First-Order Correct Bootstrap Support Adjustments for Splits that Allow Hypothesis Testing When Using Maximum Likelihood Estimation

Edward Susko*

Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada

***Corresponding author:** E-mail: susko@mathstat.dal.ca.

Associate editor: Jeffrey Thorne

Research article

Abstract

The most frequent measure of phylogenetic uncertainty for splits is bootstrap support. Although large bootstrap support intuitively suggests that a split in a tree is well supported, it has not been clear how large bootstrap support needs to be to conclude that there is significant evidence that a hypothesized split is present. Indeed, recent work has shown that bootstrap support is not first-order correct and thus cannot be directly used for hypothesis testing. We present methods that adjust bootstrap support values in a maximum likelihood (ML) setting so that they have an interpretation corresponding to P values in conventional hypothesis testing; for instance, adjusted bootstrap support larger than 95% occurs only 5% of the time if the split is not present. Through examples and simulation settings, it is found that adjustments always increase the level of support. We also find that the nature of the adjustment is fairly constant across parameter settings. Finally, we consider adjustments that take into account the data-dependent nature of many hypotheses about splits: the hypothesis that they are present is being tested because they are in the tree estimated through ML. Here, in contrast, we find that bootstrap probability often needs to be adjusted downwards.

Key words: maximum likelihood, topology test, bootstrap support, splits.

Introduction

In phylogenetic analyses, topological uncertainty is most frequently represented by bootstrap support or bootstrap probability (BP). Trees are presented with BP along edges or, equivalently, for the splits of the taxa into the two groups on either side of the edge. It is clear that BP correlates with topological uncertainty. Large BP implies that estimated splits are not dependent on one or a few sites being present. Such splits arise even when some sites are repeated and other sites are dropped. What has been less clear is how large BP needs to be in order to conclude that there is significant evidence that an edge is present in the tree. Felsenstein and Kishino (1993) raised the possibility that 1-BP might be interpreted as a P value for the test of the null hypothesis that the split is not present. For sometime, it was believed that 1-BP for splits was first-order correct as a P value (Efron et al. 1996): that approximately, with large sequence lengths, 1-BP has the properties of a P value. The theory for this conclusion was fleshed out in Efron and Tibshirani (1998) but applied to trees only indirectly by analogy. Recently, Susko (2009) showed that 1-BP for splits is not first-order correct because of the unusual nature of tree space.

For maximum likelihood (ML) estimation, a by-product of Susko (2009) was determination of the large sequence-length distribution of BP for a split. As will be illustrated here, this distribution can be used to convert a BP value into an adjusted bootstrap probability (aBP) that is first-order correct. This implies that if a split is not present in the true underlying tree, one expects an aBP of 95% or more only 5% of the time. In the present article, we outline how the results of Susko (2009) can be used to adjust BP, discuss approximations of the key information matrix quantities required for implementation, and provide examples. We also present the results of simulations that suggest that adjustments tend to be fairly constant across various substitution models, rates-across-sites parameters, frequencies, and even when amino acid data are considered instead of nucleotide data. Extreme edge-length settings, however, illustrate that there are situations where adjustments should take into account information about the estimated tree.

Assumptions

The methods for adjusting BP successively consider each internal edge of the tree. Because calculations are done separately for each edge, it suffices to consider a single edge of interest in describing the main ideas. The situation is then illustrated in figure 1. The edge pointed to with an arrow in figure 1*A* is treated as the edge of interest. In the theory for limiting distributions, the edge length corresponding to it is of length 0. The large–sequence length calculations assume that other edges are well resolved. For example, for the generating topology in figure 1*A*, the split of taxa 1, 5, and 2 is considered as poorly resolved but the split of taxa 1 and 5 from the rest is not. With large sequence lengths, the ML tree is then sure to contain the split of 1 and 5 from the rest but there will remain positive probability that it will estimate one of the topologies in figures 1*B*–1*D*. More generally, we assume that the generating tree is Topology 1 of figure 1 and that, with large sequence length,

