Syst. Biol. 57(5):814–821, 2008

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150802422308

A Justification for Reporting the Majority-Rule Consensus Tree in Bayesian Phylogenetics

MARK T. HOLDER,[1] JEET SUKUMARAN,[1] AND PAUL O. LEWIS[2]

[1]*Department of Ecology and Evolutionary Biology, University of Kansas, 1200 Sunnyside Avenue, Lawrence, Kansas 66045, USA; E-mail: mtholder@ku.edu (M.T.H.)*
[2]*Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Road, Unit 3043, Storrs, Connecticut 06269-3043, USA*

Systematists must frequently deal with substantial uncertainty in their phylogenetic estimates. Nonparametric bootstrapping (Felsenstein, 1985) and Markov chain Monte Carlo (MCMC) simulations used for Bayesian phylogenetic inference (Mau et al., 1999; Larget and Simon, 1999; Huelsenbeck and Ronquist, 2001) are two of the most popular computational approaches for assessing support for different parts of a phylogenetic tree. Both of these techniques produce large collections of trees. A majority-rule consensus tree is often used to summarize such a collection of trees. As has been discussed (e.g., in Barrett et al., 1991, and the ensuing debate), a consensus tree is a summary of a set of trees, and not necessarily an optimal estimator of the phylogeny.

Here we present a context in which the majority-rule consensus tree of samples from the posterior probability distribution over trees *can* be viewed as the optimal tree to report. We explicitly rephrase phylogenetic inference as the problem of "what tree should I publish for this group of taxa, given my data?" The majority-rule consensus tree can be shown to be the optimal tree to report if we view the cost of reporting an estimate of the phylogeny to be a linear function of the number of incorrect clades in the estimate and the number of true clades that are missing from the estimate and we view the reporting of an incorrect grouping as a more serious error than missing a clade.

The work of Berry and Gascuel (1996) on reporting results from nonparametric bootstrapping overlaps significantly with the results presented here. Berry and Gascuel (1996) present arguments from Bayesian decision theory, which is also the theoretical basis of our work. Berry and Gascuel focus on frequentist properties of estimators (type I and type II error rates) and interpret bootstrapping proportions as measures of the probability of a clade being present. In order to apply these decision rules to bootstrapping analyses, they study the correlation between bootstrap proportions for clades and the posterior probability of those clades.

BACKGROUND

Decision theory is a well-developed branch of statistics. We do not intend to provide a full review of Bayesian decision theory here; we refer the interested reader to Robert (2007) and chapter 13 of Jaynes (2003) for nice introductions to the topic. Despite the large statistical literature on decision theory, these techniques have been used relatively rarely in discussions of systematic methodology. There are some notable exceptions. Wheeler (1991) presented a decision-theory argument for choosing among trees using a 0-1 loss function (see the section on "all-or-nothing" losses below). Jermiin et al. (1997) used frequentist and Bayesian decision theory arguments in the justification for their methods for constructing a majority-rule consensus of trees with likelihoods that are close to the maximum likelihood score. Minin et al. (2003) developed a model selection methodology from decision theory, and Abdo et al. (2005) applied and extended this approach to account for uncertainty with respect to the estimated tree. Abdo and Golding (2007) recently applied decision theory to the problem of assigning new sequences to species groups in the DNA barcoding context. Steel and Székely (1999) and Steel and Székely (2002) both employed techniques from statistical decision theory.

Making a decision without complete knowledge is a situation that we all face in everyday life, and clearly a rational decision will rest both on what conditions are likely to be true and on the consequences of our decision. The concept of "what conditions are likely to be true" can be captured quite naturally by assigning a probability to any possible outcome. When we have some information about the system in question, then the posterior probability is an appropriate choice of probabilities.

There are many possible ways to quantify the consequences of our choices. Fortunately, when we are making a decision, we only need a measure of the cost of one choice relative to the cost of another choice—we do not necessarily need to have a measure with a value that has an absolute meaning. This is helpful because it is easier to formulate a system of relative costs than it would be to derive the absolute cost of each decision. A common formulation of the problem rests on specifying a *loss function*. A loss function, L, measures the cost that we would have to pay if we took a particular action. Obviously, we aim to minimize our loss. For the remainder of this paper we are concerned with the decision of what tree to report for a dataset, so our "action" can be equated with the selection of a particular tree to report. Note that the evaluation of a loss function requires an action and values of the parameters that we consider to be true—to calculate how bad it would be to report a particular

tree we need to compare the estimated tree to the true tree. Thus, our loss functions are of the form $L(T^*, T)$, where T^* is the true tree, and T is the tree that we report.

A decision rule, δ, maps data that we can observe, x, to actions. In frequentist decision theory, it is common to characterize a decision rule by the expected loss over all possible datasets. This is referred to as the risk, R, of a decision rule δ:

$$R(\theta, \delta) = \int L[\theta, \delta(x)]P(x|\theta)dx \quad (1)$$

where $P(x|\theta)$ is used to weight each loss with the probability that a dataset identical to x would occur if θ were true. Because we do not know the true value of θ when we are evaluating the risks of different decision rules, frequentist decision theory often focuses on decision rules that minimize the risk over all possible values of θ.

