Syst. Biol. 56(4):578–590, 2007
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150701485091

Summarizing a Posterior Distribution of Trees Using Agreement Subtrees

KAREN A. CRANSTON[1] AND BRUCE RANNALA[2]

[1]*Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721 USA; E-mail: cranston@email.arizona.edu*
[2]*Genome Center, University of California Davis, One Shields Avenue, Davis, California 95616, USA; E-mail: brannala@ucdavis.edu*

Abstract.—Bayesian inference of phylogeny is unique among phylogenetic reconstruction methods in that it produces a posterior distribution of trees rather than a point estimate of the best tree. The most common way to summarize this distribution is to report the majority-rule consensus tree annotated with the marginal posterior probabilities of each partition. Reporting a single tree discards information contained in the full underlying distribution and reduces the Bayesian analysis to simply another method for finding a point estimate of the tree. Even when a point estimate of the phylogeny is desired, the majority-rule consensus tree is only one possible method, and there may be others that are more appropriate for the given data set and application. We present a method for summarizing the distribution of trees that is based on identifying agreement subtrees that are frequently present in the posterior distribution. This method provides fully resolved binary trees for subsets of taxa with high marginal posterior probability on the entire tree and includes additional information about the spread of the distribution. [Agreement subtrees; Bayesian phylogenetic inference; credible sets; posterior distribution of trees; threshold accepting.]

Phylogenetic inference has long been troubled by the difficulty of performing statistical analysis on tree topologies. The topologies are discrete, categorical, and non-nested hypotheses about the species relationships. They are not amenable to standard summary analyses such as the calculation of means and variances and cause difficulties for many traditional forms of hypothesis testing (Yang et al., 1995). Although the complexity of the phylogeny as a parameter makes traditional statistical methods problematic, this is certainly not due to a lack of information contained in the phylogenies, leading us to develop creative ways of summarizing and comparing an input list of trees.

A Bayesian phylogenetic analysis (Rannala and Yang, 1997; Huelsenbeck and Ronquist, 2001) uses a Markov chain Monte Carlo (MCMC) algorithm to produce a posterior distribution of phylogenies, and we are interested in methods for summarizing this distribution. In a systematics study, such a method may be needed to graphically summarize the relationships between species on a tree. An evolutionary study may require one or more trees in order to study such topics as divergence times (Sanderson, 2002; Thorne et al., 1998), selection (Yang et al., 2005), or phylogeography (Knowles and Maddison, 2002; Hewitt, 2001). In either case, the information in the full posterior distribution of trees must be sufficiently reduced for the required purpose.

A common starting point for summarizing any distribution is to report summary statistics that include such concepts as the location and spread. A Bayesian analysis produces a set of trees with a posterior probability for each tree topology. The trees are categorical variables and a natural point estimate is therefore the mode of the distribution, or the tree topology with the highest posterior probability. This is known as the maximum a posteriori, or MAP, tree. An interval estimator is then the credible set of trees—the number of unique sampled trees whose probability sum to a certain limit, say 90% or 95%. In an ideal situation, the credible set would be small and a significant proportion of the posterior probability would be assigned to the MAP tree.

We rarely see this sort of reporting in the phylogenetics literature. Why is this the case? First, the probability of the MAP tree is often very low. This can happen if the posterior distribution is broad, giving credible sets with a large number of topologies. A large credible set is unlikely to produce a single MAP tree with high probability. In this case, the the probability of the single best tree may not be much greater than any number of other tree topologies in the posterior sample.

The second reason may be that reporting a credible sets of trees is not as intuitive as a credible set for a continuous numerical parameter. The credible set of trees defines the number of tree topologies contained in the set, giving a measure of the overall spread of the distribution. However, there is no information about the relationship between the trees or how widely they differ from one another. Contrast this to a posterior distribution of a continuous parameter with a single mode, where the credible set contains a range of values that fall between a well-defined minimum and maximum point.

Instead, the most common method for summarizing a posterior distribution of phylogenies is the majority-rule consensus (MRC) tree as a point estimate with partition probabilities at each internal node as a measure of the uncertainty. The MRC tree is constructed by combining all partitions with probability greater than 0.5 from the list of observed partitions in the posterior sample. This often results in multifurcations being introduced into the tree in order to combine low probability binary partitions into a single well-supported multifurcating node. Because the MRC is a combination of sampled partitions, it is possible (although unlikely) that the entire tree was never actually sampled during the MCMC, meaning that we cannot assign a posterior probability to the whole tree. If construction of the MRC involves collapsing nodes to produce multifurcations, then it is certain that we did not sample the tree and there is no probability for the MRC tree. Given that one of the advantages of a Bayesian analysis is the ability to assign probabilities to tree topologies, it is preferable to retain this measure of support when presenting the results.

Whether we use the MAP tree, the MRC tree, or a different consensus method, reducing the distribution to a single tree fails to adequately describe the full distribution. Although providing marginal probabilities for a tree, or a single clade, can improve the information content, other summary methods can describe aspects of the distribution not captured by such point estimates.

If the credible set of trees is large and the probability of the MAP tree is small, what information can we obtain from the distribution? This situation may occur if there is a lack of information, conflicting signals, or possibly due a lack of convergence of the MCMC. If we can eliminate the issue of convergence (which we note is a nontrivial process for many data sets), then either there is simply not enough information in the data to infer a single strongly supported tree or there is an underlying evolutionary process that does not support a single tree for the full set of taxa. In any case, additional types of analyses can elucidate information about the support for various parts of the evolutionary history even if we cannot place a high probability value on a single tree.

Summary Using Tree Pruning

One way to extract additional information from the distribution of phylogenies is to simplify the distribution until we find a well-supported MAP tree. This is akin to finding an underlying well-supported "skeleton tree" within the full posterior distribution of trees. While sampling trees using MCMC, we expect that portions of the tree will remain nearly fixed while integrating over the uncertainty. We can find this constant tree by removing uncertain taxa and leaving those taxa that are well supported by a large percentage of the input trees. This type of approach is similar to the so-called maximum agreement subtree (MAST) methods (Finden and Gordon, 1985). Given a list of trees, the maximum agreement subtree is the largest subtree that is contained in all of the input trees. This subtree (or set of subtrees) is generally found using heuristic search and even for three input trees, the unbound problem is NP-complete (Amir and Keselman, 1997). When the input trees are constrained to be rooted and binary, efficient algorithms exist for finding the MAST tree (Cole et al., 2000; Lee et al., 2005).

The MAST tree, by definition, must agree with each of the input trees. This means that any taxa with ambiguous relationships are stripped from the input trees, irrespective of the support values for the ancestral nodes. When requiring strict agreement, the resulting subtree may contain only a small subset of the taxa in the starting dataset. In the worst-case scenario, the MAST tree may be trivial, containing only two leaves.

Rather than look for a strict MAST tree contained in every one of the input trees, we instead propose a method that searches for agreement subtrees that may be present in only some of the input trees. This is similar to the body of literature describing frequent subtree mining (reviewed in Chi et al., 2005). There are two properties that differentiate phylogenies from other types of trees and networks used in more general subtree mining algorithms. The first is that phylogenies are unordered; that is, there is no information in the left-right orientation of nodes. Whether we draw a given taxon as a left or right descendant of an ancestral node does not affect the uniqueness of the tree. The second, more important, property is that ancestral nodes are only interesting when they have descendant nodes. If we remove two sister taxa from a phylogeny, then the ancestral node is no longer of interest, and we no longer include it on the tree. This differs from situations in non-phylogenetic applications where we are interested in properties of lower-level nodes even if there are no higher-level descendents. Frequency subtree mining has been addressed specifically in the context of phylogenies (Shasha et al., 2004) but only to look for common pairs of taxa in trees with different sets of taxon labels. The problem addressed in this study uses a posterior sample of trees as input, and each tree contains the same set of taxon labels.

