Syst. Biol. 64(2):243–255, 2015

DOI:10.1093/sysbio/syu099
Advance Access publication November 27, 2014

Bayesian Long Branch Attraction Bias and Corrections

EDWARD SUSKO*

Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, canada B3H, 4R2
**Correspondence to be sent to: Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4R2; E-mail: edward.susko@gmail.com*

Received 19 August 2014; reviews returned 27 September 2014; accepted 23 November 2014
Associate Editor: Laura Kubatko

Abstract.—Previous work on the star-tree paradox has shown that Bayesian methods suffer from a long branch attraction bias. That work is extended to settings involving more taxa and partially resolved trees. The long branch attraction bias is confirmed to arise more broadly and an additional source of bias is found. A by-product of the analysis is methods that correct for biases toward particular topologies. The corrections can be easily calculated using existing Bayesian software. Posterior support for a set of two or more trees can thus be supplemented with corrected versions to cross-check or replace results. Simulations show the corrections to be highly effective. [Bayesian methods; bias; long branch attraction; posterior probability; star tree.]

Suzuki et al. (2002) seem to have been the first to note that posterior probabilities for resolved topologies can occasionally be very large even when the true tree is a star tree and sequence lengths are large. The phenomenon has come to be referred to as the star-tree paradox and received further support from a number of other studies (Cummings et al. 2003; Lewis et al. 2005; Yang and Rannala 2006). Kolaczowski and Thornton (2006) raised questions about the star-tree paradox but the work of Steel and Matsen (2007) as well as Yang (2007) conclusively demonstrated that the star-tree paradox was a real phenomenon.

Susko (2008) considered star-tree paradox results in the four-taxon setting. One of the findings was that, not only are large posterior probabilities possible when the true tree is a star tree, but when that tree has two long edges and two short edges, there is a substantial chance of very large posterior probability for the tree with long edges together. Since behavior of Bayesian methods varies continuously with parameters, an implication is that when the true tree has long edges apart but a small middle edge, large posterior probability for the incorrect tree with long edges together is expected. Kolaczkowski and Thornton (2009) showed that this was indeed the case and provided further results showing a long branch attraction (LBA) bias for Bayesian posterior probabilities.

The phenomenon of LBA found in Susko (2008) and Kolaczkowski and Thornton (2009) is an example of an oft-reported bias first noted in Felsenstein (1978). In most references to LBA, however, some form of model misspecification is present (Huelsenbeck 1995; Inagaki et al. 2004; Susko et al. 2004) or methods different from likelihood and Bayesian methods, like parsimony, are considered (Felsenstein 1978; Hendy and Penny 1989). Part of what was surprising in the result coming from the star-tree paradox is that the bias occurs in the absence of model misspecification.

The current article extends star-tree paradox and LBA bias results for Bayesian methods to settings with more taxa and partially resolved trees. Extensions of Laplace approximation results of Susko (2008) provide theoretical reasons for expecting a LBA bias and indicate additional sources of potential bias. More importantly, the approximations provide motivation for several simple corrections and reasons for expecting them to work well. Simulations confirm that LBA bias is a substantial problem for Bayesian methods and that the corrections are effective at reducing or even eliminating this source of bias.

THEORY AND METHODS

Setting and Assumptions

Bias in topological estimation is defined as a tendency to estimate a particular type of tree (for instance, one with long branches together) even when it is not correct. Bias as considered here is consistent with this definition but includes additional elements. First, the true tree is usually only partially resolved: It has at least one zero-length edge. Multiple topologies then include the true tree as special cases where some of their edge lengths are set to zero. For example, every five-taxon tree includes the five-taxon star tree in Figure 1e as a special case whereas only the three trees in Figure 1j–l include the partly resolved tree as a special case. This is the setting where bias is clearest: Since each tree is equally correct, it is undesirable to frequently obtain large posterior probability in favor of any particular one of these topologies. Since behavior of methods usually varies continuously with edge lengths, biases toward a particular tree arising from a partially resolved true tree imply biases toward that tree when some other well-resolved true tree, similar to the partially resolved tree, generates the data. It is a proxy for the real setting of interest when the tree is poorly resolved. Finally, rather than focusing attention on estimation alone, bias will be considered as a difference in the distribution of posterior probabilities over those topologies that include the true tree as a special case.

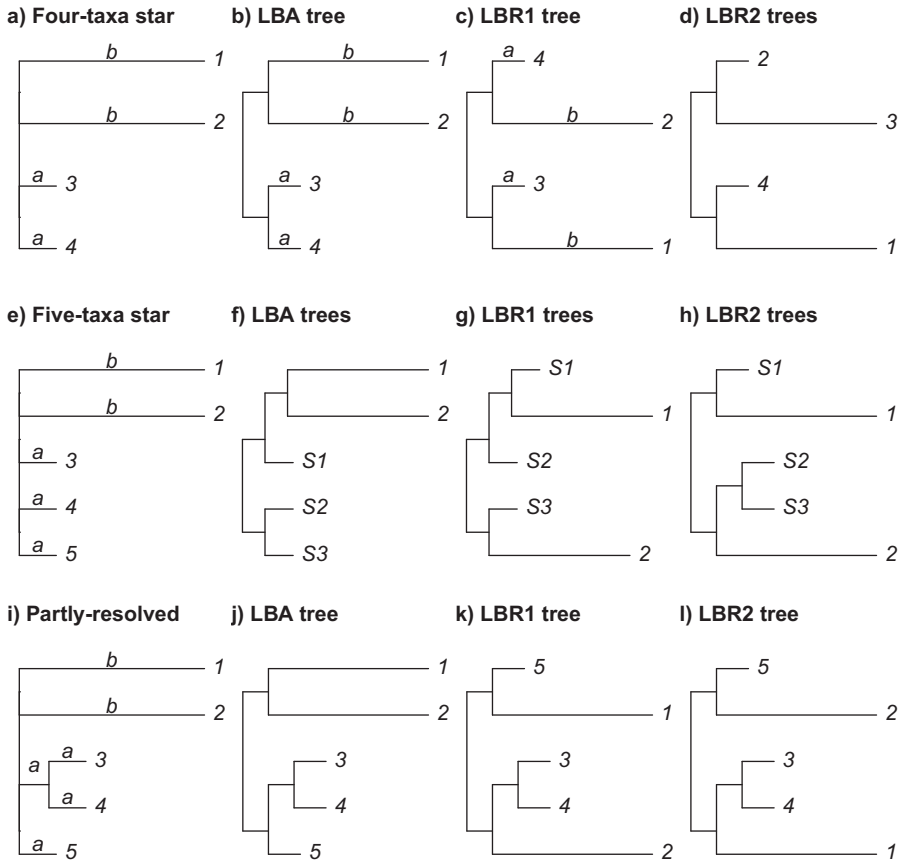

FIGURE 1. Simulations generate data from four- (Fig. 1a) and five- (Fig. 1e) taxon star trees, the partially resolved tree (Fig. 1i) and the fully resolved LBA (Fig. 1b), and LBR1 (Fig. 1c) trees. In each case, terminal edges are a mix of long (length b) and short (length a) edges. For the four-taxon settings there are three possible estimated trees, given in Figure 1b–d. Due to the symmetry in the five-taxon generating star tree, the three sets of trees in Figure 1f–h are the ones that have distinct posterior probability distributions; S1–S3 are some permutations of 3–5. In the partially resolved case, in most random generations, only the three trees in Figure 1j–l have appreciable posterior probabilities.

Some notation is required and is chosen to be consistent to a large degree with Susko (2008). Denote the possible tree topologies as $j=1,2,\ldots,$ (ordering is unimportant). Let $l_j(t)$ denote the log likelihood for topology j and edge length vector t, of length p. For any topology j that includes the true tree, edges can be ordered with zero length edge at the end so that a t_*, independent of j, denotes the edge lengths for the true tree. For example, for a true four-taxon star tree and for any of the three possible topologies in Figure 1b–d, $t_*=[t_{1*},\ldots,t_{4*},0]^T$ can represent the edge length vector for the true tree, where t_{i*} is the terminal edge length for i. Let $l_*(t_*)$ be the log likelihood for the true tree. Note that for any topology j that includes the true tree as a special case, $l_j(t_*)=l_*(t_*)$. For such a topology, let $\sqrt{n}S_{jn}$ be the vector of derivatives of $l_j(t_*)$, where n is sequence length, let $nJ_{jn}(t)$ denote the second derivative matrix of $l_j(t)$ and let $J_{jn}=J_{jn}(t_*)$.

