Title: Assessing confidence in phylogenetic trees: bootstrap versus Markov chain Monte Carlo

Author(s): Tom Burr, Justin Doak, James Gattiker, William D. Stanbro

Submitted to: 2002 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences, Las Vegas, Nevada, June 24-27, 2002

Assessing confidence in phylogenetic trees: bootstrap versus Markov chain Monte Carlo

Tom Burr, Justin Doak, James Gattiker, William D. Stanbro
Safeguards Systems Group, NIS-7, Mail Stop E541
Los Alamos National Laboratory
Los Alamos, NM 87545
Ph: (505) 665-7865, 667-3377, 665-0604, 667-6779
tburr@lanl.gov, jdoak@lanl.gov, gatt@lanl.gov, wstanbro@lanl.gov
fax: (505) 667-7626

METMBS 2002
RRP 510ME


**Corresponding author:** Tom Burr
**Presenting author:** Justin Doak
**Keywords:** phylogenetic trees, bootstrap, Bayesian, confidence measures
Assessing confidence in phylogenetic trees: bootstrap versus Markov chain Monte Carlo

Tom Burr, Justin Doak, James Gattiker, William D. Stanbro
Safeguards Systems Group, NIS-7
Los Alamos National Laboratory
Los Alamos, NM 87545

Abstract Recent implementations of Bayesian approaches are one of the largest advances in phylogenetic tree estimation in the last 10 years. Markov chain Monte Carlo (MCMC) is used in these new approaches to estimate the Bayesian posterior probability for each tree topology of interest. Our goal is to assess the confidence in the estimated tree (particularly in whether prespecified groups are monophyletic) using MCMC and to compare the Bayesian estimate of confidence to a bootstrap-based estimate of confidence. We compare the Bayesian posterior probability to the bootstrap probability for specified groups in two real sets of influenza sequences and two sets of simulated sequences for our comparison. We conclude that the bootstrap estimate is adequate compared to the MCMC estimate except perhaps if the number of DNA sites is small.

Keywords: Markov Chain Monte Carlo, Bayesian, bootstrap, monophyletic groups, phylogenetic trees

1. Introduction

There are many inference challenges in the context of evolutionary (phylogenetic) trees [1]. Examples include inferring the time to the most recent ancestor and inferring demographic history such as whether the associated population is growing or shrinking [2]. Our interest is on estimating trees using contemporary DNA sequences, and our focus is assessing confidence in aspects of the branching order, such as whether a prespecified group is monophyletic. A group (clade) is monophyletic if in tracing the taxa (the entities being grouped) back in time to its most recent common ancestor (MRCA), we do not encounter any taxa from outside the group.

A tree (Figs. 1a and 1b) describes a branching order and branch lengths. Usually, the branch lengths in the path connecting two taxa are related to the expected genetic distance between them.

Figure 1a is one of three possible unrooted trees with 4 taxa and Figure 1b is one of 15 possible rooted trees with 4 taxa. In a rooted tree there is a known direction of increasing time, radiating outward from the root. There is no known time direction in an unrooted tree. The most common way to root an unrooted tree is to use an outgroup taxon (which is distant from the study taxa but not so different that the separation among the study taxa is severely decreased).

(a) One of three possible unrooted trees.

(b) One of fifteen possible rooted trees.

Figure 1. (a) Unrooted tree with 4 taxa; (b) Rooted tree with 4 taxa.

The tree in Fig. 2 includes 21 nucleoprotein (NP) genes of the influenza virus in three
hosts. For clarity, only 21 sequences are shown from our full data set of 129, with 44 avian (A), 57 human (H) and 28 swine (S). The tree in Fig. 3 includes 17 Hemagglutinin (HA) genes from the human influenza virus. This data is available from www.flu.lanl.gov and is described in [3].

![Figure 2.](image1.png)

**Figure 2.** Rooted tree including 21 NP sequences (7 of H, 7 of S, and 7 of A).

![Figure 3.](image2.png)

**Figure 3.** Rooted tree including 17 HA sequences (8 from 1993, 8 from 1996, and rooted using a 1968 sequence).

We will compare the Bayesian and bootstrap estimates of confidence in the most likely topology for simulated 4-taxon trees under several conditions. We also compare the confidence estimates in whether the 3 species form monophyletic groups in the NP sequences and whether the 1991 HA sequences are monophyletic with respect to the 1996 sequences (and similarly for 1993, 1994, and 1995, each versus 1996).