To search for agreement subtrees in a posterior sample, we can use the posterior probabilities to score the various subtrees. The posterior sample of trees contains a count of each tree in proportion to its probability, so we prefer agreement subtrees that are present in a larger proportion of the original sampled trees. In effect, we are weighting the agreement subtrees using the sum of the posterior probabilities of the input trees that agree with the subtree. An example of this strategy is shown in Figure 1.

Summarizing the set of trees using pruned subtrees, in addition to producing estimates of the best-supported subtrees, also helps to identify so-called rogue taxa (Wilkinson et al., 2000; Sanderson, 2002). These are species that appear in multiple relationships with other taxa in the trees, and are of particular concern when more than one relationship has non-negligible probability. Removing these taxa from the posterior distribution will have a greater effect on the summed MAP tree probabilities than taxa that have well-supported relationships on the input trees. For example, given the posterior distribution of trees in Table 1, we can prune off different taxa from the trees and compare the resulting distributions.

TABLE 1. A sample posterior distribution of phylogenies.

Probability	Tree
0.6029	((((((2,(6,(16,27))),(7,(48,(43,44))))
0.1637	(((((((2,6),(16,27)),(7,(48,(43,44))))
0.0832	((((((6,(2,(16,27))),(7,(48,(43,44))))
0.0646	((((((2,(16,(6,27))),(7,(48,(43,44))))
0.0254	(((((((2,16),(6,27)),(7,(48,(43,44))))
0.0194	((((((2,(27,(6,16))),(7,(48,(43,44))))
0.0123	((((((2,(6,(16,27))),(7,(44,(43,48))))
0.0045	((((((16,(6,(2,27))),(7,(48,(43,44))))
0.0040	((((((6,(27,(2,16))),(7,(48,(43,44))))
0.0040	((((((16,(2,(6,27))),(7,(48,(43,44))))
0.0030	((((((6,(16,(2,27))),(7,(48,(43,44))))
0.0024	((((((16,(27,(2,6))),(7,(48,(43,44))))
0.0021	(((((((2,6),(16,27)),(7,(44,(43,48))))
0.0020	(((((((6,16),(2,27)),(7,(48,(43,44))))
0.0015	(((((((2,6),(16,27)),(7,(43,(44,48))))
0.0015	((((((2,(16,(6,27))),(7,(43,(44,48))))
0.0012	((((((27,(6,(2,16))),(7,(48,(43,44))))
0.0005	((((((6,(2,(16,27))),(7,(43,(44,48))))

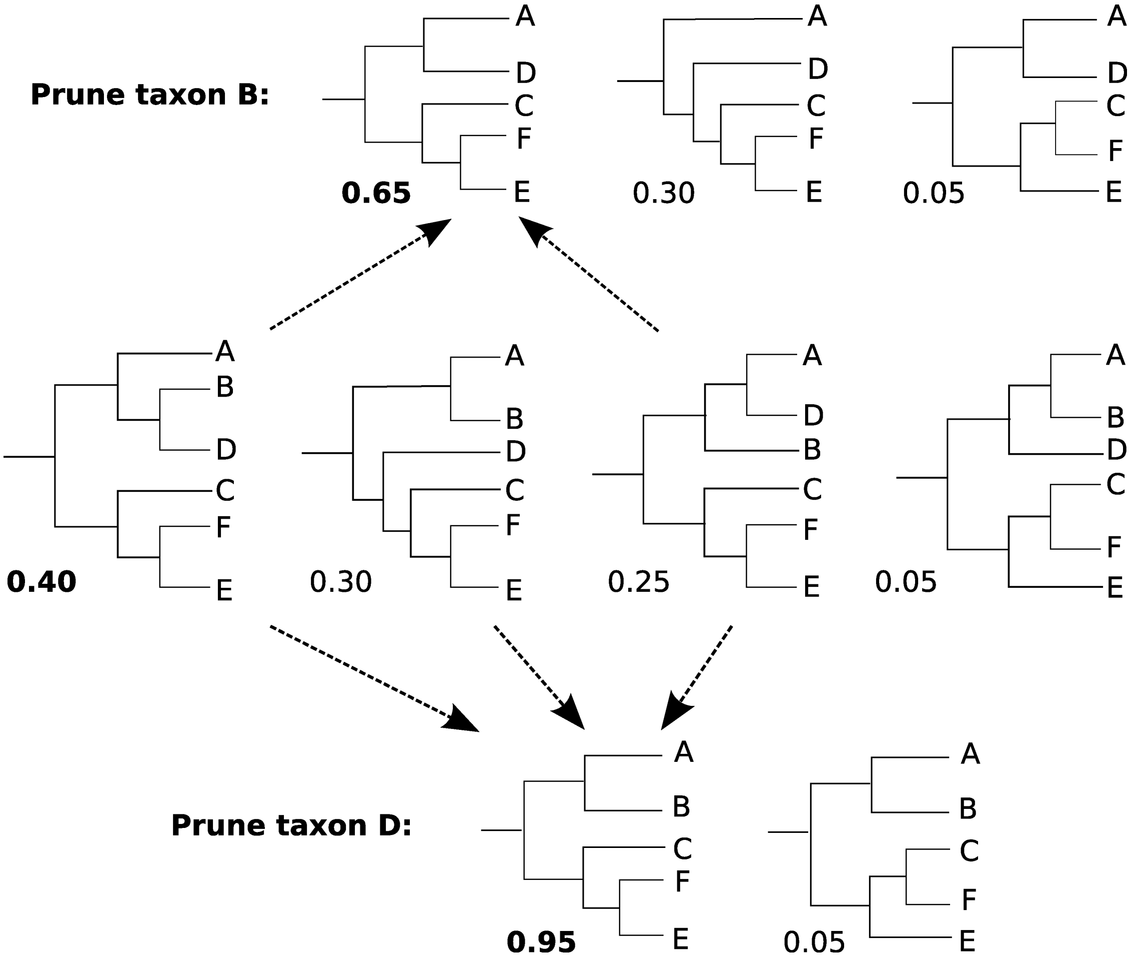


FIGURE 1. An example of how the pruning method can increase the probability of the MAP tree. The four trees in the center row represent the original posterior sample, with the MAP probability labeled in bold. We can increase the P_{MAP} be pruning off any single taxon. In this example, we prefer to prune taxon D rather than taxon B due to the higher probability of the resulting subtree. Dotted arrows indicate the trees that collapse to produce the agreement subtree.

The original distributions and those resulting from pruning taxon 2, 6, or 16 are shown in Figure 2. Although pruning any of these taxa improves the probability of the MAP tree and decreases the width of the credible sets, taxon 6 gives the greatest improvement; therefore, we prefer removal of taxon 6 over the other two taxa. This result indicates that the placement of taxon 6 on the tree is less well supported than the placement of taxon 2 or 16, which is not immediately obvious from studying the posterior distribution of phylogenies.

This can then be extended to greater numbers of pruned taxa. For a more complex distribution of trees (with more starting taxa or greater initial spread), removing a single taxon may not be sufficient to produce a subtree with high probability. If this is the case, pruning additional taxa will further collapse the input trees until a well-supported skeleton is discovered.

THEORY

Assume a vector of R tree topologies, $\tau = \{\tau_i\}$, $i = 1, 2, \ldots, R$, generated by a Bayesian MCMC phylogenetic inference algorithm. The probability of the MAP tree, M, can be calculated as follows:

$$P_{MAP}(\tau) = \frac{1}{R} \sum_{i=1}^{R} I_M(\tau_i) \tag{1}$$

where

$$I_M(\tau_i) = \begin{cases} 1 & \text{if } \tau_i = M \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Our objective is to identify k taxa (from a total of $\bar{S}$ taxa) that, when eliminated from each tree in $\bar{\tau}$, improve the support for a single best tree. We set a target posterior probability for a MAP tree based on $S - k$ taxa and then attempt to minimize k.