The Laplace approximations used to motivate the methods require some assumptions. Site patterns (character states at a site for all observed taxa) are assumed to be generated independently on a tree with non zero terminal edge-lengths, from a Markov model with non zero frequencies of character states and non zero rates of exchange. This implies that the probability of any pattern of character states at a site is non zero and avoids settings where $E[J_{jn}]$ is not positive definite; $E[J_{jn}]$ is assumed positive definite and has been found to be so in all settings considered. For any topology j, the prior, α_j, is assumed positive and the prior for the edge lengths, $\pi_j(t)$, is assumed to be continuous, positive and have a bounded derivative. This assumption is satisfied for common priors like the uniform or exponential. It can likely be relaxed but it is important that the prior have mass and be well-behaved in a neighborhood of the true tree parameters. Finally, it is assumed that the trees and edge lengths are identifiable: If two trees give the same probabilities of site patterns, no matter what the site patterns, the two trees must be equivalent. Such an assumption has been shown to be valid for conventional continuous-time Markov models as well as a number of models that allow variation in rates across sites (Chang 1996; Allman et al. 2008; 2012).

Laplace Approximations

Approximations that hold as sequence length, n, gets large are derived in Supplementary Material available on Dryad at http://dx.doi.org/10.5061/dryad.g180s. The approximations are used to motivate the methods and to provide reasons for expecting them to work well beyond the simulation settings considered. Approximations are for

$$\eta_j = (2\pi)^{-p/2} n^{p/2} \alpha_j \int_{t\geq 0} \exp[l_j(t) - l_*(t_*)]\ \pi_j(t)\ dt \qquad (1)$$

The reason η_j is of interest is that the factor $\exp[-l_*(t_*)]$ can be taken out of the integral and canceled when taking the ratio

$$\frac{\eta_j}{\sum_k \eta_k} = \frac{\alpha_j \int_{t\geq 0} \exp[l_j(t)]\ \pi_j(t)\ dt}{\sum_k \alpha_k \int_{t\geq 0} \exp[l_k(t)]\ \pi_k(t)\ dt} \qquad (2)$$

giving that the posterior probability of topology j is proportional to η_j. The main result established in Supplementary Material. (Data available on Dryad at http://dx.doi.org/10.5061/dryad.g180s) is that $\eta_j \approx 0$ when j does not include the true tree as a special case. Otherwise

$$\eta_j \approx \exp[\frac{1}{2} S_{jn} J_{jn}^{-1} S_{jn}] \alpha_j \pi_j(t_*) |J_{jn}|^{-1/2} u_j(S_{jn}) \qquad (3)$$

When $\pi_j(t_*)$ does not depend on j, as is usually the case, it cancels in calculating the ratio (2) and can be ignored in (3). Here $|J_{jn}|$ denotes the determinant of J_{jn}. The transformation $u_j(s)$ will be referred to as the boundary factor, and is calculated as follows. Let Y have a normal distribution with mean $J_{jn}^{-1}s$ and covariance matrix J_{jn}^{-1} and let Y_r denote those elements of Y that correspond to indices of t_* that are 0. Then $u_j(s)$ is calculated as $P(Y_r > 0)$.

The result (3) is a consequence of what is commonly referred to as a Laplace approximation (Tierney and Kadane 1986). In contrast to usual applications of Laplace approximation, however, where approximations are with respect to estimated parameters, approximations here are with respect to true parameters. Also, the term $u_j(S_{jn})$ is a consequence of some parameters being on the boundary of the parameter space in the true model; some edge lengths equal zero. Since parameters are usually in the interior of the parameter space, the $u_j(S_{jn})$ term is not present in usual Laplace approximations.

Bias Corrections

The Laplace approximation (3) is useful in indicating potential sources of bias and in suggesting potential corrections. It is established in Supplementary Material (data available on Dryad at http://dx.doi.org/10.5061/dryad.g180s) that S_{jn} is approximately normal with mean 0 and covariance matrix J_{jn}. Standard distributional results then give that $S_{jn}^T J_{jn}^{-1} S_{jn}$ is approximately chi-squared with p degrees of freedom. Thus, the distribution of the first factor in (3) is the same no matter which topology j is considered. Ignoring $\pi_j(t)$ and α_j, which need not depend on j, the main potential sources of bias are $|J_{jn}|^{-1/2}$ and $u_j(S_{jn})$ as these are the terms that may vary over j.

The first bias correction, referred to as the *prior correction*, uses α_j and corrects for the $|J_{jn}|^{-1/2}$ term. It would be ideal if one could set $\alpha_j \propto |J_{jn}|^{1/2}$ which would cancel the $|J_{jn}|^{-1/2}$ term in (3). However, since $J_{jn} = J_{jn}(t_*)$ depends on the unknown true edge lengths, t_*, it must be replaced with an estimate, $\alpha_j \propto |J_{jn}(\hat{t}_j)|^{1/2}$, where $\hat{t}_j$ is a consistent estimate of edge lengths for topology j; in the examples, the posterior mean edge lengths were used.

The second correction considered also corrects for the $|J_{jn}|^{-1/2}$ term. Aitkin (1991) defined the posterior Bayes factor (PBF) as an alternative to usual Bayes factors. In the context considered here, the ratio of posterior probabilities for topology j and k are replaced by $\bar{L}_j^A / \bar{L}_k^A$ where

$$\bar{L}_j^A = \int \exp[l_j(t)] \pi(t|D,j)\ dt \qquad (4)$$

and $\pi(t|D,j)$ is the posterior distribution of edge lengths t given topology j and all of the data D. The "corrected posterior" for topology j that gives this Bayes factor is proportional to $\bar{L}_j^A$ and will be referred to as the *PBF correction* or *PBF-corrected posterior*. Note that (4) is similar to the numerator for the usual posterior probability but, since $\pi_j(t)$ is replaced by $\pi(t|D,j)$, weights more heavily those edge lengths that are supported by the data.

Two approaches are given for calculating the PBF-corrected posteriors. The first assumes the software in use outputs likelihoods and trees encountered during Markov chain Monte Carlo sampling; MrBayes (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) provides such output. When this is the case, the average $\exp[l_j(t)]$, averaged over Markov chain Monte Carlo samples that gave topology j, approximates (4); if topology j never arises, L_j^A can be taken as zero. Since only ratios are ever needed, to avoid numerical difficulties with small likelihoods, $\exp[l_j(t)]$ can be replaced with $\exp[l_j(t) - l_{\max}]$ in averaging, where $l_{\max}$ is the maximum log likelihood encountered during sampling.

Simplification of (4) is required to establish that it will be effective at eliminating the bias due to $|J_{jn}|^{-1/2}$ and gives an alternative method for calculating PBF-corrected posteriors. Substituting

$$\pi(t|D,j) = \exp[l_j(t)]\pi_j(t) / \int \exp[l_j(t)]\pi_j(t)\ dt$$

in (4) gives

$$\bar{L}_j^A = \int \exp[2l_j(t)]\pi_j(t)\,dt / \int \exp[l_j(t)]\pi_j(t)\,dt = 2^{-p/2}\exp[l_*(t_*)] \times \frac{(2\pi)^{-p/2}(2n)^{p/2}\alpha_j \int \exp[2l_j(t)-2l_*(t)]\pi_j(t)\,dt}{(2\pi)^{-p/2}n^{p/2}\alpha_j \int \exp[l_j(t)-l_*(t)]\pi_j(t)\,dt} \propto \frac{\eta_j^{(2)}}{\eta_j}$$

where $\eta_j^{(2)}$ denotes the η_j in (1) corresponding to concatenating the data set with itself to create a new data set of size $2n$. Since the posterior probabilities for the original and doubled data are proportional to η_j and $\eta_j^{(2)}$, the PBF-corrected posterior is proportional to the ratio of the posterior probability for the doubled data to the posterior probability for the original data. This fact provides motivation for the PBF correction. If the data strongly support a topology j, that support should increase when the data set is concatenated with itself, leading to a large ratio of posterior probabilities. That the PBF-corrected posterior is proportional to the ratio of the posterior probability for the doubled data to the posterior probability for the original data also provides a simple, albeit intensive, way of calculating the PBF corrected posterior. Any Bayesian software implementation should provide posterior probabilities for the topologies and concatenating a data set with itself is straightforward, hence the PBF-corrected posteriors can be calculated by calculating posterior probabilities for doubled and original data, taking ratios and then normalizing.