From a Bayesian standpoint, this is unsatisfying. We have data in hand, so we have *some* information about what values of θ are likely. Furthermore, we may have prior information about what values of θ are probable. In a Bayesian framework, the posterior expected loss, ρ, can be calculated from a posterior probability distribution over θ:

$$\rho[\delta(x)] = \int L[\theta, \delta(x)]P(\theta|x)d\theta \quad (2)$$

and we can select the decision rule that has the lowest posterior expected loss. Another Bayesian approach would entail defining an integrated risk, r, as the expectation of the risk shown in Equation (1) taken over π, the prior distribution of the parameters:

$$r(\pi, \delta) = \int R(\theta, \delta)\pi(\theta)d\theta \quad (3)$$

These two approaches are equivalent because choosing the action, $\delta(x)$, that minimizes the posterior expected loss for any dataset, x, is a procedure that minimizes the integrated risk (Robert, 2007: 61–63). To paraphrase Robert (2007: 62), Bayesian decision theory argues that it is better to integrate over unknown quantities (i.e., θ) and condition on the observations (i.e., x) than to take into consideration values of x that were not observed and condition on θ as if it were known.

In phylogenetics, we integrate over uncertainty in the true tree topology; i.e., $\theta = T^*$. We can also substitute the tree topology that would be returned under a decision rule, T, rather than referring to it indirectly as the result of the decision, $\delta(x)$. Furthermore, all of the loss functions considered here ignore errors in branch length estimates. So we use the terms "tree" and "tree topology" interchangeably and perform a sum over tree topologies (rather than an integration over the space of all tree topology and branch length combinations). The posterior expected loss of a tree is thus

$$\rho(T) = \sum_{T^*} L(T^*, T)P(T^*|x) \quad (4)$$

Bayesian decision theory seeks to minimize the posterior expected loss; in Bayesian phylogenetic inference, this corresponds to reporting the tree minimizing the posterior expected loss—we refer to this tree as the *MPELT* (the Minimum Posterior Expected Loss Tree).

Thus far, we have briefly reviewed the basics of Bayesian decision theory. But how do we choose a loss function for reporting trees?

All-or-Nothing Loss Functions

An all-or-nothing approach would be to assign a loss of 1 if the tree that we report is not identical to the true tree but a loss of 0 if we report the true tree. In this case we have:

$$L_\dagger(T^*, T) = \begin{cases} 0 \text{ if } T^* = T \\ 1 \text{ if } T^* \neq T \end{cases} \quad (5)$$

where the dagger subscript (†) denotes quantities associated with this all-or-nothing loss function. The expected loss function becomes:

$$\rho_\dagger(T) = \sum_{T^*} I(T^* \neq T)P(T^*|x) \quad (6)$$

where $I(T^* \neq T)$ is an indicator function that is 1 if T^* and T are not the same and 0 if they have the same topology. Note that this is equivalent to summing the posterior probability over all trees that are not T, and, by the law of total probability, this sum of posterior probabilities is simply $1 - P(T|x)$. So

$$\rho_\dagger(T) = 1 - P(T|x). \quad (7)$$

So under a simple all-or-nothing loss function, the tree with the maximum posterior probability is the MPELT. This result was first presented by Wheeler (1991). In other contexts, this loss function is often referred to as a 0-1 loss. However, in phylogenetics one could view splits or topologies as the focus of inference. 0-1 loss functions could be applied to either. Thus, we will use the name "all-or-nothing loss" for a 0-1 loss function on trees and "per-branch loss" for a 0-1 loss function applied to splits (next section).

Per Branch Loss Function

Although the previous result is intuitive and justifies reporting the tree with the highest posterior probability, it has the drawback of penalizing slightly incorrect trees just as much as ridiculously poor estimates of the tree topology. In reality, most systematists would prefer an estimate that is close to the truth over a tree with no

correct clades. To model this preference, we could assign a loss function that assigns a penalty for each clade in the true tree that is missing and another penalty for each clade in the reported tree that is not present in the true tree.

Note that in the case of unrooted trees, we do not know whether a group of taxa on one side of an internal branch is a monophyletic clade or a paraphyletic grouping. In such cases, we should refer to the groupings of taxa as splits rather than clades. "Split" refers to the partition of the taxa that would be created if we cut a tree in two by removing a branch. Because of this tight connection between terms "split" and "branch," we will abuse terminology slightly and use the terms interchangeably in this paper.