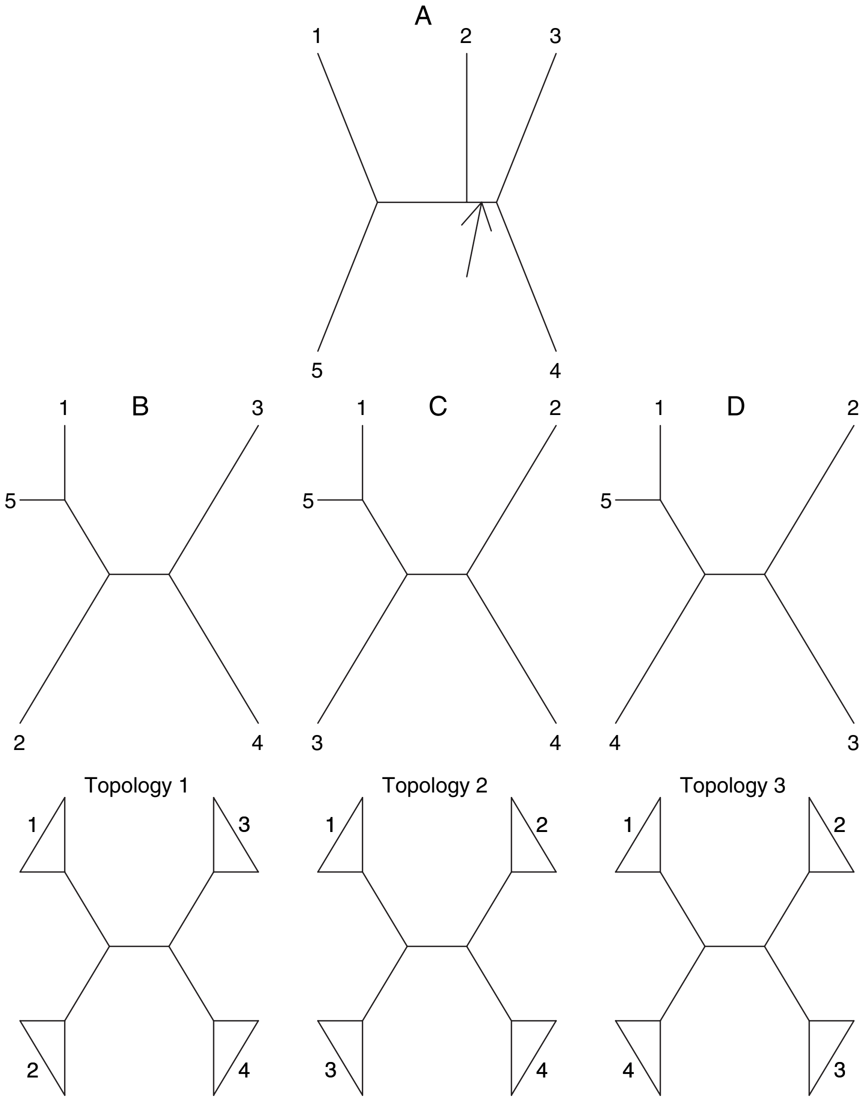

Fig. 1. (*A*) The split separating taxa 1, 5, and 2 from the rest is of interest, leading to the three competing topologies in (*B*–*D*), all of which have taxa 1 and 5 separated from the rest. More generally, there are three competing topologies associated with the split of interest, with the subtrees, indicated by triangles, all considered fixed.

all the splits in the subtrees 1–4 are correctly estimated. The competing topologies are then Topologies 1–3.

What is desired here is that 1-aBP has an interpretation as a P value. Formally, when calculated under the null hypothesis that the split is not present, the probability that 1-aBP is less than α should be α; for first-order correctness, this interpretation should be approximately correct as sequence length gets large. The restriction that the length of the edge of interest is 0 may seem unusual, but it is consistent with hypothesis testing in more standard settings. For instance, in testing the null hypothesis that a mean is less than or equal to 0, P values are calculated assuming that the mean is 0. This is because a mean of 0 is the closest parameter, under the null hypothesis, to the alternative hypothesis space. Similarly here, a tree with zero-length edge of interest, with other edges being well resolved, is at the boundary between the null and alternative hypotheses.

There are trees on a boundary of tree space that satisfy the null hypothesis but have more unresolved edges than just the edge of interest. When the true tree has more than one unresolved edge, the limiting results for BP are likely different than those of Susko (2009), particularly if at least one of the additional unresolved edges neighbors the edge of interest. Because the true tree is unknown, the natural choice is the closest tree to the estimated tree that satisfies the null hypothesis. If the tree has no zero-length edges, the closest tree to it is the one that sets the middle edge to 0 and leaves all other edges unchanged. This can be made formal by considering tree distances like the branch-length distances of Kuhner and Felsenstein (1994) or those defined in Billera et al. (2001).

The model assumed throughout is a conventional continuous-time Markov chain model where evolution is independent and according to the same process across sites. Results do not directly apply to models that allow different processes or dependence across sites, such as models that incorporate protein structure like that of Robinson et al. (2003) and the autocorrelated rates models of Felsenstein and Churchill (1996). Although the result may not apply when differing processes occur at different sites, changes in process that arise in an independent fashion from some distribution are allowed. These include the commonly used gamma rates-across-sites model of Yang (1994) where different rates apply to different sites but arise in an independent fashion from a gamma distribution.

More specifically, for a test of any edge, the assumptions are that the true generating tree contains only that one edge of zero length and that a conventional continuous-time Markov chain describes the process of evolution along edges. The requirement is that derivatives, of any order, of the likelihoods be obtainable and that pattern probabilities be positive. This is assured for any general time reversible (GTR) model where all the rates of exchange and frequencies of characters are positive. It is also assumed that the information matrix or expected second derivative matrix of the negative log-likelihood is invertible. This can easily be checked for any parameter setting and has been true for all cases I have encountered. Finally, it is assumed that for a given topology, edge lengths are identifiable. That is, no two sets of differing edge lengths give the same probabilities for every possible alignment. Chang (1996) shows that equal-rates models are identifiable, and for the gamma rates-across-sites model of Yang (1994), Allman et al. (2008) have established identifiability.

Notation

In order to present the methods and discuss issues of implementation, we need to recall the notation of Susko (2009). Let $l(\boldsymbol{t}_0;j)$ denote the likelihood for topology j in figure 1. Here $\boldsymbol{t}_0$ denotes the true generating edge lengths for the tree. We will denote the first partial derivatives of the log likelihood, $l(\boldsymbol{t};j)$ evaluated at $\boldsymbol{t}_0$, as S_j and $\mathbf{S}_\mathrm{e}$, where $\mathbf{S}_\mathrm{e}$ is a row vector giving the partial derivatives with respect to all the external edges and S_j denotes the derivative at the middle edge of interest. Since edge-lengths are always restricted to be positive, the derivative S_j is only well-defined when limits are taken from above. Nevertheless, the same rules of differentiation can be applied to find it. Also, because the middle edge is of length 0, the $\mathbf{S}_\mathrm{e}$ components are the same regardless of which topology j is used to calculate them.