Let the taxon indices be $\bar{S} = \{1, \ldots, S\}$, where S is the total number of taxa included in the phylogenetic analysis. A potential subset of taxa is $S_k = s_1, \ldots, s_k$ where

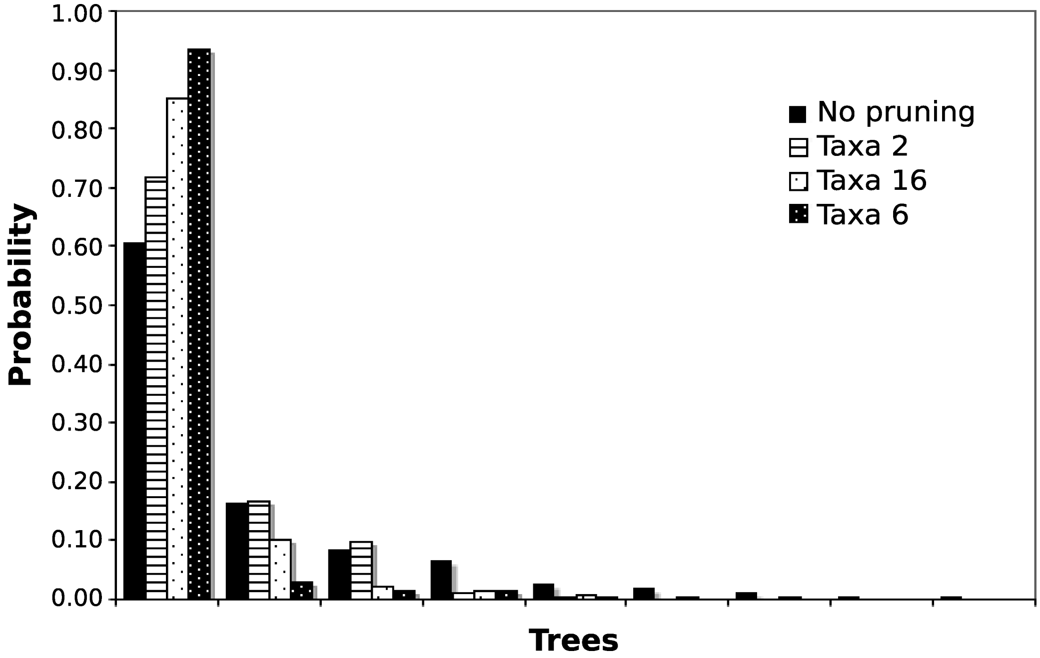


FIGURE 2. Posterior distribution changes after pruning. We show only the 10 most probable trees (others have negligible probability). The three patterned series show the effects of pruning off a single taxon from the posterior distribution. In this particular case, pruning taxon 6 gives a MAP tree with the highest probability, 0.95.

$s_j \in \{1, \ldots, S\}$ where $s_j \neq s_k$ for $j \neq k$. We then prune each of the taxa in this subset from each of the input trees to obtain a new set of trees, $\tau[S_k] = \{\tau_i[S_k]\}$, where $\tau_i[S_k]$ is the subtree obtained by removing the set of taxa S_k from tree, τ_i.

Once we have the pruned list of trees, which is a sample from the distribution of subtrees, we can find the mode of the distribution:

$$P_{MAP}(\tau[S_k]) = \frac{1}{R}\sum_{i=1}^{R} I_M(\tau_i[S_k]) \tag{3}$$

We start with a small value for k, prune various combinations S_k and see which sets most improve the MAP tree probability. For even a moderate number of taxa, the number of distinct sets is enormous—equivalent to the number of combinations of k items chosen from S items. An exhaustive search is possible for at least $k < 3$, but as k increases, the number of possible combinations is far too large, especially as S, the number of taxa in the input trees, increases. Therefore, we developed two stochastic search algorithms to search for sets of taxa that, when pruned, improve the probability of the MAP subtree. One of the algorithms is a Metropolis-Hastings search and the other uses Threshold Accepting. Both use the same general strategy of selecting a set of k taxa, S_k, to prune from the tree, perturbing the set to create S'_k, and determining if this new set improves the probability of the MAP tree as compared to the original set. If removing a particular set of taxa improves P_{MAP}, then we keep this set for the next iteration. If not, then we sometimes keep the new set, according to a set of rules that differs between the two algorithms. We describe the details of each method below.

Metropolis-Hastings Algorithm

The Metropolis-Hastings (MH) algorithm uses the relative probability of the new MAP tree against the former MAP probability as an objective function. The use of MH in this situation is solely for the purpose of optimizing the objective function, rather than inferring the posterior probability distributions of the parameters of interest, as would be the case in a Bayesian framework.

The MH algorithm is implemented as follows:

1. Set number of species removed = k
2. Choose a subset of k species, S_k, from the total list of S species
3. For each unique tree in the sampled set, remove the species in S_k from the tree
4. Calculate the MAP score for this set of removed taxa: $P_{MAP}(\tau[S_k])$
5. Choose number of iterations, i, based on total number of possible subsets
6. Start the MH loop and run for i iterations:
   (a) Create a new list of k species = S'_k, by removing one or more species from S_k and replacing them with an equal number of species from the remaining list of $S - k$ taxa
   (b) Remove the species in S'_k from the original list of trees
   (c) Calculate $P_{MAP}(\tau[S'_k])$
   (d) If $\left(\frac{P_{MAP}(\tau[S'_k])}{P_{MAP}(\tau[S_k])}\right) >$ uniform(0,1) then accept the new

```
      list; else discard and keep the old list
7. If $max(P_{MAP}) < limit$, then $k = k + 1$ and repeat from 1; else quit
```

The measure $P_{MAP}(\tau[S_k])$ is not a true probability under this post hoc sampling model, but functions as the empiric objective function for comparing taxon sets. Proposing the new list of k species to remove, S'_k, involves moving a number of taxa from the current list, S_k, into a holding vector of unremoved taxa and then moving an equivalent number of taxa from the holding vector into S'_k. The number of taxa moved in each step is chosen from a Poisson distribution with rate $0.5(k - 1)$.

The Metropolis-Hastings (MH) ratio consists only of comparing the objective function, which is the MAP tree score for the given sets of taxa: $P_{MAP}(\tau[S_k])$. The probability of accepting a "worse" taxa set is then proportional to the MH ratio.

When the algorithm discovers a new optimum, we discard any saved results and store the new subtree as well as the list of pruned taxa. We keep all of the sets of taxa (and the resulting pruned subtrees) that have the P_{MAP} equal to the optimum value.

Threshold Accepting

In Threshold Accepting (Dueck and Scheuer, 1990), a new state is always accepted if it is within a certain distance, or threshold, of the current state. The threshold is relatively large at the start of the algorithm, allowing exploration in a large region of the sample space and movement between local optima. As the algorithm progresses, the threshold is progressively lowered until the method only accepts solutions that are extremely close to the current solution. Threshold Accepting (TA) is related to the simulated annealing algorithm (Kirkpatrick et al., 1983). In simulated annealing, acceptance is based on the function $e^{\triangle E/T}$, where $\triangle E$ is the difference in objective functions between the two states and T is a temperature parameter. TA simplifies the simulated annealing strategy by always accepting if $\triangle E$ is within a certain threshold. In our method, we accept a new set of taxa if the difference $|P'_{MAP} - P_{MAP}|$ is less than the threshold. For our purposes, this strategy works well because the P_{MAP} values are constrained to the range [0,1] and the threshold can be set to a well-defined difference in probability. TA also eliminates the cost of exponentiation in simulated annealing and random number generation present in simulated annealing and the MH strategy.

