Syst. Biol. 55(1):116–121, 2006

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150500481648

Fundamental Differences Between the Methods of Maximum Likelihood and Maximum Posterior Probability in Phylogenetics

BODIL SVENNBLAD,[1] PER ERIXON,[2] BENGT OXELMAN,[2] AND TOM BRITTON,[3]

[1]*Department of Mathematics, Uppsala University, Box 480, SE-751 06 Uppsala, Sweden; E-mail: bodil.svennblad@math.uu.se*
[2]*Department of Systematic Botany, Uppsala University, Norbyvägen 18D, SE-75236 Uppsala, Sweden; E-mail: per.erixon@ebc.uu.se (P.E.); bengt.oxelman@ebc.uu.se (B.O.)*
[3]*Department of Mathematics, Stockholm University, SE-106 91 Stockholm, Sweden; E-mail: tom.britton@math.su.se*

Abstract.—Using a four-taxon example under a simple model of evolution, we show that the methods of maximum likelihood and maximum posterior probability (which is a Bayesian method of inference) may not arrive at the same optimal tree topology. Some patterns that are separately uninformative under the maximum likelihood method are separately informative under the Bayesian method. We also show that this difference has impact on the bootstrap frequencies and the posterior probabilities of topologies, which therefore are not necessarily approximately equal. Efron et al. (Proc. Natl. Acad. Sci. USA 93:13429–13434, 1996) stated that bootstrap frequencies can, under certain circumstances, be interpreted as posterior probabilities. This is true only if one includes a noninformative prior distribution of the possible data patterns, and most often the prior distributions are instead specified in terms of topology and branch lengths. [Bayesian inference; maximum likelihood method; Phylogeny; support.]

Phylogenetic methods based on likelihood aim to find the best topology by maximizing the likelihood function with respect to topology and branch lengths (maximum likelihood method, e.g., Felsenstein, 1981) or by comparing posterior probabilities for the different possible topologies (Bayesian inference, e.g., Rannala and Yang, 1996). Regardless of the method of inference, a measure of confidence, or support, is often desired for the estimated topology. Felsenstein (1985) suggested a nonparametric bootstrap procedure to obtain such a measure, a procedure that is commonly used with the maximum likelihood method. In Bayesian inference, the posterior probabilities are the support values.

Efron et al. (1996) noted that bootstrap frequencies can be interpreted as posterior probabilities under certain circumstances. Those sentences have been frequently cited (e.g., Larget and Simon, 1999; Cummings et al., 2003; Simmons et al., 2004), claiming bootstrap support values and Bayesian posterior probabilities to be approximately equal. Other authors have noticed empirically that Bayesian posterior probabilities for the best supported clades are significantly higher than corresponding nonparametric bootstrap frequencies, (e.g., Suzuki et al., 2002; Wilcox et al., 2002; Alfaro et al., 2003; Douady et al., 2003; Erixon et al., 2003). We will, in Appendix 1, show that the statement of Efron et al. (1996) is true under the conditions therein given. In this paper, however, we will show that those conditions are violated in general phylogenetic inference.

METHODOLOGY FOR ESTIMATING PHYLOGENETIC TREES

We are interested in finding the phylogenetic tree, τ, of s species from aligned DNA sequences of length N. That is, we have an $s \times N$ data matrix, $\boldsymbol{x}$, where each row represents the DNA sequence for one of the species.

The columns in $\boldsymbol{x}$ are assumed to be independent and identically distributed (iid) where each column is one of $k = 4^s$ possible ones. Denote the possible columns by $X_1, X_2, \ldots, X_k$ where, e.g., $X_1' = \{A, A, ..., A\}$, $X_2' = \{C, C, ..., C\}$ etc. The proportions of $X_1, X_2, ..., X_k$, generated from the true phylogeny τ with branch lengths $\boldsymbol{b}^{(\tau)}$ are $\boldsymbol{p} = (p_1, p_2, ..., p_k)$, where $\sum p_i = 1$. Each $\{\tau, \boldsymbol{b}^{(\tau)}\}$ induces a different $\boldsymbol{p}$ vector. The $\boldsymbol{p}$-space is a continuous space that can be divided into different subspaces corresponding to the different topologies.