We can express a loss function based on the number of correct clades succinctly if we use $\mathcal{B}(T)$ to represent the set of internal branches in tree T:

$$L_\star(T^*, T) = \alpha\left\{\sum_{b\in\mathcal{B}(T)} I[b \notin \mathcal{B}(T^*)]\right\} + (1-\alpha)\left\{\sum_{b\in\mathcal{B}(T^*)} I[b \notin \mathcal{B}(T)]\right\} \quad (8)$$

where $I[b \notin \mathcal{B}(T^*)]$ is an indicator function that is 1 if b is a branch *not* found in T^*. α is the cost of each false positive (a branch in the reported tree that is not in the true tree), whereas $(1-\alpha)$ is the cost of a false negative (a branch in the true tree that is missing from reported tree). If $\alpha > 0.5$, then the loss associated with a false-positive branch is higher than the cost of missing a branch that is in the true tree. The asterisk subscript $(*)$ denotes quantities associated with this per branch loss function. The constraint $0 \leq \alpha \leq 1$ guarantees that the neither the false-positive loss nor the false-negative loss are negative. Without this constraint, our loss function might confer rewards for errors. Note that if $T = T^*$, the loss will be zero. Berry and Gascuel (1996) pointed out that this form of the loss function can be seen as a loss based on a generalization of the Robinson and Foulds (1981) distance. In fact, when $\alpha = (1-\alpha) = 0.5$, this loss function is equivalent to one-half the Robinson and Foulds distance between the true tree and the reported tree.

The formula for the posterior expected loss of a tree looks daunting:

$$\rho^*(T) = \sum_{T^*} P(T^*|x)\left(\alpha\left\{\sum_{b\in\mathcal{B}(T)} I[b \notin \mathcal{B}(T^*)]\right\} + (1-\alpha)\left\{\sum_{b\in\mathcal{B}(T^*)} I[b \notin \mathcal{B}(T)]\right\}\right), \quad (9)$$

but can be reorganized by pulling the summations over the branches to the outside and recognizing that $\sum_{b\in\mathcal{B}(T^*)} I[b \notin \mathcal{B}(T)] = \sum_{b\notin\mathcal{B}(T)} I[b \in \mathcal{B}(T^*)]$:

$$\rho^\star(T) = \alpha \sum_{b\in\mathcal{B}(T)}\sum_{T^*} P(T^*|x) I[b \notin \mathcal{B}(T^*)] + (1-\alpha)\sum_{b\notin\mathcal{B}(T)}\sum_{T^*} P(T^*|x) I[b \in \mathcal{B}(T^*)] \quad (10)$$

In Bayesian phylogenetics, we frequently refer to the posterior probability of a split. This can be estimated as the proportion of all trees that contain a particular branch in a sample generated by MCMC using software such as MrBayes (Ronquist and Huelsenbeck, 2001) or BEAST (Drummond and Rambaut, 2007). In the context of rooted trees, these quantities can be referred to as clade posterior probabilities.

The posterior probability of split b is defined as:

$$P(b|x) = \sum_{T^*} P[T^*|x] I(b \in \mathcal{B}(T^*)]. \quad (11)$$

If we use $\mathcal{B}$ to represent the set of all possible splits for the taxa under consideration, then we can rearrange Equation (10):

$$\rho^\star(T) = \alpha \sum_{b\in\mathcal{B}(T)} [1 - P(b|x)] + (1-\alpha)\sum_{b\notin\mathcal{B}(T)} P(b|x) \quad (12)$$

$$= \alpha \sum_{b\in\mathcal{B}(T)} [1 - P(b|x)] + (1-\alpha)\left[\sum_{b\in\mathcal{B}} P(b|x) - \sum_{b\in\mathcal{B}(T)} P(b|x)\right] \quad (13)$$

$$= (1-\alpha)\sum_{b\in\mathcal{B}} P(b|x) + \sum_{b\in\mathcal{B}(T)} \{\alpha\,[1 - P(b|x)]\} - \sum_{b\in\mathcal{B}(T)} [(1-\alpha)P(b|x)] \quad (14)$$

We introduce the constant $K = (1-\alpha)\sum_{b\in\mathcal{B}} P(b|x)$ to simplify the equations because K does not depend on T. This substitution yields:

$$\rho^\star(T) = K + \sum_{b\in\mathcal{B}(T)} [\alpha - P(b|x)]\,. \quad (15)$$

The tree that minimizes Equation (15) is, by definition, an MPELT. How can we find this tree, or set of trees? The general answer to this seems difficult, but we can make progress if we consider different components of the loss function in isolation.

No False-Negative Loss

Setting $\alpha = 1.0$ means that $1 - \alpha = 0.0$, so there is no false-negative penalty. Such a loss function implies that

we do not mind if we miss a branch; returning a less-resolved tree is just as good as returning the true tree. The posterior expectation of the loss becomes:

$$\rho_+(T) = \sum_{b \in \mathcal{B}(T)} [1 - P(b|x)] \tag{16}$$

Unsurprisingly, this odd loss function results in unhelpful behavior: regardless of the dataset, the star tree (tree with no internal branches) has the minimum possible posterior expected loss (i.e., zero). Trees with internal branches that have posterior probability 1.0 will also attain the minimal posterior expected loss, because then the sum in Equation (16) over splits in the tree only includes terms that contribute nothing. But under standard models and priors, no split will have a posterior of exactly 1.0 (though the MCMC estimate of the posterior probability may be 1.0 for some splits), thus this loss function would always prefer the star tree. Clearly we prefer resolved trees to the star tree (if we have support for the branches), so a false-negative loss of 0.0 is inappropriate.