The TA algorithm is implemented as follows:

```
1. Set number of species removed $= k$
2. Choose a subset of $k$ species, $S_k$, from the total list of $S$ species (number of possible subsets $= \binom{S}{k}$)
3. For each unique tree in the sampled set, remove the species in $S_k$ from the tree
4. Calculate the MAP score for this set of removed taxa: $P_{MAP}(\tau[S_k])$
5. Choose number of iterations, $i$, based on total number of possible subsets
6. Start TA loop for $i$ iterations:
   (a) Set threshold $= t$ and decrement $= \Delta t$
   (b) Start loop for this threshold:
       i. Create a new list of $k$ species $= S'_k$, by removing one or more species from $S_k$ and replacing them with an equal number of species from the remaining list of $n$ taxa
       ii. Remove the species in $S'_k$ from the original list of trees
       iii. Calculate $P_{MAP}(\tau[S'_k])$
       iv. If $[(P_{MAP}(\tau[S'_k]) - P_{MAP}(\tau[S_k])] > -t$ then accept the new list; else discard and keep the old list
   (c) $t = t - \Delta t$ (if $t = 0$, exit loop)
7. If $max(P_{MAP}) < limit$, then $k = k + 1$ and repeat from 1; else quit
```

Changing the threshold sequence allows for fine-tuning of the algorithm to the particular data set being analyzed. Depending on the starting P_{MAP} and the breadth of the distribution, different initial threshold and decrement values will be appropriate. The starting threshold should be chosen so that the initial acceptance rate is approximately 80% (Dueck and Scheuer 1990). The acceptance rate should then decrease as the threshold decreases. If the acceptance rate is too high, then we accept too many moves and the procedure behaves more like a random search, without moving towards an optimum. If the acceptance rate is too low, the likelihood of getting trapped in a local optimum increases. Although convergence of the TA method is not guaranteed in a theoretical sense (in terms of its asymptotic performance), this does not seem to affect its finite-time performance, which can be quite good (Althofer and Koschnick, 1991; Jacobson and Yucesan, 2004).

Implementation and Output

The method is implemented as MAPminer, a C++ program that takes as input either a Nexus trees block containing a posterior distribution of phylogenies (such as the *.trprobs summary file from MrBayes) or any file containing a list of unweighted phylogenetic trees in newick format (such as the *.t output file from MrBayes). In the latter case, the program will calculate the posterior probabilities of the input trees from the unweighted list before beginning the pruning algorithm. The trees can be rooted or unrooted. It is also possible to pool samples from multiple runs. The user specifies the probability limit for the MAP tree ($prlimit$) and the maximum number of taxa to remove (max_k). The program exits when it reaches $prlimit$ or max_k, whichever comes first. The number of iterations and the length of the burn-in can also be adjusted, and specific taxa can be excluded from pruning (in cases where you are interested in the most well-supported subtree that contains a particular taxon or group of taxa).

For each value of k (number of taxa removed), the program outputs the list of subtrees with maximal P_{MAP}. There may be more than one subtree with the same maximum value for P_{MAP} if more than one set of taxa can be pruned to give the same probability for the MAP tree. For each taxon, the method summarizes the frequency that the taxon is removed from each of the best subtrees. For example, if half of the best subtrees lack a given taxon, then the frequency for that taxon is 0.50.

The running time of the algorithm depends on the number of taxa in the input trees, the number of unique input trees and the shape of the distribution. A more disperse distribution contains a larger number of unique trees to search and will also likely require a larger number of taxa to be pruned, meaning that the runtime will be longer than for a more sharply peaked distribution. Although we did not study this explicitly, the analysis time for the pruning algorithm is likely to be positively correlated with the amount of time required for the initial phylogenetic inference.

METHODS

To illustrate use of the method, we analyzed the posterior distribution of trees from analysis of both simulated and empirical data. For simulated data, we generated five different phylogenies of 50 taxa using a birth-death process (speciation rate = 8.5, extinction rate = 0.5, and sampling frequency = 0.01). For each phylogeny, we simulated 5000 sites under the Jukes Cantor model of evolution using the evolver package of PAML (Yang, 1997). Phylogenetic inference was performed with MrBayes version 3.1.2 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) using the known evolutionary model and a birth-death prior on branch lengths. We note that in this study, the details of the phylogenetic inference method are not critical, as the goal is simply to produce a distribution of trees for the postrun analysis, rather than infer the phylogenies themselves.

Using the output from the phylogenetic inference, we ran MAPminer using both algorithms in order to summarize the posterior distribution. The limit on number of taxa to prune was set at 10 (20% of the total taxa) and the desired limit for the probability of the MAP tree was 95%. For small numbers of removed taxa ($k = 3$ and $k = 4$), we also performed an exhaustive search in order to compare the true frequent subtrees with those found by the stochastic searches.

The empirical data was the posterior distribution of phylogenies for a data set of 85 Carnivore species (Fulton and Strobeck, 2006). We note that we did not perform the phylogenetic analysis, instead obtaining the MrBayes output files directly from the authors of the original paper and using these files as input for our MAPminer program.

MH Settings

Initial testing using the MH algorithm with posterior samples of topologies from the simulated data displayed an extremely high acceptance rate (greater than 90%). In light of the data, this high acceptance rate is expected. The objective function uses the posterior probabilities, which are proportional to the likelihood of the trees. We know a priori that the list of input trees only contains trees selected as reasonable by the original phylogenetic inference. The range of likelihood values for these trees is much smaller than for the full tree space, so comparing the posterior probilities based on these likelihoods should very often accept a proposed state. However, this makes the algorithm inefficient, as we are almost sampling randomly from the possible taxon sets, rather than moving towards an optimal set. In order to increase the sensitivity of the method, we altered the acceptance procedure so that we accept if:

$$\left(\frac{P_{MAP}(\tau[S'_k])}{P_{MAP}(\tau[S_k]}\right)^x > U(0,1) \tag{4}$$

where x is a small integer. This has the effect of exaggerating the differences between P_{MAP} values and reducing the acceptance rate. In a standard MCMC application, altering the ratio in such a way would prevent sampling from the chain (similar to the inability to sample from a heated chain in Metropolis-coupled MCMC [Geyer, 1991]). In this application, however, the goal of the MH algorithm is to search for optimal subtrees, not to sample from the space of subtrees, so the modification simply has the effect of making the search more efficient. When comparing the MH algorithm to the TA algorithm, we used 20,000 iterations with x values of 3, 4, and 5.

TA Settings

The TA algorithm requires an initial choice of threshold and a decrement value. Using the posterior samples from the simulated data sets, we tested starting threshold values of $t = 0.1$ and $t = 0.2$. Either starting value produced acceptance rates in the appropriate range. For the first set of analyses, we used a $t = 0.1$ with a decrement value, $\Delta t = 0.025$. When comparing the TA algorithm with the MH algorithm, we used three combinations of starting threshold and decrement values, (t, Δt): (0.1, 0.025), (0.2, 0.05), and (0.1, 0.02). The number of iterations was 4000 or 5000 per threshold decrement so that the total number of iterations was 20,000 (equivalent to the number used for the MH method, so that the two algorithms could be fairly compared).

RESULTS

Phylogenetic Inference

The phylogenetic inference results from the five simulated data sets are summarized in Table 2. Each posterior sample contained 10,000 total tree topologies, with the first 1000 discarded as burn-in based on plots of the log-likelihood and tree length. Changing the burn-in to 5000 did not significantly alter the size of the credible set or the probability of the MAP tree. For the five posterior samples, the probability of the MAP trees ranged from 0.055 to 0.887 and the size of the 95% credible set varied

TABLE 2. Summary of phylogenetic inference. For each simulated data set, we report the probability of the MAP tree, the size of the 50%, 90%, and 95% credible sets, the number of nodes in the MRC tree with less than 100% posterior probability and the number of multifurcations present in the MRC tree.

Data set	P_{MAP}	Size of credible sets (50%, 90%, 95%)	Nodes <100%	Multifurcations
1	0.206	(23, 34, 58)	4	1
2	0.136	(42, 59, 89)	5	1
3	0.887	(2, 4, 11)	5	0
4	0.055	(131, 178, 267)	6	1
5	0.072	(204, 286, 434)	9	1

from 11 to 434 trees. Data set 3 produced a narrow credible set of topologies with one fairly well-supported tree, whereas the results from data sets 4 and 5 are much more dispersed with a very low probability on the MAP tree. The other two data sets fall in between these extremes. The majority-rule consensus trees for each data set contain at least 4 nodes with uncertain resolution, and all but data set 3 contain a multifurcation in the MRC tree. These results give a sufficiently variable sets of input trees for testing the pruning algorithms.

