Syst. Biol. 57(4):602–612, 2008
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150802302468

On the Distributions of Bootstrap Support and Posterior Distributions for a Star Tree

EDWARD SUSKO

Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada B3H 3J5; E-mail: susko@mathstat.dal.ca

Abstract.— Several authors have recently noted that when data are generated from a star topology, posterior probabilities can often be very large, even with arbitrarily large sequence lengths. This is counter to intuition, which suggests convergence to the limit of equal probability for each topology. Here the limiting distributions of bootstrap support and posterior probabilities are obtained for a four-taxon star tree. Theoretical results are given, providing confirmation that this counterintuitive phenomenon holds for both posterior probabilities and bootstrap support. For large samples the limiting results for posterior probabilities are the same regardless of the prior. With equal-length terminal edges, the limiting distribution is similar but not the same across different choices for the lengths of the edges. In contrast to previous results, the case of unequal lengths of terminal edges is considered. With two long edges, the posterior probability of the tree with long edges together tends to be much larger. Using the neighbor-joining algorithm, with equal edge lengths, the distribution of bootstrap support tends to be qualitatively comparable to posterior probabilities. As with posterior probabilities, when two of the edges are long, bootstrap support for the tree with long branches together tends to be large. The bias is less pronounced, however, as the distribution of bootstrap support gets close to uniform for this tree, whereas posterior probabilities are much more likely to be large. Our findings for maximum likelihood estimation are based entirely on simulation and in contrast suggest that bootstrap support tends to be fairly constant across edge-length choices. [Bootstrap support; posterior probability; star tree.]

Downloaded from sysbio.oxfordjournals.org by guest on January 24, 2011

When data are generated from a four-taxon star topology, most people's intuition would lead them to expect statistical measures of uncertainty to show little or no support for any one of the resolved topologies (Fig. 1). Suzuki et al. (2002) seem to have been the first to have noticed that posterior probabilities for resolved topologies could occasionally be quite large. More surprisingly, they found that this phenomenon, which has come to be referred to as the star tree paradox, persists with large sequence lengths.

A number of authors took note of and followed up on the observations in Suzuki et al. (2002). Cummings et al. (2003) considered the relationship between bootstrap support and posterior probabilities with a four-taxon star topology, finding that values were, by and large, comparable and that both showed considerable variation. Because they used sequence lengths of 1000, the possibility was left open that bootstrap support and posterior probabilities would eventually converge upon some value. Lewis et al. (2005) conducted the first focused study on the issue, finding that variation in posterior probabilities persisted even with very large sequence lengths of 100,000. They noted that the paradox arose more generally in simpler problems such as coin-tossing experiments. As a remedy, they suggested placing unresolved and resolved trees on an equal footing by placing positive prior probability on unresolved trees. The work of Yang and Rannala (2006) expanded on some of the themes in Lewis et al. (2005). They gave further examples of simple problems where analogous issues arose and pointed out that such problems had a long history (Lindley, 1957; Shafer, 1982). They also extended the scope of remedies to include priors that place increasing density near unresolved trees.

Kolaczowski and Thornton (2006) argued that there is no star tree paradox. To examine limiting behavior, they conducted repeated MCMC analyses using expected pattern frequencies for star tree models as data and found posteriors close to 1/3 with little variation. They also found, however, that variance in posterior probabilities was not diminished even with sequence lengths as large as 10^7 and posteriors larger than 0.95 were occasionally found with large sequence lengths. Steel and Matsen (2007) proved that no matter how large the threshold, there was positive probability that any of the posteriors would be greater than it. Yang (2007) recently reconsidered the paradox as well. Using simulations and Laplace approximations similar to the ones here, the notion of a fixed limit for posterior probabilities was dismissed and, as in the earlier works of Lewis et al. (2005) and Yang and Rannala (2006), posterior probabilities with increasing mass near the star tree were put forward as a remedy.

These results are elaborated on here. Some of the reasons for continuing interest in the issue is to gain a more complete understanding of this surprising result and to see how far the results can be extended; extension here is primarily to differing edge lengths with additional results for equal edge lengths. Because star trees are the most extreme case of a poorly resolved tree, the approximations also provide a better understanding of the large sequence length properties of commonly used statistical measures of topological uncertainty when topologies are not well resolved. The interest in the paradox is, however, unusual in being a frequentist study of Bayesian probabilities; the generating star tree is fixed in analyses. Huelsenbeck and Rannala (2004) and Yang and Rannala (2006) make the important point, relevant to the study here, that Bayesian posteriors are not designed to have frequentist properties. Posterior probabilies are not supposed to yield 5% "type I errors," when data are generated from fixed parameters; indeed the notion of a type I error is not clearly relevant in a Bayesian context. Although it may thus be no surprise when Bayesian probabilities do not have certain properties in a frequentist simulation setting, this does not imply that what properties they *do* have in such settings are not of interest. They are of interest to researchers who

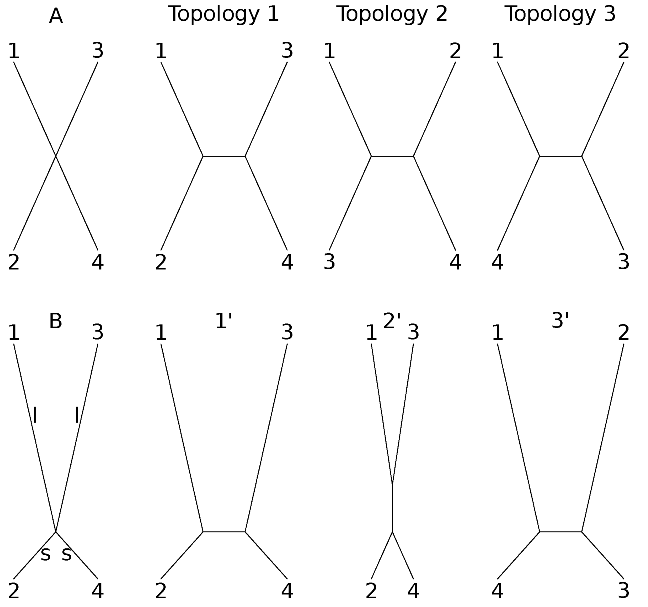

FIGURE 1. The star topology, with zero internal edge length, is given in A. The three resolved topologies are labeled 1 to 3. In some examples, for the generating star tree, two of the edges are long (l) and the other two are short (s), as in B. In this case the estimated trees for topologies 1 to 3 will usually be similar to 1′-3′.

are increasingly presented with, and need to make comparisons between, results presented according to both frequentist and Bayesian paradigms.

Characterizations of the distribution of posterior probabilities in a four-taxon case will be obtained here. The approach is similar to Yang (2007) in that Laplace approximations to posterior numerator terms are used. The limiting distribution turns out to be expressible as a transformation of dependent uniform and chi-squared random variables or, alternatively, as transformations of normal random variables. One of the merits of the theoretical approach is that it indicates that the results are not dependent on the prior.

Consistent with previous findings, even as sequence lengths get arbitrarily large, posterior probabilities do not converge to a fixed value—in particular, not to the anticipated fixed value of 1/3—but rather have a limiting distribution with positive variance. In contrast to most previous studies, consequences of unequal edge lengths in the tree are considered. When two of the edges for the star tree are long and the others are short, it turns out that the posterior probabilities for the topology with the long branches together tend to be much larger.

The main other measure of statistical uncertainty used in practice is bootstrap support. Although the rationales for the two measures are quite different, because they both arise in practice, there is considerable interest in whether they are comparable and in when and why they tend to be different. Our theoretical results for bootstrap support are limited to the more tractable case of the neighbor-joining algorithm (Saitou and Nei, 1987). The distribution of bootstrap support tends to be comparable to posterior probabilities for equal edge lengths. As the lengths of the long edges increase, much like posterior probabilities, bootstrap support for the long-branches-together tree tends to become larger, eventually getting close to a uniform distribution.

Simulation was required to obtain the results for likelihood methods. Surprisingly, the distribution is comparable to that of posterior probabilities with equal edge lengths but it remains relatively stable as the long edges increase.