No False-Positive Loss

A loss function that assigns no penalty to returning a branch that does not actually exist would be quite bizarre as well. If the true tree contained a hard polytomy then most biologists would *not* be just as happy with a method that returned an arbitrary resolution of the polytomy as they would be with one that returned the true tree. The behavior under this loss function is more difficult to analyze; the posterior expected loss becomes:

$$\rho_-(T) = K - \sum_{b \in \mathcal{B}(T)} P(b|x) \tag{17}$$

where the minus subscript (−) recognizes the fact that this loss function only penalizes false negatives. If all splits have non-zero posterior probability, then the set of MPELTs for this loss function will be fully resolved, because resolving a polytomy will add a branch b to $\mathcal{B}(T)$, and this will always decrease the loss by $P(b|x)$ relative to a tree with a polytomy. Intuitively, a tree with high posterior probability will probably contain splits with high posterior probability, so perhaps the tree that maximizes the posterior probability also minimizes the loss shown in Equation (17). The counterexample in Figure 1 shows that this correspondence is not true, in general—the tree with maximum posterior probability does *not* necessarily minimize the posterior expectation of the loss. The table shows a contrived set of posterior probabilities for trees of 5 taxa and the resulting split posterior probabilities. The posterior probabilities are expressed in a general form as the variables p, q, and r. The tree probabilities must satisfy the law of total probability, so $p + q + 2r = 1$ in this example. If $p > q > r > 0$, then the tree $((A, B), C, (D, E))$ maximizes the posterior probability. However if the inequality $p < r + q$ is also true, then the tree $((A, E), C, (B, D))$ minimizes the posterior expected loss under the no false-positive loss function given in Equation (17). At least one set of posterior probabilities ($p = 0.27, q = 0.25, r = 0.24$) satisfies these constraints. Less artificial examples, in which all of the

Tree	Posterior Probability	Example
$((A, B), C, (D, E))$	p	0.27
$((A, E), C, (B, D))$	q	0.25
$((B, D), A, (C, E))$	r	0.24
$((A, E), D, (B, C))$	r	0.24
all other trees	0	0
Split	**Posterior Probability**	**Example**
$AE\|BCD$	$q + r$	0.49
$BD\|ACE$	$q + r$	0.49
$AB\|CDE$	p	0.27
$DE\|ABC$	p	0.27
$BC\|ADE$	r	0.24
$CE\|ABD$	r	0.24
all other splits	0	0
Tree	**Posterior Expected Loss**	**Example**
$((A, B), C, (D, E))$	$K - 2p\beta$	$K - 0.54$
$((A, E), C, (B, D))$	$K - 2(q + r)\beta$	$K - 0.98$

FIGURE 1. Tables showing the posterior probabilities for four five-taxon trees, the resulting split posteriors, and the variable portion of the posterior expected loss for each of the trees under the loss function shown in Equation (17). If $p > q > r$ and $p < r + q$, the MAP tree is $AB|C|DE$, but the tree minimizing the posterior expected loss under Equation (17) is $AE|C|BD$. The tree probabilities must also satisfy the law of total probability, so $p + q + 2r = 1$, but there are combinations of probabilities (such as $p = 0.27, q = 0.25, r = 0.24$) that satisfy all of these constraints.

trees have a non-zero posterior probability can also be constructed.

Examination of Equation (17) reveals that we are seeking the tree with the maximal sum of split probabilities. This is a form of the maximum-weighted split compatibility problem, which is known to be NP-hard (Day and Sankoff, 1986). Thus, a general, efficient algorithm for finding the solution does not exist, but in many cases it may be feasible to find such a tree by creating a greedy resolution of the 50% majority-rule consensus tree (as discussed in the excellent review of consensus methods by Bryant, 2003).

The asymmetric median consensus tree (Phillips and Warnow, 1996) is defined in terms of minimizing the sum of the weights of splits present in the collection of trees but missing in the consensus tree. For a fixed set of split weights, this criterion is identical to maximizing the sum of split weights that are present in the consensus tree. When split weights are interpreted as posterior probabilities, then this task is identical to finding the tree that minimizes the posterior expectation of the loss given in (17). Thus, if the posterior distribution is approximated using MCMC then the MPELT under the loss given in (17) will be identical to the asymmetric median consensus tree of the trees sampled during MCMC.

Conservative, Per Branch Loss Functions

It seems prudent to prefer a conservative estimation procedure. When we are uncertain of a grouping on the tree, we prefer to report a soft polytomy for that portion of the tree. This summary is certainly understood by systematists who routinely interpret polytomies as statements of uncertainty (rather than hypotheses of simultaneous divergence into more than two species). We can accomplish this by using a loss function that penalizes incorrect branches in the reported tree more than missing branches. In other words, choosing a loss function in which $\alpha > 0.5$.