Using PBF-corrected posteriors effectively eliminates the bias due to $|J_{jn}|^{-1/2}$. To see this note that the Laplace approximation (3) applies for $\eta_j^{(2)}$ with $\sqrt{2n}S_{jn}^{(2)}$ and $2nJ_{jn}^{(2)}$, denoting the first and second derivatives of the log likelihood, $2l_j(t_*)$, for the doubled data. Thus $\sqrt{2n}S_{jn}^{(2)}/2$ and $2nJ_{jn}^{(2)}/2$ are the first and second derivatives of $l_j(t_*)$, which were denoted $\sqrt{n}S_{jn}$ and nJ_{jn}. This gives that $S_{jn}^{(2)} = \sqrt{2}S_{jn}$ and $J_{jn}^{(2)} = J_{jn}$. Substituting in (3) and taking ratios gives that the PBF-corrected posterior is proportional to

$$\eta_j^{(2)}/\eta_j \approx \frac{\exp[S_{jn}^T J_{jn}^{-1} S_{jn}]\alpha_j \pi_j(t_*)|J_{jn}|^{-1/2} u_j(\sqrt{2}S_{jn})}{\exp[\frac{1}{2}S_{jn}^T J_{jn}^{-1} S_{jn}]\alpha_j \pi_j(t_*)|J_{jn}|^{-1/2} u_j(S_{jn})} = \exp[\frac{1}{2}S_{jn}^T J_{jn}^{-1} S_{jn}]\frac{u_j(\sqrt{2}S_{jn})}{u_j(S_{jn})} \quad (5)$$

which no longer involves $|J_{jn}|^{-1/2}$. Note that PBF correction also partially corrects the boundary factor, since $u_j(\sqrt{2}S_{jn})$ should correlate well with $u_j(S_{jn})$, giving a ratio that is closer to constant than $u_j(S_{jn})$ is.

The final correction adjusts the prior correction for the boundary factor, $u_j(S_{jn})$. Correcting for the boundary factor is unnecessary when there is a single zero-length edge in the true tree. It follows similarly as in Susko (2008) that $u_j(S_{jn})$ has an approximate uniform distribution, irrespective of the topology j. When there is more than one zero-length edge in the true tree, correction is necessary and difficult due to the complex dependence of $u_j(S_{jn})$ on S_{jn}, a random quantity. Since S_{jn} is approximately normal with mean 0, $u_j(0)$ is a reasonable guess for what $u_j(S_{jn})$ is without knowing S_{jn}. Thus the third correction considered, referred to as the *boundary correction*, uses $\alpha_j \propto |J_{jn}(\hat{t}_j)|^{1/2}\hat{u}_j(0)^{-1}$. Here $\hat{u}_j(0)$ is an estimate of $u_j(0)$. In calculating $\hat{u}_j(0)$ in examples, J_{jn} was replaced by $J_{jn}(\hat{t}_j)$ and the boundary set r was taken as all internal edges.

Direct use of the corrections provides new corrected "posteriors" or measures of support for the topologies. These can be converted to measures of support for a split by summing over all topologies that have the split present.

Simulation Settings and Bayesian Implementation

Simulations were for a range of four- and five-taxon settings. Attention is focused on four- and five-taxon settings partly to insure that results are not due to convergence difficulties of Markov chain Monte Carlo methods. For each parameter setting, 1000 data sets were generated using Seq-Gen (Rambaut and Grassley 1997), each having 1000 sites from the Jukes–Cantor model (Jukes and Cantor 1969).

Generating trees are given in Figure 1. In the four-taxon settings to investigate bias and its correction, the generating tree was a star tree (Fig. 1a) having two long edges of length b and two short edges of length a. To illustrate the implications of bias for resolved trees, data were also generated from the LBA (Fig. 1b) and LBR1 (Fig. 1c) trees, with long edges separated, again with short edges of length a and long edges of length b. However, the middle edge length was allowed to vary away from 0. The LBA and LBR prefixes are intended as evocative of long branch attraction and long branch repulsion trees, respectively.

In five-taxon settings, to investigate biases and their correction, generation was considered from a star tree (Fig. 1e) with two long edges of length b and three short edges a. Generation from a partially resolved tree (Fig. 1i) was also considered. In this case, the resolved internal edge, leading to a cherry with two short edges, was set to the same value, a, as all of the other short edges. For this generating tree, there is a relatively small chance that trees without the resolved edge will be estimated. Although such cases are summarized, primary interest is in the distribution of posterior support for the LBA (Fig. 1j), LBR1, (Fig. 1k) and LBR2 (Fig. 1l) trees.

For simulations from a five-taxon star tree, there are fifteen possible trees that will be estimated. However,

because of the symmetry present in the generating trees, some of the estimated trees are sure to have identical distributions of posterior probabilities. For instance, if Tree 1 is the LBA tree (Fig. 1f) with $S1$–$S3$ set to 3–5 and Tree 2, the LBA tree with $S1$–$S3$ set to 4,3, and 5, then, for any P, the long-run frequency with which the posterior probability for Tree 1 is larger than P is exactly the same as the corresponding frequency for Tree 2. The sets of trees giving the same distributions of posterior probabilities are indicated in Figure 1f–h, where $S1$–$S3$ are any rearrangement of taxa 3,4, and 5. There are three LBA trees, corresponding to the ways of choosing the $S2$ and $S3$ taxa. There are six LBR1 trees and the remaining six trees are LBR2 trees. Consequently, results are more concisely presented as means over those trees having the same distributions of posterior probabilities; for instance, the mean frequency that a posterior probability was larger than a fixed threshold. Similarly, a number of splits have the same distributions of posterior probabilities. The split of taxa 1 and 2 from the others, 12|345, is labeled the LL split because it has two long edges together and has a different distribution of posterior probabilities than any other split. However, the distribution of posterior probabilities for any two splits with two short edges together (eg., 34|125 or 35|124) will be the same. Such splits are labeled SS splits. Finally, splits with a short and a long edge together (an LS split) will yield the same distributions of posterior probabilities.

MrBayes 3.1.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) was used to obtain posterior probabilities. That the Jukes–Cantor model is the correct substitution model was treated as known in the prior (`nst=1, statfreqpr=fixed(equal)`). A total of 100,000 MCMC generations were run with a sampling frequency of 10 and 25 burn-in samples (`ngen=100000, samplefreq=10, sumt burnin=25`). All other parameters were set to default values.

Results

Four-Taxon Simulation Results

Figure 2 gives the frequencies with which posterior probabilities exceeded a threshold for a number of the four-taxon settings and indicates that a substantial LBA bias occurs with Bayesian methods that is reduced to varying degree by the corrections introduced here. The uncorrected cases in the first row show a substantial bias in favor of the LBA tree when the short edge lengths are $a=0.05$. When sequence length is 1000, there is roughly a 40% chance that the posterior probability for the LBA tree will be larger than 0.8 and this event occurs only slightly less frequently when sequence length is as large as 25,000. Bias is present but not as substantial when short edges $a=0.5$ are only half as long as the long edges. There is $<20\%$ chance of a posterior probability >0.8. The prior correction corrects the bias effectively when sequence length is 25,000. When sequence length is 1000, however, while the bias decreases substantially, an LBA bias remains and is particularly prominent when $a=0.05$. The PBF correction is very effective in all cases with a slight overcorrection, whereby the LBA tree is more likely to have small posterior probabilities.

The LBA bias of Bayesian posterior probabilities depends substantially on the ratio of short and long edge lengths as is illustrated in Figure 3. Across settings, there is a 60–80% chance the LBA tree will be estimated when the short edge length, a, is 10% of the long edge length, b, but the LBA tree is only estimated $\sim$40% of the time when $a/b=0.5$. The LBA bias also depends on how long the long edge is. When $b=0.5$ and $n=1000$, there is roughly a 60% chance of estimating the LBA tree with an edge length that is 10% of b. This probability goes up to 80% when $b=1$. The prior and PBF corrections do well at reducing the frequency with which the LBA tree is estimated. For short edges that are more than 10% of the long edge lengths, LBA trees arise $<50\%$ of the time. The prior correction tends to undercorrect and the PBF correction tends to overcorrect, however, with the biases in correction being worst when sequence lengths are smaller ($n=1000$) and when the short edge is a smaller fraction of the long edge.

The LBA bias when some edges of a tree are unresolved implies biases will arise as the edge lengths that are zero in the incompletely resolved tree are increased. Figure 4 shows a tendency to estimate the LBA tree even when the true generating tree is the LBR1 tree of Figure 1. With a sequence length of 1000, the LBA tree is much more likely to be estimated when the middle edge length in the true tree is small and is still more likely to be estimated than the true tree when middle edge length is as large as 0.02. With larger sequence lengths of 25,000, there is still a tendency to estimate the LBA tree with small middle edge lengths but the frequency with which the true tree is estimated increases quickly, with a $>80\%$ chance that the true tree is estimated when its middle edge length is 0.025. With sequence length of $n=1000$, the prior correction gives a $<60\%$ chance of the LBA tree being estimated. Still, the tendency for the prior correction to under-correct the LBA bias causes it to have a relatively small probability of estimating the correct tree; $<60\%$, even with middle edge lengths as large as 0.025. With $n=25,000$, the prior correction is very effective. The true tree is always more likely to be estimated than the LBA tree but not much more likely when middle edge length approaches zero. The PBF correction does an effective job of correcting across settings although its tendency to overcorrect causes it to have a larger than desired frequency of estimating the LBR1 tree (40%) when the middle edge length is zero, $n=1000$ and $a=0.05$.