The data matrix $\boldsymbol{x}$ is a sample from the true phylogeny. When assuming sites to be iid, the data matrix $\boldsymbol{x}$ can instead be represented by a vector $\boldsymbol{n} = (n_1, n_2, \ldots n_k)$, where $\sum n_i = N$ and each n_i is the number of columns in $\boldsymbol{x}$ that equals X_i. The vector $\boldsymbol{p}$ can then be estimated by $\hat{\boldsymbol{p}} = (\hat{p}_1, \hat{p}_2, \ldots \hat{p}_k)$, which depends on $\boldsymbol{n}$. Depending on the method used for phylogenetic inference and choice of model of evolution, each $\boldsymbol{n}$ vector gives an estimate of the topology. The discrete $\boldsymbol{n}$ space is therefore divided in different regions, R_i, giving $\hat{\tau} = \tau_i$ and there might also be regions where the estimate is not unique.

In likelihood based methods the information in data is contained in the likelihood function, $L(\tau_i, \boldsymbol{b}^{(\tau_i)}, \boldsymbol{\theta}|\boldsymbol{n})$. The likelihood is the probability of data, $\boldsymbol{n}$, $f(\boldsymbol{n}|\tau_i, \boldsymbol{b}^{(\tau_i)}, \boldsymbol{\theta})$, given the topology, τ_i, corresponding branch lengths, $\boldsymbol{b}^{(\tau_i)}$ and model of evolution (with parameters $\boldsymbol{\theta}$), but seen as a function of τ_i, $\boldsymbol{b}^{(\tau_i)}$ and $\boldsymbol{\theta}$.

In the maximum likelihood approach, $L(\tau_i, \boldsymbol{b}^{(\tau_i)}, \boldsymbol{\theta}|\boldsymbol{n})$ is maximized for each τ_i with respect to $\boldsymbol{b}^{(\tau_i)}$ and $\boldsymbol{\theta}$. The estimated topology $\hat{\tau}_{ML}$ is the topology τ_i with the largest maximized likelihood.

In the Bayesian approach, prior distributions for τ_i, $\boldsymbol{b}^{(\tau_i)}$, and $\boldsymbol{\theta}$ have to be specified. The posterior probability can then be calculated by Bayes' theorem. In the method of maximum posterior probability the estimate of the topology, $\hat{\tau}_{MPP}$, is the topology τ_i with the largest posterior probability, $\pi(\tau_i|\boldsymbol{n})$ (not the majority-rule consensus tree as given by, e.g., Huelsenbeck and Ronquist, 2001). To summarize, the estimate of the true phylogeny

is, depending on the method of inference,

$$\hat{\tau} = \begin{cases} \hat{\tau}_{ML} & = argmax_i\{\max_{\boldsymbol{b},\boldsymbol{\theta}} L(\tau_i, \boldsymbol{b}^{(\tau_i)}, \boldsymbol{\theta}|\boldsymbol{n})\} \\ \hat{\tau}_{MPP} & = argmax_i\{\pi(\tau_i|\boldsymbol{n})\} \end{cases} \quad (1)$$

Differences between Maximum Likelihood and Maximum Posterior Probability

Efron et al. (1996) stated that, "The bootstrap probability that $\hat{\tau}^* = \hat{\tau}$ is almost the same as the aposteriori probability that $\tau = \hat{\tau}$ starting from an *uninformative* prior density on $\boldsymbol{p}$", where $\hat{\tau}^*$ is the estimate of the topology for a bootstrap replicate. In Appendix 1 this statment is proven. Note, however, that the parameter used is the 4^s-dimensional vector $\boldsymbol{p}$, where s is the number of taxa. As explained earlier, the method of maximum posterior probability does not use the $\boldsymbol{p}$ vector as a parameter. The parameters used with this method are the topology τ_i, corresponding branch lengths $\boldsymbol{b}^{(\tau_i)}$, and the parameters of the model of evolution used, $\boldsymbol{\theta}$. The posterior probability for the topology is

