Topology-Bayes versus Clade-Bayes in Phylogenetic Analysis

Ward C. Wheeler* and Kurt M. Pickett†

*Division of Invertebrate Zoology, American Museum of Natural History, NY; and †Department of Biology, University of Vermont

Several features of currently used Bayesian methods in phylogenetic analysis are discussed. The distinction between Clade-Bayes and Topology-Bayes is presented and illustrated with an empirical example. Three problems with Bayesian phylogenetic methods—exaggerated clade support, inconsistently biased priors, and the impossibility of hypothesis testing of cladograms—are shown to be the result of using a Clade-based Bayesian approach. Topology-based Bayesian methods do not share these shortcomings.

Introduction

Bayesian methods in phylogenetics have a 30-year history, tracing back at least to Farris (1973) (discussed in Felsenstein 2004). Explicit attempts to use empirical priors and to create Bayesian estimators of topology were attempted as early as Wheeler (1991), and recent techniques and implementations have been reviewed by Huelsenbeck et al. (2001). The purpose of this discussion is to examine the application of recent Bayesian techniques to the problem of estimating cladistic relationships among terminal taxa—that is, their branching pattern.

As an optimality criterion to choose among candidate topologies, posterior probability is a fine option, but current use (as in Huelsenbeck and Ronquist 2003) does not follow this path. The entities whose posterior probability is most frequently estimated are clades. The set of clades with posterior probability >one-half is presented as the Bayesian hypothesis of phylogenetic topology. Indeed, many proponents of the methodology have asserted repeatedly that these 50% majority rule consensus topologies reflect the probability of the truth of clade identity (Huelsenbeck et al. 2002), and hundreds of authors employing the methods have repeated this assertion as a justification for the methodological choice. Such an approach differs from that of examining the relative merits of alternative topologies directly, resulting in 3 problems: 1) exaggerated clade support, 2) inconsistently biased priors, and 3) the impossibility of topology hypothesis testing. Here, we discuss these issues and show that these problems with Bayesian phylogenetics are not inherent to Bayesianism per se but to the particular path that has been taken by the majority of the community. We show that the adoption of the Bayesian optimality position—supported by Rannala and Yang (1996), though not adopted by most practitioners—abrogates these problems.

Topology-Bayes and Clade-Bayes

Topology-Bayes

A Topology-Bayesian estimator of a phylogenetic hypothesis is that topology (T_t) which has the maximum posterior probability (assuming all other topologies have equal cost of error or "loss"). Here, we describe a tree T as composed of a set of vertices V and edges E:

$$T = (V, E). \qquad (1)$$

Bayes' theorem allows for the calculation of the probability of the hypothesis—here, the topology—given the data, $\mathrm{pr}(T_i|D)$. This is desirable as phylogeneticists begin with data and usually seek the best estimate of the topology. If investigating the posterior probability of a topology $\mathrm{pr}(T_i|D)$, given topological "prior" probabilities $pr(T_i)$, Bayes' theorem can be written:

$$\mathrm{pr}(T_i|D) = \frac{\mathrm{pr}(T_i) \cdot \mathrm{pr}(D|T_i)}{\sum_{j \in T} \mathrm{pr}(T_j) \cdot \mathrm{pr}(D|T_j)}. \qquad (2)$$