METHODS

The Limiting Distribution of Posterior Probability

Similarly as in Yang, Laplace approximation (cf. Tierney and Kadane, 1986) is used to obtain the limiting distribution of the posterior distributions. In contrast to Yang, where approximations were based on expansion about the maximum likelihood estimator, the approximations used here are about the true star tree parameters. This makes approximation easier. For instance, one of the issues that Yang needed to deal with was the possible nonsingularity of the information matrix on random generations. Only one information matrix, the one corresponding to expansion about the true parameter values, is used here so that singularity either holds or it doesn't, once and for all. The derivation of the limiting distribution is given in Appendix A.

Assumptions used in obtaining the result are as follows. It is assumed that the terminal edges in the generating tree are all positive. It is also assumed that substitutions along an edge are according to a Markov model with nonzero frequencies of character states and nonzero rates of exchange. Together these assumptions imply that probabilities of any pattern of character states at a site are nonzero. It is also assumed that, for a given topology, edge lengths are identifiable: no two sets of edge lengths give exactly the same pattern probabilities. All of the examples use the Jukes-Cantor model (Jukes and Cantor, 1969), which, with positive terminal edges, satisfies the above assumptions. At least in a neighborhood of the true edges, the prior, $\pi(t)$, for the edge lengths is assumed to be continuous, positive, and have a bounded derivative. This assumption is satisfied for common priors like the uniform or exponential.

To establish some notation, label the four taxon topologies $j = 1$, 2, and 3 according to the neighbor, $j+1$, of 1 in Figure 1. Let $t_* = [t_{1*}, \ldots, t_{4*}, 0]$ denote the generating edge lengths for the star topology. Let $\sqrt{n}S_{jn}$ be the gradient of the log-likelihood for the jth topology, $l_j(t)$, evaluated at t_*, having rth element $\frac{\partial}{\partial t_r} l_j(t_*)$, and let I_j denote the covariance matrix of S_{jn}; $\sqrt{n}S_{jn}$ is sometimes referred to as the score function and nI_j are the expected information matrix, where n is the number of sites. As mentioned above, I_j is assumed positive definite.

The limiting distribution of posterior probability for the jth topology is the same as that of

$$\frac{\exp\left(\frac{1}{2}X_j^2\right)\ |I_j|^{-1/2}\ U_j}{\sum_k \exp\left(\frac{1}{2}X_k^2\right)\ |I_k|^{-1/2}\ U_k,} \qquad (1)$$

Downloaded from sysbio.oxfordjournals.org by guest on January 24, 2011

where the X_j^2 have chi-squared distributions with 5 degrees of freedom and the U_j have uniform distributions. The derivation is given in Appendix A, which explicitly indicates the nature of the dependence between the X_j^2 and U_j. Because the distributions of the X_j^2 and U_j are the same, regardless of the topology, the main factor leading to a potential bias of posteriors in favor of a particular topology is $|I_j|$. When the star tree has equal edge lengths, the I_j are all the same and there is no bias, but when some of the edges are of different length, the I_j become different and the distribution of the posterior for one of the topologies becomes skewed towards relatively larger values.

To briefly outline the reasons for the limiting distribution, the posterior for topology j can be expressed as $\eta_j / \sum_k \eta_k$, where

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

A Laplace approximation at the true edge lengths t_* gives the approximation

$$\eta_j \propto |I_j|^{-1/2} \exp\left[\frac{1}{2} S_{jn}^T I_j^{-1} S_{jn}\right] \Phi\left(\left[I_j^{-1} S_{jn}\right]_5 \Big/ \sqrt{\left[I_j^{-1}\right]_{55}} \right), \qquad (3)$$

where $\Phi(z) = P(Z \leq z)$ for a standard normal random variable Z. The second factor gives the U_j and $X_j^2 = S_{jn}^T I_j^{-1} S_{jn}$, the limiting distributions of which follow from likelihood theory.

As is shown in Appendix A, the posteriors can alternatively be characterized as having a limiting distribution that is the same as a known transformation of a normal random vector. In the results reported below, repeated simulation of normal random vectors and then transformation as described in Appendix A was used to obtain the distributions. Note that simulation here is much simpler than the usual approach. Usual approaches require repeat generations of large sequence alignments and then repeated generations within a Markov chain Monte Carlo (MCMC) algorithm, for each of these sequence alignments, to get posterior probabilities.

The Limiting Distribution of Bootstrap Support

In obtaining distributions for bootstrap support, attention will be restricted to the neighbor-joining algorithm and the Jukes-Cantor model (Jukes and Cantor, 1969); as before, nonzero terminal edge lengths are assumed. Although simulation results for maximum likelihood estimation will indicate that the distributions do not hold generally across different methods of estimation, I expect that they can be extended to other substitution models. Focus here is upon bootstrap support for topology 1, having taxa 1 and 2 as neighbors; the results imply the distributions for the other topologies, through a permutation of the taxon labels.

For the neighbor-joining algorithm, topology 1 is estimated if

$$\begin{aligned} Z_1 &= d_{23} + d_{14} - d_{12} - d_{34} > 0 \\ Z_2 &= d_{24} + d_{13} - d_{12} - d_{34} > 0 \end{aligned} \qquad (4)$$

where d_{ij} is the estimated distance for pair i and j. Let $G(v_1, v_2; \rho)$ denote the probability that two standard normal random variates with correlation ρ are greater than v_1 and v_2, respectively. Let ρ_* denote the limiting correlation between Z_1 and Z_2 in (4); an expression for ρ_* is given in Appendix B. The main result is that the distribution of bootstrap support is the same as the distribution of

$$G(V_1, V_2; \rho_*), \qquad (5)$$

where V_1 and V_2 are standard normal and also have correlation ρ_*. The derivation is given in Appendix B.

Special cases where explicit forms for the distribution can be obtained illustrate how the distribution changes as a function of ρ_*. In the case that $\rho_* = 1$, G gives the probability that two perfectly correlated standard normal random variables are greater than V_1. Because they are perfectly correlated, this is the same as the probability that either one of them is greater than V_1. In short, the limiting distribution of bootstrap support is the same as $1 - \Phi(V_1)$, where Φ is the standard normal cumulative distribution function. It can be shown that this random variable has a uniform distribution.

When the correlation is 0, the random variables are independent and it turns out that

$$G(V_1, V_2; \rho_*) = [1 - \Phi(V_1)]\,[1 - \Phi(V_2)].$$

Because V_1 and V_2 are independent in this case as well, bootstrap support has the distribution of the product of two independent uniform random variables. It can be shown that the probability that bootstrap support is less than v is then approximately $v - v\log(v)$. Compared to the uniform distribution, this distribution gives much smaller probabilities of large bootstrap support.

What the special cases suggest is that as correlation between Z_1 and Z_2 increases, the probability of large bootstrap support gets larger and attains a maximum when the correlation is 1. For the Jukes-Cantor model, this special case turns out to be of particular interest.

Results

With equal edge lengths, distributions of posteriors are the same across different topologies and comparable but different across choices of edge lengths. Moreover, there is a nonnegligible probability of posterior probabilities being large, regardless of the sequence length. This is illustrated in Figure 2, which gives the limiting survivor functions for different choices of edge lengths.

Downloaded from sysbio.oxfordjournals.org by guest on January 24, 2011

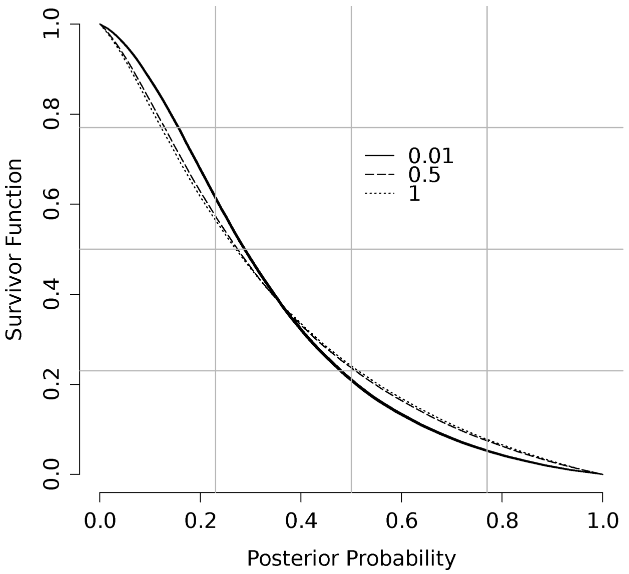

FIGURE 2. The limiting survivor function or frequency with which posterior probabilities are larger than p, plotted against p (posterior probability). The survivor functions are given for topology 1 and equal but differing terminal edge lengths. Because simulation was used to obtain the survivor functions, pointwise 95% bounds, using the same line types, have been included in the case that the equal edges are 0.01 and are barely visible because of the large number (10,000) of simulations.