$$\pi(\tau_i|\boldsymbol{n}) = \frac{\int\cdots\int \pi\left(\boldsymbol{n}|\tau_i, \boldsymbol{b}^{(\tau_i)}, \boldsymbol{\theta}\right)\pi\left(\tau_i, \boldsymbol{b}^{(\tau_i)}, \boldsymbol{\theta}\right)d\boldsymbol{b}^{(\tau_i)}d\boldsymbol{\theta}}{\pi(\boldsymbol{n})}, \quad (2)$$

where $\pi(\tau_i, \boldsymbol{b}^{(\tau_i)}, \boldsymbol{\theta})$ is the prior for those parameters. Even though flat priors for topology and branch lengths are used, that is not equivalent to a flat prior for $\boldsymbol{p}$, which simply is not considered in current implementations of Bayesian phylogenetic inference.

Denote the regions in the $\boldsymbol{n}$-space where the method of maximum likelihood gives estimates $\hat{\tau} = \tau_i$ with R_i (see Appendix 1). Denote the corresponding regions where the method of maximum posterior probability gives estimates $\hat{\tau} = \tau_i$ with R_i'. In Efron et al. (1996), a distance-based method was used to obtain the regions R_i. When stated that bootstrap support values could be interpreted as posterior probabilities, those regions remained unchanged. Hence, in their example, $R_i \simeq R_i'$. When comparing the maximum likelihood method and the method of maximum posterior probability, this means that it must hold that $\hat{\tau}_{ML}$ should be equal to $\hat{\tau}_{MPP}$ for almost all $\boldsymbol{n}$ vectors with $\sum n_i = N$.

To examine whether $\hat{\tau}_{ML} = \hat{\tau}_{MPP}$, a situation as simple as possible but where more than one topology is possible has been used; i.e., a four-taxon case under the Jukes-Cantor model (Jukes and Cantor, 1969). Many of the 256 possible columns contribute to the likelihood for τ and $\boldsymbol{b}^{(\tau)}$ in exactly the same way, so we say they have the same "pattern." For example, $AACC$ contributes to the likelihood in the same way as $CCTT$ (under the Jukes-Cantor model), whereas $ACAC$ is another pattern contributing differently. For the Jukes-Cantor model and four taxa, there are 15 different patterns. The possible different patterns of data at a site for four taxa are summarized in Table 1. If a specific pattern on its own gives a unique estimate of the topology we say it is *separately informative*.

TABLE 1. With Jukes-Cantor model of evolution and with four taxa there are 15 different patterns contributing to the likelihood in different ways.

Pattern no.	Pattern	Description
1	$XXYY$	Groups of two nucleotides equal
2	$XYXY$	within the group but not equal
3	$XYYX$	between groups.
4	$XXXY$	One nucleotide differs from the
5	$XXYX$	rest.
6	$XYXX$	
7	$YXXX$	
8	$XXYZ$	One group with two equal
9	$YZXX$	nucleotides, the other two differ
10	$XYXZ$	from the group and from each
11	$XYZX$	other.
12	$YXXZ$	
13	$YXZX$	
14	$XYZU$	All nucleotides different.
15	$XXXX$	All nucleotides equal.

Assume data for taxa $\{a, b, c, d\}$ consists of one single column of pattern 8, $XXYZ$. In Appendix 2 it is shown analytically that for this pattern the maximized likelihoods for $\tau_1 = \{(a, b), (c, d)\}$, $\tau_2 = \{(a, c), (b, d)\}$, and $\tau_3 = \{(a, d), (b, c)\}$ are indistinguishable.