The information matrix that determines the distribution of BP is decomposed as

$$\begin{bmatrix} I_\mathrm{e} & \boldsymbol{I}_{j\mathrm{e}}^T \\ \boldsymbol{I}_{j\mathrm{e}} & I_j \end{bmatrix}. \tag{1}$$

All the entries give the expected values of $-1/n$ multiplied by the matrix of second partial derivatives of the log-likelihood, where n is the number of sites. For $\mathbf{I}_\mathrm{e}$, these are second partial derivatives for external edges (all edges other than the edge of interest)

$$[I_\mathrm{e}]_{kl} = -E\left[\frac{\partial^2}{\partial t_k \partial t_l} l(\boldsymbol{t}_0;j)\right] \Big/ n, \tag{2}$$

where t_k and t_l are the kth and lth external edge. Similarly,

$$[\boldsymbol{I}_{j\mathrm{e}}]_l = -E\left[\frac{\partial^2}{\partial t_j \partial t_l} l(\boldsymbol{t}_0;j)\right] \Big/ n \tag{3}$$

gives the second partial derivative for the length of the edge of interest and the lth external edge. The final entry

$$I_j = -E\left[\frac{\partial^2}{\partial t_j^2} l(\boldsymbol{t}_0;j)\right] \Big/ n \tag{4}$$

gives the second derivative for the middle edge length.

Let $p_k(\boldsymbol{t};j)$ denote the probability of site pattern k calculated for topology j using edge lengths $\boldsymbol{t}$. Note that because $\boldsymbol{t}_0$ contains a zero-length edge of interest, $p_k(\boldsymbol{t}_0;j)$, which we will denote as p_k, is the same no matter which topology j is used. The key variance–covariance matrix used for calculation of the distribution of BP is

$$\begin{aligned}[\Sigma^c]_{ij} =& [I_i^c \cdot I_j^c]^{-1/2} \\ & \times \left[\sum p_k^{-1} \frac{\partial}{\partial t_i} p_k(\boldsymbol{t}_0;i) \frac{\partial}{\partial t_j} p_k(\boldsymbol{t}_0;j) - \boldsymbol{I}_{ie} I_\mathrm{e}^{-1} \boldsymbol{I}_{j\mathrm{e}}^T\right].\end{aligned} \tag{5}$$

Here partial derivatives are being taken with respect to lengths of the edge of interest for the two topologies i and j. The values I_i^c and I_j^c are defined through

$$I_j^c = I_j - \boldsymbol{I}_{j\mathrm{e}} I_\mathrm{e}^{-1} \boldsymbol{I}_{j\mathrm{e}}^T.$$

Finally, we define the standardized score for the jth topology as

$$V_{jn}^c = [I_j^c]^{-1/2} [S_j - \boldsymbol{I}_{j\mathrm{e}} I_\mathrm{e}^{-1} \boldsymbol{S}_\mathrm{e}^T], \tag{6}$$

which are the key summary statistics that determine the ML topology.

Bootstrap Support Probability Calculation

The details of the limiting properties of BP are given in Susko (2009) but the basic ideas can be outlined as follows. It turns out that, up to terms that become small for large sequence length, the difference in maximized likelihoods is determined by the standardized scores in equation (6). Specifically, the difference in maximized log-likelihoods for topologies j and k is

$$\begin{aligned} l(\hat{\boldsymbol{t}}_j;j) - l(\hat{\boldsymbol{t}}_k;k) \approx & [V_{jn}^c]^2 I\{V_{jn}^c \geqslant 0\}/2 \\ & - [V_{kn}^c]^2 I\{V_{kn}^c \geqslant 0\}/2, \end{aligned} \tag{7}$$

where the notation $I\{x \geqslant 0\}$ indicates a value that is 1 or 0 according to whether $x \geqslant 0$ or not. What we see is that topology j will have a larger likelihood than topology k if $V_{jn}^c \geqslant 0$ and $V_{jn}^c \geqslant V_{kn}^c$. Thus, j will give the estimated topology if, in addition, $V_{jn}^c \geqslant 0$ and $V_{jn}^c \geqslant V_{kn}^c$, for all other topologies k.

Although the standardized score vector, $\boldsymbol{V}_n^c$, is determined from an alignment, for large sequence lengths, it has an approximate normal distribution with mean $\mathbf{0}$ and variance–covariance matrix Σ^c: a $N(\mathbf{0}, \Sigma^c)$ distribution. Equation (7) indicates that, for large sequence lengths, the ML estimate of the topology is determined by the standardized scores. Thus, in terms of the relevant quantities for tree determination, random generation of an alignment from the true tree is approximately the same as random generation of $\boldsymbol{V}_n^c$ from a $N(\mathbf{0}, \Sigma^c)$ distribution.