The survivor function gives the probabilities of having posterior probabilities larger than the corresponding value on the x-axis; it is equal to 1 minus the cumulative distribution function. Although unconventional, this interpretation is of particular interest in the present setting due to the surprisingly nonnegligible limiting frequencies of large posteriors. Because it is known that the distributions are the same for each topology, the survivor functions are plotted for the first topology alone.

The survivor functions were obtained through repeated simulation of normal vectors, W, with mean 0 and variance-covariance matrix Σ having entries given by (A1) to (A3) in Appendix A. Transformation (A4) then gives the posteriors. A total of 100,000 simulations were conducted to obtain each curve. Because simulation was used to obtain the survivor functions, pointwise 95% bounds have been included in the case that the equal edge lengths are 0.01 to give an indication of possible effects of simulation; they are visible only as a thicker line. Although the survivor functions are comparable for different choices of equal edge lengths, they differ as well.

To check the large sample approximations against the results that one would obtain using MCMC techniques and fixed sequence lengths, MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) was used. A total of 1000 simulations were conducted at each parameter setting with sequence length 25,000 from the Jukes-Cantor substitution model. 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 10,000 MCMC generations were run with a sampling frequency of 10 and 25 burn-in samples (`ngen=10000, samplefreq=10, sumt burnin=25`). All other parameters were set to default values. The results are given in Figure 3 and indicate a very close fit between theoretical and actual distribution in the case that all edge lengths are equal.

Consider now the case where two of the terminal edges are long and two are short. Specifically, assume that the edges leading to 1 and 3 are long and that those leading to 2 and 4 are short. As discussed earlier, because the X_j^2 and U_j in (1) have the same distribution regardless of the topology j, the main reason that a bias may enter into the posteriors is that the $|I_j|^{-1/2}$ may be different. From (1) a rough measure of the posteriors that can be expected is what I refer to as the information ratio:

$$\frac{|I_2|^{-1/2}}{|I_1|^{-1/2}+|I_2|^{-1/2}+|I_3|^{-1/2}}. \qquad (6)$$

For any fixed choice of long branches that I have considered, as the ratio of short to long edge lengths increases from 0 to 1, the information ratio decreases from 1 to 1/3. This indicates that there will be a strong bias towards high posterior probabilities for the long-branch-attract topology 2 in the case that the ratio of short to long edge lengths is small. The bias towards high posteriors for the long-branch-attract topology 2 suggested by the information ratios is confirmed in Figure 3.

The distribution of bootstrap support for topology 1 using the neighbor-joining algorithm is determined by the correlation of Z_1 and Z_2 in (4). When the correlation is 1, the distribution is uniform and as the correlation decreases, the distribution becomes more skewed towards smaller values. I have calculated the correlations over a large range of long (l) and short (s) edge lengths. When $l = s$ the correlation is 0.5, but, as l gets relatively large, the correlation approaches 1 for the tree with two long branches together and 0 for the tree with long branches apart. As discussed in the previous subsection, the distribution of bootstrap support is uniform when the correlation is 1 and the probability of bootstrap support less than v is $v - v\log(v)$ when the correlation is 0. Figure 4 gives the limiting distributions of bootstrap support and the estimated survivor functions when sequence length is 10,000 based on 1000 simulations from the Jukes-Cantor model for the same edge-length settings and topologies as were considered in Figure 3. There is a very close fit between the theoretical and finite sequence-length distributions regardless of edge-length setting. As expected based on the calculated limiting correlations, distributions are very close to uniform (linear survivor functions) in cases when two of the edges are long and two are short.

Theoretical limiting distributions were difficult to obtain for bootstrap support using maximum likelihood estimation and thus relied entirely on simulation. The DNAML routine in the PHYLIP package (Felsenstein, 1989, 2004) was used to obtain maximum likelihood estimates. A total of 1000 simulations were conducted for each edge-length setting, each with sequence length

Downloaded from sysbio.oxfordjournals.org by guest on January 24, 2011

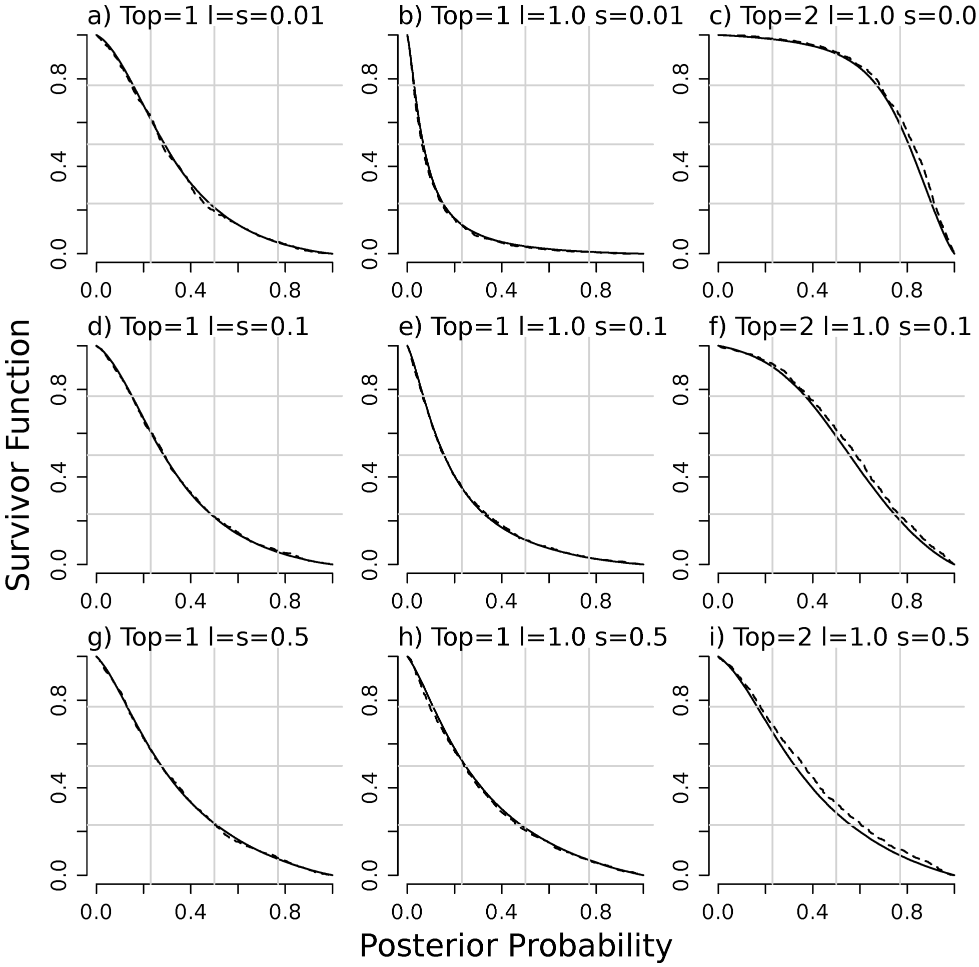

FIGURE 3. The limiting survivor functions of posterior probabilities of topologies (top) are plotted for several choices of edge lengths (solid line). *l* gives the edge lengths for edges leading to 1 and 3, and *s* gives the edge lengths for edges leading to 2 and 4. Plotted as well are the estimated survivor functions with sequence length 25,000 from 1000 simulations using MrBayes to obtain posteriors (dashed line). The distribution for topology 1 is the only one plotted in the case of equal edge lengths; the distributions are the same for all topologies. The distributions for topologies 1 and 2 are plotted in the case of two long and two short branches; the distributions for topologies 1 and 3 are the same because of the symmetry of the problem.

Downloaded from sysbio.oxfordjournals.org by guest on January 24, 2011

10,000; similar results were obtained with sequence lengths of 25,000. In contrast to bootstrap support for distance methods, the survivor functions are similar across all edge-length settings (Fig. 5) and, in fact, are similar to the limiting distributions of posteriors and bootstrap support for the neighbor-joining algorithm when the generating tree has all equal edge lengths (Fig. 6).