Although the LBA bias of uncorrected posterior probabilities is generally undesirable, if the true tree actually is the LBA tree, as in Figure 5, correcting the bias toward it can lead to poorer performance. With $b=1$ and $a=0.05$, the LBA tree is estimated $>80\%$ of the time when the middle edge length is zero and the frequency of estimation increases as the middle edge length increases.

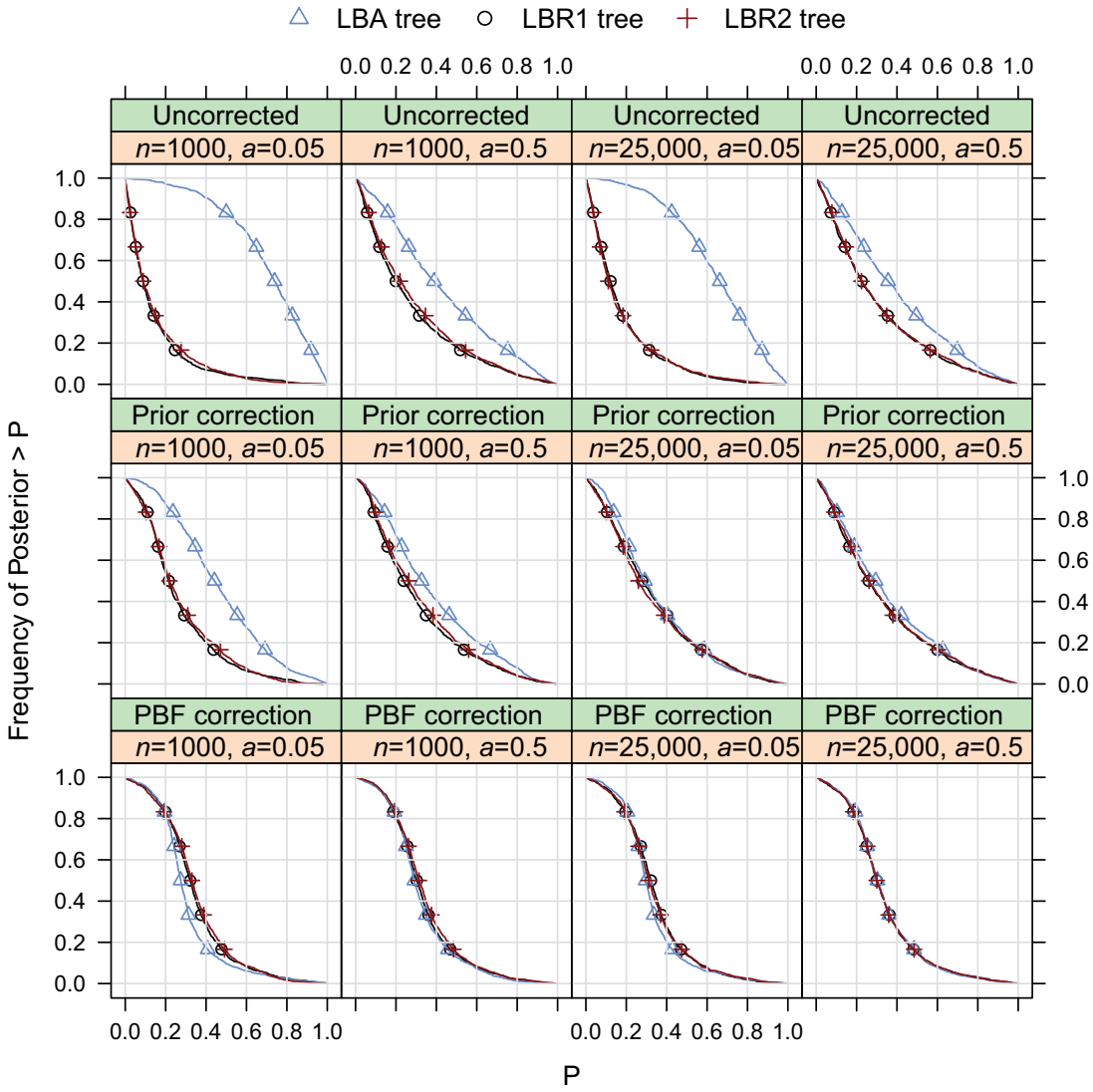

FIGURE 2. The frequencies with which posterior probabilities exceeded a threshold P in simulations from a four-taxon star tree with long edge length $b=1$. Here n is the sequence length and short edges are of length a. The y-axis value for any panel gives the frequency with which the posterior for the given topology exceeded the x-axis value.

For the prior and PBF corrections, the frequency with which the correct LBA tree is estimated is smaller but increases as the middle edge length increases. With a sequence length of 25,000, except for middle edge lengths <0.005, the LBA tree is more likely to be estimated. With $n=1000$, because the prior correction was an undercorrection it always estimates the correct LBA tree more frequently. The consequence of PBF overcorrection, however, is that the frequency of correct LBA estimation remains low (<40%) even with middle edge lengths of 0.025.

Five-Taxon Simulation Results

The LBA bias with four taxa is suggestive of a more general phenomenon that occurs when there is a mix of long and short edges in a tree. This is confirmed in the five-taxon results reported in Figure 6. The frequency with which posterior probabilities exceed a threshold for the LBA tree is comparable to but less than the same frequency when data are generated from the four-taxon tree with the same lengths of short and long edges (Figure 2). This may be due to lengthening the total distance from the split of primary interest to some terminal nodes via the additional well-resolved internal edge of length a in Figure 1i. Figure 6 gives results only for the three trees in Figure 1j–l. These were the only trees that arose with appreciable posterior probability; with $n=1000$ and $a=0.05$ there were 100 cases where the sum of posterior probabilities for all other trees was larger than 0.1, with $a=0.5$ there was one case and for all other settings there were no such cases. Once again, with $n=1000$, the prior correction undercorrects the LBA bias substantially when $n=1000$ and the PBF correction

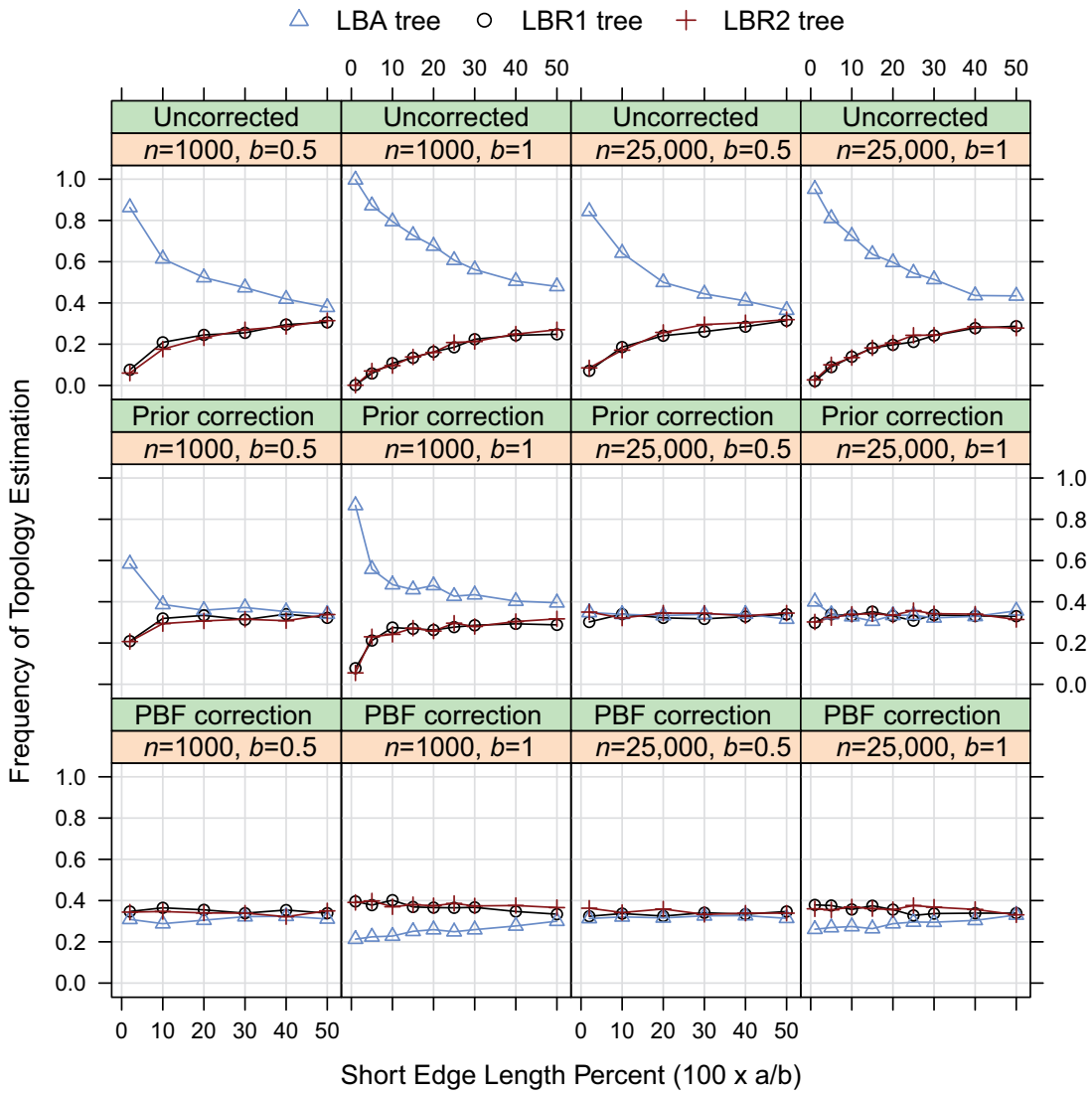

FIGURE 3. The frequencies with which topologies are estimated as a function of the ratio of short to long edge lengths in simulations from a four-taxon star tree with long edge lengths $b=0.5$ or 1.

overcorrects slightly. Both corrections are effective with the larger sequence length.