For a fixed alignment, the large sequence-length approximations to the bootstrap distribution mirror the approximations above. In a similar fashion, the difference in likelihoods for topologies j and k, for a bootstrap alignment, is approximated by equation (7) but with the entries of $\boldsymbol{V}_n^c$ being replaced by those of $\boldsymbol{V}_n^{c*}$, the standardized score for the bootstrap alignment. Although the standardized score vector, $\boldsymbol{V}_n^{c*}$, is determined from a bootstrap alignment, it too has an approximate normal distribution. The variance–covariance matrix for this normal distribution is still Σ^c but the mean is $\boldsymbol{V}_n^c$; it depends upon the alignment that bootstrapped data are generated from. Equation (7) indicates that, for large sequence lengths, the ML estimate of the topology for the bootstrapped sample is determined by the standardized scores, $\boldsymbol{V}_n^{c*}$. Thus, in terms of the relevant quantities for tree determination, random generation of a bootstrap alignment from a fixed alignment depends on the fixed alignment only through its standardized scores $\boldsymbol{V}_n^c$ and is approximately the same as random generation of $\boldsymbol{V}_n^{c*}$ from a $N(\boldsymbol{V}_n^c, \Sigma^c)$ distribution.

As a consequence of the approximations discussed above, we can obtain the probability of observing BP larger than x for the jth topology through the following steps:

1. Generate a large number of trivariate random vectors $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_B$ from a $N(\mathbf{0}, \Sigma^c)$ distribution.
2. For each $\boldsymbol{V}_b$, BP for the jth split is approximated by the probability that $[\boldsymbol{V}_b^*]_j$ is the largest element of $\boldsymbol{V}_b^*$, given that at least one of the $\boldsymbol{V}_b^*$ is positive. This probability is calculated for $\boldsymbol{V}_b^*$ having a $N(\boldsymbol{V}_b, \Sigma^c)$.
3. The proportion of BP $\geqslant x$ among cases where at least one component of $\boldsymbol{V}_b$ was positive gives an approximation to the probability that BP is at least as large as x under the null hypothesis that the split is not present.

Adjusted Bootstrap Support

To convert BP to a first-order correct P value, we can use the distribution of BP. The quantity aBP that we use is the limiting probability, under the null hypothesis, of BP being larger than the actual observed BP obtained for the split of interest. It can be calculated through Steps 1–3 above for any x. The idea here is similar to Beran (1988) where it was shown that transforming a test statistic into a new test statistic by substituting it into its large-sample distribution function can often give a more precise uniform distribution approximation to the resulting P values. Here, we seek not a more precise approximation but that 1-aBP be interpretable as a P value for large sequence lengths.

Let $F(x)$ denote the limiting cumulative distribution of BP: $F(x) = P(\text{BP} \leqslant x)$. Because, in the limiting distribution, BP is a continuous probability transformation of a multivariate normal random vector (see Step 2 of the previous section), $F(x)$ will be a continuous distribution function. The quantity aBP can be expressed in terms of $F(x)$ as

$$\text{aBP} = 1 - F(\text{BP}). \tag{8}$$

A continuous cumulative distribution function, like F, has an increasing inverse function: there is a function F^{-1} with the property that if $y = F(x)$, for $0 \leqslant y \leqslant 1$, then $F^{-1}(x) = y$. Because it is increasing, $F(x) < y$ if and only if $x < F^{-1}(y)$. Thus,

$$\begin{aligned} P(1 - \text{aBP} < \alpha) &= P(F(\text{BP}) < \alpha) \\ &= P(\text{BP} < F^{-1}(\alpha)) \\ &\approx F(F^{-1}(\alpha)) = \alpha, \end{aligned}$$

which establishes that, for large sequence lengths, 1-aBP can be interpreted as a P value.

Approximation Issues

The calculation of aBP requires the information matrix in equation (1). This matrix is calculated by taking expectations at the true generating edge lengths $\boldsymbol{t}_0$. In practice, this is not possible because the true edge lengths are unknown. However, with large sequence lengths, the distribution is still approximately correct if $\boldsymbol{t}_0$ is replaced by the estimated edge lengths for each external edge and 0 for the internal edge. A second practical difficulty is that expectations require summation over all possible site patterns. For nucleotide data and a small number of taxa, this is feasible. For instance, with eight taxa, there are 65,536 patterns to sum over. With larger numbers of taxa or amino acid data, however, the number of patterns make exact calculation infeasible. Approximation of the information matrix can be accomplished in two ways. Both approaches involve replacing expectations of the second partial derivatives in equations (2)–(4) by observed values for some data set. For instance, in equation (4), the expectation is replaced by $-\frac{\partial^2}{\partial t_j^2} l(\hat{\boldsymbol{t}}; j)/n$ for a data set. This derivative can be expressed as

$$-\sum_k \hat{p}_k \frac{\partial^2}{\partial t_j^2} \log[p_k(\hat{\boldsymbol{t}}_0; j)],$$

where $\hat{\boldsymbol{t}}_0$ indicate the ML edge lengths but with the middle edge set to 0. In addition, $\hat{p}_k$ is the proportion of times the pattern k arose in the data set. Because $\hat{\boldsymbol{t}} \approx \boldsymbol{t}_0$ for large sequence lengths, if $\hat{p}_k \approx p_k$, this should give a good approximation to equation (4). The two approaches differ in the data set used.

1. Simulation: Generate a large number of sites from the fitted model. This approach will allow one to approximate equations (2)–(4) arbitrarily well because the $\hat{p}_k$ can be made arbitrarily close to the p_k.
2. Observed information: Use the data that were used to fit the model. The resulting matrix is sometimes referred to

Table 1. BP, aBP, and aBPo (aBP using the observed information matrix) for the Three Splits from the HIV Data That Did Not Have 100% BP.