Because the denominator of the right side of equation (2) is a constant, the Topology-Bayesian estimator will maximize $\mathrm{pr}(T_i) \cdot \mathrm{pr}(D|T_i)$, the product of the prior probability of the topology and the probability of the data, given the topology. Therefore, if the topological priors are uniform, the Topology-Bayesian estimator will be proportional to the maximum likelihood of the topology given the data (i.e., $\mathrm{pr}(T|D) \propto L(T|D)$) (Edwards 1992). In other words, the ordering of topologies from most to least optimal is the same for both the likelihood and the posterior probability, even if the absolute values differ. Thus, the maximization of the likelihood of the topology also maximizes the posterior probability (see eq. 2). When we say maximum likelihood, here we mean the maximum integrated likelihood; we use this term in the usual sense to mean that if the topology and the so-called nuisance parameters are generated from a prior distribution, they can be integrated out, thus guaranteeing that the maximum integrated likelihood topology is equivalent to the posterior probability topology (see Steel and Penny 2000). Although different from the more typical maximum relative likelihood, in which nuisance parameters are adjusted to maximize the likelihood of the topology, relative likelihood is merely an approximation of integrated likelihood, and Markov Chain Monte Carlo (MCMC) methods permit the calculation (albeit, heuristic) of integrated likelihood directly. This equivalence of the relative ordering of integrated likelihood and posterior probability topologies, however, is only guaranteed if the priors for the hypothesis under consideration are uniform. If uniform topological priors are used, and the optimal topology is selected as the best estimate of the phylogeny, as explained above, the most likely topology is guaranteed to be the tree of maximum posterior probability. If, however, this topology is not the hypothesis of interest, but the frequency of clades across a group of trees is the goal, the guaranteed proportionality is lost. Much of the argumentation in favor of Bayesian MCMC methods has asserted that results will be proportional to the likelihood

(e.g., Huelsenbeck et al. 2001); as we show here, this is only true if the Topology-Bayesian approach is adopted.

Clade-Bayes

The Clade-Bayesian estimator is that set $V_c = \{V_{c,1}, V_{c,2}, ...\}$ of clades that have a posterior probability >one-half. Because any pair of clades from V_c cannot conflict (because, by definition, there is a positive probability that some tree contains both clades), the set V_c forms a tree, which we will denote T_c. This is the Clade-Bayes topology. Unlike T_t, T_c has no associated optimality value even though each clade in T_c does. Most current Bayesian phylogenetic analyses (e.g., MrBayes) produce T_c (Huelsenbeck and Ronquist 2003). Consider the posterior probability clade on a topology, given data:

$$\operatorname{pr}(V_{c,i}|D) = \frac{\operatorname{pr}(V_{c,i}) \cdot \operatorname{pr}(D|V_{c,i})}{\sum_{j \in V_c} \operatorname{pr}(V_{c,j}) \cdot \operatorname{pr}(D|V_{c,j})}. \qquad (3)$$

Although equation (3) represents the posterior probability for a single clade on a given tree, this is not how clade probabilities are determined in popular implementations. Instead, a clade is assigned the probability of the tree to which it belongs. As tree probability is merely the frequency of a given tree across the retained visited trees, clade probability is the cumulative frequency of a clade of given membership across all trees (from the retained pool) that contain it (regardless of the taxic rearrangements inside the clade, and regardless of the position the clade takes in the trees). This, however, causes some undesirable effects. If uniform topological priors are stipulated as a means of representing the prior distribution on clades, that clade prior distribution V_c is not uniform (Pickett and Randle 2005). Indeed, some clades will be profoundly less probably than others, and it is impossible to assign a label-invariant (i.e., standard) distributional topological prior, uniform or otherwise, that will induce uniformity on V_c (Steel and Pickett 2006). As such, even if $\operatorname{pr}(D|V_{c,i})$ is maximal (of all clades, e.g.), because that value would be multiplied by its prior probability (which may be quite low under uniform topological priors), it may not correspond to the clade of maximum posterior probability. Simply put, the proportionality and relative ordering of likelihood and posterior probability are no longer guaranteed. In any given case, depending on the data and the particular prior employed, the clade of maximum likelihood may also be the clade of maximum posterior probability, but there is no requirement that this be so.

Given that T_t and T_c are different entities, they need not agree. In such a case, T_t and T_c would have clades in conflict and T_c would not represent the Topology-Bayes estimator or the maximum a posteriori probability (MAP) estimate of Rannala and Yang (1996). This could occur, for example, if several suboptimal (i.e., lower posterior probability) topologies share a group not found in the "best" tree. Below, we provide an example.