Tree Pruning, Simulated Data

We first present pruning results for topologies from the simulated data, using the TA algorithm in MAPminer. Subtrees with $P_{MAP} > 0.95$ were found for only two of the five posterior samples before reaching the maximum number of taxa to remove ($k = 15$). The results are summarized in Table 3. Figure 3 illustrates detailed results for data set 1, where the pruning method discovered well-supported subtrees.

It is possible to summarize the taxa present in the subtrees. If the method finds a single subtree, then the output includes that subtree and the list of taxa removed from the tree. If there is more than one equivalent subtree, the output lists the subtrees and the taxa removed to produce each subtree. For multiple subtrees, the output also includes the fraction of subtrees that do not contain each taxon. If a taxon is always absent, then it has lower overall support in the original distribution of phylogenies.

Multiple equivalent subtrees result when different combinations of taxa produce subtrees with the same probability. This is the case when multiple taxa have the same resolution within the tree (in terms of the marginal posterior probabilities of clades). This can happen with distantly related taxa, or it can happen for taxa in the same clade when the poor resolution is on the ancestral node. To separate these two issues, we added an optional preprocessing step to the algorithm. Before starting the pruning, this step identifies all internal nodes that have 100% posterior probability on both the node itself and on all descendent internal nodes. Then, it collapses these nodes (and all descendent nodes), replacing them with a single marker node. This is justified because removing a taxon descendent to one of these internal nodes has no effect on the probability of the MAP tree (removing such a taxon cannot collapse any of the input trees, because 100% of the input trees contain the same pattern). An example of this strategy is shown in Figure 4. The preprocessing step also greatly increases the speed of the algorithm, because it excludes taxa from the search that cannot improve P_{MAP}. The disadvantage is that the resulting subtrees are not of the same size for a given value of k, because removing a marker node is treated as one taxon, whereas we are actually removing the entire clade. This issue could be eliminated by weighting the taxa chosen for removal in terms of the number of descendent nodes in the noncollapsed trees, although the proposal method with such a weighting is not straightforward.

TABLE 3. Summary of pruning results from the five posterior samples of phylogenies. The algorithm ran until reaching a final P_{MAP} of greater than 0.95 or until removing the maximum number of allowable taxa (equal to 15 in this case).

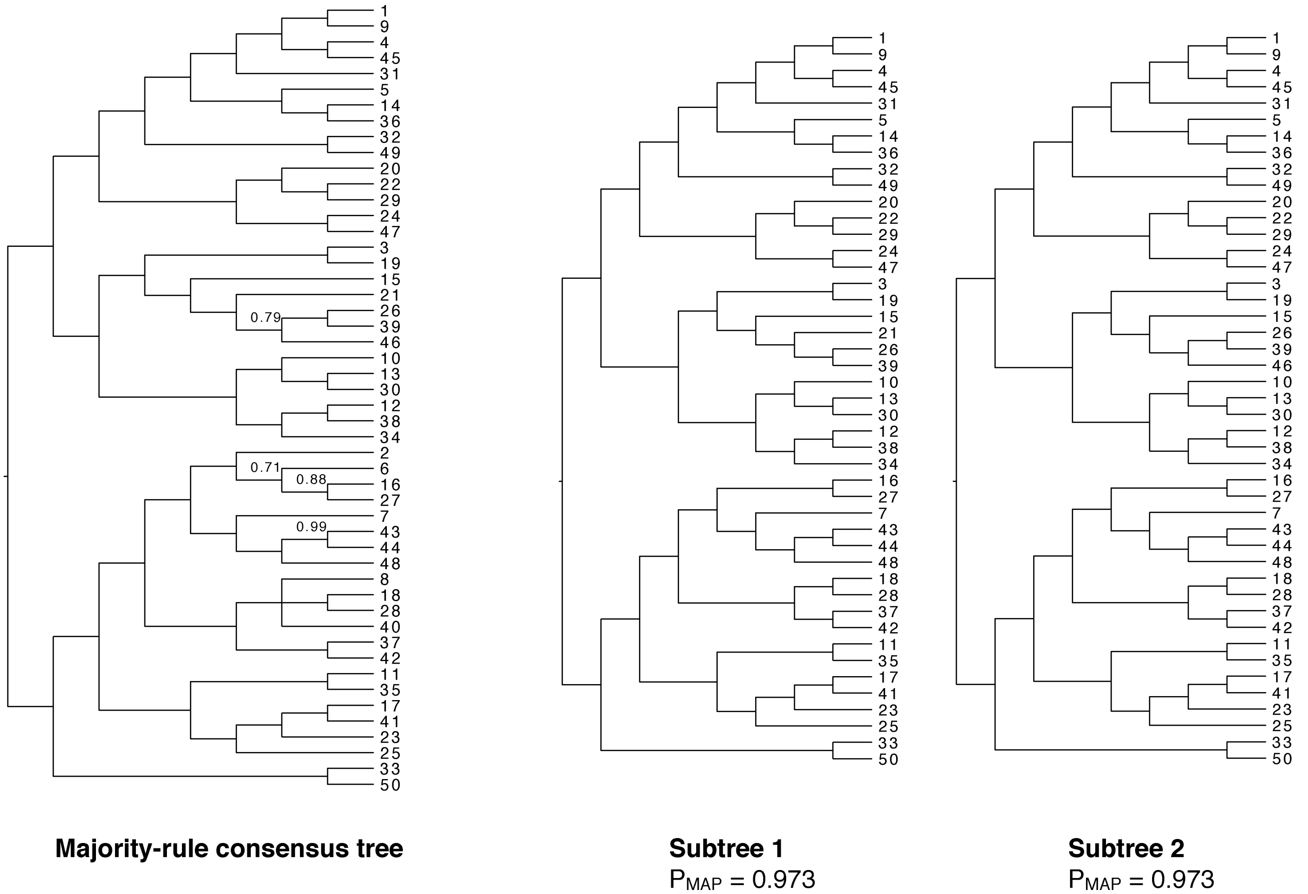
Data set	Starting P_{MAP}	Final P_{MAP}	Taxa removed	Equivalent subtrees
1	0.206	0.9726	5	2
2	0.136	0.6604	15	1
3	0.887	0.9575	4	2
4	0.055	0.6964	15	1
5	0.072	0.9334	15	4

We suggest performing an initial search with a relatively small number of iterations and a large upper limit on k, the number of taxa removed. From these results, determine a smaller range of k values that give probabilities near the desired range. Then, perform an intensive search, with a larger number of iterations in the smaller range of k values. The algorithm should run until no further changes are observed in the optimal P_{MAP} or in the number of equivalent subtrees.

Comparison of Algorithms

We compared the performance of the MH and TA algorithms, using each posterior distribution of trees and the described range of implementation parameters for each method. Performance was judged based on three nested criteria: (1) the maximum P_{MAP} tree found; (2) the number of subtrees with this probability; and (3) the number of iterations required to find the optimal solution. The best method found subtrees with the highest P_{MAP} and the largest number of equivalent subtrees in the smallest number of iterations. For each of the input posterior distributions, the number of best subtrees for a given number of taxa ranges from 1 to 18. The results are summarized in Table 4. The TA algorithm was far more efficient in both finding an optimal P_{MAP} and finding the largest number of subtrees with that P_{MAP}. Of the five data sets, the TA algorithm found the optimal solution in all five cases. The time requirement for each of the algorithms was very similar (data not shown).