Split	BP (%)	aBP (%)	aBPo (%)
*A*1,*B*,*D*|*A*2,*E*1,*E*2	76	90	90
*A*2,*B*,*D*|*A*1,*E*1,*E*2	14	31	31
*A*1,*A*2|*B*,*D*,*E*1,*E*2	10	25	25

as the observed information matrix. In theory, if the data are consistent with the model, this should give a good approximation. Moreover, in the derivations in Susko (2009), it was often the case that the quantities (2)–(4) were obtained as approximations to the observed information matrix rather than the other way around, suggesting that the observed information matrix might actually give better approximations to the distribution of BP.

Large-sample properties are preserved using either the first or the second approach which suggests that they should generally give similar results. A number of studies have been conducted in other contexts to compare the performance of variance estimates based on observed and expected information matrices. Efron and Hinkley (1978) show that variances determined from observed information provide better approximations to conditional variances and thus often are better at constructing confidence intervals with appropriate coverage properties both conditionally and unconditionally. Wang et al. (2002) consider the variance estimation of additional model parameters in a phylogenetic setting; in their case, a transition/transversion parameter and a substitution rate. Although they find comparable performance in some settings, they find that, particularly when edge-length estimation is not adjusted for, variances based on expected information tend to be closer to actual variances.

Examples

We now consider some examples. In all of these, $B = 10,000$ normal variates were generated to obtain aBP values. The package PAML (Yang 1997, 2007) was used for ML fitting, and 1,000 bootstrap samples were used to obtain BP values.

As a first example, we consider the HIV data set considered previously in Goldman et al. (2000). It has six homologous sequences, each with 2,000 sites from the *gag* and *pol* genes for isolates of HIV-1 subtypes. There are two A subtypes $A1$ and $A2$, two E subtypes $E1$ and $E2$, and B and D subtypes. We fit trees with a GTR model, observed frequencies as stationary frequencies and eight rate categories in a gamma rates-across-sites model. The subtypes B and D grouped together with 100% BP as did $E1$ and $E2$. The only significant uncertainty is how the $A1$ and $A2$ subtypes split with the rest. There are three possibilities that are listed in table 1 along with BP and aBP values.

It is valuable to contrast the results with those of Goldman et al. (2000). They used the SOWH test, named after the authors, Swofford et al. (1996), that originally described it. For the HIV data, this is roughly the same as that a particular split in table 1 is correct. The particular hypothesis considered in Goldman et al. (2000) is that the correct tree is the one that groups the two A subtypes together: $A1, A2|E1, E2, B, D$. The P value from the SOWH test for this hypothesized split is 0.002. In contrast, the aBP value for the split grouping the A subtypes is 25%. Alternatively, the P value for the test of the hypothesis that the split is not present is 0.75. There is no significant evidence against this hypothesis. The main reason for the difference in conclusions is that SOWH tests whether there is significant evidence "against" the split, whereas small 1-aBP provides significant evidence "for" a split. Considering the split $A1, B, D|A2, E1, E2$, 1-aBP is 0.10, giving a borderline failure to reject the hypothesis that some other split might be correct. Thus, the overall conclusion from aBP is that no one of the three alternative splits is strongly supported enough to draw a firm conclusion.

Table 2. BP, aBP, and aBPo (aBP using the observed information matrix) for Three Splits from the Mammalian Mitochondrial Data. The Labeling Scheme Is H = *Homo sapiens* (human), P = *Phoca vitulina* (harbor seal), B = *Bos taurus* (cow), O = *Oryctolagus cuniculus* (rabbit), M = *Mus musculus* (mouse), and D = *Didelphis virginiana* (opossum).

Split	BP (%)	aBP (%)	aBPo (%)
MD|*OHBP*	94	99	98
OMD|*HBP*	51	73	73
HMD|*OBP*	40	64	64

The next example that we consider is the mammalian mitochondrial data considered previously in Shimodaira and Hasegawa (1999) and Goldman et al. (2000). The amino acid alignment had 3,414 sites and 6 taxa. Using the labeling scheme of Shimodaira and Hasegawa (1999), these are H = *Homo sapiens* (human), P = *Phoca vitulina* (harbor seal), B = *Bos taurus* (cow), O = *Oryctolagus cuniculus* (rabbit), M = *Mus musculus* (mouse), and D = *Didelphis virginiana* (opossum). ML fits were obtained using the mtREV model (the general time-reversible model for mtDNA-encoded proteins of Adachi and Hasegawa 1996) with data set frequencies as stationary frequencies and a gamma rates-across-sites model with eight rate categories. Because these are amino acid data, the number of possible site patterns is large (6.4×10^7). Thus, simulation was used with 100,000 sites to approximate expected information matrices. Some of the resulting values of BP and aBP are given in table 2.

The split of cow and harbor seal from the rest had BP of 100%. The aBP for the split of opposum and mouse from the rest is 99% providing significant evidence that this is a correct split. This leaves as uncertain the relative placements of rabbit and human. The aBP values for the two choices are given in table 2 as 73% and 64% indicating insufficient support for resolution of these relative placements. The conclusions are consistent with the conclusions of the Shimodaira-Hasegawa and Kishino-Hasegawa tests reported in Shimodaira and Hasegawa (1999) but differ from the conclusions of the SOWH reported in Goldman et al. (2000) who rejected the tree (P value < 0.001) with