DISCUSSION

The results indicating that bootstrap support and posteriors tend to be comparable when edge lengths are equal, but differ substantially when long edges are present, are similar to the findings in Cummings et al. (2003), who found the biggest differences in what they referred to as the two branch corner. The Cummings et al. (2003) star tree simulations found bootstrap support and posterior probabilities to be comparable. Because the simulation had equal expected (simulations did not use constant edges) external edge lengths, the findings are consistent with the ones here.

In the case of equal edge lengths, the distribution of posterior probabilities differed across choices of edge lengths. This finding differs somewhat from Yang (2007). However, the direct results there were for a three-taxon rooted tree. The analogue of $\sqrt{n}S_{jn}$ in the three-taxon problem is the gradient of the likelihood, now $l_j(t_1, t_0)$; edge lengths t_0 and t_1 are as in figure 2 of Yang (2007). The information matrix, I_j, can still be taken as the covariance matrix of S_{jn}. With these changes in interpretation, the argument in Appendix A for the approximations to the posterior can be repeated to give (3). Substantial simplification of the score and information is possible in this three-taxon problem and can be used to show that the limiting distribution of the posterior for topology j is

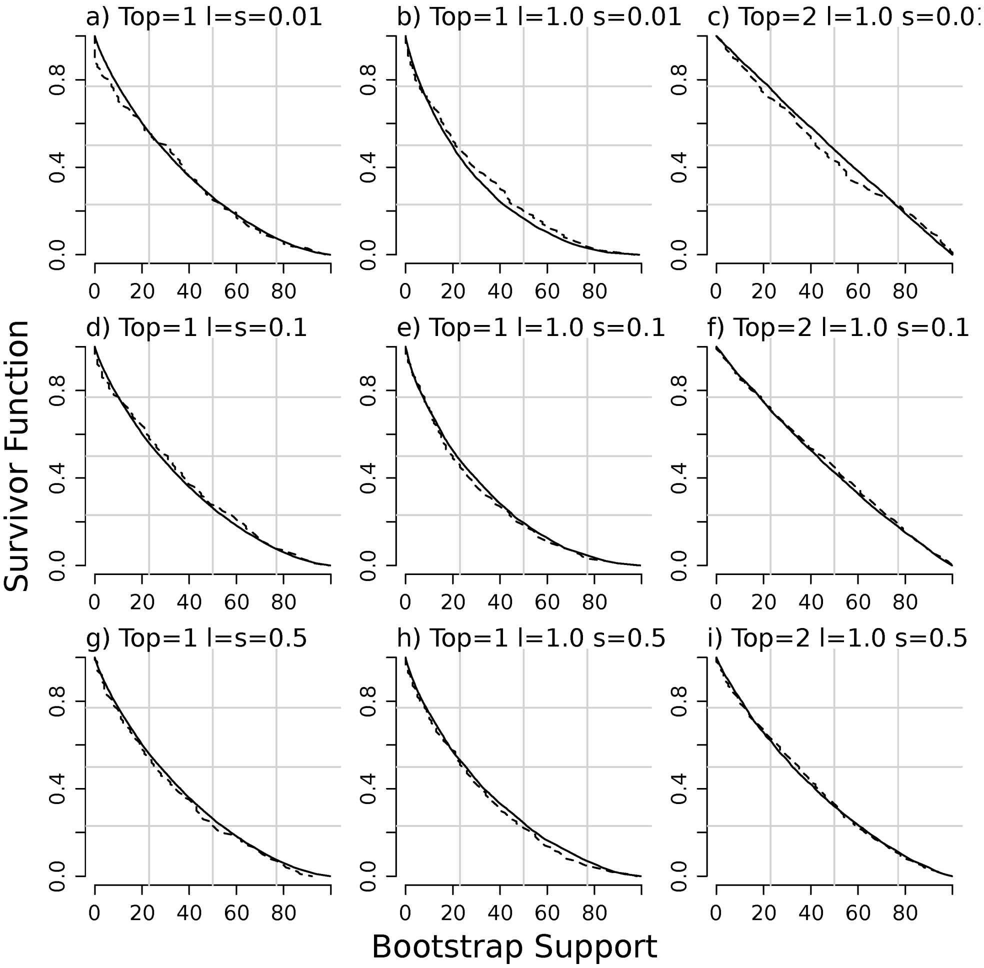

FIGURE 4. The limiting survivor functions of bootstrap support of topologies (top) for the neighbor-joining algorithm (solid line) and the estimated survivor function for sequence length 10,000 based on 1000 simulations from the Jukes-Cantor model for each edge-length setting (dashed line). Edge length and topology settings are the same as in Figure 3.

the same as that of

$$\frac{\exp\left[Z_j^2/2\right]\Phi(Z_j)}{\sum_k \exp\left[Z_k^2/2\right]\Phi(Z_k)}, \tag{7}$$

where $Z_3 = -(Z_1 + Z_2)$ and $[Z_1, Z_2]$ has a multivariate normal distribution with mean 0, variances equal to 1, and covariance $-1/2$. This distribution does not depend upon the edges and thus provides a mathematical argument for the conclusion in Yang (2007) that the distribution does not vary across edge lengths.

Most common edge-length priors used in practice, including the exponential and uniform priors, satisfy the conditions required for the limiting distributions. It is interesting to note that the limiting distribution of posterior probability does not depend on the prior, except in as much as it must satisfy the initial assumptions. The star-tree paradox is a general one that is not due to, nor influenced by, the choice of priors.

Lewis et al. (2005) pointed out that placing positive prior probability at the star topology will cause the prior for the star topology to converge to 1. Yang and Rannala (2006) and Yang (2007) extended this result to priors that place prior density near the star topology that increases quickly enough with n. If density near the star topology increases without bound, but slowly enough, posteriors for each of the topologies are expected to converge to $1/3$. Not surprisingly, because different distributions result, priors such as these do not satisfy the assumptions here. Although such solutions provide insight into how comparison of models of differing dimension can create unusual behavior, it is exactly the type of prior assignment that has been one of the primary criticisms of Bayesian approaches: choosing priors biased towards a particular point of interest.

Another assumption that was made was that edge lengths are identifiable parameters: no two sets of edge lengths give exactly the same pattern probabilities. If this is not the case, integration over regions of t that are far from the generating t_* cannot be ignored with large sequence lengths, as they are in Appendix A. It is likely that the distributions and properties of posterior support will be quite different. For models that do not involve

Downloaded from sysbio.oxfordjournals.org by guest on January 24, 2011

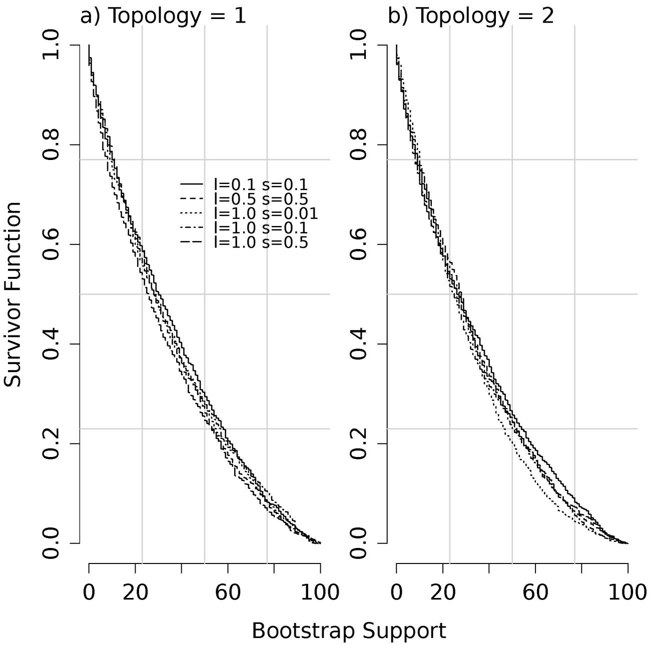

FIGURE 5. The estimated survivor functions for bootstrap support using maximum likelihood estimation. Results are based on 1000 simulations of sequences of length 10,000 from the known Jukes-Cantor model for each edge-length setting. Edge lengths to taxa 1 and 3 are denoted by l, those leading to 2 and 4 are denoted by s. The distributions for topologies 1 and 2 are plotted; the distributions for topologies 1 and 3 are the same because of the symmetry of the problem.

rates across sites, as is the case here, identifiability assumptions have been shown to hold in Chang (1996). For models that do involve rates across sites, however, Steel et al. (1994) illustrate that identifiability does not generally hold. Recent work (Allman and Rhodes, 2008; Allman et al., 2008) indicates, however, that for common rates across sites models, like the gamma model of Yang (1994), whereas there may be choices of generating t_* that are not identifiable, there will usually be many for which identifiability holds.