We will demonstrate that when $\alpha > 0.5$ the $(100 \times \alpha)\%$ majority-rule consensus tree of the collection of trees from an MCMC sample will minimize the posterior expected loss. A parallel result was first presented by Berry and Gascuel (1996) in their discussion of which clades to include when summarizing trees from nonparametric bootstrapping. This majority-rule consensus tree is defined to be the tree that is composed of all splits that occur in over $(100 \times \alpha)\%$ of the trees in the collection. In most cases the set of MPELTs will contain only this one tree. If some splits occur in exactly $(100 \times \alpha)\%$ of the trees, then the set of MPELTs will contain the $(100 \times \alpha)\%$ majority-rule consensus and other trees that resolve this consensus tree by adding splits that have estimated posterior probabilities of exactly α. We will refer to the resolution of the $(100 \times \alpha)\%$ majority-rule consensus tree that includes all splits that occur in exactly $100 \times \alpha\%$ of the input trees as the $\geq(100 \times \alpha)\%$ majority-rule consensus tree.

To prove this conclusion, we can characterize the splits contained in the MPELT when $\alpha > 0.5$. We can derive one necessary condition for a split that is contained in an MPELT by comparing a MPELT, T, with a tree that is identical to it except for the fact that one branch, s, has been collapsed; this tree will be denoted T/s. By definition of the MPELT, we have the constraint that:

$$\rho^\star(T) - \rho^\star(T/s) \leq 0 \tag{18}$$

Under our loss function we can restate Equation (15) with these two trees in mind:

$$\rho^\star(T/s) = K + \sum_{b \in \mathcal{B}(T/s)} [\alpha - P(b|x)] \tag{19}$$

$$\rho^\star(T) = K + \alpha - P(s|x) + \sum_{b \in \mathcal{B}(T/s)} [\alpha - P(b|x)] \tag{20}$$

Thus, we can rearrange the inequality in (18) to yield:

$$\alpha \leq P(s|x) \tag{21}$$

This places a lower bound on the posterior probability of any split that is in the MPELT.

Note that this constraint is a *necessary* condition for a split to be present in an MPELT—if it is not met, then the T/s will have a lower posterior expected loss than T, so T will not be a MPELT. We have not shown that Inequality (21) is a *sufficient* condition for a split to be included in any (or every) tree that is a MPELT.

Here, we are concerned with cases in which $\alpha > 0.5$. This is fortunate because the set of splits with posterior probability greater than 50% is guaranteed to be pairwise compatible and therefore compatible (Buneman, 1971). Thus it is possible for a tree to contain every split that satisfies Inequality (21)—in fact, this tree is simply the $\geq(100 \times \alpha)\%$ majority-rule consensus tree of the posterior distribution over trees. This guarantees that no split that is not in the $\geq(100 \times \alpha)\%$ majority-rule consensus tree can be in any tree that is an MPELT. Such a split, y, would have a posterior probability lower than α, and examination of Equations (19) and (20) reveals that the tree T/y would have a lower posterior expected loss than a tree, T, which does contain the split y. So we do not need to consider trees that are incompatible with, or are refinements of, the $\geq(100 \times \alpha)\%$ majority-rule consensus tree.

If the posterior probablity of a split is greater than α, then Equations (19) and (20) show that the posterior expected loss of a tree that contains the split will be lower than a tree that has the corresponding branch collapsed. Thus, every split found in the $(100 \times \alpha)\%$ majority-rule consensus tree will be found in every tree in the set of MPELT. If $P(s|x)$ is exactly equal to α, then split s is exactly on the cutoff for inclusion and this split will not be in every tree in the MPELT set. In this case, T/s and T will both be in the MPELT set. This situation will be rare. Because α and $P(s|x)$ are continuous variables, they will almost never be exactly equal. It is possible that our MCMC-based estimates of $P(s|x)$ will be equal to α. Ignoring these rare cases, we can say that the $(100 \times \alpha)\%$ majority-rule consensus tree will correspond to the MPELT under this loss function.

If no nontrivial split has a posterior probability greater than or equal to α, then the star tree will minimize the posterior expected loss.

Per Branch Loss Functions That Emphasize Power

In the most common context of reporting a phylogeny for a group, it seems appropriate to use a loss similar to the one described in the previous section—a loss function with $\alpha > 0.5$. In some cases we may be more interested in reporting any split that seems plausible. For example, one might want to constrain parts of the tree because it is not computationally feasible to explore all of tree space. In such a context, we might want to make sure that our constraints are not ruling out splits that might be present in the true tree. Therefore, the penalty for missing a branch would be higher than the penalty for including an extra branch.

When $\alpha \leq 0.5$, the two incompatible splits can each satisfy the necessary condition for a split to be in the MPELT (Inequality (21)). If we insist on returning a tree, then we must search through trees to find one that minimizes Equation (15); the algorithms introduced by Susko (2006) for finding collections of trees with split weights above a threshold may provide inspiration for an algorithm for finding the MPELT in this case.

More importantly, in a situation in which we want to penalize missing branches more than false branches, it may not be helpful to restrict ourselves to reporting a tree. Summarizing all of the splits that satisfy Inequality (21) in a consensus network may be the more appropriate route to take (see Huson and Bryant, 2006, for a helpful overview of these approaches).