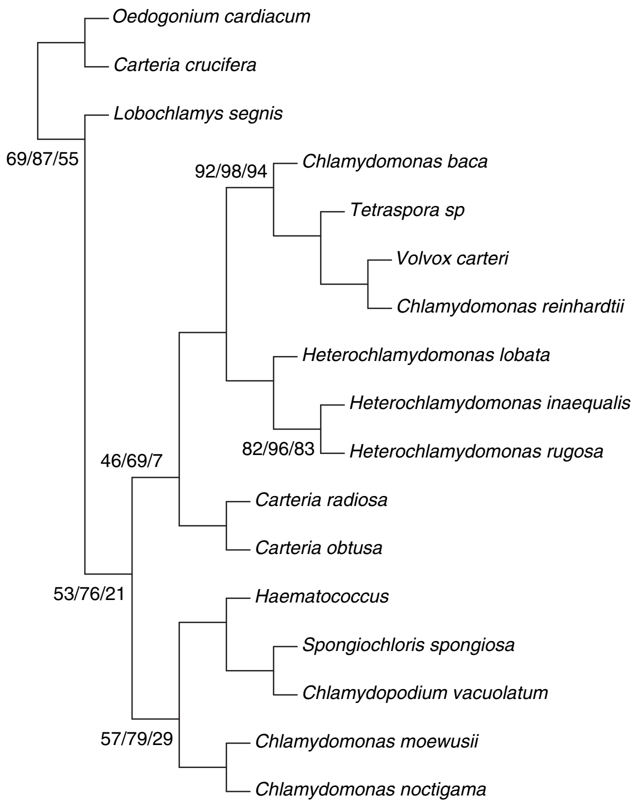

FIG. 2. The ML Chlamydomonadales topology with interior edge labels, ordered from left to right: BP, aBP, and aBP with an additional ML adjustment. For ease of illustration, internal edges were taken as all equal and larger than terminal edges. Internal edges without labels had 100% BP.

the split grouping rabbit, seal, and cow together and the split from human, mouse, and opposum.

The final example considers 17 taxa coming from an analysis of the Chlamydomonadales clade reported in Lewis et al. (2005). The alignment had 3,341 nucleotide sites from combined 18S and 28S ribosomal genes. We repeated the analysis with an Hasegawa–Kishino–Yano model (Hasegawa et al. 1985) with data set frequencies as stationary frequencies and eight rate categories in the gamma rates-across-sites distribution. The ML tree is given in figure 2 with BP and aBP. Although there are only four character states in this case, the larger number of taxa make exact information matrix calculation infeasible. Simulated data were used to approximate the information matrix with 100,000 sites. Once again, aBP and aBP based on the observed information gave similar values. Based upon the aBP values, two of the six splits with BP less than 100% are significant at the 5% level. In Lewis et al. (2005), it was reported that all the splits had Bayesian posterior probabilities larger than 0.93. Although the adjustment to BP brings these values closer to the posterior probabilities, it does not bring them close enough for them to be truly comparable.

Adjustment for ML Estimation of a Split

In principle, the hypotheses of interest should be data independent. In many phylogenetic studies this is the case. For instance, the Coelomata and Ecdysozoa hypotheses of Dopazo H and Dopazo J (2005) were the consequence of inconclusive prior studies, and tree fitting was done to test these hypotheses. In practice, however, it is quite common that an ML tree will be presented with BP and that some of the splits considered would not have been a priori hypotheses. In this case, a hypothesis about splits is data dependent: it is being considered because it is in the ML tree, which depends on the same data being used to test the hypothesis. The limiting results for BP are not applicable in such cases as they assume a fixed split as a hypothesis. For instance, for a fixed split, theory predicts that under the null hypothesis that it is not present, BP will be less than 10% more than 20% of the time. Although this makes sense for a fixed split of interest, it is unlikely that such a split will be in the ML tree when its BP is less than 10%. The theory presented in Susko (2009) can be adjusted to accommodate the data-dependent nature of the hypothesis.

The argument for Steps 1–3, giving probabilities for BP, indicates that BP for the j th topology is well approximated by the probability that V_j^{c*} is nonnegative and larger than the rest of the V_i^{c*}, where the $\boldsymbol{V}^{c*}$ have a multivariate $N\left(\boldsymbol{V}_n^c, \Sigma^c\right)$ distribution. Here $\boldsymbol{V}_n^c$ denotes a standardized score (eq. 6) and, with high probability, j is the ML split if V_{nj} is positive and larger than the other entries of $\boldsymbol{V}_n$. To obtain the probability of observing BP larger than x for the j th split, given that it is the ML split, we simply restrict attention to cases where V_{nj} is positive and larger than the other entries of $\boldsymbol{V}_n$:

1. Generate a trivariate normal random vectors $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_B$ from a $N\left(\boldsymbol{0}, \Sigma^c\right)$ distribution.
2. Among cases where $\left[\boldsymbol{V}_b\right]_j$ is positive, and larger than the rest of the entries of $\boldsymbol{V}_b$, approximate BP for the j th topology as the probability that $\left[\boldsymbol{V}_b\right]_j^*$ is the largest element of $\boldsymbol{V}_b^*$, given that at least one of the $\boldsymbol{V}_b^*$ is positive, where $\boldsymbol{V}_b^*$ is generated from a $N\left(\boldsymbol{V}_b, \Sigma^c\right)$ distribution.
3. The proportion of BP $\geqslant x$ gives an approximation to the probability that BP is at least as large as x under the null hypothesis that the split is not present.

The difference between this and the previous method of obtaining the probability of BP larger than x is in the restriction imposed in Step 2: we only consider cases where $\left[\boldsymbol{V}_b\right]_j$ is positive and larger than the rest of the entries of $\boldsymbol{V}_b$. In other words, based on the limiting approximations, we only consider cases where topology j was the ML topology.