The next section provides additional background. Section 3 gives details regarding our models of evolution. Section 4 describes our comparison experiments on simulated data using 4 taxa. Section 5 describes the two influenza data sets. Section 6 gives results.

2. Background

A phylogenetic tree consists of a branching order (topology) and branch lengths and can be represented in any of several standard forms. The Newick format (a common tree format) for the tree in figure 1a is ((1,2),(3,4)) and for figure 1b is (((1,2),(3)),(4)). These are the types of tree formats that we use to collect summaries of many trees. The Newick format represents rooted trees, but rooting the tree introduces complications we can avoid for our purposes here. Therefore, we arbitrarily root the unrooted trees but consider ((1,2),(3,4)) to be the same unrooted tree as (1, (3,4), 2), and as (3,(1,2),4).

The topology and branch lengths are each estimated. The concept of uncertainty in branch lengths is straightforward, but the measure of closeness of the estimated topology to the true topology is a large topic. Several metrics have been proposed with a tendency to favor metrics that penalize mistakes near the root or center of the tree more than mistakes near the tips [4]. To focus this work we will consider only the topology and only whether prespecified groups are monophyletic. We will compare bootstrap estimates of confidence based on maximum likelihood (ML) [6] to a Bayesian estimate that uses Markov Chain Monte Carlo (MCMC as implemented in BAMBE [5]).

Felsenstein [7] introduced the bootstrap for assessing confidence in estimated trees. Two studies raised the question of whether the discrete aspect of the topological space would render bootstrap estimates unreliable [8,9]. References [10-12] responded to [8] and [9] and concluded that although the standard bootstrap could be improved with a very computationally intensive double-stage
bootstrap, the bootstrap is still valid (unbiased and not too imprecise). The basic bootstrap strategy is to sample each site with replacement as bootstrap sample 1. Apply the tree estimation algorithm to each of (typically) 100 to 1000 bootstrap samples and summarize results over all bootstrap samples. Several complications arise in practice. The number of unrooted topologies for $T$ taxa is

$$\prod_{i=3}^{T} (2i - 5)$$

which grows prohibitively large for more than approximately 100 taxa. Therefore, unless there are only a few taxa, existing ML algorithms do not evaluate all topologies and choose the topology having maximum likelihood. Instead, heuristic methods limit the search to a small number of candidate topologies. Sophisticated stochastic searches such as those using simulated annealing [13] improve matters, but do not remove the complication that the tree-finding algorithm cannot evaluate all topologies. In contrast, the BAMBE strategy uses MCMC to search the posterior distribution (all topologies considered equally likely in the prior probability) over branch lengths and topologies. Second, the discrete nature of the decision space for choosing topology 1 versus topology 2 means that the estimated topology cannot change in a smooth way as the input sequences vary. Newton [10] established a large deviation result for the bootstrap empirical distribution in a finite sample space, thereby validating bootstrapping in some phylogenetic inference settings. Previous to this result, existing theory did not support the bootstrap because of the discrete nature of the space of tree topologies and because of the types of questions involved.

Both [8] and [9] reported empirical evidence on simulated data that bootstrap proportions provide unbiased but imprecise estimates of repeatability but biased estimates of accuracy. If the phylogenetic estimation method is consistent (the estimated topology converges to the true topology as more DNA sites are used) the bootstrap appeared to underestimate high accuracies and overestimate low accuracies. Results reported in [8] and [9] indicated an empirically observed bootstrap bias (in the direction of underestimating confidence) in simulated data in addition to overdispersion. Results in [10] and [11] presented theoretical results to refute the bias claim. However, both [10] and [11] were forced to use theoretical arguments, largely invoking a simpler situation involving sampling from a normal distribution and inferring whether its mean was positive on the basis of the sample mean. We therefore believe that this experiment comparing Bayesian confidence measures to bootstrap confidence measures is valuable. Such a comparison was not available when [8-11] were published.

Efron et. al. [11] showed that the discrete space aspect does not invalidate the bootstrap in this context and that it does not lead to a systematic over or under estimate of confidence, which contradicted the results in [8] and [9]. Efron et. al. [11] also suggested modifications involving multiple stage bootstrapping (bootstrapping the bootstrap samples which can become computationally prohibitive) to improve the bootstrap. Regarding the “method must be consistent” assumption, [14] showed that use of a wrong model leads to underdispersion. As an example, “long branches attract,” which means the long branches tend to be grouped together regardless of whether that is the correct topology. This phenomenon will impact any method so we will deliberately avoid it in simulated data by using the same model to estimate the tree as we do to simulate the sequences. However, we can never be sure it is not in effect with real data because real data never follow any model exactly. Recently, likelihood ratio tests have been used to select models [1], but even in the best of real-data cases, we expect at least some degree of model misspecification.