Although the results are restricted to four taxa, they are suggestive of behavior with larger trees. For more general trees, the results may be encouraging in that they suggest edge-length priors are not likely to have a large influence on a Bayesian analysis when sequence length is large. On the other hand, the high probability of large posteriors when both long and short edge lengths are present in the generating tree suggests that large posteriors for splits should be of concern when long and short edge lengths are present in a tree. The results also provide some insight into some of the reasons that bootstrap support does not always correlate well with posterior probabilities. For a poorly resolved split, with equal edges, similar distributions are expected but, in cases of long and short branches, posterior probabilities for the long-branch-together split tend to be larger.

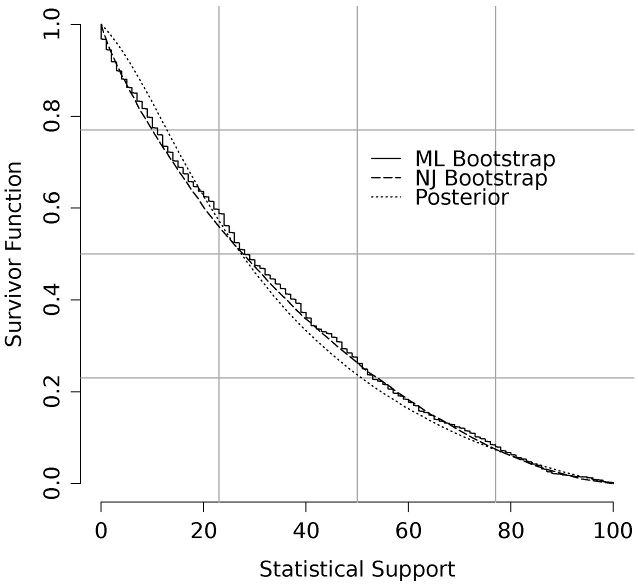

FIGURE 6. The estimated survivor function for bootstrap support using maximum likelihood estimation (ML bootstrap), the theoretical limiting survivor of bootstrap support for the neighbor-joining algorithm (NJ bootstrap), and the limiting survivor function of posterior probability when the generating star tree has equal edge lengths of length 0.5.

ACKNOWLEDGMENTS

I would like to thank Andrew Roger, Mike Steel, Ziheng Yang, and an anonymous referee for valuable comments. This research was supported by a Discovery grant from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Allman, E. S., and J. A. Rhodes. 2008. Identifying evolutionary trees and substitution parameters for the general Markov model with invariable sites. Math. Biosci. 211:18–33.

Allman, E. S., C. Ané, and J. A. Rhodes. 2008. Identifiability of a Markovian model of molecular evolution with gamma-distributed rates. Adv. Appl. Prob. 40:229–249.

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

Cummings, M. P., S. A., Handley, D. S., Myers, D. L., Reed, A. Rokas, and K. Winka. 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Syst. Biol. 52:477–487.

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. Department of Genome Sciences, University of Washington, Seattle.

Huelsenbeck, J. P., and B. Rannala. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Syst. Biol. 53:904–913.

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

Jukes, T. H., and C. R. Cantor. 1969. Pages 21–123 *in* Mammalian protein metabolism. Academic Press, New York.

Kiefer, J., and J. Wolfowitz. 1956. Consistency of the maximum likelihood estimates in the presence of infinitely many nuisance parameters. Ann. of Math. Stat. 27:887–906.

Kolaczowski, B., and J. W. Thornton. 2006. Is there a star tree paradox? Mol. Biol. Evol. 23:1819–1823.

Downloaded from sysbio.oxfordjournals.org by guest on January 24, 2011

Lehman, E. L. 1983. Theory of point estimation. Wiley, New York.

Lewis, P. O., M.T. Holder and K.E. Holsinger. 2005. Polytomies and Bayesian phylogenetic inference. Syst. Biol. 54:241–253.

Lindley, D. V. 1957. A statistical paradox. Biometrika 44:187–192.

Pawitan, Y. 2001. In all likelihood: Statistical modelling and inference using likelihood. Oxford University Press, Oxford UK.

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

Saitou, N., and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing evolutionary trees. Mol. Biol. Evol. 4:406–425.

Shafer, G. 1982. Lindleys paradox. J. Am. Stat. Assoc. 77:325–334.

Steel, M., and F. A. Matsen. 2007. The Bayesian star paradox persists for long finite sequences. Mol. Biol. Evol. 24:1075–1079.

Steel, M., L. A. Székely, and M. D. Hendy. 1994. Reconstructing trees when sequence sites evolve at variable rates. J. Comp. Biol. 1:153–163.

Suzuki, Y., Glazko, G. V. and M. Nei. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. USA 99:16138–16143.

Tierney L, and J. B. Kadane. 1986. Accurate approximations for posterior moments and marginal densities. J. Am. Stat. Assoc. 81:82–86.

Wald, A. 1949. Note on the consistency of the maximum likelihood estimate. Ann. of Math. Stat. 20:595–601.

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. 2007. Fair-Balance paradox, star-tree paradox, and Bayesian phylogenetics. Mol. Biol. Evol. 24:1639–1655.

Yang, Z., and B. Rannala. 2006. Branch-Length prior influences Bayesian posterior probability of phylogeny. Syst. Biol. 54:455–470.

First submitted 7 January 2008; reviews returned 26 February 2008; final acceptance 28 April 2008
Associate Editor: Paul Lewis

Appendix A: The Limiting Distribution of Posterior Probability

A more complete statement of the limiting distribution result for posterior probabilities follows; it is the one that is used in the repeated simulations of normals.

Let $p_k(t; j)$ denote the pattern probabilities for the pattern k, for the jth topology and edge lengths $t = [t_1, \dots, t_4, t_5]$; k is the pattern for an alignment column, for instance, $k = xxyy$. Here t_5 denotes the internal edge length and t_l the terminal edge length leading to l. Let $t_* = [t_{1*}, \dots, t_{4*}, 0]$ denote the generating edge lengths for the star topology. Let $p_k = p_k(t_*)$ denote the pattern probabilities for the star topology. Similarly, let $\hat{p}_k$ denote the proportion of times pattern k arose in the original alignment.

Let $W = [W_1, \dots, W_7]^T$ be $N(0, \Sigma)$ (normal distributed with mean 0 and covariance matrix Σ) where Σ is given by the following relationships for $i,\ j \leq 4,\ l, l' \leq 3$

$$\mathrm{cov}(W_i, W_j) = \sum p_k^{-1} \frac{\partial}{\partial t_i} p_k(t_*) \frac{\partial}{\partial t_j} p_k(t_*) \tag{A1}$$

$$\mathrm{cov}(W_i, W_{4+l}) = \sum p_k^{-1} \frac{\partial}{\partial t_i} p_k(t_*) \frac{\partial}{\partial t_5} p_k(t_*; l) \tag{A2}$$

$$\mathrm{cov}(W_{4+l'}, W_{4+l}) = \sum p_k^{-1} \frac{\partial}{\partial t_5} p_k(t_*; l') \frac{\partial}{\partial t_5} p_k(t_*; l) \tag{A3}$$

Let $V_j = [W_1, \dots, W_4, W_{4+j}]^T$ and let I_j denote the variance-covariance matrix of V_j. The limiting distribution of the posterior for the jth topology is then the same as that of

$$\frac{\exp\left(\frac{1}{2} V_j^T I_j^{-1} V_j\right)\ |I_j|^{-1/2}\ \Phi\Big(\big[I_j^{-1} V_j\big]_5 / \sqrt{\big[I_j^{-1}\big]_{55}}\Big)}{\sum_k \exp\left(\frac{1}{2} V_k^T I_k^{-1} V_k\right)\ |I_k|^{-1/2}\ \Phi\Big(\big[I_k^{-1} V_k\big]_5 / \sqrt{\big[I_k^{-1}\big]_{55}}\Big)}, \tag{A4}$$

where $|I_j|$ denotes the determinant of I_j and $\Phi(z) = P(Z \leq z)$ for $Z \sim N(0, 1)$.