Now consider the method of maximum posterior probability for pattern $XXYZ$. The priors for topology and branch lengths are considered independent; that is,

$$\pi\left(\tau_i, \boldsymbol{b}^{(\tau_i)}\right) = \pi(\tau_i)\prod_{j=1}^{5}\pi\left(b_j^{(\tau_i)}\right).$$

Assume flat priors for topology, $\pi(\tau_i) = 1/3$, as well as for branch lengths, $\pi(b_j^{(\tau_i)}) = 1/M$, that is the prior for a branch length is uniform on the interval $[0, M]$ with expected value $M/2$. The priors will then only be scaling factors and the main contribution to the posterior distribution is the integral of the likelihood. Those integrals can be calculated analytically:

$$\begin{aligned} I_1 &= \int_0^M\cdots\int_0^M L\left(\tau_1, \boldsymbol{b}^{(\tau_1)}|\boldsymbol{n}\right)d\boldsymbol{b}^{(\tau_1)} \\ &= [M^5 + 2M^3(1-e^{-M})^2 - 4M^2(1-e^{-M})^3 \\ &\quad - 3M(1-e^{-M})^4 + 4(1-e^{-M})^5]/4^4 \quad (3) \\ I_2 &= \int_0^M\cdots\int_0^M L(\tau_2, \boldsymbol{b}^{(\tau_2)}|\boldsymbol{n})d\boldsymbol{b}^{(\tau_2)} \\ &= [M^5 - 2M^3(1-e^{-M})^2 + M(1-e^{-M})^4]/4^4. \quad (4) \end{aligned}$$

For $M > 0$ the difference between Equations (3) and (4) is

$$\begin{aligned} [4M^3(1-e^{-M})^2 &- 4M^2(1-e^{-M})^3 - 4M(1-e^{-M})^4 \\ &+ 4(1-e^{-M})^5]/4^4 = [(1-e^{-M})^2 \\ &\times (M-(1-e^{-M}))^2(M+(1-e^{-M}))]/4^3. \end{aligned} \quad (5)$$

All the factors of (5) are strictly positive, and hence the difference is strictly positive and $I_1 > I_2$.

The posterior probability of τ_i, $\pi(\tau_i|\boldsymbol{n})$, is obtained from (2). To calculate the denominator $\pi(\boldsymbol{n})$ in (2) one has to integrate over the branch lengths $\boldsymbol{b}$ and the parameter(s) $\boldsymbol{\theta}$ for all possible topologies and sum those, a complicated numerical task. However, if it is possible to calculate the numerator in (2) for all possible topologies, the posterior probability for τ_i can be obtained by dividing the numerator in (2) for τ_i with the sum of the numerators for all topologies. Then $\pi(\tau_i|\boldsymbol{n})$ can be calculated analytically from (3) and (4) as

$$\pi(\tau_i|\boldsymbol{n}) = \frac{I_i}{I_1 + I_2 + I_3},$$

where $I_3 = I_2$. Because $I_1 > I_2$, the posterior probability for τ_1 will be larger than for τ_2 and τ_3 and hence, using flat priors, $\hat{\tau}_{MPP} = \tau_1$ for pattern $XXYZ$. The posterior probability will tend to 1/3 as M, the length of the interval of the prior for branch lengths, increases. For any finite M though, τ_1 will be the unique estimate. Using an exponential distribution with mean $\frac{1}{\lambda}$ as prior for branch lengths but still using uniform distribution as prior for topology, the posterior probabilities can be written as $\pi(\tau_i|\boldsymbol{n}) = A_i/(A_1 + A_2 + A_3)$ where

$$A_1 = \frac{1}{\lambda^5} + \frac{2}{\lambda^3 y^2} - \frac{4}{\lambda^2 y^3} - \frac{3}{\lambda y^4} + \frac{4}{y^5},$$

$$A_2 = A_3 = \frac{1}{\lambda^5} - \frac{2}{\lambda^3 y^2} + \frac{1}{\lambda y^4},$$

where $y = (\lambda + 1)$. Pattern $XXYZ$ is separately informative for the maximum posterior probability method with those priors too, because $A_1 > A_2$ for all $\lambda > 0$.

Analogous to the analysis of pattern $XXYZ$ it can be shown that the two methods of inference differ for many of the 15 patterns. For the maximum likelihood method only patterns 1, 2, and 3 are separately informative. These are separately informative for the maximum posterior probability method also, but so are patterns 8 to 13 (see Table 2).

Even though a site is separately uninformative, it contributes to the likelihood function and therefore cannot be ignored. The fundamental differences between the two methods of inference, summarized in Table 2, have impact on the regions R_i and R_i', as the next example will show.

