, p. 341 - 371. In D. H. Erwin and S. L. Wing (eds.), Paleobiology Memoir 26 (Suppl. to No. 4)

WAGNER, P. J. 2000c. The quality of the fossil record and the accuracy of phylogenetic inferences about sampling and diversity. Systematic Biology 49:65 - 86.

WAGNER, P. J. 2001. Rate heterogeneity in shell character evolution among lophospiroid gastropods. Paleobiology 27:290 - 310.

WAGNER, P. J., AND D. H. ERWIN. 1995. Phylogenetic tests of speciation hypotheses, p. 87 - 122. *In* D. H. Erwin and R. L. Anstey (eds.), New approaches to studying speciation in the fossil record. Columbia University Press, New York.

WANG, S. C. 2001. Quantifying passive and driven large-scale evolutionary trends. Evolution 55:849 - 858.

WANG, S. C., AND P. J. EVERSON. 2007. Confidence intervals for pulsed mass extinction events. Paleobiology 33:324-336.

WIENS, J. J., R. M. BONETT, AND P. T. CHIPPINDALE. 2005. Ontogeny discombobulates phylogeny: paedomorphosis and higher-level salamander relationships. Systematic Biology 54:91-110.

WRIGHT, S. 1931. Evolution in Mendelian populations. Genetics 16:97 - 159.

YANG, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Molecular Biology and Evolution 10:1396 - 1401.

YANG, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular Evolution 39:306 - 314.

YANG, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends in Ecology and Evolution 11:367-372.

YANG, Z., AND B. RANNALA. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Molecular Biology and Evolution 14:717 - 724.

APPENDIX I: GLOSSARY

Symbol	Definition
Ψ	A tree topology (i.e., a cladogram) indicating recency of common ancestry
ψ	A phylogeny indicating recency of common ancestry and additional information about the types of relationships and time separating them
δ	The rate of character evolution along each branch of a phylogenetic tree
d	The date of a node
s	The sampled portion of a branch
t	The unsampled portion of a branch implied by the phylogenetic tree
k	The number of character states of a character
p_{ij}	The probability of beginning in state i and ending in state j after an interval of time
$\mathbf{P}$	A k x k matrix of p_{ij}
Γ	The gamma distribution
α	Shape parameter of the Γ distribution
β	The “rate” parameter of the Γ distribution
λ	The rate of taxonomic origination
μ	The rate of taxonomic extinction
m	The number of opportunities for change (e.g., speciation events under a punctuational model)
θ	The rate of preservation

APPENDIX II - THE EFFECT OF ANCESTOR-DESCENDANT HYPOTHESES ON LIKELIHOOD COMPLEXITY

Given any one cladistic topology Ψ, the likelihood functions for trees with only sister-taxon hypotheses (e.g., ψ_3 in the text) are more complex than are the likelihood functions for trees with explicit ancestor-descendant hypotheses (e.g., ψ_4 in the text). Combining Equation 2 from the text, the likelihood for phylogenies such as ψ_3 given a binary character is:

$$
\begin{aligned}
\mathrm{L}[\psi_3|0111] \propto\ & \boldsymbol{([p_{\mathrm{E00}}\times\{p_{\mathrm{A00}}\times p_{\mathrm{B01}}\}]\times[p_{\mathrm{F00}}\times\{p_{\mathrm{C01}}\times p_{\mathrm{D01}}\}])} + \\
& \boldsymbol{([p_{\mathrm{E01}}\times\{p_{\mathrm{A10}}\times p_{\mathrm{B11}}\}]\times[p_{\mathrm{F00}}\times\{p_{\mathrm{C01}}\times p_{\mathrm{D01}}\}])} + \\
& \boldsymbol{([p_{\mathrm{E00}}\times\{p_{\mathrm{A00}}\times p_{\mathrm{B01}}\}]\times[p_{\mathrm{F01}}\times\{p_{\mathrm{C11}}\times p_{\mathrm{D11}}\}])} + \\
& ([p_{\mathrm{E01}}\times\{p_{\mathrm{A10}}\times \boldsymbol{p_{\mathrm{B11}}}\}]\times[p_{\mathrm{F01}}\times\{\boldsymbol{p_{\mathrm{C11}}}\times p_{\mathrm{D11}}\}]) \\
& + \boldsymbol{([p_{\mathrm{E10}}\times\{p_{\mathrm{A00}}\times p_{\mathrm{B01}}\}]\times[p_{\mathrm{F10}}\times\{p_{\mathrm{C01}}\times p_{\mathrm{D01}}\}])} + \\
& \boldsymbol{([p_{\mathrm{E11}}\times\{p_{\mathrm{A10}}\times p_{\mathrm{B11}}\}]\times[p_{\mathrm{F10}}\times\{p_{\mathrm{C01}}\times p_{\mathrm{D01}}\}])} + \\
& \boldsymbol{([p_{\mathrm{E10}}\times\{p_{\mathrm{A00}}\times p_{\mathrm{B01}}\}]\times[p_{\mathrm{F11}}\times\{p_{\mathrm{C11}}\times p_{\mathrm{D11}}\}])} + \\
& ([p_{\mathrm{E11}}\times\{p_{\mathrm{A10}}\times \boldsymbol{p_{\mathrm{B11}}}\}]\times[p_{\mathrm{F11}}\times\{\boldsymbol{p_{\mathrm{C11}}}\times p_{\mathrm{D11}}\}])
\end{aligned}
$$

where p_{Xij} gives the probability of change from state i to state j for branch X (see Equation 1 in the main text). Note that the likelihood is proportional to this function because the complete function should include stasis within lineages (if it is observed), which will be a constant among all trees at any given rate of change.

Because the ancestors are unsampled, this equation allows ancestral taxa (E, F and G) to freely vary across either state 0 or state 1. However, phylogenies such as tree 4 (ψ_4) that posit sampled ancestors eliminate several terms (in bold above). Now, we assume that B=E and C=F. As we observe the states for B and C (in both cases, 1), $p_{\mathrm{E\bullet 0}}=p_{\mathrm{F\bullet 0}}=0$. All parts of the equation multiplied by either of these terms now are eliminated. Moreover, now $p_{\mathrm{B11}}=p_{\mathrm{C11}}=1.0$ because the phylogenies posit that it is known that these taxa began and ended with state 1. Those terms also can be removed. This leaves us with a much simpler equation:

$$
\begin{aligned}
\mathrm{L}[\psi_4|0111] \propto\ & ([p_{\mathrm{E01}}\times p_{\mathrm{A10}}]\times[p_{\mathrm{F01}}\times p_{\mathrm{D11}}]) \\
& + ([p_{\mathrm{E11}}\times p_{\mathrm{A10}}]\times[p_{\mathrm{F11}}\times p_{\mathrm{D11}}])
\end{aligned}
$$

Another way to think of this is that phylogenies such as ψ_3 unite eight possible evolutionary histories whereas phylogenies such as ψ_4 unite only two possible evolutionary histories.

A corollary of this is that the reduction on complexity increases as the number of states increases. Consider the same two trees and the same distribution of character states, but now assume that there is a third state that we do not observe among these four taxa. Now, our likelihood function with only sister-taxa becomes:

$$
\begin{aligned}
\mathrm{L}[\psi_3|0111] \propto\ & \boldsymbol{([p_{\mathrm{E00}}\times\{p_{\mathrm{A00}}\times p_{\mathrm{B01}}\}]\times[p_{\mathrm{F00}}\times\{p_{\mathrm{C01}}\times p_{\mathrm{D01}}\}])} + \\
& \boldsymbol{([p_{\mathrm{E01}}\times\{p_{\mathrm{A10}}\times p_{\mathrm{B11}}\}]\times[p_{\mathrm{F00}}\times\{p_{\mathrm{C01}}\times p_{\mathrm{D01}}\}])} + \\
& \boldsymbol{([p_{\mathrm{E02}}\times\{p_{\mathrm{A20}}\times p_{\mathrm{B21}}\}]\times[p_{\mathrm{F00}}\times\{p_{\mathrm{C01}}\times p_{\mathrm{D01}}\}])} + \\
& \boldsymbol{([p_{\mathrm{E00}}\times\{p_{\mathrm{A00}}\times p_{\mathrm{B01}}\}]\times[p_{\mathrm{F01}}\times\{p_{\mathrm{C11}}\times p_{\mathrm{D11}}\}])} + \\
& ([p_{\mathrm{E01}}\times\{p_{\mathrm{A10}}\times \boldsymbol{p_{\mathrm{B11}}}\}]\times[p_{\mathrm{F01}}\times\{\boldsymbol{p_{\mathrm{C11}}}\times p_{\mathrm{D11}}\}]) + \\
& \boldsymbol{([p_{\mathrm{E02}}\times\{p_{\mathrm{A20}}\times p_{\mathrm{B21}}\}]\times[p_{\mathrm{F01}}\times\{p_{\mathrm{C11}}\times p_{\mathrm{D11}}\}])} + \\
& \boldsymbol{([p_{\mathrm{E00}}\times\{p_{\mathrm{A00}}\times p_{\mathrm{B01}}\}]\times[p_{\mathrm{F02}}\times\{p_{\mathrm{C21}}\times p_{\mathrm{D21}}\}])} + \\
& \boldsymbol{([p_{\mathrm{E01}}\times\{p_{\mathrm{A10}}\times p_{\mathrm{B11}}\}]\times[p_{\mathrm{F02}}\times\{p_{\mathrm{C21}}\times p_{\mathrm{D21}}\}])} + \\
& \boldsymbol{([p_{\mathrm{E02}}\times\{p_{\mathrm{A20}}\times p_{\mathrm{B21}}\}]\times[p_{\mathrm{F02}}\times\{p_{\mathrm{C21}}\times p_{\mathrm{D21}}\}])} \\
& + \boldsymbol{([p_{\mathrm{E10}}\times\{p_{\mathrm{A00}}\times p_{\mathrm{B01}}\}]\times[p_{\mathrm{F10}}\times\{p_{\mathrm{C01}}\times p_{\mathrm{D01}}\}])} + \\
& \boldsymbol{([p_{\mathrm{E11}}\times\{p_{\mathrm{A10}}\times p_{\mathrm{B11}}\}]\times[p_{\mathrm{F10}}\times\{p_{\mathrm{C01}}\times p_{\mathrm{D01}}\}])} + \\
& \boldsymbol{([p_{\mathrm{E12}}\times\{p_{\mathrm{A20}}\times p_{\mathrm{B21}}\}]\times[p_{\mathrm{F10}}\times\{p_{\mathrm{C01}}\times p_{\mathrm{D01}}\}])} + \\
& \boldsymbol{([p_{\mathrm{E10}}\times\{p_{\mathrm{A00}}\times p_{\mathrm{B01}}\}]\times[p_{\mathrm{F11}}\times\{p_{\mathrm{C11}}\times p_{\mathrm{D11}}\}])} + \\
& ([p_{\mathrm{E11}}\times\{p_{\mathrm{A10}}\times \boldsymbol{p_{\mathrm{B11}}}\}]\times[p_{\mathrm{F11}}\times\{\boldsymbol{p_{\mathrm{C11}}}\times p_{\mathrm{D11}}\}]) + \\
& \boldsymbol{([p_{\mathrm{E12}}\times\{p_{\mathrm{A20}}\times p_{\mathrm{B21}}\}]\times[p_{\mathrm{F11}}\times\{p_{\mathrm{C11}}\times p_{\mathrm{D11}}\}])} + \\