Example Comparison

The fact that topologies of maximum likelihood need not correspond to the Clade-Bayes tree has been considered recently (Svennblad et al. 2006). But the fact that the Topology-Bayes approach and the Clade-Bayes approach can result in topological differences has received almost no attention. Here, we demonstrate this potential disparity using empirical data.

Consider a case of arthropod morphological data with 54 taxa and 303 characters (Giribet et al. 2001). If we begin with uniform topological prior probabilities, then the T_t will be the maximum likelihood topology. Any model will suffice; here we use the no common mechansim (NCM) model of Tuffley and Steel (1997). Because MrBayes (Huelsenbeck and Ronquist 2003) does not seek the optimal trees (see below), we calculated T_t (the Topology-Bayes or MAP estimate) using POY3 (Wheeler et al. 1996–2005), which searches and saves all identified optimal solutions, given a criterion. The 7 characters treated as additive in the original analysis of Giribet et al. (2001) are treated as nonadditive here. T_c was estimated using MrBayes (Huelsenbeck and Ronquist 2003), implementing the same model of evolution. Searches were performed with the following options:

POY3: buildsperreplicate 10, replicates 10, treefuse, likelihood, likelihoodroundingmultiplier 10,000. This implements 10 random replicates with 10 Wagner builds per replicate followed by tree-bisection-and-regrafting (TBR) branch-swapping and tree fusing (Goloboff 1999) within and between replicates; the NCM (Tuffley and Steel 1997) model of evolution was employed, rounding likelihoods to 5 decimal places.

MrBayes: lset parsmodel = yes mcmc ngen = 10,000,000 samplefreq = 500 nchains = 4. This implements the NCM (Tuffley and Steel 1997) model of evolution with 2 simultaneous runs of 4 chains and 10,000,000 generations each, saving every 500th topology visited to file.

Ten cladograms of −log likelihood 773.02908 were found by POY3. Their strict consensus (T_t) is shown in figure 1a. The T_c produced by MrBayes is shown in figure 1b (−log likelihood of the best binary resolution of this consensus cladogram is 773.38128). Overall, the 2 trees are quite similar. They differ, however, in several major taxonomic groupings. Foremost among these is that the Chelicerata are monophyletic in T_c and paraphyletic (pycnogonids basal) in T_t.

The 2 topologies are similar, but not identical. Implementation issues aside, this example clearly demonstrates a case where $T_t \neq T_c$.