For the HIV data, the ML estimation had only one split with BP less than 100%: the split $A1, B, D | A2, E1, E2$ that had 76% BP and 90% aBP. When correction is made for this being the ML split, however, the aBP is 69%. For the mammalian mitochondrial data, two splits occurred with BP less than 100%. These were the $MD | OHBP$ splits and $OMD | HBP$ splits. The BP for these splits were 94% and 51% and the aBP values were 99% and 73%. In contrast, the aBP values after correcting for ML estimation are 95% and 18%. The conclusions remain the same after correction for ML estimation although the relative support has changed substantially. For

Table 3. Median aBP Values (in %) for BP Values of 10%, . . . , 90% for Selected Simulation Settings with Four Taxa. The Medians Are Over 1,000 Simulated Values of the Other Parameters in the Model.

Edge Lengths	Rate Variation	Data Type	BP 10	20	30	40	50	60	70	80	90
Random	Gamma	Nucleotide	25	41	53	64	73	80	87	93	97
Extreme	Gamma	Nucleotide	28	45	58	69	77	85	91	95	98
Extreme	Equal rates	Nucleotide	33	52	66	76	84	90	95	98	99
Extreme	Gamma	Amino acid	27	43	56	67	75	83	89	94	98
Extreme	Equal rates	Amino acid	29	46	60	70	79	86	92	96	99
Adjusted for ML Estimation											
Random	Gamma	Nucleotide	0	0	0	2	15	38	59	77	91
Extreme	Gamma	Nucleotide	0	0	0	5	21	45	66	83	94
Extreme	Equal rates	Nucleotide	0	0	2	15	38	62	80	91	98
Extreme	Gamma	Amino acid	0	0	0	3	18	42	62	80	93
Extreme	Equal rates	Amino acid	0	0	0	6	24	49	69	85	95

the Chlamydomonadales data, only the split with the largest BP (among those with BP less than 100%) is close to significant when adjustment is made for these being the ML splits.

Simulations

In the examples, given a similar value of BP, the adjustments were quite similar across the different examples even though the models being fit, the number of taxa, and even the type of data (amino acid or nucleotide) varied substantially. For instance, the BP of 69% for the Chlamydomonadales data became an aBP of 86%. For the HIV data, the most similar BP of 76% became 90%. For the Chlamydomonadales data, the BPs of 53% and 57% became 75% and 78%. For the mammalian mitochondrial data, a comparable BP of 51% became an aBP of 73%. To investigate further how much variation in adjustments we should expect across parameter settings, we repeatedly simulated parameter settings and obtained the aBP values corresponding to BP values of 10%, 20%, . . . , 90%. Note that each of these parameter settings can be thought of as possible ML fits from sequence data because this is the only information from the original data that is required for aBP calculations.

We considered simulations for trees of four and five taxa as well as two from trees with six taxa; one simulation with a split of interest that had three taxa on either side and the other simulation where the split of interest had two taxa on one side and four on the other. We simulated from an F84 model (Felsenstein and Churchill 1996) with nucleotide data and from a JTT model (Jones et al. 1992) for amino acid data. We also considered simulation from a gamma rates-across-sites model and an equal-rates model. Thus, altogether there are 16 simulation settings: 4 trees × 2 data types (nucleotide or amino acid) × 2 rate settings (gamma rates-across-sites or equal-rates). For each of these settings, we randomly generated the remaining parameters 1,000 times. Edge lengths for the tree were independently generated from an exponential distribution with a mean edge length of 0.5. Stationary frequencies were generated from a Dirichlet distribution with all parameters set to 1; this implies a uniform probability density function. In the nucleotide simulations, an additional K parameter giving the transition–transversion ratio is required. This was generated from a uniform distribution with values between 0.1 and 10. For the simulations involving rates across sites, the α parameter for the gamma model was also generated from a uniform distribution with values between 0.1 and 10.

The results were remarkably consistent across settings. The median aBP values over the 1,000 random generations of other parameters are reported on the first line of table 3 for four taxa, nucleotide data, and a gamma rates-across-sites model. The medians for the other 16 settings were almost identical. Within a setting, for any given BP, the difference between the 99th percentile and 1st percentile of the aBP values was always less than 5%.

To investigate whether there may be more extreme settings that are likely to show additional variation in aBP values, we fixed edge lengths at more extreme values. Considering the three topologies given in the bottom row of figure 1, the total distance from the split of interest to the terminal nodes in subtrees 1 and 3 were fixed at 2.5. The total distance to terminal nodes in subtrees 2 and 4 was 0.01. aBP was calculated for Topology 2 in the figure.

Once again, within a setting, among the 1,000 random generations, there was little variation: the difference between the 99th percentile and 1st percentile of the aBP values was always less than 5%. Consequently, only median values are reported in table 3. For a given setting of data type (nucleotide or amino acid) or rate variation model (gamma or equal rates), there was little variation across trees and so results are reported for four taxa. We do, however, see variation across data types and rate variation model. In all cases, the more extreme edge-length setting causes an increase in aBP.

As a final setting to consider, we simulated from the Chlamydomonadales topology of figure 2, with 17 taxa. To have taxa separated from the split of interest by a number of internal nodes, the edge with BP of 57% was selected as the edge of interest. Simulation was from an F84 model with a gamma rates-across-sites correction. Edge lengths, stationary frequencies, and K parameter were generated as in the four-taxon case. The results were almost identical to the random/gamma/nucleotide setting in table 3 for four taxa. As in the four-taxon case, the difference between the 1st and 99th percentile was less than 5%. The median adjusted bootstrap values were within 1% of the values in table 3.