Discussion

We have examined the implications of viewing the reporting of phylogenetic estimates from the standpoint of statistical decision theory. This leads to (yet) another phylogenetic optimality criterion: a preference for the trees with the minimum posterior expected loss. In particular, we propose a simple, cautious loss function that is appropriate for the routine task of reporting a phylogeny estimated by a Bayesian analysis. This loss function (described in detail above Conservative, Per Branch Loss Functions) expresses a preference for trees with as few incorrect branches as possible, but the function also penalizes estimates that omit a branch that is present in the true tree. Furthermore, the false-positive penalty is larger than the false-negative penalty.

If such a per branch loss function is used, then the tree that minimizes the posterior expected loss will be the majority-rule consensus tree of samples from the posterior probability distribution over trees—exactly the type of summary that many systematists already use. This loss function is a generalization of the Robinson and Foulds distance, and, as Berry and Gascuel (1996) point out, the $(100 \times \alpha)\%$ majority-rule tree is naturally associated with this distance. Previously this consensus tree has often been viewed as merely a summary tool and not as an optimal estimate of the tree under any criterion.

If we prefer to be more conservative, then we can make the cost of extra branches in a tree higher than the cost associated with a missing branch. Implementing such a more conservative loss function which penalizes incorrect groupings even more strongly amounts to merely raising the cutoff for the majority-rule consensus tree. For instance, if the cost of a false positive is nine times higher than the cost of a false negative, then the cutoff becomes 90%; thus only the 90% majority-rule tree would be presented. The idea that including an incorrect branch in an estimated tree is a more serious mistake than omitting a branch is certainly not new (see Berry and Gascuel, 1996; Phillips and Warnow, 1996, for example). As noted above, in the unlikely event that a grouping has posterior probability that corresponds exactly to the cutoff for inclusion in the majority-rule consensus tree, than these splits can also be included to produce more trees that also minimize the posterior expected loss. So the MPELT set under this loss function might include more-resolved versions of the majority-rule consensus in addition to the majority-rule consensus itself.

From this decision-theoretic standpoint, the 50% majority-rule tree is an elegant summary of the posterior distribution over trees because it allows readers to use their own level of aversion to questionable groupings by simply looking at the tree and ignoring branches with support lower than their own cutoff. Once again, this perspective justifies a common practice among systematists: the 50% majority-rule tree is often presented with an asterisk or other symbol highlighting the branches that exceed a cutoff which the authors feel comfortable viewing as strong support. Much of the discussion in the papers then centers around these strongly supported clades.

It is important to note that the loss function described here is meant to reflect the decision of which tree to take as a phylogeny worth reporting and discussing. We have not attempted to exhaustively sample the universe of potential loss functions. Thus, we do not claim to have derived *the* correct loss function for any Bayesian phylogenetic analyses. In other contexts, very different loss functions may be appropriate, or it may not be helpful to consider losses in conjunction with tree estimation. For instance, many uses of Bayesian phylogenetics treat the tree as a nuisance parameter. In such cases, there is not a need to compress the posterior distribution into a summary—the entire sample of trees from the posterior distribution is helpful in characterizing the uncertainty in our estimates of the phylogeny.

The tree that has the highest posterior probability can also be a tree that minimizes the posterior expected loss, but it does not *have* to be the MPELT. If we were forced to bet on a single tree topology and we would lose our bet if any part of it is incorrect, then our loss function would be an all-or-nothing statement about the tree. In such a case, the tree with the maximum posterior probability is always the tree that minimizes the posterior expected loss. In most contexts our loss function should not be all-or-nothing: reporting a tree that is mostly correct should not cost as much as reporting a tree that is completely wrong. In the simple case of a per branch loss

function that penalizes false positives more than false negatives, the tree with the maximum posterior probability will only be the MPELT if it is identical to the majority-rule consensus tree (using the appropriate cutoff). Of course, there are many other possible forms of loss functions that we have not considered here. There are many ways of quantifying how different two trees are. Starting from the idea that we would prefer to report trees that are close to the true tree, one could derive a set of loss functions for every different measure of tree-to-tree dissimilarity. Thus there may be many forms of loss functions for which the tree with the highest posterior probability is guaranteed to be a MPELT.

Wheeler and Pickett (2008) recently criticized the practice of reporting the majority-rule consensus from a Bayesian MCMC simulation as leading to "exaggerated clade support, inconsistently biased priors, and the impossibility of hypothesis testing of cladograms." We will not attempt to address all of their arguments here, but we do note that their paper concludes that the majority-rule consensus tree "may perhaps be regarded as statements of support but not as best-supported scientific hypotheses of phylogenetic relationships." This statement is in keeping with a tradition of systematists treating consensus trees as useful summaries but not estimators in the truest sense. For example, Swofford (1991) states that "consensus trees are simple statements about areas of agreement among trees; they should not be interpreted as phylogenies," because a literal interpretation of a consensus tree as a phylogeny would imply that any polytomies present in the consensus represent the simultaneous origin of more than two species. Miyamoto (1985) and Carpenter (1988) further caution against some uses of consensus trees for summarizing a collection of most parsimonious trees, on the grounds that the consensus tree will (often) have a worse fit to the data than any of the most parsimonious trees. The example given by Barrett et al. (1991) serves as a warning against treating the consensus tree from analyses of subsets of the data as a "safe" statement of branches that will be present in analysis of the full data.