In examples considered so far, a single unresolved edge is present in the true tree. In that case the boundary factor, $u_j(S_{jn})$, in the Laplace approximation (3) has the same distribution for each topology j and does not induce a bias. When there is more than one unresolved edge, the boundary factor can have a substantial impact. In the case that two edge lengths are zero, the approximate distribution of $u_j(S_{jn})$, treating J_{jn} as fixed and equal to $E[J_{jn}]$ as a further approximation, depends only upon the corresponding correlation coming from the covariance matrix $E[J_{jn}]^{-1}$. Summary quantities for these distributions are given in Table 1. In brief, smaller boundary factors are expected when the correlation is highly negative and they are expected to be less variable. Table 2 gives the correlations that are expected from the J_{jn}^{-1} matrix with large sequence lengths for the three types of trees that arise when data are generated from the five-taxon star tree of Figure 1e; they were calculated from $E[J_{jn}]^{-1}$. The correlations for the LBR2 trees are much less than for the LBA tree which are comparable but smaller than the correlations for the LBR1 tree. Thus for the LBR2 trees, which have the smallest correlations, small boundary factors are expected in (3) leading to smaller posterior probabilities. This is confirmed in Figure 7 which gives the mean frequencies with which the three types of trees were estimated. The LBA bias is still predominant in the uncorrected case but the LBR2 trees are much less likely to be estimated than LBR1 trees. That this is due to the truncation factor is evident by contrasting the effects of the prior and boundary corrections when $n=25,000$. The prior correction is effective at eliminating the LBA bias but the frequency of LBR2 estimation is still low. The boundary correction, which simply adds a factor to the prior correction, is

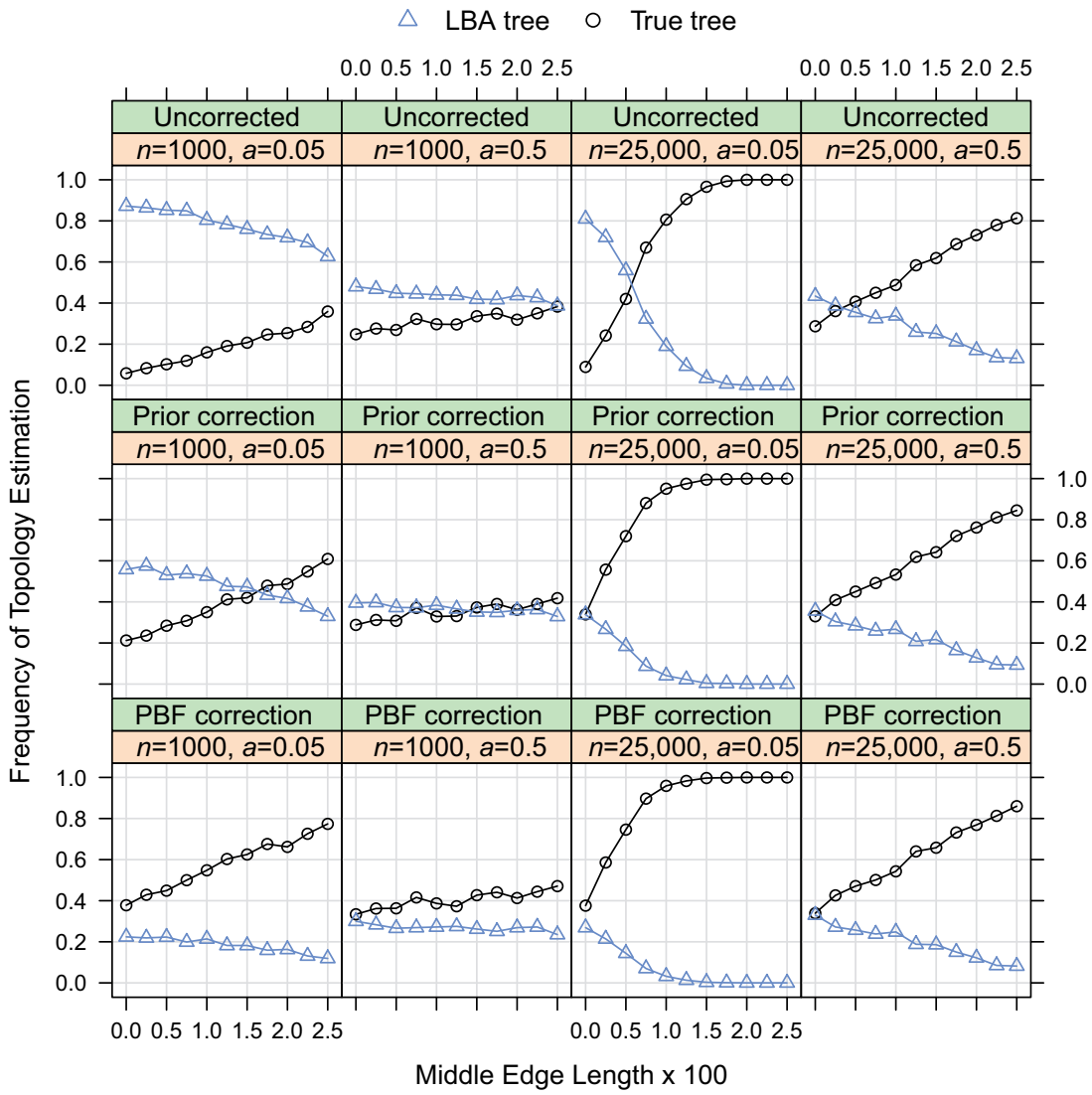

FIGURE 4. The frequencies with which topologies are estimated as a function of middle edge length in simulations from the four-taxon LBR1 tree of Figure 1c with long edge length $b=1$.

effective at making each type of tree equally likely. The PBF correction is much less affected by the boundary factor.

Biases of posterior probabilities for topologies imply biases for splits. In the uncorrected case in Figure 8, it is much more likely that the posterior probability of the split with the long edges together will exceed any given threshold than the other posterior probability of the other types of splits. The performance of PBF and prior corrections is similar as in all other examples considered. Although both corrections are effective at reducing bias, the prior correction undercorrects and the PBF correction is a slight overcorrection.

DISCUSSION

The simulations illustrated a clear LBA bias for Bayesian methods, even in the absence of model misspecification. Extension to five-taxon cases illustrate that other biases (effectively caused by the boundary factor) can cause one tree to be favored over another without having long branches together. There is an intuitive explanation for the LBA bias. The posterior for topology j is proportional to $\int \exp[l_j(t)]\pi_j(t)\ dt$. Assuming, as is usually the case, that $\pi_j(t)$ is not highly concentrated, the effective region of integration for the numerator term is those sets of edge lengths where the likelihood is relatively large. For each topology, it is reasonable to expect that the effective regions of integration for the terminal edge lengths will be comparable. If data is generated from, for instance, the four-taxon star tree in Figure 1a, it is the effective region of integration for the middle edge lengths that will differ. With small a edge lengths, there will be substantial signal in the data that the taxa 3 and 4 are closely related and $l_j(t)$ will only be relatively large when the edge