Discussion

The tools presented here provide more interpretable BP values. Software is available at http://www.mathstat.dal.ca/ tsusko. The main program aBPn uses a control file similar to that of the package PAML. The implementation includes a number of widely used nucleotide and amino acid substitution models. The program does not obtain ML estimates or obtain BP for these models, which can be obtained from the packages PHYLIP (Felsenstein 1989, 2004), TREE-PUZZLE (Schmidt et al. 2002), and PAML. Instead, the program takes as input a Newick tree file with BP as labels and outputs a Newick tree file, with the same topology, but with labels changed to aBP values. Because of the similarity of conversion of BP to aBP values across different examples and simulated settings, as a rough approximation to aBP values for models not covered by the software, one might use the median values of the first row of table 3 as a guide. However, as the other rows of table 3 indicate, this may not give appropriate adjustments when there is a mix of long and short branches in the tree.

The limiting results used to adjust BP correspond to a single unresolved edge and, because the true edge lengths are unknown, use the estimated edge lengths for all external edges. With large sequence lengths, the estimated edge lengths will be close to the actual edge lengths and thus will result in first-order correct P values. More problematic is the possibility that the null hypothesis is true but there is more than one unresolved edge. The distribution of BP is not currently available in this case, but it is reasonable to expect that it will be different from the distribution with a single unresolved edge. I conjecture that using this limiting distribution would give adjusted BP values larger than the aBP values described here. To see this, consider the extreme case that the generating tree is a star tree. For a given split, because there are so many more possible splits that might arise in the estimated tree, it seems likely that BP will tend to be smaller than if that split were the only unresolved split in the generating tree. This suggests that aBP likely provides a conservative P value if, in fact, the true generating tree has multiple unresolved edges; for instance, the probability of obtaining aBP larger than, say, 95% will be less than 5% in such cases.

The aBP values obtained after adjustment for ML estimation differ substantially from those that do not make an adjustment. The distinction between a priori hypothesis and data-dependent hypothesis can get blurry. For instance, in the mitochondrial mammalian data, the split of mouse and opposum from the rest might very well have been an a priori hypothesis of interest that happened to end up in the ML tree. The fact that a split ends up in the ML tree need not imply that an additional adjustment is required. On the other hand, for splits that one is surprised to find in an ML tree, the aBP value adjusted for ML estimation provides a cautionary note on accepting the hypothesis that the split actually is present. Under the hypothesis that it is not present, table 3 indicates that it is not nearly as improbable that one will obtain large BP.

Acknowledgment

I would like to thank Andrew Roger for valuable comments. This research was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada.

References

Adachi J, Hasegawa M. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J Mol Evol.* 42:459–468.

Allman ES, Ané C, Rhodes JA. 2008. Identifiability of a Markovian model of molecular evolution with gamma-distributed rates. *Adv Appl Probab.* 40:229–249.

Beran R. 1988. Prepivoting test statistics: a bootstrap view of asymptotic refinements. *J Am Stat Assoc.* 83:687–697.

Billera LJ, Holmes SP, Vogtman K. 2001. Geometry of the space of phylogenetic trees. *Adv Appl Math.* 27:733–767.

Chang JT. 1996. Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency. *Math Biosci.* 137:51–37.

Dopazo H, Dopazo J. 2005. Genome-scale evidence of the nematode-arthropod clade. *Genome Biol.* 6:R41.

Efron B, Halloran E, Holmes S. 1996. Bootstrap confidence levels for phylogenetic trees. *Proc Natl Acad Sci. U S A* 93:7085–7090.

Efron B, Hinkley DV. 1978. Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika* 65:457–487.

Efron B, Tibshirani R. 1998. The problem of regions. *Ann Stat.* 26: 1687–1718.

Felsenstein J. 1989. PHYLIP—phylogeny inference package (version 3.2). *Cladistics* 5:164–166.

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

Felsenstein J, Churchill GA. 1996. A hidden Markov model approach to variation among sites in rate of evolution. *Mol Biol Evol.* 13:93–104.

Felsenstein J, Kishino H. 1993. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. *Syst Biol.* 42: 193–200.

Goldman N, Thompson JP, Rodrigo AG. 2000. Likelihood-based tests of topologies in phylogenetics. *Syst Biol.* 49:652–670.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol.* 22:160–174.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci.* 8:275–282.

Kuhner MK, Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol.* 11:459–468.

Lewis PO, Holder MT, Holsinger KE. 2005. Polytomies and Bayesian phylogenetic inference. *Syst Biol.* 54:241–253.

Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL. 2003. Protein evolution with dependence among codons due to tertiary structure. *Mol Biol Evol.* 20:1692–1704.

Schmidt HA, Strimmer K, Vingron M, von Haesler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18:502–504.

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol.* 16:1114–1116.

Susko E. 2009. Bootstrap support is not first order correct. *Syst Biol.* 58:211–223.

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. 1996. Phylogentic inference. In: Hillis DM, Moritz C, Mable BK, editors. Molecular systematics. Sunderland (MA): Sinauer. p. 407–514.

Wang Q, Salter LA, Pearl DK. 2002. Estimation of evolutionary parameters with phylogenetic trees. *J Mol Evol.* 55:684–695.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol.* 39:306–314.

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

Yang Z. 2007. PAML 4: a program for phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.