The posterior for topology j can be alternatively expressed as the $\eta_j / \sum_k \eta_k$, where

$$\eta_j = n^{-1/2} \int_{t \geq 0} \exp[l_j(t) - l_j(t_*)] \pi(t)\, dt. \tag{A5}$$

The limiting distribution of the posterior is then determined by the limiting distribution of $[\eta_1, \eta_2, \eta_3]$. Unless otherwise stated, convergence means convergence of distributions. The reader interested in a overview of the argument might read the first few sentences of the sections starting in bold below.

The integral over $||t - t_*|| > \epsilon$. Because the log likelihood will only be large for t near the generating t_* for large n, it can be shown that for any $\epsilon > 0$,

$$\eta_j = n^{-1/2} \int_{\{t \geq 0, ||t - t_*|| \leq \epsilon\}} \exp[l_j(t) - l_j(t_*)] \pi(t)\, dt + R_{\epsilon n}, \tag{A6}$$

where $R_{\epsilon n} \to 0$.

This follows from the consistency of ML estimation for the problem (Wald, 1949; Kiefer and Wolfowitz, 1956). These results require that we be able to extend the definition of $p_k(t; j)$ to values of t where one or more of the edge lengths are infinite, and that limits of $p_k(t; j)$ coincide with this definition as one or more of the edge-lengths diverge in a corresponding manner. We can do this by independently calculating probabilities of patterns for (groups of) taxa separated by an infinite edge length. Because for the original edge lengths, t_*, probabilities of patterns will not be independent for any two sets of taxa, the extended definition will give rise to pattern probabilities that differ from the true pattern probabilities. Because it has been assumed that pattern probabilities from any $t \neq t_*$ will differ from the true pattern probabilities, identifiability of edge lengths holds more generally on this extended parameter space that includes infinite edge lengths.

Because all expectations here involve finite sums, because derivatives of any order of $p_k(t; j)$ with respect to edge lengths can be taken, and because all of the $p_k > 0$, it is not difficult to show that all other regularity conditions required for consistency are satisfied. Equation (2.12) of Kiefer and Wolfowitz (1956) then applies giving that, with probability 1, for $\epsilon > 0$, there will be an $0 < h < 1$ with

$$\max_{||t - t_*|| > \epsilon} \exp[l_n(t) - l_n(t_*)] \leq h^n$$

for all large n. It follows that

$$R_{\epsilon n} = n^{-1/2} \int_{\{t \geq 0, ||t - t_*|| > \epsilon\}} \exp[l_j(t) - l_j(t_*)] \pi(t)\, dt$$
$$\leq h^n \int_{\{t \geq 0\}} \pi(t)\, dt = h^n \to 0.$$

An approximation for $\exp[l_j(t) - l_j(t_*)]$. Let

$$[J_{jn}]_{kl} = -\frac{\partial^2}{\partial t_k \partial t_l} l_j(t_*)/n \tag{A7}$$

Downloaded from sysbio.oxfordjournals.org by guest on January 24, 2011

and let $\sqrt{n}S_{jn}$ denote the gradient of $l_j(t)$ at t_*. A Taylor's series expansion gives

$$l_j(t) - l_j(t_*) = \sqrt{n}S_{jn}^T(t - t_*) - \frac{n}{2}(t - t_*)^T J_{jn}(t - t_*) + R_{jn}. \quad \text{(A8)}$$

where the remainder term R_{jn} in (A8) can be expressed as

$$R_{jn} = n \cdot \sum_{l,m,p} \left[\sum_k \hat{p}_k \frac{\partial^3}{\partial t_l \partial t_m \partial_p} \log[p_k(\tilde{t}; j)](t_l - t_{*l})(t_m - t_{*m})(t_p - t_{*p}), \right. \quad \text{(A9)}$$

with $\tilde{t}$ a value satisfying $||\tilde{t} - t^*|| < \epsilon$.

A bound on the remainder term. There is a constant K such that, with probability 1, for large n, for all $||t - t_*||| < \epsilon$,

$$|R_{jn}| \leq nK||t - t_*||^3. \quad \text{(A10)}$$

The third-order partial derivatives of $\log[p_k(t; j)]$ are sums of terms (raised to a positive power) that have partial derivatives of $p_k(t; j)$ in the numerator and $p_k(t; j)$ in the denominator. For Markov substitution models, partial derivatives of $p_k(t; j)$ are bounded as function of t. Because the terminal edge lengths in t_* are positive, for ϵ sufficiently small and $||t - t_*|| < \epsilon$, t will have positive terminal edge lengths as well. For Markov substitution models with nonzero frequencies and rates of exchange, this is sufficient for the pattern probabilities, $p_k(t; j)$ to be positive. We can thus conclude that there is a K such that, for all patterns k and all $||t - t^*|| < \epsilon$,

$$\left| \frac{\partial^3}{\partial t_l \partial t_m \partial_p} \log[p_k(t; j)] \right| \leq K.$$

It then follows that

$$\sum_k \hat{p}_k \left| \frac{\partial^3}{\partial t_l \partial t_m \partial_p} \log[p_k(t; j)] \right| \leq \sum_k \hat{p}_k K = K$$

Considering (A9) this gives (A10).

Large-sample likelihood theory. Two results used in what follows come from the large-sample likelihood theory (c.f. Lehman, 1983; Pawitan, 2001, chapter 9). These results are that S_{jn} converges in distribution to a V_j that has a $N(0, I_j)$ distribution and that J_{jn} converges to I_j. The matrix I_j is assumed positive definite (whether it is or not can be figured out in any particular case and was always true in our examples). The large-sample theory requires that regularity conditions be satisfied. Because all expectations here involve finite sums, because derivatives of any order of $p_k(t; j)$ with respect to edge lengths can be taken and because all of the $p_k > 0$, it is not difficult to show that these regularity conditions are satisfied.

The integral over $n^{-2/5} < ||t - t_*|| < \epsilon$. The large-sample likelihood theory results and the bound (A10) can be used to show that $\exp[l_j(t) - l_j(t_*)]$ is small enough for $n^{-2/5} < ||t - t_*|| < \epsilon$ that (A6) can be refined to

$$\eta_j = n^{-1/2} \int_{\{t \geq 0, ||t - t_*|| \leq n^{-2/5}\}} \exp[l_j(t) - l_j(t_*)]\pi(t)\, dt + R_{In}, \quad \text{(A11)}$$

where $R_{In} \to 0$.

Substituting the bound (A10) in (A8) gives that for $n^{-2/5} < ||t - t_*||$

$$\frac{l_j(t) - l_j(t_*)}{n||t - t_*||^2} \leq n^{-1/2} S_{jn}^T w/||t - t_*|| - \frac{1}{2} w^T J_{jn} w + K||w||^2||t - t_*||$$

$$\leq n^{-1/10} S_{jn}^T w - \frac{1}{2} w^T J_{jn} w + K||w||^2||t - t_*||, \quad \text{(A12)}$$

where $w = (t - t_*)/||t - t_*||$. Because S_{jn} converges to a normal random variable, $n^{-1/10} S_{jn}$ converges to 0. Because J_{jn} converges to a positive definite matrix, the second term will be bounded above by a negative number. The third term will be small as long as $||t - t_*||$ is. Thus for $\epsilon > 0$ sufficiently small and all $n^{-2/5} < ||t - t_*|| < \epsilon$, there is a $\delta > 0$ such that $\{l_j(t) - l_j(t_*)\}/(n||t - t_*||^2) \leq -\delta$. Thus

$$\begin{aligned} R_{In} &= R_{\epsilon n} + n^{-1/2} \int_{\{t \geq 0, n^{-2/5} \leq ||t - t_*|| \leq \epsilon\}} \exp[l_j(t) - l_j(t_*)]\pi(t)\, dt \\ &\leq R_{\epsilon n} + n^{-1/2} \int_{\{t \geq 0, n^{-2/5} \leq ||t - t_*|| \leq \epsilon\}} \exp[-n||t - t_*||^2 \delta]\pi(t)\, dt \\ &\leq R_{\epsilon n} + n^{-1/2} \exp[-n^{1/5}\delta] \int_{\{t \geq 0, n^{-2/5} \leq ||t - t_*|| \leq \epsilon\}} \pi(t)\, dt \to 0. \end{aligned}$$