3. Substitution Models
We consider aligned sequences of DNA data with no gaps, such as the example with 4 taxa in Table 1.
An evolutionary model [15] specifies the probability per unit time of a given nucleotide mutating to another nucleotide. A fully parameterized 4-by-4 substitution probability matrix would have 12 freely varying entries, but most current analyses and software restrict the number of free parameters to 1 to 5. One of the most common currently used is a 5-parameter model (F84) that specifies positive $\mu$, $\kappa$ and the four nonnegative nucleotide frequencies $\pi_A$, $\pi_C$, $\pi_G$, and $\pi_T$ (that sum to 1). The parameter $\mu$ determines the average rate of change (usually assumed to be constant over time and to be the same for each taxa). Typically, the number of changes per unit time is assumed to be Poisson distributed with mean $\mu$. The parameter $\kappa$ allows for purine-to-purine (A to/from G or C to/from T) or pyrimidine-to-pyrimidine mutations (transitions) to be different than purine-to-pyrimidine or pyrimidine-to-purine (transversions). Although the $\pi$s can be estimated using the observed nucleotide frequencies, the best way to estimate $\mu$ and $\kappa$ requires access to an outgroup taxa.

Recently, it has been demonstrated that allowing $\mu$ to vary across site can be important [16]. Typically, $\mu$ is assumed to have a gamma distribution with the parameter $\gamma$ determining its variation across sites. We will assume that $\mu$ is constant across sites in all cases rather than complicate our comparison by having MCMC and ML+bootstrap use different schemes to estimate $\gamma$. There are several reasons that models are likely to be misspecified to some degree in practice [15]. However, we know the exact model in our simulated data so we can specify the correct model.

One difficulty in evaluating related publications is that methods perform differently according to how closely the assumed model agrees with the actual model. For example, a maximum likelihood method using the “wrong model” leads to biases, which favor recovering certain favored topologies (such as long branches tending to group together). If the true topology happens to agree with one of these “favored” topologies then the variation among ML fits on bootstrap samples will be unrealistically low.

Here, we use the same model in BAMBE and DNAMLk ([18] and see Appendix) to estimate the phylogeny and in the SeqGen code to generate the data for a given topology. We use the model denoted F84 with no rate heterogeneity and a constant substitution rate over time (“molecular clock”). Therefore, our results with the simulated data represent a safe comparison and our results on real data are as safe as possible with real data.

### 4. Simulated Data

We consider 4-taxon trees with different branch lengths. For each of 2 sets of branch lengths, we:

(a) simulated 6 sets of 4 sequences using SeqGen [17];

(b) used BAMBE to estimate the probability of each of the 3 possible topologies;

(c) used DNAMLk in PHYLIP [18] to find the maximum likelihood topology for each of 100 to 1000 bootstrap samples;

(d) compared the probability for the true topology based on the MCMC-based estimate to the ML + bootstrap based estimate.

For each of the 2 sets of relative branch lengths, the true topology was ((1,2),(3,4)) and because we estimate unrooted trees, the other two possible topologies are ((1,3), (2,4)) and ((1,4), (2,3)). We used $n = 300, 200, and 100$ sequences and replicated the SeqGen process twice. For each simulated data set, we applied both estimation schemes twice, for a total of 12 evaluations per set.

<table>
<thead>
<tr>
<th>Taxa Code</th>
<th>Mutually Aligned DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CCCGATCAAATT...</td>
</tr>
<tr>
<td>2</td>
<td>CACGCTCAAATT...</td>
</tr>
<tr>
<td>3</td>
<td>TGCGAACAAATT</td>
</tr>
<tr>
<td>4</td>
<td>TGCGAACAAATT...</td>
</tr>
</tbody>
</table>

Table 1: Example DNA sequence data.
5. Real Data
We consider the influenza sequences used in Figures 2 and 3. To avoid having all high 0.999 or higher confidence, we subsampled the DNA sites to have from 979 to 35 sites, depending on the case. We also varied the subset of taxa present. For example, with HA, we compared the 7 1991 taxa to the 8 1996 taxa, and the 7 1993 taxa to the 8 1996 taxa, etc., and similarly for the NP sequences. As an aside, the Bayesian estimate of confidence for both the NP (Figure 2) and HA (Figure 3) clades is 100%. The bootstrap estimate is 98% and 99%, respectively.