It is worth noting that the maximum posterior probability topology from MrBayes has a −log likelihood of 773.03, which differs from the score reported by POY3 (773.02908) only in rounding. MrBayes reported only 2 trees of this score; it may have visited the other 8 topologies of highest probability but because the program is designed to create T_c, it does not seek to save all optimal topologies. In fact, one of the 2 optimal topologies visited by MrBayes was found before the end of the burn-in in our second run (tree 1905000) and so would not even be included in a majority rule calculation of the postburn-in topologies. However, there would be no burn-in period and no reason to abandon any of the visited topologies, when searching for the optimal solution.
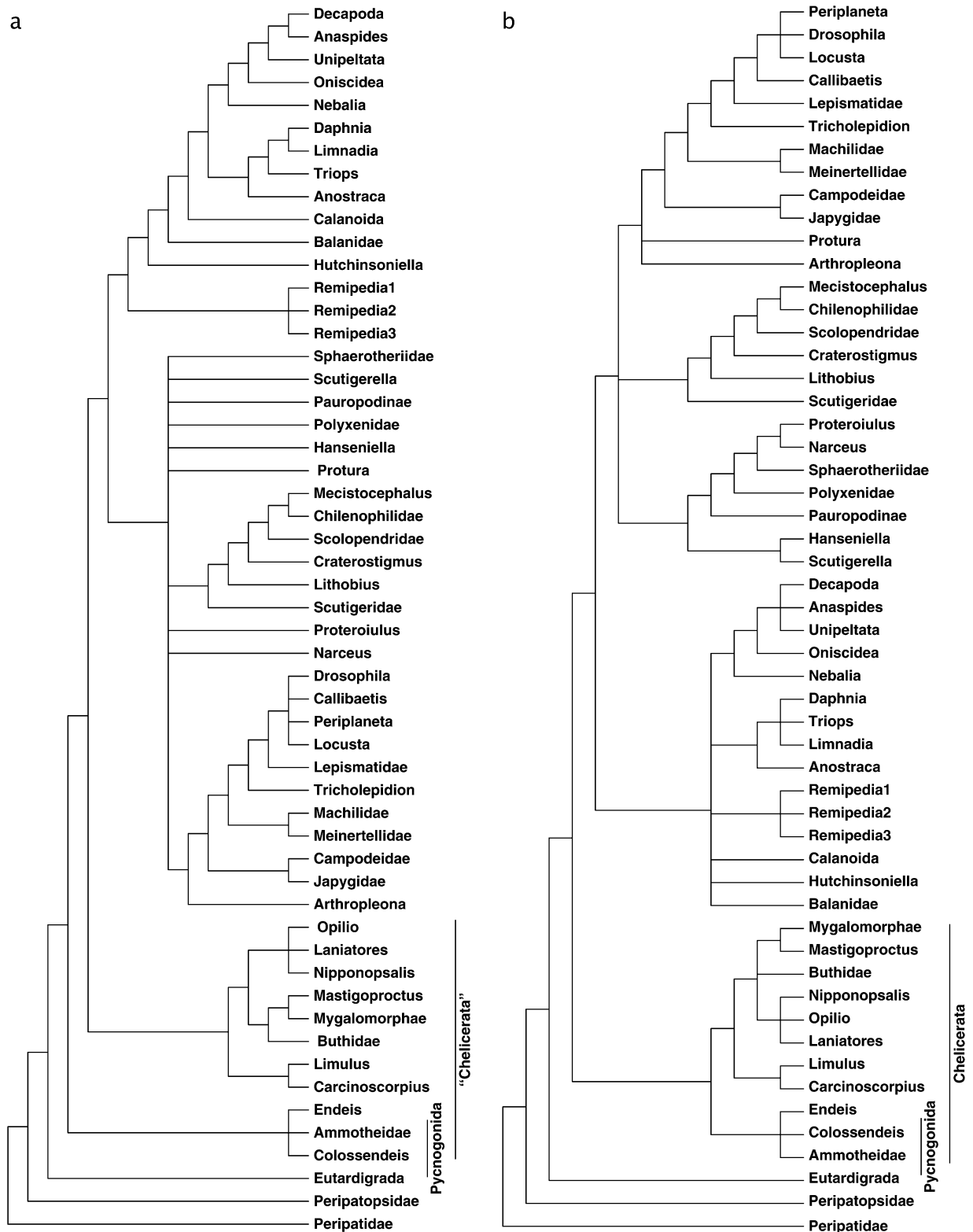
Fig. 1.—Bayesian cladograms of arthropod data with chelicerate and pycnogonid taxa marked: (*a*) T_t with –log likelihood of 773.02908, (*b*) T_c with –log likelihood of the best binary resolution 773.38128

Table 1
Simple Data Set of 4 Taxa (A–D) and 4 Binary Characters

Taxon		Characters		
D	0	0	0	1
C	0	0	1	0
B	1	1	0	0
A	1	1	1	1

Doubling Behavior: Clade-Bayes and Support

One of the properties of all statistical methods is an increase in levels of support with a multiplication of identically distributed data. In other words, larger data sets with the same proportional balance of data in favor of, in opposition to, and indifferent to a hypothesis will assign higher support. Consider the data of table 1. Under NCM (Penny et al. 1994; Tuffley and Steel 1997, p. 587. eqs. 12 and 13), the maximum posterior probability cladogram is that of figure 2 at 0.5 (eq. 4).

$$\text{pr}\big(T_{((A,B)(C,D))}\big) = \frac{\frac{1}{3} \cdot 2^{-10}}{\left(\frac{1}{3} \cdot 2^{-10}\right) + \left(\frac{1}{3} \cdot 2^{-11}\right) + \left(\frac{1}{3} \cdot 2^{-11}\right)} = \frac{1}{2}.$$