Here, we advocate the use of the majority-rule consensus tree as an optimal summary in the context of a per branch loss function. Our results do not conflict with all of the points raised by these authors. For example, Swofford's (1991) point about treating polytomies in a majority-rule consensus trees as soft polytomies applies to the summaries that we favor. Nor do our results imply that the majority-rule consensus will fit the data better (in the sense of higher posterior probability or likelihood, or lower parsimony score) than other trees. Rather, the decision-theory framework gives us an argument for viewing the majority-rule tree as more than merely *a* summary. It can be seen as the *optimal* summary. In fact, if we need to report one tree and accept the tenets of the loss function described above (a per branch loss with $\alpha > 0.5$), then the $(100 \times \alpha)\%$ majority-rule tree is superior (in terms of expected loss) to the tree that has the highest posterior probability. If one were to view hard polytomies as impossible and assign polytomies a prior probability of 0, then all trees with polytomies would have a posterior probability of 0. Even in this context, a majority-rule consensus that has polytomies can *still* be viewed as the optimal tree to report from a decision-theoretic viewpoint. Despite the fact that the tree has no chance of being a completely correct representation of the phylogeny, it does the best job of conveying groupings of which the analysis is confident while avoiding weakly supported groups.

We agree with the statement by Wheeler and Pickett (2008) that the majority-rule consensus tree is a "statement of support" rather than the tree topology that has the highest probability of being a completely correct representation of the evolutionary history of the group. This does not imply that we agree with most of their objections to the majority-rule consensus tree. For example, Wheeler and Pickett (2008) mention the fact that prior probabilities of different-sized clades are not necessarily equal in Bayesian analyses (except in cases of trees with very few taxa); this fact has been mentioned by authors including Pickett and Randle (2005), Randle and Pickett (2006), and Yang (2006: 176). Although this fact may make some systematists reluctant to use clade posterior probabilities, we refer readers to the work of Steel and Pickett (2006) and Velasco (2007), which demonstrate that the nonuniform priors are the direct consequence of unproblematic statements about uncertainty with respect to the tree shape. The priors are fundamental aspects of probability statements on trees and do not indicate a problem with the Bayesian approach to phylogenetic inference. Interested readers should also consult Brandley et al. (2006).

We note that there are other contexts in which the majority-rule tree can be viewed as an optimal tree. A median tree refers to the tree closest to all members of collection of trees, in the sense that it has the smallest sum of distances to all of the trees in the collection. Barthélemy and McMorris (1986) showed that the 50% majority-rule consensus of a collection of trees is the median tree when the symmetric distance is used as the metric for comparing trees (note that if the number of trees is even then the set of median trees may contain trees that resolve the 50% majority-rule tree by adding splits which occur in exactly half of the input trees). McMorris (1990) extended that work and pointed out that if we treat the splits that are present in a collection of trees as data, then we can use a simple model to calculate a likelihood for any summary tree. In McMorris's model there is a probability p that a split will occur in an input tree if the split is present in the summary tree. For each split present in an input tree but absent in the summary tree, the probability $1 - p$ is used in the likelihood. For any value of p in the range $0.5 < p < 1.0$, McMorris demonstrated that 50% majority-rule tree is the summary that maximizes the likelihood. The model is hard to justify as good description of which splits are likely to occur in a collection of estimated trees (for one thing the presence of each split is treated as an independent datum in McMorris model). Recently, Steel and Rodrigo (2008) have proposed a more realistic model of errors in tree topologies in the context

of their ML supertree methodology. Although the interpretations of majority-rule consensus trees given by Barthélemy and McMorris (1986) and McMorris (1990) highlight interesting properties of this consensus technique, we feel that the decision-theoretic interpretation presented here provides a more intuitive interpretation of the role of a majority-rule consensus of a sample from a Bayesian analysis. Berry and Gascuel (1996) also found this decision-theoretic perspective helpful in the context of reporting the results of bootstrapping.

ACKNOWLEDGMENTS

The authors wish to thank Jack Sullivan, Olivier Gascuel, Mike Steel, Vincent Ranwez, and an anonymous reviewer for their helpful suggestions and pointers to other relevant literature that were omitted in the first version of the manuscript. The anonymous reviewer pointed out the connection to asymmetric median consensus trees. Thanks to Tandy Warnow for pointing out to us that finding the tree that maximizes the sum of split probabilities is merely an instance of maximum weighted compatibility. The authors thank the NSF grant DEB-0732920 to M.T.H. and NSF grant DBI-0306047. M.T.H. also thanks the Isaac Newton Institute for Mathematical Sciences in Cambridge (UK), Vincent Moulton, Mike Steel, Daniel Huson (who organized Phylogenetics Programme at the INI), and the University of Kansas for travel funds to attend the INI.

REFERENCES

Abdo, Z., and G. B. Golding. 2007. A step toward barcoding life: A model-based, decision-theoretic method to assign genes to preexisting species groups. Syst. Biol. 56:44–56.