For $k = 3$ and $k = 4$ (where k = number of removed taxa), we also compared the two algorithms with the results from an exhaustive search. This gave five data sets × 2 values $k = 10$ comparisons. In 8 out of 10 comparisons, the MH algorithm found the optimal solution, measuring



Majority-rule consensus tree

Subtree 1
P_{MAP} = 0.973

Subtree 2
P_{MAP} = 0.973

FIGURE 3. Majority-rule consensus tree and best agreement subtrees for simulated data set 1. Nodes without explicit posterior probabilities on the MRC tree have probability 1.00. The starting MAP tree probability was only 0.206, despite relatively high posterior probabilities on the MRC tree. Taxa 2, 6, 8, and 18 are pruned to produce both subtrees, as well as taxon 21 for subtree 1 and taxon 46 for subtree 2.

P_{MAP} to four significant digits. The TA algorithm found the optimal solution in 9 out of 10 comparisons. In the analyses where the stochastic and exhaustive search results did not match, the difference in P_{MAP} was less than 0.002. These results are summarized in Table 5.

Empirical Data

The original phylogenetic analysis of the Carnivora data set produced a consensus tree with good resolution between the major groups but with a lack of resolution within groups (Fulton and Strobeck, 2006). Figure 5 shows the well-resolved relationship between the family-level groups. The posterior distribution of phylogenies contained 20,000 total sampled trees from two Metropolis-coupled MCMC chains, with the first half of each chain discarded as burn-in. The MRC tree contains five multifurcations and 13 nodes with marginal posterior probability less than 0.95. The probability of the MAP tree is only 0.001 and the 50%, 90%, and 95% credible sets contain 3471, 7472, and 7972 trees each.

Exploratory pruning analysis (with various threshold annealing parameters and a small number of iterations) indicated that subtrees with probablities near 50% could be found by pruning approximately 30 taxa from the tree. This is a large percentage of the total taxa but is

TABLE 4. Comparison of MH and TA algorithms for five posterior samples of trees, using three different sets of analysis parameter for each algorithm. The *k* value is the largest number of taxa removed for that analysis (with an upper limit of 10). For each algorithm, the table lists the optimum P_{MAP} found by each algorithm. The number in parentheses is the number of subtrees with the given probability. The best result (shown in bold) is the one that first maximizes P_{MAP} and then the number of subtrees. Details about settings for each algorithm are in the text.

Data set	*k*	MH1	MH2	MH3	TA1	TA2	TA3
1	5	0.9189 (1)	0.9643 (1)	0.9726 (1)	0.9726 (2)	**0.9726 (3)**	0.9726 (2)
2	10	0.5569 (1)	0.6233 (1)	0.6233 (1)	0.6233 (18)	0.6233 (1)	**0.6283 (4)**
3	5	0.9545 (1)	0.9452 (3)	0.9545 (1)	0.9545 (2)	0.9545 (2)	**0.9575 (1)**
4	10	0.4885 (1)	0.4891 (1)	0.5275 (1)	**0.5863 (3)**	0.5246 (1)	0.5712 (1)
5	10	0.6977 (1)	0.6977 (1)	0.6393 (1)	**0.8620 (1)**	0.6824 (1)	0.7816 (1)

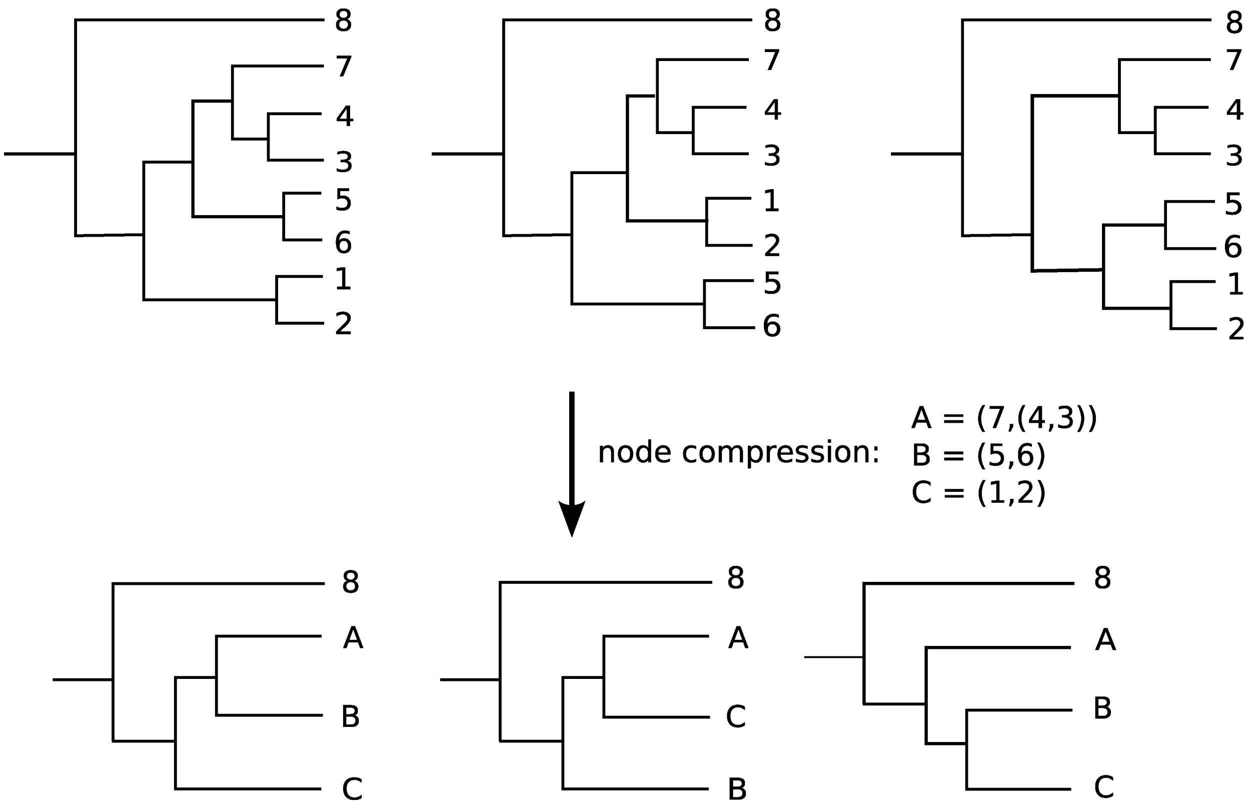


FIGURE 4. Preprocessing step that compresses nodes with posterior probability of 100%.

to be expected given the very broad initial distribution. We then performed a more extensive analysis, with total number of pruned taxa ranging from 25 to 35 and a larger number of iterations. This data set was much more sensitive to the TA settings for initial threshold and increment than were the simulated data sets. With an initial P_{MAP} of 0.0006, the starting threshold needed to be 0.01 or lower for reasonable results. We performed eight separate analyses with starting thresholds ranging from 0.005 to 0.02 and increments that were 0.1% of the starting threshold.

The best result was a subtree with $P_{MAP} = 0.9558$ after the removal of 28 taxa (leaving a tree with 57 taxa). In three independent analyses, MAPminer found seven unique subtrees of size 57 with probabilities greater than 0.95. The original tree and the subtrees are shown in Figure 6. Of the 85 original taxa, there were 19 that were always absent from the subtrees and 50 that were present in all of the high probability subtrees. The 16 remaining taxa were present in some but not all of the seven subtrees. The 19 "always absent" taxa are definite candidates for further sequencing efforts or for removal from the data set before additional phylogenetic inference.

TABLE 5. Comparison of stochastic and exhaustive search strategies for the five posterior samples of trees. Similar to Table 4, each cell contains the optimal P_{MAP} and the number of equivalent subtrees in parentheses. Cells in bold are those analyses that did not find the optimal solution.