The integral over $||t - t_*|| \leq n^{-2/5}$. For $||t - t_*|| \leq n^{-2/5}$, the bound (A10) on the remainder term for (A8) gives that $|R_{jn}| \leq Kn^{-1/5}$. It follows that

$$\begin{aligned} \eta_j &= n^{-1/2} \int_{\{t \geq 0, ||t - t_*|| \leq n^{-2/5}\}} \exp[l_j(t) - l_j(t_*)]\pi(t)\, dt + R_{In} \\ &= n^{-1/2} C_n \int_{\{t \geq 0, ||t - t_*|| < n^{-2/5}\}} \exp\left\{ \sqrt{n} S_{jn}^T (t - t_*) - \frac{n}{2}(t - t_*)^T J_{jn}(t - t_*) \right\} \\ &\quad \times \pi(t)\, dt + R_{In}. \end{aligned}$$

where $\exp[-n^{-1/5}K] \leq C_n \leq \exp[n^{-1/5}K]$ converges to 1. A change of variables to $u = \sqrt{n}(t - t_*)$ gives

$$\eta_j = C_n \int_{\{-\sqrt{n}t_* \leq u < n^{1/10}\}} \exp\left\{ S_{jn}^T u - \frac{1}{2} u^T J_{jn} u \right\} \pi(t_* + u/\sqrt{n})\, du + R_{In}. \quad \text{(A13)}$$

Bounding the integral for $||u|| > n^y$. Because $\exp\{S_{jn}^T u - \frac{1}{2} u^T J_{jn} u\}$ is small for large $||u||$,

$$\int_{||u|| \geq n^y} \exp\left\{ S_{jn}^T u - \frac{1}{2} u^T J_{jn} u \right\} \pi(t_* + u/\sqrt{n})\, du \to 0$$

for any value $y > 0$. Consequently, any terms in (A13) involving integration over values of $||u|| > n^y$ can be ignored or included. They merely add a remainder term to R_{In} that still converges to 0.

For $||u|| \geq n^y$,

$$\left(S_{jn}^T u - \frac{1}{2} u^T J_{jn} u \right) / ||u||^2 = S_{jn}^T w/||u||^2 - \frac{1}{2} w^T J_{jn} w,$$

where $w = u/||u||$. Because S_{jn} is bounded the first term converges to 0. Since J_{jn} converges to I_j positive definite, the second term is bounded above by a negative number. It follows that for $||u|| \geq n^y$

$$\exp\left\{ S_{jn}^T u - \frac{1}{2} u^T J_{jn} u \right\} \leq \exp[-||u||^2 K'] \leq \exp[-n^{2y} K']$$

Downloaded from sysbio.oxfordjournals.org by guest on January 24, 2011

for some constant $K' > 0$. Thus

$$\int_{||u|| \geq n^y} \exp\left\{ S_{jn}^T u - \frac{1}{2} u^T J_{jn} u \right\} \pi(t_* + u/\sqrt{n})\, du$$
$$\leq \exp[-n^{2y} K'] \to 0.$$

Replacing the factor due to the prior.

$$\eta_j = C_n \pi(t_*) \int_{\{-\sqrt{n}t_* \leq u < n^{1/10}\}} \exp\left\{ S_{jn}^T u - \frac{1}{2} u^T J_{jn} u \right\} du + R_{In}. \quad \text{(A14)}$$

For $||u|| \leq n^{1/4}$, the bounded derivative of the prior at t_* implies that for some K_*, $|\pi(t_* + u/\sqrt{n}) - \pi(t_*)| \leq K_* n^{-1/4}$ This gives that

$$|\int_{\{-\sqrt{n}t_* \leq u < n^{1/10}, ||u|| < n^{1/4}\}} \exp\left\{ S_{jn}^T u - \frac{1}{2} u^T J_{jn} u \right\} \{\pi(t_* + u/\sqrt{n}) - \pi(t_*)\}\, du|$$
$$\leq K_* n^{-1/4} \int \exp\left\{ S_{jn}^T u - \frac{1}{2} u^T J_{jn} u \right\} du. \quad \text{(A15)}$$

The integral can be evaluated by expressing it as a multivariate normal integral:

$$\int \exp\left\{ S_{jn}^T u - \frac{1}{2} u^T J_{jn} u \right\}$$
$$= (2\pi)^{p/2} \left| J_{jn}^{-1} \right|^{1/2} \exp\left[\frac{1}{2} S_{jn}^T J_{jn}^{-1} S_{jn} \right]$$
$$\times \int \exp\left\{ -\frac{1}{2} \left[u - J_{jn}^{-1} S_{jn} \right]^T J_{jn} \left[u - J_{jn}^{-1} \right] \right\} \frac{1}{(2\pi)^{p/2} \left| J_{jn}^{-1} \right|^{1/2}}\, du$$
$$= (2\pi)^{p/2} \left| J_{jn}^{-1} \right|^{1/2} \exp\left[\frac{1}{2} S_{jn}^T J_{jn}^{-1} S_{jn} \right]. \quad \text{(A16)}$$

The first term converges to $|I_j^{-1}|^{1/2}$ and the second is bounded in probability. Together with (A15) and the fact that the integral over $||u|| > n^{1/4}$ can be added or ignored, this gives (A14).

A normal integral expression. Because the contribution to the integral converges to 0, the region $u > n^{1/10}$ can be added to the region of integration in (A14). For the terminal edges, Because $t_{*j} > 0$, $-\sqrt{n}t_{*j}$ will be less than $-n^{1/4}$ for n large. Thus the regions $u_j < -\sqrt{n}t_{*j}$ can be added to the region of integration in (A14). For the middle edge, however, $t_{*5} = 0$ and so the constraint $u > -\sqrt{n}t_*$ in (A14) gives $u_5 > 0$ for this component. With these changes to the region of integration and with an alternative expression of the integrand as in (A16),

$$\eta_j = R_{In} + C_n \exp\left[\frac{1}{2} S_{jn}^T J_{jn}^{-1} S_{jn} \right] \pi(t_*) (2\pi)^{p/2} \left| J_{jn}^{-1} \right|^{1/2}$$
$$\times \int_{u_5 \geq 0} \exp\left\{ -\frac{1}{2} \left[u - J_{jn}^{-1} S_{jn} \right]^T J_{jn} \left[u - J_{jn}^{-1} S_{jn} \right] \right\}$$
$$\times \frac{1}{(2\pi)^{p/2} |J_{jn}^{-1}|^{1/2}}\, du. \quad \text{(A17)}$$

The integral (A17) gives the probability that $Y_5 > 0$, where $Y \sim N(J_{jn}^{-1} S_{jn}, J_{jn}^{-1})$. If Y is normal then so is Y_5 and its mean and variance are determined from the mean and variance-covariance matrix of Y. Specifically, the mean is $[J_{jn}^{-1} S_{jn}]_5$ and $[J_{jn}^{-1}]_{55}$. Standard normal calculations give that the probability that $P(X > 0) = \Phi(\mu/\sigma)$ when $X \sim N(\mu, \sigma)$ Replacing J_{jn} with its limiting value I_j, the final approximation is

$$\eta_j = R_{In} + C_n \exp\left[\frac{1}{2} S_{jn}^T I_j^{-1} S_{jn} \right]$$
$$\pi(t_*) (2\pi)^{p/2} \left| I_j^{-1} \right|^{1/2} \Phi\left(\left[I_j^{-1} S_{jn} \right]_5 \Big/ \sqrt{\left[I_j^{-1} \right]_{55}} \right). \quad \text{(A18)}$$

An extension of the large-sample likelihood theory. The first four terms of S_{jn} are the (rescaled) derivatives of the log likelihood for the terminal edge lengths. When evaluated at the star tree value t_* they are the same regardless of the topology. Thus the distinct derivatives can be ordered as

$$W_n = ([S_{1n}]_1, [S_{1n}]_2, [S_{1n}]_3, [S_{1n}]_4, [S_{1n}]_5, [S_{2n}]_5, [S_{3n}]_5)^T.$$

The covariances for the W_n entries are not difficult to obtain and are given in (A1) to (A3). The central limit theorem then gives that W_n converges in distribution to the W described before (A4).