(4)

In this case, $T_c = T_t$ and the posterior probability of group A + B ($V_{(AB)}$) is equal to 0.5. The evidence in favor of AB is equal to that against. If we were to replicate these data n times, with exactly the same balance of evidence for and against AB, $\text{pr}(V_{AB})$ would increase with n until arbitrarily close to 1 (eq. 5).

$$\text{pr}(V_{AB}) = \frac{1}{1 + 2^{1-n}}.$$

(5)

This same behavior will be observed in a bootstrap, jackknife analysis or other resampling approach to support. Similarly, the same behavior would be observed for any data, contrived or otherwise, that are duplicated identically, regardless of the model employed; our example is presented to permit exact calculations. Although reasonable on statistical grounds, even when support levels are very close to unity, there are still as many observations contradicting the grouping as supporting. The point here is that this behavior is an ineluctable outcome of data duplication. It is important to note that the "inflation" of support values described here is not related to the higher values seen in Bayesian over bootstrap or jackknife support (Cummings et al. 2003; Erixon et al. 2003; Simmons et al. 2004), reported from T_c. It is also worth noting that if the T_t approach is adopted, the support inflation seen in T_c becomes moot.

Priors: Uniform, Biased, and Empirical

A central issue surrounding Bayesian techniques is the choice of appropriate prior probabilities. For the purposes here, we can divide these into 3 types: uniform, biased, and empirical. Uniform or ignorance priors are usually employed when there is no useful preexisting information on the entity to be estimated. Phylogenetic analysis tends
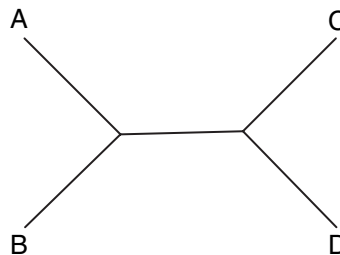

FIG. 2.—Maximum posterior probability (0.5) cladogram of data from table 1 under the NCM (Tuffley and Steel, 1997) model of substitution. The 2 alternate cladograms have posterior probabilities of 0.25.

to adhere to uniform topological priors, and discussions tend to rely on current evidence to draw conclusions in the admirable feeling that the investigator cannot say which hypotheses are more probable a priori. Biased priors attach greater or lesser initial probability to entities based on non-uniform distributions. In many cases of Bayesian analysis, this is not undesirable. It may be well known that processes follow certain distributions and profitable use can be made of them. In phylogenetic analysis, they are less well regarded because they are not based on biological information. Empirical priors are based, unsurprisingly, on previous experience, data, and knowledge. In general, empirical priors are unobjectionable because they allow the use of previous empirical results. Some strict Bayesians object to the use of data as priors when the data themselves are not, in fact, temporally and ontologically prior to the other data under consideration. This objection derives from the fundamental difference with frequentist statistics: the Bayesian view that prior events predict future events. We note this strict interpretation here only to point out that as phylogeneticists begin using empirical data as prior assertions, this philosophical objection may loom, especially if the prior data did not occur first, even if they were observed first and thus formed the prior phylogenetic viewpoint (as would be the case, e.g., if morphology was used as priors for molecular data; morphology is not temporally antecedent to DNA and thus would violate this strict Bayesian interpretation of appropriate prior data).

Priors are attached to the entities to be estimated and hence play different roles for T_t and T_c. As mentioned above, phylogenetic Bayesian methods generally employ uniform priors; the most commonly used is the "proportional-to-distinguishable-arrangements" distribution (Rosen 1978). This is straightforward with T_t because each topology (as in eq. 3) can be assigned the inverse of number of topologies. This cannot be done for T_c.