TABLE 2. Separately informative patterns favoring topologies $\tau_1 = \{(a, b), (c, d)\}$, $\tau_2 = \{(a, c), (b, d)\}$ and $\tau_3 = \{(a, d), (b, c)\}$ for the methods of maximum likelihood and maximum posterior probability, respectively. The enumeration of the patterns follows Table 1.

	Estimate of topology		
Method	τ_1	τ_2	τ_3
ML	1	2	3
MPP	1, 8, 9	2, 10, 13	3, 11, 12

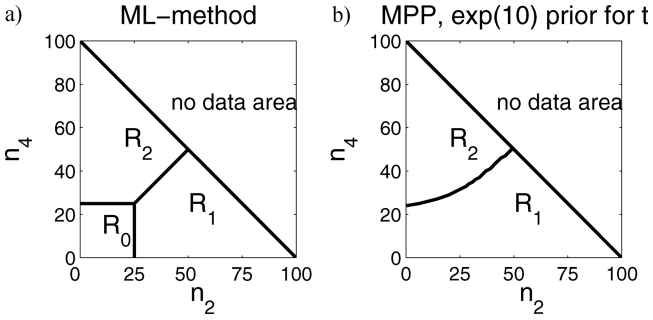

FIGURE 1. Regions R_i, where $\hat{\tau} = \tau_i$ for the maximum likelihood method (a) and Bayesian inference (b) with $n_1 + n_2 + n_8 = 100$. The number of columns in data matrix that are equal to pattern 1 is on the x-axis, number of pattern 2 on the y-axis, and for each dot in the triangle bounded by the axes and $n_1 + n_2 = 100$ we have $n_8 = 100 - n_1 - n_2$.

Example

Suppose $N = 100$ and data only consists of patterns 1, 2, and 8 (e.g., $n_1 = 45$, $n_2 = 35$, $n_8 = 20$, $n_i = 0$ for $i = 3, \ldots, 7, 9, \ldots, 15$). For this data set $\hat{\tau}_{ML} = \hat{\tau}_{MPP} = \tau_1$. From Table 2 we see that patterns 1 and 2 are separately informative, favoring topologies τ_1 and τ_2, respectively, for both methods of inference. Pattern 8 is, as we have shown analytically, separately informative for the maximum posterior probability method, favoring topology τ_1, but not for the maximum likelihood method. Columns of this pattern still contribute to the likelihood and are therefore taken into account even in the maximum likelihood method. A consequence of this is the region R_0 in Figure 1, where no unique estimate of topology can be found. The regions in Figure 1 have been determined by using PAUP*v4.0b10 (Swofford, 2002) for all possible $\boldsymbol{n}$-vectors with $n_i = 0$ for all i except $n_1 \geq 0$, $n_2 \geq 0$ and $n_8 \geq 0$ with $\sum n_i = 100$ for the maximum likelihood method. For the Bayesian inference, R_i' was determined by using MrBayes3.04 (Huelsenbeck and Ronquist, 2001) with an exponential prior with intensity 10 (mean 1/10) for branch lengths and uniform prior for topology. Figure 1 shows that $R_i \neq R_i'$ and the bootstrap support value for $\hat{\tau}_{ML}$ is smaller (0.8715 when approximated by bootstrap frequencies from PAUP*) than the posterior probability for $\hat{\tau}_{MPP}$ (which is approximated by MrBayes to 1.0).

With an original data set far away from the borders between different regions $\{R_i\}$ and $\{R_i'\}$, respectively (e.g., $n_1 = 70$, $n_2 = 10$, and $n_8 = 20$ in Figure 1), the bootstrap support value and the posterior probabilities will both be high, tending to 1, even though $R_i \neq R_i'$. This is due to the fact that the probability of a bootstrap replicate ending up in a region other than R_i is then very small.

CONCLUDING REMARKS

We have shown that the often cited statement of Efron et al. (1996) about approximate equivalence between bootstrap support values and posterior probabilities is true when the only parameter of interest is $\boldsymbol{p}$ and a noninformative prior is used. This parameter is not considered

in the current implementations of Bayesian inference of phylogenies where the parameters topology, τ, branch lengths $\boldsymbol{b}^{(\tau)}$, and model parameters $\boldsymbol{\theta}$ are considered, nor in the maximum likelihood method.