6. Results
Figure 4 is the MCMC-based probability estimate vs the ML + bootstrap based probability estimate (of the probability of either the true topology (simulated data) or a prespecified group (real data)). The simulated data for case (a) used ((T1:0.3:T2:0.3):0.1,T3:0.3,T4:0.4) and for case (b) used (((T1:0.1:T2:0.2):0.05,T3:01,T4:0.2), where the branch lengths are specified following the taxa (T1 is taxa 1) designation.

If we assume there is no difference in the MCMC and ML + bootstrap estimates then we can compute a reference distribution against which to compare the observed sum of differences for each of the four cases in Figure 4. This reference distribution is shown in Figure 5. Qualitatively, the conclusion is the same as in Figure 4: Case 1 exhibits a tendency for the ML + bootstrap estimate to be higher than the MCMC estimate, but there is no evidence of a difference for the other 3 cases. Formally, the “p-value” is approximately 1% or less for case 1 for testing equality of the two methods whether we fit a regression as in Figure 4 and test the hypothesis: slope = 1, or use the reference distribution (obtained from randomly assigning the labels MCMC or ML+bootstrap for all 2^n possibilities where n is the number of comparisons) as in Figure 5.

7. Conclusions/Summary
Although [8] and [9] reported results suggesting that ML + bootstrap will understate the confidence in the most likely topology, [10] and [11] reinterpreted the results in [8] and [9] and provided a theoretical basis in a simpler setting than tree estimation to defend ML + bootstrap. This study provided a direct comparison of the ML + bootstrap based probabilities to probabilities obtained from MCMC via BAMBE and except for cases having few DNA sites, we also conclude that ML + bootstrap is adequate in this context.

Other issues arise in discussing the performance of the ML + bootstrap estimates compared to the MCMC-based Bayesian estimates. Reference [9] repeated the simulation of DNA sequences for each of 21 trees (topology and branch lengths) many times to study two interpretations of bootstrap proportions: repeatability and accuracy. The repeatability was defined as the probability that a specified group would be found in repeated analyses of independent samples of sequences (includes the sampling implicit in the randomness in realized substitutions due to the evolutionary model). The accuracy was defined as the probability that a group is contained in the true tree. Reference [11] discussed the variation of the bootstrap estimate around its mean and of the bootstrap estimate around the true topology.

A serious problem in assessing ML + bootstrap is that the ML method will become biased for small numbers of DNA sites. This bias can result in higher precision around the wrong answer. This is a candidate explanation for the outlier points in Figure 4a. The two points near (0.7, 0.4) occurred for the case of 100 DNA sites and possible bias in ML is a possible explanation.

Finally, it became clear during this study that when we simulate new data from the same model as in cases 1 and 2, we can get drastically different measures of evidence for the true topology (ranging from near 1 to near 0). We mention this because of the
Figure 4. Probability of true topology (simulated data) or of prespecified group (real data) using MCMC vs ML+bootstrap for (1) Simulated tree 1; (b) Simulated tree 2; (3) HA data; (4) NP data.

Figure 5. The sum of $p_{\text{MCMC}} - p_{\text{ML+BS}}$ compared to reference distribution for the same 4 cases (a)-(d). The arrow indicates the observed value of the sum of the differences between the MCMC based estimate and the ML + bootstrap estimate.
common misconception [9] that the ML + bootstrap strategy mimics the following process: a hypothetical second realization of the genetic sampling processes that generate the data plus fitting the ML to this new data. Instead, either MCMC or ML + bootstrap approximate the probability distribution for each topology given the genetic data. There could be drastically different confidence in the estimated topology in other realizations of the genetic sampling process.

References

Appendix. Model Details
We used Seqgen [17] to simulate data with inputs: seq. length = 100, 200, or 300; branch lengths as specified in the text; substitution model F84; transition/transversion ratio = 2; πA= πC = πG = πT = 0.25, no heterogeneity in substitution rate across sites and a molecular clock (constant substitution rate).
This same model was then assumed in both BAMBE and DNAMLk (we needed to choose an evolutionary model that both BAMBE and DNAMLk could accommodate). For all of the BAMBE runs, we checked for convergence of the Markov Chain by increasing the burnin time until the topology probabilities converged. For the DNAMLk runs, we jumbled the input order of the taxa because the code does not guarantee that the solution found is the global maximum likelihood. Also, we used a small number of taxa in each case, so we can be reasonably confident that the overall performance was the same as if the maximum likelihood solution were guaranteed to be found for each run.