Steel and Pickett (2006) have shown that uniform priors cannot be constructed for clades ($V_{c,i}$). In short, clades of size 2 or $n - 2$ (for n taxa) will have higher probabilities than those of intermediate size. This can result in huge prior disparities (orders of magnitude) among clades (Pickett and Randle 2005). In absence of any data, some groups will be favored over others. This is not a problem that occurs with Topology-Bayesian analysis (T_t). The problem only arises when topological priors (and their resultant clade priors) are used to estimate clade posteriors, as in Clade-Bayesian (T_c) analyses.
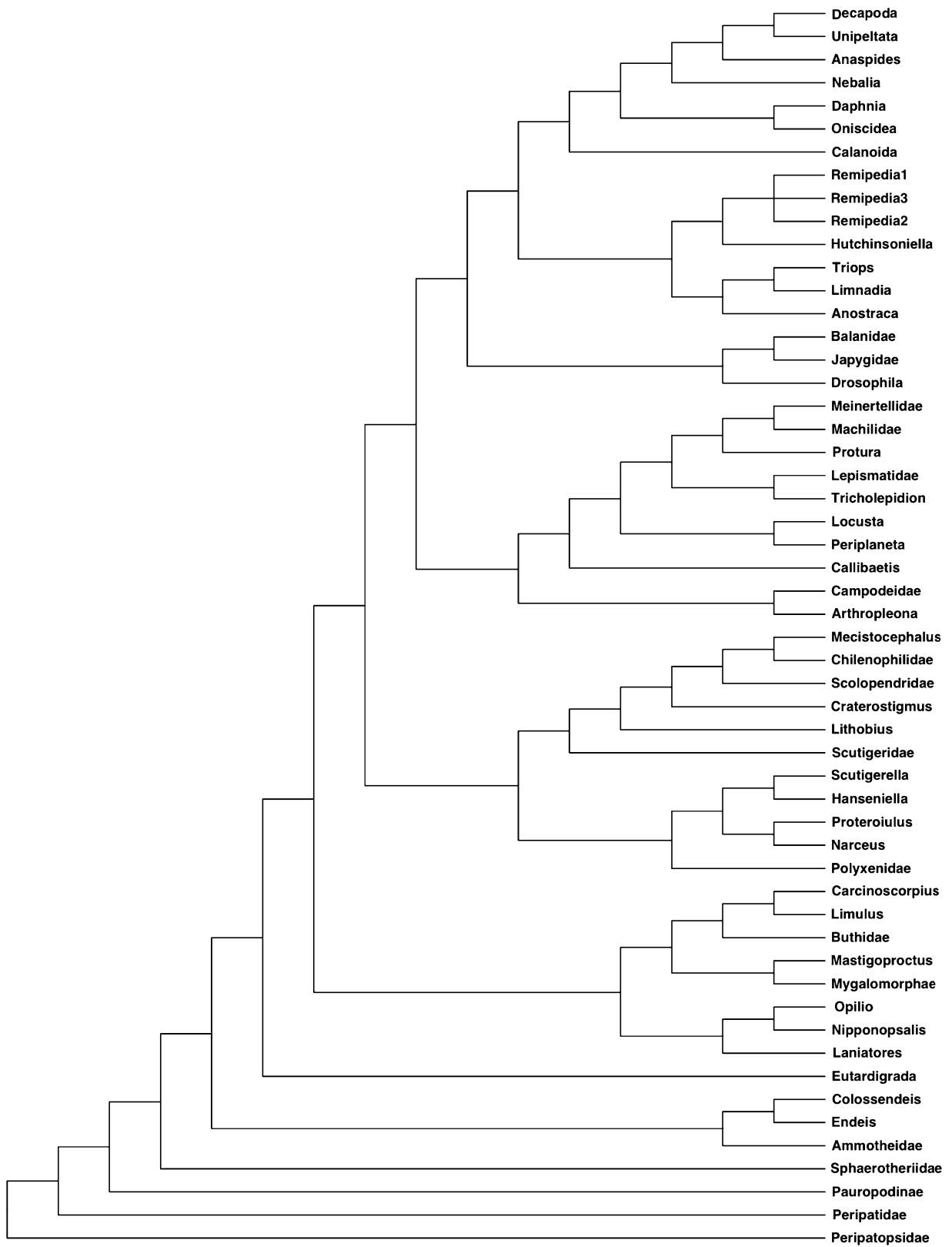
FIG. 3.—Arthropod cladogram with maximum posterior probability using morphologically estimated prior probabilities (under NCM) and molecular likelihood values calculated via the methods in Wheeler (2006).