For the four-taxon case with Jukes-Cantor model of evolution, the 256 different possible data columns can be reduced to 15 different patterns contributing to the likelihood function in different ways. We have shown analytically that some of those patterns are separately informative for the method of maximum posterior probability but not for the method of maximum likelihood.

The differences between the methods remain when more taxa are considered. In the five-taxon case, $\{a, b, c, d, e\}$, there are $4^5 = 1024$ different possible data columns that can, under the Jukes-Cantor model, be reduced to 51 different data patterns. For example, for pattern $XXYYZ$, seven topologies have the same maximized likelihood, all of them obtained on a border where at least one internal branch is of length 0; i.e., the tree collapse (see Appendix 2 for an analogous case with four taxa). Some of the topologies collapse to the same, not fully resolved, tree. For this specific pattern, four different collapsed trees are obtained. The groups $\{a, b\}$ and $\{c, d\}$ exist in three of them, respectively. The tree where the group is absent does not contradict the group. Hence, with the method of maximum likelihood, this pattern is not informative for topology but is informative for some groups. The method of maximum posterior probability, on the other hand, gives a unique estimate of the topology, $(a, b, (d, (e, f)))$, and hence also of groups. Numerically, by using PAUP*v40b10 and MrBayes3.04 we have found that none of the 51 patterns gives a unique estimate of the topology with maximum likelihood but 15 patterns give fully resolved trees with maximum posterior probability. This is also the case for the 187 different data patterns of the six-taxon case. The posterior probabilities of groups are, for six taxa and more, also strongly influenced by unequal priors for groups that follow a flat prior for topologies (Pickett and Randle, 2005), which may further explain the observed disparities between bootstrap values and posterior probabilities.

The fundamental differences between the methods of maximum likelihood and maximum posterior probabilities that we have demonstrated influence the regions of the $\boldsymbol{n}$ space where the estimates equal a specific topology. Those regions influence the bootstrap support values and posterior probabilities differently, and they should therefore not be expected to be approximately equal. How much of the observed differences between maximum likelihood bootstrap values and posterior probabilities that can be attributed to this factor is not clear but need to be further studied.

ACKNOWLEDGEMENTS

We are grateful to Roderic D. M. Page, Jeff Thorne, and an anonymous reviewer for constructive criticism of an earlier version of this manuscript. This work was supported by the Linnaeus Centre for Bioinformatics, Uppsala University, and the Swedish Research Council.

REFERENCES

Alfaro, M. E., S. Zoller, and F. Lutzoni. 2003. Mol. Biol. Evol. 20:255–266.
Berger, J. O. 1980. Statistical decision theory, foundations, concepts and methods. Springer-Verlag, New York.
Bernardo, J. M., and A. F. M. Smith. 1994. Bayesian theory. John Wiley & Sons, New York.
Box, G. E., and G. C. Tiao. 1973. Bayesian inference in statistical analysis. Addison-Wesley Publishing Company.
Cummings, M. P., S. A. Handley, D. S. Myers, D. L. Reed, A. Rokas, and K. Winka. 2003. Syst. Biol. 52:477–487.
Douady, C. J., F. Delsuc, Y. Boucher, W. F. Doolittle, and E. J. P. Douzery. 2003. Mol. Biol. Evol. 20:248–254.
Efron, B. 1982. SIAM CBMS-NSF Monograph 38.
Efron, B., E. Halloran, and S. Holmes. 1996. Proc. Natl. Acad. Sci. USA 93:13429–13434.
Erixon, P., B. Svennblad, T. Britton, and B. Oxelman. 2003. Syst. Biol. 52:665–673.
Felsenstein, J. 1981. J. Mol. Evol. 17:368–376.
Felsenstein, J. 1985. Evolution 39:783–791.
Huelsenbeck, J. P., and F. Ronquist 2001. Bioinformatics 17:754–755.
Jukes, T. H., and C. R. Cantor 1969. Pages 21–132 *in Mammalian protein metabolism*. (H. N. Munro., ed.). Academic Press, New York.
Larget, B., and Simon, D. L. 1999. Mol. Biol. Evol. 16:750–759.
Pickett, K. M., and Randle, C. P. 2005. Mol. Phyl. Evol. 34:203–211.
Rannala, B., and Yang, Z. 1996. J. Mol. Evol. 43:304–311.
Simmons, M. P., Pickett, K. M., and Miya, M. 2004. Mol. Biol. Evol. 21:188–199.
Steel, M. 1994. Syst. Biol. 43:560–564.
Suzuki, Y., Glazko, G. V., and Nei, M. 2002. Proc. Natl. Acad. Sci. USA 99:16138–16143.
Swofford, D. L. 2002. PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.
Wilcox, T. P., Zwickl D. J., Heath T. A., and Hillis, D. M., 2002. Mol. Phylogenet. Evol. 25:361–371.