Data set	k	Exhaustive	MH	TA
1	3	0.7000 (1)	0.7000 (1)	0.7000 (1)
1	4	0.9143 (2)	0.9143 (2)	0.9143 (2)
2	3	0.3946 (4)	0.3946 (4)	0.3946 (4)
2	4	0.5154 (2)	0.5154 (2)	0.5154 (2)
3	3	0.9405 (2)	0.9405 (2)	0.9405 (2)
3	4	0.9454 (2)	**0.9452 (1)**	**0.9435 (2)**
4	3	0.2987 (2)	0.2987 (2)	0.2987 (2)
4	4	0.3883 (4)	0.3883 (4)	0.3883 (4)
5	3	0.2578 (8)	0.2578 (8)	0.2578 (8)
5	4	0.3303 (8)	**0.3303 (6)**	0.3303 (8)

This empirical data set illustrates an extreme case. There was a very low probability on the MAP tree and several thousand trees in the 90% credible set. As could be expected, the discovery of a well-supported skeleton tree within the distribution required the removal of a relatively large number of taxa compared to the results we saw in the simulated data sets. However, the algorithm was still able to discover several very well-supported subtrees containing 2/3 of the original taxa in the data set.

DISCUSSION

One of the advantages of Bayesian inference is that it produces, not just a point estimate, but a full posterior distribution for the parameters of interest. The posterior probability of particular parameter value (or particular phylogeny) gives a mathematically well-defined and intuitive measure of the support for that value. One

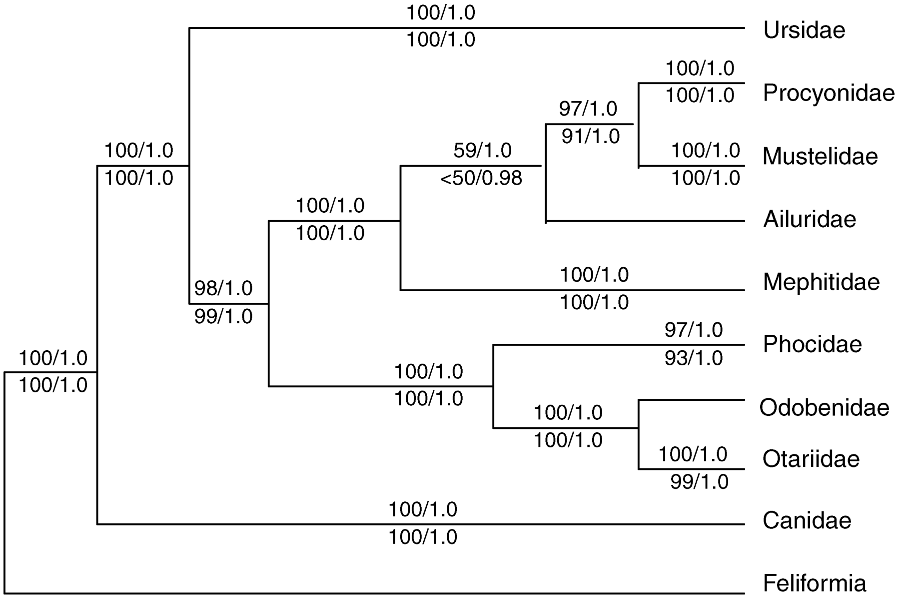


FIGURE 5. Family-level subtree for the 85-taxa Carnivora phylogeny. The first number above each branch is the posterior probability for the full data set (below the branch are the support values for a second analysis with fewer genes). Reproduced from Fulton and Strobeck (2006).

obvious summary statistic for a distribution trees is then the most probable tree, the MAP tree. Another commonly used summary is the majority-rule consensus tree. In either case, relying solely on a single point estimate does not adequately describe the full posterior distribution. The MAP tree often has low overall probability. When reporting a MRC tree, high partition probabilities on a majority of internal nodes in the MRC tree do not imply that there is a strongly supported MAP tree and a narrow distribution of trees. Even with most partition probabilities approaching 100%, there may still be a very large number of unique trees in the credible set.

The ideal result for a Bayesian phylogenetic inference would be a single well-supported tree, defined by a high posterior probablity. In reality, many data sets return a large credible set of trees and no single tree with high probability. Our tree pruning method provides a list of the largest well-supported subtrees that exist within the posterior distribution of phylogenies. The quality of the subtrees is determined by the sum of the posterior probabilities of the input trees that agree with a given subtree. By pruning taxa from the input trees, we can search for optimal agreement subtrees and produce a modified posterior distribution of phylogenies with narrower credible sets and higher probabilities on the MAP trees. The MAP tree (the most probable tree) is the most natural point estimate for summarizing a posterior distribution and is more natural in a Bayesian context than using consensus trees. Bayesian methods provide us with probabilities for entire trees, while the use of consensus techniques causes the whole-tree probabilty to be unreported or lost.

We implement and compare two different algorithms for the subtree search. With the MH methodology, proposed solutions are accepted in proportion to a Metropolis-Hastings ratio of the proposed and current objective functions. Our second algorithm is the threshold accepting (TA) algorithm, which accepts proposed solutions that are within a certain threshold of the current solution, and then progressively lowers the threshold until it searches only in the region of the optimal solution. It too uses a Metropolis acceptance step, but with a different objective function and probability of acceptance. TA is strictly an optimization algorithm, rather than a sampling algorithm that can provide a picture of the underlying distribution. The performance of the TA algorithm was superior, both when compared directly to the MH algorithm and when both methods were compared to an exhaustive search (the "true" result).

The possible outcomes of this type of analysis are threefold. First, the method can produce a single subtree with the specified probability, meaning that the posterior distribution of trees contains one well-supported skeleton tree and a unique set of taxa to prune. Second, there may be multiple subtrees with the same probability (or very similar probabilities). This result is most likely when the unresolved nodes are deeper in the tree, so that resolution involves removing entire clades of taxa rather than individual taxa. A preprocessing node compression step can separate the two explanations. Finally, the method may reach the upper limit on the number of taxa to remove and exit without finding any well-supported subtrees. There can be two reasons for this result. The data may simply not be informative enough to support even a subtree within the distribution, in which case a reevaluation of the input data may be required. The other possibility is nonconvergence of the Markov chain Monte Carlo in the original phylogenetic inference. We encourage users to ensure that the phylogenetic inference method has converged with respect to the log likelihood, model parameters, and other output.

FIGURE 6. MRC tree of the Carnivora. Solid arrows mark taxa that are removed in all 57-taxon subtrees with probability greater than 0.95. Dashed arrows mark taxa marked that are removed from some, but not all of the subtrees. Joined arrows indicate clades with 100% posterior probability, but uncertain position of the group as a whole.

Rather than use the probability of the MAP tree as an end point, it would also be possible to use the size of the credible set. Pruning taxa from the input trees increases the mode of the distribution and also the spread. We would then run the analysis until the number of trees in the 95% credible set was less than a specified limit (or until we reached the maximum number of allowed pruned taxa). This may be a more useful strategy if the results from the Bayesian phylogenetic inference are being used in a program that takes a set of trees as input.

This pruning method shares some properties with the Reduced Consensus methods (Wilkinson, 1994, 1996) for improving bootstrap values on trees. The Reduced Consensus methods create a profile of subtrees based on common *n*-taxon statements (rooted trees) or partitions (unrooted trees) in the original set of trees. The original method (Wilkinson, 1994) was strict, requiring agreement between all input trees, but a later majority-rule method (Wilkinson, 1996) allowed less than 100% bootstrap support on the subtrees. Our method operates with entire subtrees, which are more informative than *n*-taxon statements or partitions. This was previously recognized as a better solution (Sanderson and Schaffer, 2002). Using entire subtrees means that we can place support values both on the full tree and on partitions within the tree. In addition, the methods proposed by Wilkinson have "quite severe limitations on the numbers of taxa and numbers of trees that can be analyzed" (Wilkinson, 1996). Our method can accept thousands of input trees with at least 100 taxa. Also, by limiting our output to the largest agreement subtrees, we avoid the problem of exponential growth of the number of trees in the Reduced Consensus profile with increasing number of taxa in the input trees (Bryant, 1997).