Abdo, Z., V. Minin, P. Joyce, and J. Sullivan. 2005. Accounting for uncertainty in the tree topology has little effect on the decision-theoretic approach to model selection in phylogeny estimation. Mol. Biol. Evol. 22:691–703.

Barrett, M., M. Donoghue, and E. Sober. 1991. Against consensus. Syst. Zool. 40:486–493.

Barthélemy, J., and F. McMorris. 1986. The median procedure for n-trees. J. Classif. 3:339–334.

Berry, V., and O. Gascuel. 1996. On the interpretation of bootstrap trees: Appropriate threshold of clade selection and induced gain. Mol. Biol. Evol. 13:999–1011.

Brandley, M., A. Leache, D. Warren, and J. M. Guire. 2006. Are unequal clade priors problematic for Bayesian phylogenetics? Syst. Biol. 55:138–146.

Bryant, D. 2003. A classification of consensus methods for phylogenetics. Pages 163–184 *in* Bioconsensus, volume 61. (M. F. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts, eds.). American Mathematical Society, Providence, Rhode Island.

Buneman, P. 1971. The recovery of trees from measures of dissimilarity. Pages 387–395 *in* Mathematics in the Archaeological and Historical Sciences (F. R. Hodson, D. G. Kendall, and P. Tautu, eds.). The Royal Society of London and the Academy of the Soc. Rep. of Romania Edinburgh University Press, Edinburgh.

Carpenter, J. M. 1988. Choosing among equally parsimonious cladograms. Cladistics 4:291–296.

Day, W. H. E. and D. Sankoff. 1986. Computational complexity of inferring phylogenies by compatibility. Syst. Zoo. 35:224–229.

Drummond, A. J. and A. Rambaut. 2007. BEAST v1.4. http://beast.bio. ed.ac.uk/.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791.

Huelsenbeck, J. P., and F. R. Ronquist. 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Huson, D., and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254–267.

Jaynes, E. T. 2003. Probability theory: The logic of science. Cambridge University Press, New York.

Jermiin, L., G. Olsen, K. Mengersen, and S. Easteal. 1997. Majority-rule consensus of phylogenetic trees obtained by maximum-likelihood analysis. Mol. Biol. Evol. 14:1296–1302.

Larget, B., and D. L. Simon. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16:750–759.

Mau, B., M. A. Newton, and B. Larget. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics 55:1–12.

McMorris, F. R. 1990. The median procedure for n-trees as a maximum likelihood method. J. Classif. 7:77–80.

Minin, V., Z. Abdo, P. Joyce, and J. Sullivan. 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst. Biol. 52:674–683.

Miyamoto, M. M. 1985. Consensus cladograms and general classifications. Cladistics 1:186–189.

Phillips, C., and T. Warnow. 1996. The asymmetric median tree—A new model for building consensus trees. Disc. Appl. Math. 71:311–335.

Pickett, K. M., and C. P. Randle. 2005. Strange Bayes indeed: Uniform topological priors imply non-uniform clade priors. Mol. Phylogenet. Evol 34:203–211.

Randle, C. P., and K. M. Pickett. 2006. Are nonuniform clade priors important in Bayesian phylogenetic analysis? A response to Brandley et al. Syst. Biol. 55:147–151.

Robert, C. P. 2007. The Bayesian choice: From decision-theoretic foundations to computational implementation, 2nd edition. Springer Verlag, New York.

Robinson, D. F., and L. R. Foulds. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Ronquist, F. R., and J. P. Huelsenbeck. 2001. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:754–755.

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

Steel, M., and A. G. Rodrigo. 2008. Maximum-likelihood supertrees. Syst. Biol. 57:243–250.

Steel, M. A., and L. A. Székely. 1999. Inverting random functions. Ann. Combin. 3:103–113.

Steel, M. A., and L. A. Székely. 2002. Inverting random functions II: explicit bounds for discrete maximum likelihood estimation, with applications. SIAM J. Disc. Math 15:562–575.

Susko, E. 2006. Using minimum bootstrap support for splits to construct confidence regions for trees. Evol. Bioinformatics 2:1–15.

Swofford, D. L. 1991. When are phylogeny estimates from molecular and morphological data incongruent? Pages 295–333 *in* Phylogenetic analysis of DNA sequences (M. M. Miyamoto and J. Cracraft, eds.). Oxford University Press, New York.

Velasco, J. D. 2007. Why non-uniform priors on clades are both unavoidable and unobjectionable. Mol. Phylogenet. Evol. 45:748–749.

Wheeler, W. C. 1991. Congruence among data sets: A Bayesian approach. Pages 334–346 *in* Phylogenetic analysis of DNA sequences (M. M. Miyamoto and J. Cracraft, eds.). Oxford University Press, New York.

Wheeler, W. C., and K. Pickett. 2008. Topology-Bayes versus Clade-Bayes in phylogenetic analysis. Mol. Biol. Evol. 25:447–453.

Yang, Z. 2006. Computational molecular evolution. Oxford University Press, New York.

First submitted 30 November 2007; reviews returned 11 February 2008; final acceptance 6 May 2008
Associate Editor: Olivier Gascuel