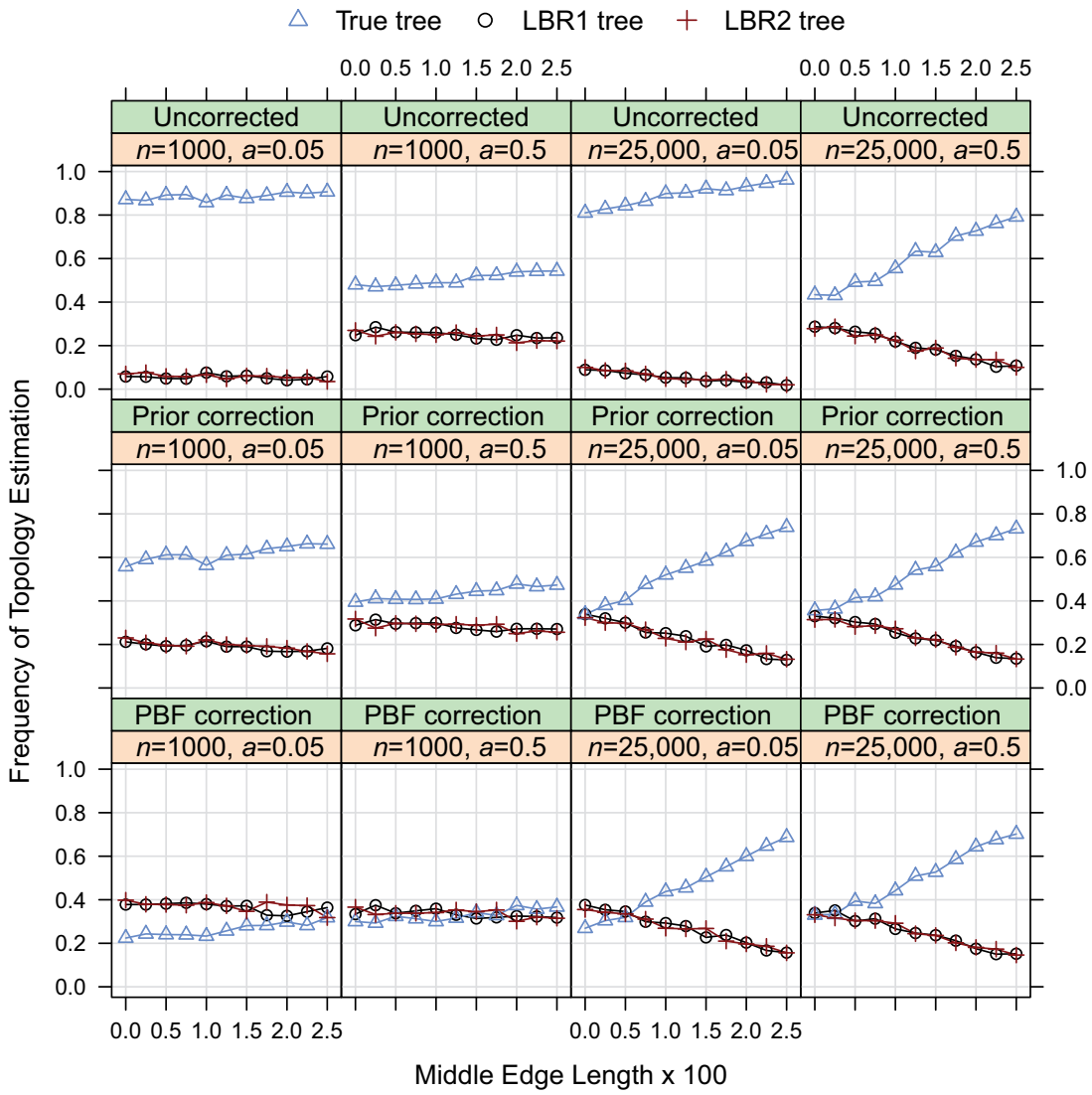

FIGURE 5. The frequencies with which topologies are estimated as a function of middle edge length in simulations from the four-taxon LBA tree of Figure 1b with long edge length $b=1$.

lengths are consistent with this. For the LBR1 and LBR2 trees, the only way to accomplish this is to have a small middle edge. The effective region of integration for the middle edge is thus small. For the LBA tree, however, the distance between taxa 3 and 4 depends only on their terminal edge lengths. Thus, there will be a much larger set of middle edge lengths for which $l_j(t)$ is relatively large. The effective region of integration for the middle edge is relatively large and the numerator for the LBA topology can be expected to be larger than for the other two.

The intuitive explanation for the LBA bias can be related to the $|J_{jn}|^{-1/2}$ factor of (3) and maximum likelihood estimation. The implication of the argument above is that there is more uncertainty about the middle edge length when data are fitted to the LBA tree than to the LBR trees. Thus, the variances of the estimated ML middle edge lengths are larger for the LBA tree than the LBR with variances for the other edge lengths being more comparable. The matrix J_{jn}^{-1} is a valid approximation to the covariance matrix of the ML edge lengths under topology j and $|J_{jn}|^{-1/2} = |J_{jn}^{-1}|^{1/2}$ will be relatively large when some of the entries of J_{jn}^{-1} are relatively large, which is exactly what is expected for the LBA tree. More generally, topologies that give large variances to some subset of internal edges are likely to have relatively large $|J_{jn}|^{-1/2}$ and there will be a bias toward them. Since there is stronger signal for shorter distances than longer distances, topologies that group together similar sequences will have more highly variable internal edge lengths leading to the more dissimilar sequences. Their effective regions of integration will consequently be larger.

The goal here has been to reduce bias toward any particular tree among well-resolved trees. Examining performance when data are generated from a partially

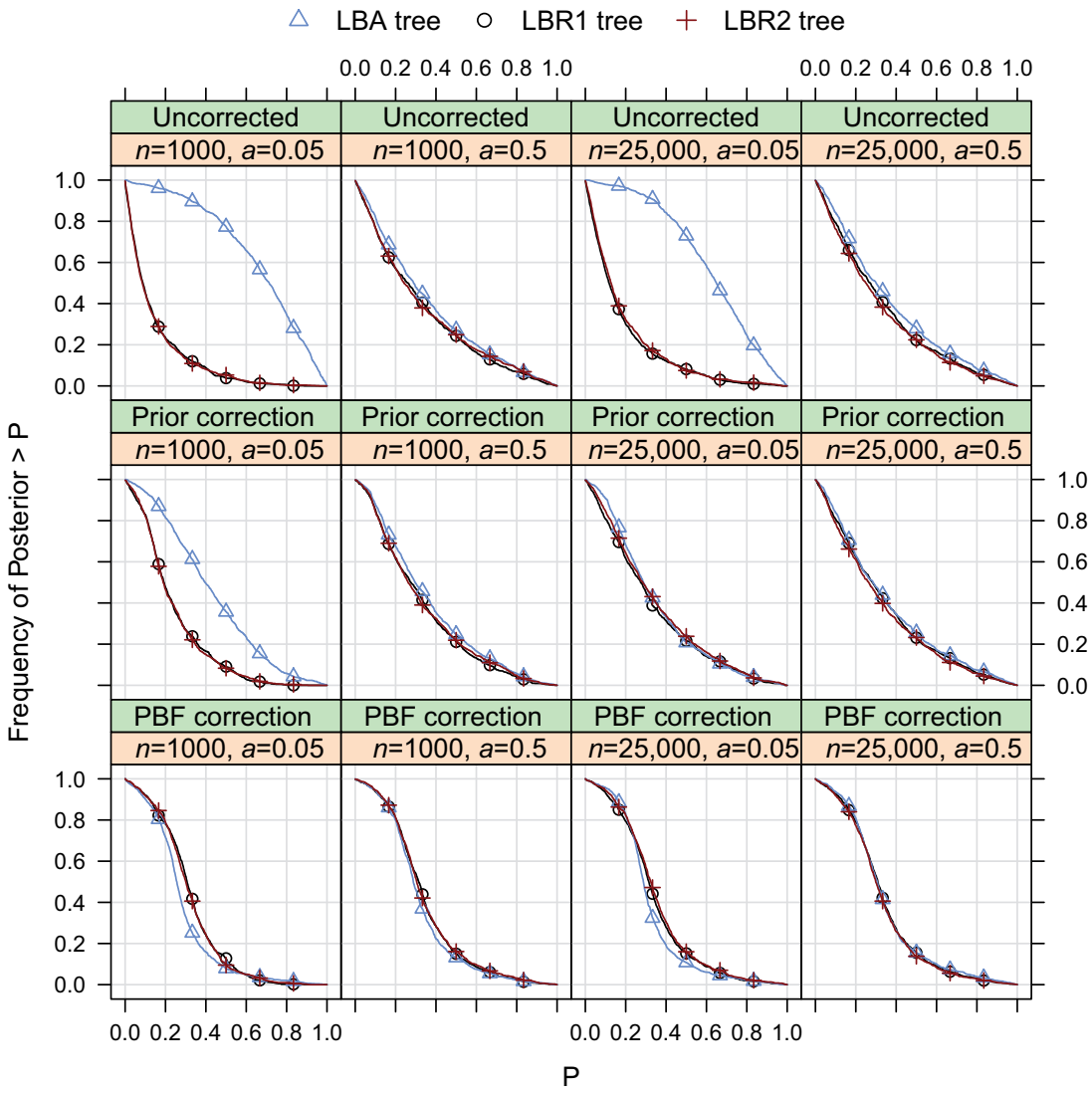

FIGURE 6. The frequencies with which posterior probabilities exceeded a threshold P in simulations from the partially resolved five-taxon tree of Figure 1i with long edge length $b=1$.