First submitted 1 February 2005; reviews returned 6 April 2005; final acceptance 11 August 2005
Associate Editor: Rod Page

APPENDIX 1

In Efron et al. (1996), an example where the data $\boldsymbol{x}$ are a two-dimensional normal vector with expectation vector $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and with identity covariance matrix is used. It is noted that if $\boldsymbol{\mu}$ a priori could be anywhere in the plane with equal probability, the a posteriori distribution of $\boldsymbol{\mu}$, given $\hat{\boldsymbol{\mu}} = \boldsymbol{x}$, is also normally distributed with expectation vector $\hat{\boldsymbol{\mu}}$ and identity covariance matrix, exactly the same as the bootstrap distribution of $\hat{\boldsymbol{\mu}}^\star$, using parametric bootstrap. Referring to this example it is further stated that "The bootstrap probability that $\hat{\tau}^* = \hat{\tau}$ is almost the same as the aposteriori probability that $\tau = \hat{\tau}$ starting from an *uninformative* prior density on $\boldsymbol{p}$", where $\hat{\tau}^*$ is the estimate of the topology for a bootstrap replicate and $\boldsymbol{p}$ is the vector of proportions of the possible data patterns $X_1, X_2, \ldots, X_{4^s}$, where s is the number of taxa and, e.g., $X'_1 = \{A, A, \ldots, A\}$, $X'_2 = \{C, C, \ldots, C\}$, etc. We will now show that the statement is true if a noninformative prior is used for $\boldsymbol{p}$ under the assumption that the method of inference gives the correct tree.

The data, $\boldsymbol{x}$, is an $s \times N$ data matrix of aligned DNA sequences. This data matrix is equivalent with drawing N columns from the true phylogeny where each possible column $X_1, X_2, \ldots, X_k$ has probability $p_1, p_2, \ldots, p_k$ of being drawn, where $k = 4^s$. The data matrix $\boldsymbol{x}$ can then be represented by a vector $\boldsymbol{n} = (n_1, \ldots n_k)$ where $\sum n_i = N$ and each n_i is the number of columns in $\boldsymbol{x}$ that equals X_i. The probability of data or, equivalently, the probability of $\boldsymbol{n}$, the likelihood, then follows a multinomial distribution.

In the nonparametric bootstrap method used in phylogenetic inference (Felsenstein, 1985), a bootstrap sample of the same size as the original data is obtained by drawing columns from the matrix $\boldsymbol{x}$ with replacement. The bootstrap replicate $\boldsymbol{n}^* = (n_1^*, n_2^*, \ldots, n_k^*)$ follows, regardless of the method of inference used, a multinomial distribution

with probability function f defined by

$$f(n_1^*, n_2^*, \ldots n_k^*) = \binom{N}{n_1^*\, n_2^* \cdots n_k^*} \hat{p}_1^{n_1^*} \hat{p}_2^{n_2^*} \cdots \hat{p}_k^{n_k^*}, \quad (6)$$

where $\hat{p}_i = n_i/N$, with means $E(n_i^*) = N\hat{p}_i = n_i$, variances $V(n_i^*) = N\hat{p}_i(1-\hat{p}_i) = (n_i(N-n_i))/N$ and covariances $C(n_i^*, n_j^*) = -N\hat{p}_i\hat{p}_j = -(n_i n_j)/N$.