We want to emphasize the important distinction between performing a phylogenetic analysis without a given taxon and the post-analysis pruning method described here. The addition of taxa to a phylogenetic inference problem is known to improve accuracy of the inference (Rannala et al., 1998; Greybeal, 1998; Zwickl and Hillis, 2002). Some studies have explicitly compared the accuracy of trees inferred from a subset of taxa with pruned trees derived from the inference of the full set of taxa (Rosenberg and Kumar, 2001; Pollock et al., 2002). Although the magnitude of the effect is disputed (Rosenberg and Kumar, 2003; Hillis et al., 2003), the studies do indicate that the pruned trees have lower error rates than trees analyzed with only a subset of the data.

Although there is information gained from each taxon in the original phylogenetic inference, the inclusion of some taxa may disproportionately complicate the postrun analysis. This may be particularly worrisome if the added taxa are not the ones of greatest interest to the study. For example, taxa may have been added in an attempt to break up long branches, or simply because the sequences were available. As the number of sequences in public databases continues to grow, it is ever easier to use larger taxon sets to infer trees. The postinference summary method described here allows all available taxa to be included in the original phylogenetic analysis. The tree pruning algorithm then identifies the taxa whose placement is not well supported in the posterior sample of tree topologies. This information can be used to direct future efforts; for example, obtaining additional sequence for the most unstable taxa.

Our aim in this paper is to present a novel method for summarizing the posterior distribution of phylogenies and to encourage developers and users of Bayesian phylogenetic inference to investigate a variety of methods. A summary method may simply involve reducing the distribution to one single tree as a point estimate, such as the MAP tree or the MRC tree. In contrast, we can summarize the entire distribution in a network structure, which retains every partition relationship present in the full distribution at the expense of a more complex interpretation (for example, Huson and Bryant, 2006). Between these two extremes, there is great potential for other methods that balance simplicity of interpretation with maximal information content in ways that are appropriate to the desired application of the phylogenetic results.

The MAPminer method, based on frequent agreement subtrees within the posterior sample of topologies, provides individual well-supported binary trees that can be easily reported or input into other software packages for secondary analyses. A posterior distribution with wide credible sets requires a larger number of taxa to be pruned from the trees in order to discover a well-supported agreement subtree within the distribution. The absence of well-supported subtrees indicates a lack of information in the posterior distribution of phylogenies. This method currently summarizes only the topology information, but we are working to also include information about the distributions of branch lengths. One benefit of this particular postrun analysis is that it allows the original inference of the phylogeny to proceed with all of the available data, yet allows the summary to contain only the results that describe well-supported binary trees.

ACKNOWLEDGEMENTS

The authors would like to thank T. Fulton for kindly providing the output files from phylogenetic analysis of the Carnivore data. Suggestions from Marc Suchard, Joe Felsenstein, and Paul Lewis greatly improved a previous version of this manuscript. Support for this research was provided by Canadian Institute for Health Research grant MOP 44064 and National Institute of Health grant HG01988 to B. Rannala and a University of Alberta Dissertation Fellowship to K. Cranston. MAPminer is available from http://rannala.org.

REFERENCES

Althofer, I., and K.-U. Koschnick. 1991. On the convergence of Threshold Accepting. Appl. Math. Optim. 24:183–195.

Amir, A., and D. Keselman. 1997. Maximum agreement subtree in a set of evolutionary trees: Metrics and effcient algorithms. SIAM J. Comput. 26:1656–1669.

Bryant, D. 1997. Building trees, hunting for trees and comparing trees: Theory and methods in phylogenetic analysis. Ph.D. thesis, University of Canterbury.

Chi, Y., R. R. Muntz, S. Nijssen, and J. H. Kok. 2005. Frequent subtree mining—An overview. Fundam. Inform. 66:161–198.

Cole, R., M. Farach-Colton, R. Hariharans, T. Przytycka, and M. Thorup. 2000. An O(n log n) algorithm for the maximum agreement subtree problem for binary trees. SIAM J. Comput. 30:1385–1404.

Dueck, G., and T. Scheuer. 1990. Threshold accepting: A general purpose optimization algorithm. J. Comput. Phys. 90:161–175.

Finden, C. R., and A. D. Gordon. 1985. Obtaining common pruned trees. J. Classif. 2:255–276.

Fulton, T. L., and C. Strobeck. 2006. Molecular phylogeny of the Arctoidea (Carnivora): Effect of missing data on supertree and supermatrix analyses of multiple gene data sets. Mol. Phylogenet. Evol. 41:165–181.

Geyer, C. 1991. Markov chain Monte Carlo maximum likelihood. Pages 156–163 *in* Computing science and statistics: Proceedings of the 23rd Symposium on the Interface (E. Keramidas, ed.). Interface Foundation, Fairfax Station.

Greybeal, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst. Biol. 47:9–17.

Hewitt, G. M. 2001. Speciation, hybrid zones and phylogeography—Or seeing genes in space and time. Mol. Ecol. 10:537–549.

Hillis, D. M., D. D. Pollock, J. A. McGuire, and D. J. Zwickl. 2003. Is sparse taxon sampling a problem for phylogenetic inference? Syst. Biol. 52:124–126.

Huelsenbeck, J. P. and F. 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.

Jacobson, S. H., and E. Yucesan. 2004. Global optimization performance measures for generalized hill climbing algorithms. J. Global Optim. 29:173–190.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. Science 220:671–680.

Knowles, L. L., and W. P. Maddison. 2002. Statistical phylogeography. Mol. Ecol. 11:2623–2635.

Lee, C.-M., L.-J. Hung, M.-S. Change, C.-B. Shen, and C.-Y. Tang. 2005. An improved algorithm for the maximum agreement subtree problem. Inform. Process. Lett. 94:211–216.

Pollock, D. D., D. J. Zwickl, J. A. McGuire, and D. M. Hillis. 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst. Biol. 51:664–671.

Rannala, B., J. P. Huelsenbeck, Z. Yang, and R. Nielsen. 1998. Taxon sampling and the accuracy of large phylogenies. Syst. Biol. 47:702–710.

Rannala, B., and Z. Yang. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. Mol. Biol. Evol. 14:717–724.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Rosenberg, M., and S. Kumar. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc. Natl. Acad. Sci. USA 98:10751–10756.

Rosenberg, M., and S. Kumar. 2003. Taxon sampling, bioinformatics, and phylogenomics. Syst. Biol. 52:119–124.

Sanderson, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. Mol. Biol. Evol. 19:101–109.

Sanderson, M. J., and H. B. Schaffer. 2002. Troubleshooting molecular phylogenetic analysis. Ann. Rev. Ecol. Syst. 33:49–72.

Shasha, D., J. Wang, and S. Zhang. 2004. Unordered tree mining with applications to phylogeny. Pages 708-719 *in* Data engineering, 2004. Proceedings of 20th International Conference on Data Engineering.

Thorne, J. L., H. Kishino, and I. S. Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15:1647–57.

Wilkinson, M. 1994. Common cladistic information and its consensus representation: Reduced Adams and reduced cladistic consensus trees and profiles. Syst. Biol. 43:343–368.

Wilkinson, M. 1996. Majority-rule reduced consensus trees and their use in bootstrapping. Mol. Biol. Evol. 13:437–444.

Wilkinson, M., J. L. Thorley, and P. Upchurch. 2000. A chain is no stronger than it s weakest link: Double decay analysis of phylogenetic hypotheses. Syst. Biol. 49:754–776.

Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555–556.

Yang, Z., N. Goldman, and A. E. Friday. 1995. Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. Syst. Biol. 44:384–399.

Yang, Z., W. S. W. Wong, and R. Nielsen. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22:1107–1118.

Zwickl, D., and D. Hillis. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51:588–598.

First submitted 29 September 2006; reviews returned 26 February 2007; final acceptance 8 April 2007
Associate Editor: Paul Lewis