& \boldsymbol{([p_{\mathrm{E10}}\times\{p_{\mathrm{A00}}\times p_{\mathrm{B01}}\}]\times[p_{\mathrm{F12}}\times\{p_{\mathrm{C21}}\times p_{\mathrm{D21}}\}])} + \\
& \boldsymbol{([p_{\mathrm{E11}}\times\{p_{\mathrm{A10}}\times p_{\mathrm{B11}}\}]\times[p_{\mathrm{F12}}\times\{p_{\mathrm{C21}}\times p_{\mathrm{D21}}\}])} + \\
& \boldsymbol{([p_{\mathrm{E12}}\times\{p_{\mathrm{A20}}\times p_{\mathrm{B21}}\}]\times[p_{\mathrm{F12}}\times\{p_{\mathrm{C21}}\times p_{\mathrm{D21}}\}])} \\
& + \boldsymbol{([p_{\mathrm{E20}}\times\{p_{\mathrm{A00}}\times p_{\mathrm{B01}}\}]\times[p_{\mathrm{F20}}\times\{p_{\mathrm{C01}}\times p_{\mathrm{D01}}\}])} + \\
& \boldsymbol{([p_{\mathrm{E21}}\times\{p_{\mathrm{A10}}\times p_{\mathrm{B11}}\}]\times[p_{\mathrm{F20}}\times\{p_{\mathrm{C01}}\times p_{\mathrm{D01}}\}])} + \\
& \boldsymbol{([p_{\mathrm{E22}}\times\{p_{\mathrm{A20}}\times p_{\mathrm{B21}}\}]\times[p_{\mathrm{F20}}\times\{p_{\mathrm{C01}}\times p_{\mathrm{D01}}\}])} + \\
& \boldsymbol{([p_{\mathrm{E20}}\times\{p_{\mathrm{A00}}\times p_{\mathrm{B01}}\}]\times[p_{\mathrm{F21}}\times\{p_{\mathrm{C11}}\times p_{\mathrm{D11}}\}])} + \\
& ([p_{\mathrm{E21}}\times\{p_{\mathrm{A10}}\times \boldsymbol{p_{\mathrm{B11}}}\}]\times[p_{\mathrm{F21}}\times\{p_{\mathrm{C11}}\times \boldsymbol{p_{\mathrm{D11}}}\}]) + \\
& \boldsymbol{([p_{\mathrm{E22}}\times\{p_{\mathrm{A20}}\times p_{\mathrm{B21}}\}]\times[p_{\mathrm{F21}}\times\{p_{\mathrm{C11}}\times p_{\mathrm{D11}}\}])} + \\
& \boldsymbol{([p_{\mathrm{E20}}\times\{p_{\mathrm{A00}}\times p_{\mathrm{B01}}\}]\times[p_{\mathrm{F22}}\times\{p_{\mathrm{C21}}\times p_{\mathrm{D21}}\}])} + \\
& \boldsymbol{([p_{\mathrm{E21}}\times\{p_{\mathrm{A10}}\times p_{\mathrm{B11}}\}]\times[p_{\mathrm{F22}}\times\{p_{\mathrm{C21}}\times p_{\mathrm{D21}}\}])} + \\
& \boldsymbol{([p_{\mathrm{E22}}\times\{p_{\mathrm{A20}}\times p_{\mathrm{B21}}\}]\times[p_{\mathrm{F22}}\times\{p_{\mathrm{C21}}\times p_{\mathrm{D21}}\}]).}
\end{aligned}
$$

In contrast, the likelihood function for trees with two ancestor-descendant hypotheses adds only one more term:

$$
\begin{aligned}
\mathrm{L}[\psi_4|0111] \propto\ & ([p_{\mathrm{E01}}\times p_{\mathrm{A10}}]\times[p_{\mathrm{F01}}\times p_{\mathrm{D11}}]) \\
& + ([p_{\mathrm{E11}}\times p_{\mathrm{A10}}]\times[p_{\mathrm{F11}}\times p_{\mathrm{D11}}]) \\
& + ([p_{\mathrm{E21}}\times p_{\mathrm{A10}}]\times[p_{\mathrm{F21}}\times p_{\mathrm{D11}}]).
\end{aligned}
$$