TABLE 1. The mean, median, and standard deviation (SD) of the large sample distributions of boundary correction terms when the correlation between the normal components is ρ.

ρ	Mean	Median	SD
−0.95	0.05	0.03	0.055
−0.50	0.17	0.11	0.164
0.00	0.25	0.19	0.221
0.50	0.33	0.28	0.255
0.95	0.45	0.43	0.286

TABLE 2. The correlations that are expected from the J_{jn}^{-1} matrix when data are generated from the five-taxon star tree in Figure 1 with $b=1$.

Tree	$a=0.05$	$a=0.5$
LBA	−0.059	−0.159
LBR1	−0.005	−0.124
LBR2	−0.624	−0.561

resolved tree has been used as a device to define bias. Partially resolved trees are topologies with some subset of internal edge lengths set to zero and such edge lengths are not considered a priori much more or less likely than most other sets of internal edge lengths; the edge length prior $\pi_j(t)$ allows them to be somewhat more or less likely but not so much so that they have positive prior probability as opposed to prior density. Including partially resolved trees as topologies in their own right has been put forward as a solution to the star-tree paradox (Yang 2007) and it is tempting to consider it as an alternative solution to the difficulties noted here. Including partially resolved trees will make it less likely that LBA trees will be estimated but only because partially resolved trees will then have

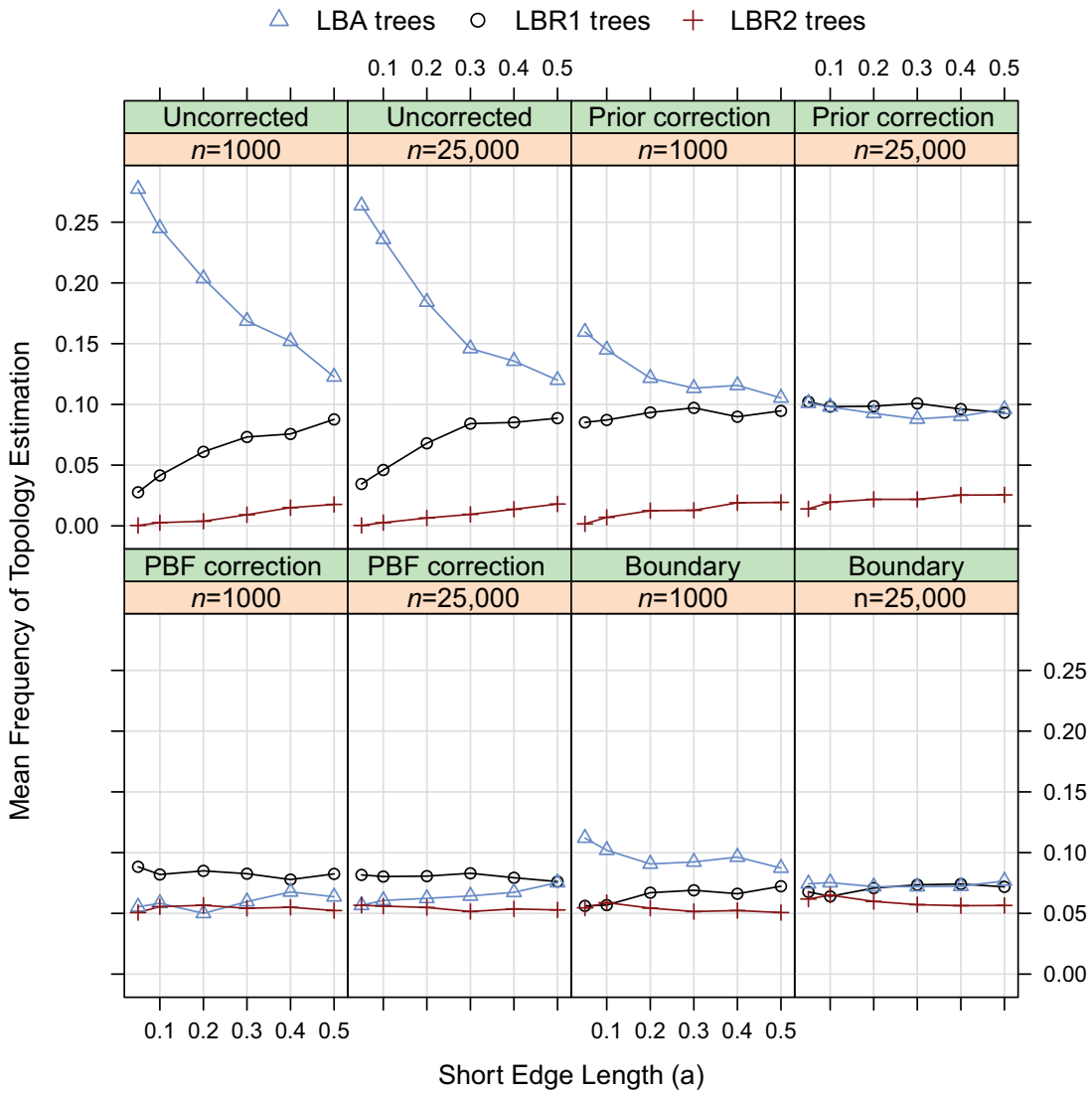

FIGURE 7. The mean frequencies with which topologies are estimated as a function of short edge lengths in simulations from a five-taxon star tree with long edge length $b = 1$.

nonnegligible posterior probability. The relative support for LBA trees, among well-resolved trees, will remain the same. The results here imply that it is more likely that either a LBA tree or a partially resolved tree will be estimated than the true tree when the true tree is an alternative well-resolved tree but with short internal edges. Moreover, there is reason to believe partially resolved trees will be estimated too frequently when the true tree is well resolved. In simple statistical settings involving model comparison, Lindley's paradox (Bartlett 1957; Lindley 1957) refers to the result that, irrespective of the data, relative support for a simple model to a complex model can become arbitrarily large as the variance of noninformative priors for parameters grows. The analogy here is that support for partially resolved trees (analogues of simple models) relative to well-resolved trees becomes arbitrarily large as the variance of edge length priors grows. Thus, implementations that included partially resolved trees as separate topologies would need to exercise greater caution in the choice of edge length priors to avoid large artefactual support for partially resolved trees.

As expected from theory, both the PBF correction and prior correction are effective with large sequence lengths. With smaller sequence lengths, however, the PBF correction was more effective than the prior correction with a slight tendency to overcorrect. By contrast, prior correction with smaller sequence lengths exhibited similar albeit less substantial biases to the uncorrected case. For both types of correction, it is unclear at present why there is a systematic tendency for correction in a particular direction.

The boundary correction was included partly to illustrate that the boundary factor of (3) is indeed a source of bias and can, in theory, be corrected. In addition to the increased complexity of calculation, practical