The limiting distribution. Since $J_{jn} \to I_j$ and the S_{jn} are obtained from the W_n in the same manner that the V_j are obtained from W, the numerator (A18) converges in distribution to the numerator in (A4) multiplied by $\pi(t_*)$. It then follows that the limiting distribution is as given in (A4).

APPENDIX B: THE LIMITING DISTRIBUTION OF BOOTSTRAP SUPPORT

There are two distinct probability distributions in bootstrapping. One is the probability distribution associated with the generation of data: characterized by the Jukes-Cantor Markov substitution model on the tree. The other is the bootstrap probability distribution. It is dependent or conditional upon the data and is the probability distribution bootstrap samples are drawn from. The limiting distribution is obtained in two steps by first obtaining an approximation to the bootstrap probability of topology 1 for fixed data and then allowing the data set to vary and considering what this approximation converges to.

First, consider bootstrap probabilities for fixed data. The vector giving the proportions of differences in a bootstrap sample, $p^{(b)}$, is obtained by independently sampling n sites from the original alignment and determining the proportions of differences for the pairs. It can be expressed as

$$p^{(b)} = n^{-1} \sum_{i=1}^{n} \delta^{(i)}, \quad \text{(B1)}$$

where $\delta^{(i)} = [\delta_{12}^{(i)}, \ldots, \delta_{34}^{(i)}]^T$ and has $\delta_{kl}^{(i)} = 1$ or 0 according to whether the pair of taxa k and l have a difference or not at the ith bootstrap sampled site. The reason for expressing $p^{(b)}$ as a mean is that it makes clear that the central limit theorem applies and the bootstrap distribution of $p^{(b)}$ is approximately normal. This is the distribution for fixed data and so the means and variances for it have to be calculated from the bootstrap distribution that yields site pattern according to their frequencies in the fixed alignment. Calculations give the mean of $p_{ij}^{(b)}$ as $\hat{p}_{ij}$, the proportion of differences for the pair i and j. Let $\hat{p}_{ij;kl}$ denote the proportion of sites where taxa i and j have differing nucleotides *and* taxa k and l have differing nucleotides. Let $p = [p_{12}, p_{13}, \ldots, p_{34}]^T$, let w denote a vector with the entries $p_{ij;kl}$. Then the variance-covariance matrix is calculated as $\Sigma_p(\hat{p}, \hat{w})$ where

$$\Sigma_p(\hat{p}, \hat{w})_{ij,kl} = [\hat{p}_{ij;kl} - \hat{p}_{ij}\hat{p}_{kl}]/n. \quad \text{(B2)}$$

The real bootstrap distribution of interest is not for $p^{(b)}$ but rather Z_1 and Z_2 in (4). For the Jukes-Cantor model, distance estimates depend

Downloaded from sysbio.oxfordjournals.org by guest on January 24, 2011

upon the data only through the observed proportions of differences:

$$d_{ij} = d_{ij}(\hat{p}) = -\frac{3}{4}\log\left[1 - \frac{4}{3}\hat{p}_{ij}\right]$$

and so Z_1 and Z_2 are continuous transformations, $z_1(p^{(b)})$ and $z_2(p^{(b)})$ of $p^{(b)}$. From (B2) it follows that the variance of $\sqrt{n}(p^{(b)} - \hat{p})$ is bounded. Thus

$$z(p^{(p)}) \approx z(\hat{p}) + z'(\hat{p})(p^{(b)} - \hat{p}). \tag{B3}$$

Here

$$z'(p) = \begin{bmatrix} d'_{23}(p) + d'_{14}(p) - d'_{12}(p) - d'_{34}(p) \\ d'_{23}(p) + d'_{14}(p) - d'_{12}(p) - d'_{34}(p) \end{bmatrix}$$

where $d'_{ij}(p) = [1 - 4p_{ij}/3]^{-1}$ and more precisely, (B3) reads as

$$\sqrt{n}[z(p^{(p)}) - z(\hat{p})] = z'(\hat{p})[\sqrt{n}(p^{(b)} - \hat{p})] + R_n,$$

where R_n converges to 0. For fixed data, the right-hand side of (B3) is a linear transformation of $p^{(b)}$, which is approximately normal, and hence the right-hand side is approximately normal as well. Its mean is computed as $z(\hat{p})$ and its variance covariance matrix is $\Sigma(\hat{p}, \hat{w})$ where

$$\Sigma(\hat{p}, \hat{w}) = z'(\hat{p})\Sigma_p(\hat{p}, \hat{w})z'(\hat{p})^T. \tag{B4}$$

It follows that, for fixed data and large n, the bootstrap probability that topology 1 is estimated is well approximated by $P(Z_1 > 0, Z_2 > 0)$ where $[Z_1, Z_2]^T$ has a multivariate distribution with mean $z(\hat{p})$ and variance-covariance matrix $\Sigma(\hat{p}, \hat{w})$. Expressing this probability in terms of standardized normal random variables gives that, for a fixed data set and large n, the bootstrap probability that topology 1 is estimated is well approximated by

$$\begin{aligned} P(Z_1 > 0, Z_2 > 0) &= P\left(\frac{Z_1 - z_1(\hat{p})}{\sqrt{\Sigma(\hat{p}, \hat{w})_{11}}} > -\frac{z_1(\hat{p})}{\sqrt{\Sigma(\hat{p}, \hat{w})_{11}}},\right. \\ &\qquad \left.\frac{Z_2 - z_2(\hat{p})}{\sqrt{\Sigma(\hat{p}, \hat{w})_{22}}} > -\frac{z_2(\hat{p})}{\sqrt{\Sigma(\hat{p}, \hat{w})_{22}}}\right) \\ &= G\left(-\frac{z_1(\hat{p})}{\sqrt{\Sigma(\hat{p}, \hat{w})_{11}}}, -\frac{z_2(\hat{p})}{\sqrt{\Sigma(\hat{p}, \hat{w})_{22}}}; \rho(\hat{p}, \hat{w})\right), \end{aligned} \tag{B5}$$

where $\rho(\hat{p}, \hat{w}) = \Sigma(\hat{p}, \hat{w})_{12}/\sqrt{\Sigma(\hat{p}, \hat{w})_{11}\Sigma(\hat{p}, \hat{w})_{22}}$.

The expression (B5) gives an approximation to the bootstrap probability for fixed $\hat{p}$. Now allow the data to vary. The situation is very similar to the one that led to (B5) but instead of sampling from the bootstrap distribution that yields site pattern according to their frequencies, sequences are generated from the pattern probabilities implied by the Jukes-Cantor model and the tree. The central limit theorem still applies to $\hat{p}$ but now its mean is calculated as p, the vector of *probabilities* that taxa i and j have differing nucleotides, and its variance-covariance matrix is $\Sigma_p(p, w)$, where $\hat{p}$ and $\hat{w}$ have been replaced by their corresponding probabilities p and w. The approximation (B3) still applies, but with $p^{(b)}$ replaced by $\hat{p}$ and $\hat{p}$ replaced by p: $z(\hat{p}) \approx z(p) + z'(p)(\hat{p} - p)$. The generating model is the star tree, however, and for this model, $z(p) = 0$; i.e., the linear transformation of distances in (4) are exactly equal to 0. Thus $z(\hat{p}) \approx z'(p)(\hat{p} - p)$ and similarly as in the argument that led to (B5) we obtain that $z(\hat{p})$ is approximately normal but with mean 0 and variance-covariance matrix $\Sigma(p, w)$. It follows that $-z_1(\hat{p})/\sqrt{\Sigma(\hat{p}, \hat{w})_{11}}$ and $-z_2(\hat{p})/\sqrt{\Sigma(\hat{p}, \hat{w})_{22}}$ converge in distribution to V_1 and V_2, which are normal with mean 0 and unit variance as well as correlation

$$\rho_* = \Sigma(p, w)_{12}/\sqrt{\Sigma(p, w)_{11}\Sigma(p, w)_{22}}.$$

Because $\rho(\hat{p}, \hat{w})$ converges to ρ_*, the approximation (B5), to the bootstrap probability that topology 1 is estimated, converges to

$$G(V_1, V_2; \rho_*),$$

where V_1 and V_2 are standard normal and also have correlation ρ_*.

Downloaded from sysbio.oxfordjournals.org by guest on January 24, 2011