Empirical priors on topologies are an underexplored area. Wheeler (1991) tried to do this using morphological data to create priors that were then combined with likelihoods based on molecular data. Using an explicit model of evolution, we can employ this approach using the probability of a topology given a set of morphological data to approximate an empirical prior (eq. 6). Of course, any model of evolution might be invoked for the data that yield the topology prior. Here, we employ NCM.

$$\mathrm{pr}(T_i) = \frac{\mathrm{pr}\left(D_{\mathrm{morph}}\middle|T_i\right)}{\sum_{j \in T} \mathrm{pr}\left(D_{\mathrm{morph}}\middle|T_j\right)}. \qquad (6)$$

There is no reason why such empirically derived priors could not be constructed using a variety of data sources, provided that they are applied over a distribution of topologies.

Arthropod Example of Empirical Priors

If we use the estimation of empirical priors of equation (6), we can extend this with likelihoods based on molecular data (D_{mol}) for the same taxa to:

$$\mathrm{pr}(T_i|D_{\mathrm{mol}}) = \frac{\mathrm{pr}\left(D_{\mathrm{morph}}\middle|T_i\right) \cdot \mathrm{pr}\left(D_{\mathrm{mol}}\middle|T_i\right)}{\sum_{j \in T} \mathrm{pr}\left(D_{\mathrm{morph}}\middle|T_j\right) \cdot \mathrm{pr}\left(D_{\mathrm{mol}}\middle|T_j\right)}. \qquad (7)$$

To maximize $\mathrm{pr}(T_i|D_{\mathrm{mol}})$, we need only maximize the numerator of equation (7) (the denominator is again a constant). As $L(T|D)$ is proportional to $\mathrm{pr}(D|T)$, the topology with greatest posterior probability will be that which maximizes:

$$L\left(T_i\middle|D_{\mathrm{morph}}\right) \cdot L\left(T_i\middle|D_{\mathrm{mol}}\right). \qquad (8)$$

The value of equation (7) can be optimized via a search over topologies by standard techniques. Wheeler (2006) performed a likelihood-based analysis of combined arthropod data (including the morphology cited above), and the result that maximized the value of equation (8) is shown in figure 3. This cladogram is based on the data of Giribet et al. (2001) with 54 taxa, 303 morphological characters, and 8 molecular loci. A single addition sequence + TBR was used in this illustration. The S6GF5 model of Wheeler (2006) was applied (general time reversible + gaps) using direct optimization Wheeler (1996) on unaligned sequences with 2 Γ classes and invariant sites (θ) parameters estimated from the sequence data. Morphological data analyzed as above. The total $-$log likelihood reported by POY3 was 115,417.38.

Hypothesis Testing

The central act of phylogenetic analysis is establishing the relative merits of 2 hypotheses. This can be done on a variety of grounds (= optimality criteria) and as long as the comparisons are transitive, a best solution can be found. Parsimony, likelihood, and Topology-Bayes all do this by default. In each case, an optimality value (cost, likelihood, and posterior probability) is assigned to each cladogram and reported by the investigator. This value is used to compare and test pairs of hypotheses. No such comparison can made for Clade-Bayes (T_c) trees in the form that

they are usually reported. Although it is true that any tree-shaped object upon which characters are plotted can be assigned an optimality score (and thus, the posterior probability of any given Clade-Bayes tree could be calculated as a Topology-Bayes hypothesis, but therefore abandoning the Clade-Bayes approach), investigators rarely, if ever, calculate or report such optimality scores. As such, few, if any of the reported Clade-Bayes trees from the literature can be compared with subsequent analyses of the same data, which may give different results. Thus, it is impossible to say that any Clade-Bayes topology is superior or inferior to any other, unless subsequent investigators compute the optimality score of the Clade-Bayes trees as topologies. As with jackknife or bootstrap trees, the strength of clade support is presented by the investigator, but not the cladogram optimality. It is also worth noting that any tree that is less resolved than the optimal binary tree—whether Clade-Bayes, jackknife, bootstrap, or other consensus tree—is most likely less optimal (in this case lower posterior probability, but strictly they could be equal) because the consensus is most likely due to character conflict. Hence, Clade-Bayes trees may perhaps be regarded as statements of support but not as best-supported scientific hypotheses of phylogenetic relationships.