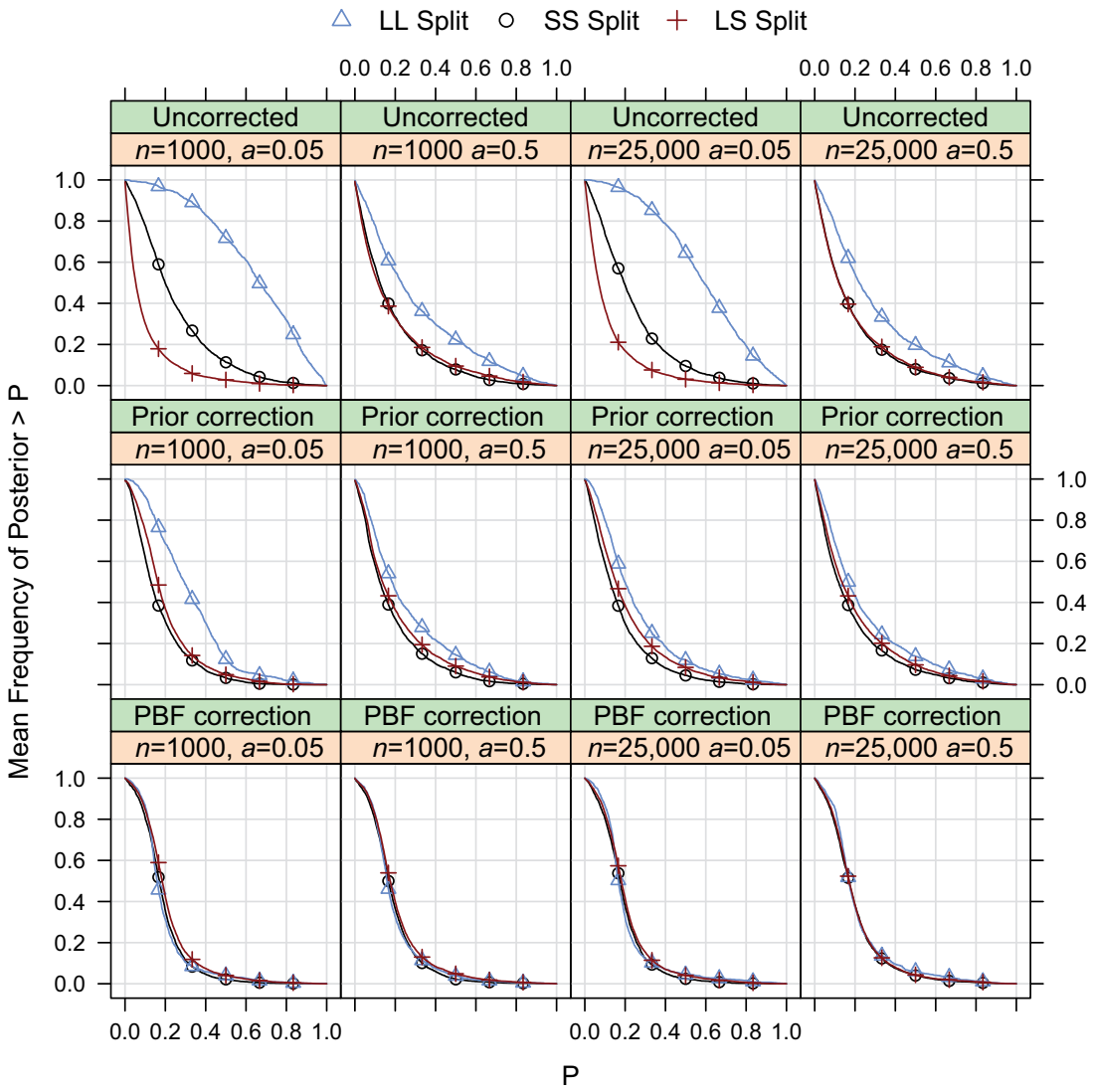

FIGURE 8. The mean frequencies with which posterior probabilities for splits exceeded a threshold P in simulations from a five-taxon star tree with long edge length $b=1$.

difficulties include that the indices of the edge lengths in t_* that are zero are not known. Potential strategies for adjusting for this include setting edge lengths below a certain threshold to zero or setting internal edge lengths to zero so that a set of well-supported trees are equivalent. In any case, the boundary factor bias did not seem as serious as the $|J_{jn}|^{-1/2}$ bias and the PBF correction makes some degree of automatic adjustment for it.

Two methods have been given for calculating PBF corrections. When allowed by the Bayesian implementation, the method that averages likelihoods is to be preferred to the method using data-doubling. First, data-doubling requires double the computation. In addition, when a topology is poorly supported, its posterior will be small for both the original and doubled data. Thus, $\eta_j^{(2)}/\eta_j$ can be reasonably estimated for well-supported topologies but the ratio will be unstable for poorly supported topologies. In the event that data-doubling is the only available option, topologies with very small η_j should likely be discarded from consideration at the outset. Software that implements most of the methods is available at

`http://www.mathstat.dal.ca/~tsusko`

The PBF correction can be viewed as replacing η_j with $\eta_j^{(2)}/\eta_j$ in the posterior $\eta_j/\sum_k \eta_k$. More generally, for $f>0$, let

$$\eta_j^{(f)} = (2\pi)^{-p/2}(fn)^{p/2}\alpha_j \int \exp\{f[l_j(t)-l_*(t_*)]\}\pi_j(t)\,dt$$

Then a similar argument as the one that gave (5) gives that

$$\eta_j^{(f)}/\eta_j \approx \exp[(f-1)S_{jn}^T J_{jn}^{-1} S_{jn}/2]u_j(\sqrt{f}S_{jn})/u_j(S_{jn})$$

Thus, replacing η_j with $\eta_j^{(f)}/\eta_j$ when $f > 1$ or $\eta_j/\eta_j^{(f)}$ when $0 < f < 1$ should similarly be effective at reducing bias. Optimal choice of f is a topic of current research. In the context of model selection, replacement of η_j by $\eta_j/\eta_j^{(f)}$ when $0 < f < 1$ gives rise to the fractional Bayes factor of O'Hagan (1995).

None of the corrections completely follows the Bayesian paradigm. The prior correction comes closest. The prior $\alpha_j \propto |J_{jn}(\hat{t}_j)|^{1/2}$ is not a valid prior because it depends on data. However, as sequence length gets large, $|J_{jn}(\hat{t}_j)|^{1/2} \approx |E[J_{jn}(t_*)]|^{1/2}$ which is a valid data-independent prior. Thus, the prior correction is asymptotically Bayesian. Closely connected to the prior correction is the Jeffreys prior (Jeffreys 1945), which could, in theory, provide an alternative correction. The Jeffreys prior for edge lengths is $\pi_j(t) \propto |E[J_{jn}(t)]|^{1/2}$. If it were used as an edge length prior, it can be seen from (3) that, since $|J_{jn}|^{1/2} \approx |E[J_{jn}(t_*)]|^{1/2} \propto \pi_j(t_*)$, $\pi_j(t_*)$ cancels the $|J_{jn}|^{-1/2}$ term. While the Jeffreys prior could, in principle, be included in Markov chain Monte Carlo implementations, repeated calculation of $|E[J_{jn}(t)]|$ makes it prohibitive and there are potential difficulties due to it being a potentially improper prior. The PBF corrected posteriors are not Bayesian even asymptotically. For the pragmatic, however, such concerns are less important than that the corrections are effective at eliminating an important source of bias.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.g180s.

Funding

This research was supported by a Discovery grant from the Natural Sciences and Engineering Research Council of Canada.

Acknowledgments

The author thank Jeet Sukumaran, Joe Bielawski, and Andrew Roger for valuable comments and discussion.

References

Aitkin M. 1991. Posterior Bayes factors. J. Roy. Stat. Soc. Ser. B. 53:111–142.
Allman E.S., Rhodes J.A., Sullivant S. 2012. When do phylogenetic mixture models mimic other phylogenetic models. Syst. Biol. 61:1049–1059.
Allman E.S., Ané C, Rhodes J.A. 2008. Identifiability of a Markovian model of molecular evolution with gamma-distributed rates. Adv. Appl. Prob. 40:229–249.
Bartlett M.S. 1957. A comment on D.V. Lindley's statistical paradox. Biometrika. 44:3–4.
Chang J.T. 1996. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. Math. Biosci. 137:51–37.
Cummings M.P., Handley S.A., Myers D.S., Reed D.L., Rokas A., Winka K. 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Syst. Biol. 52:477–487.
Felsenstein J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. 27:27–33.
Hendy M.D., Penny D. 1989. A framework for the study of evolutionary trees. Syst. Zool. 38:297–309.
Huelsenbeck J. 1995. Performance of phylogenetic methods in simulation. Syst. Biol. 44:17–48.
Huelsenbeck J.P., Ronquist F. 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.
Inagaki Y., Susko E., Fast N.M., Roger A.J. 2004. Covarion shifts cause a long branch attraction artifact that unites microsporidia and archaebacteria in EF-1α phylogenies. Mol. Biol. Evol. 21:1340–1349.
Jeffreys H. 1945. An invariant form for the prior probability in estimation problems. Proc. R. Soc. Lond. Ser. A. 186:453–461.
Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. New York: Academic Press. p. 21–132.
Kolaczowski B., Thornton J.W. 2006. Is there a star tree paradox? Mol. Biol. Evol. 23:1819–1823.
Kolaczkowski B., Thornton J.W. 2009. Long-branch attraction bias and inconsistency in bayesian phylogenetics. PloS ONE 4:e7891.
Lewis P.O., Holder M.T., Holsinger K.E. 2005. Polytomies and Bayesian phylogenetic inference. Syst. Biol. 54:241–253.
Lindley D. 1957. A statistical paradox. Biometrika. 44:187–192.
O'Hagan, A. 1995. Fractional Bayes factors for model comparison. J. Roy. Stat. Soc. Ser. B. 57:99–138.
Rambaut A., Grassly N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–8.
Ronquist F., Huelsenbeck J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.
Steel M., Matsen F.A. 2007. The Bayesian star paradox persists for long finite sequences. Mol. Biol. Evol. 24:1075–1079.
Suzuki Y., Glazko G.V., Nei M. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci USA 99:16138–16143.
Susko E. 2008. On the distributions of bootstrap support and posterior distributions for a star tree. Syst. Biol. 57:602–612.
Susko E., Inagaki Y., Roger A.J. 2004. On inconsistency of the neighbour joining method and least squares estimation when distances are incorrectly specified. Mol. Biol. Evol. 29:1629–1642.
Tierney L., Kadane J.B. 1986. Accurate approximations for posterior moments and marginal densities. J. Amer. Stat. Assoc. 81:82–86.
Yang Z. 2007. Fair-Balance paradox, star-tree paradox, and Bayesian phylogenetics. Mol. Biol. Evol. 24:1639–1655.
Yang Z., Rannala B. 2006. Branch-length prior influences Bayesian posterior probability of phylogeny. Syst. Biol. 54:455–470.