Bayesian methods can have a place in systematic analysis, but this position must be based on the relative quality of topologies, not their constituent parts. This requires the use of the Topology-Bayes approach advocated here.

Literature Cited

Cummings MP, Handley SA, Myers DS, Reed DL, Rokas A, Winka K. 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Syst Biol. 52:477–487.

Edwards A. 1992. Likelihood. Baltimore (MD): Johns Hopkins University Press.

Erixon P, Svennblad B, Britton T, Oxelman B. 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. systematic biology. Syst Biol. 52:665–673.

Farris JS. 1973. A probability model for inferring evolutionary trees. Syst Zool. 22:250–256.

Felsenstein J. 2004. Phylip (phylogeny inference package) version 3.6. Distributed by the author. Seattle (WA): Department of Genome Sciences, University of Washington.

Giribet G, Edgecombe GD, Wheeler WC. 2001. Arthropod phylogeny based on eight molecular loci and morphology. Nature. 413:157–161.

Goloboff P. 1999. Analyzing large data sets in reasonable times: solutions for composite optima. Cladistics. 15:415–428.

Huelsenbeck JP, Larget B, Miller RE, Ronquist F. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst Biol. 51:673–688.

Huelsenbeck JP, Ronquist F. 2003. MrBayes: Bayesian inference of phylogeny. 3.0 ed [Internet]. documentation. http://mrbayes.csit.fsu.edu/.

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

Penny D, Lockhart PJ, Steel MA, Hendy MD. 1994. The role of models in reconstructing evolutionary trees. In: Scotland RW, Siebert DJ, Williams DM, editors. Models in phylogeny reconstruction. Vol. 52 of Systematics Association. Oxford: Clarendon Press. p. 211–230.

Pickett KM, Randle CP. 2005. Strange bayes indeed: uniform topological priors. Mol Phylogenet Evol. 34:203–211.

Rannala B, Yang Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol. 43:304–311.

Rosen DE. 1978. Vicariant patterns and historical explanation in biogeography. Syst Zool. 27:159–188.

Simmons M, Pickett K, Miya M. 2004. How meaningful are Bayesian support values? Mol Biol Evol. 21:188–199.

Steel M, Penny D. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol Biol Evol. 17: 839–850.

Steel M, Pickett KM. 2006. On the imposibility of uniform priors on clades. Mol Phylogenet Evol. 39:585–586.

Svennblad B, Erixon P, Oxelman B, Britton T. 2006. Fundamental differences between the methods of maximum likelihood and maximum posterior probability in phylogenetics. Syst Biol. 55:116–121.

Tuffley C, Steel M. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull Math Biol. 59:581–607.

Wheeler WC. 1991. Congruence among data sets: a Bayesian approach. In: Miyamoto MM, Cracraft J, editors. Phylogenetic analysis of DNA sequences. London: Oxford University Press. p. 334–346.

Wheeler WC. 1996. Optimization alignment: the end of multiple sequence alignment in phylogenetics? Cladistics. 12:1–9.

Wheeler WC. 2006. Dynamic homology and the likelihood criterion. Cladistics. 22:157–170.

Wheeler WC, Gladstein DS, De Laet J. 1996–2005. POY version 3.0 [Internet]. New York: American Museum of Natural History. Available from: http://research.amnh.org/scicomp/projects/poy.php documentation (current version 3.0.11). documentation by D. Janies and W.C. Wheeler. commandline documentation by J. De Laet and W.C. Wheeler.