Now consider the Bayesian framework. The prior for $\boldsymbol{p}$ should, according to Efron et al. (1996), be noninformative. This term is not uniquely defined in the literature. For example, Berger (1980) says "*a non-informative prior, by which is meant a prior which contains no information about the parameter (or more crudely which favors no possible values of the parameter over others)*." Bernardo and Smith (1994) describe the idea of a non-informative prior as representing *ignorance*. Jeffreys' rule for multiparameter problems, described, e.g., in Box and Tiao (1973), states that the prior distribution for a set of parameters is approximately noninformative if "*the prior distribution for a set of parameters is taken to be proportional to the square root of the determinant of the information matrix*," which relies on invariance under parameter transformation.

The three different "definitions" of non-informative priors for $\boldsymbol{p}$ given above can all be represented by a symmetric Dirichlet distribution

$$D(\boldsymbol{p}|\boldsymbol{\theta}) = \frac{\Gamma(\sum_j \theta_j)}{\prod_j \Gamma(\theta_j)} p_1^{\theta_1 - 1} p_2^{\theta_2 - 1} \cdots p_k^{\theta_k - 1}, \quad (7)$$

where $\boldsymbol{\theta} = (1, \ldots 1)$ represents a flat (uniform) prior, $\boldsymbol{\theta} = (\frac{1}{2}, \ldots \frac{1}{2})$ represents Jeffery's prior, and letting $\boldsymbol{\theta} \to (0, \ldots, 0)$, as in Efron (1982), represents "prior ignorance." Because the Dirichlet distribution is a conjugate prior to the multinomial distribution, the posterior distribution of $\boldsymbol{p}$ is $D(\boldsymbol{p}|\boldsymbol{\theta} + \boldsymbol{n})$ with

$$\text{means } E(p_i) = \frac{n_i + \theta}{N + \theta k}, \quad (8)$$

$$\text{variances } V(p_i) = \frac{(n_i + \theta)(N - n_i + \theta(k-1))}{(N + \theta k)^2(1 + N + \theta k)}, \quad (9)$$

$$\text{covariances } C(p_i, p_j) = -\frac{(n_i + \theta)(n_j + \theta)}{(N + \theta k)^2(1 + N + \theta k)}. \quad (10)$$

From the means, variances, and covariances following (6) it is seen that for the bootstrap replicate, where $\hat{p}_i^* = n_i^*/N$,

$$E(\hat{p}_i^*) = E(n_i^*/N) = n_i/N, \quad (11)$$

$$V(\hat{p}_i^*) = V(n_i^*/N) = n_i(N - n_i)/N^3, \quad (12)$$

$$C(\hat{p}_i^*, \hat{p}_j^*) = C(n_i^*/N, n_j^*/N) = -n_i n_j/N^3. \quad (13)$$

For very large N and k small, that is for a lot of data on only a few taxa, (8) $\simeq$ (11), (9) $\simeq$ (12), and (10) $\simeq$ (13) for each of the priors suggested. As we assume N to be large, the multivariate central limit theorem makes the bootstrap distribution and the posterior distribution for $\boldsymbol{p}$ approximately equal.

Recall that for the data, $\boldsymbol{n}$, $\sum n_i = N$. Each possible $\boldsymbol{n}$-vector satisfying $\sum n_i = N$, gives an estimate of the topology. The discrete $\boldsymbol{n}$-space is therefore divided into different regions R_i, with $\hat{\tau} = \tau_i$. There might also be regions where the estimate is not unique. Denote these regions with R_0. The bootstrap support value α is the probability that the bootstrap replicate $\boldsymbol{n}^*$ ends up in the same region as $\boldsymbol{n}$. That is, the probability that the bootstrap replicate gives the same estimate of the topology as the original data set. If the division of the $\boldsymbol{n}$-space in regions R_i (giving estimate $\hat{\tau} = \tau_i$) would be known explicitly, then the bootstrap support value α could be calculated exactly by (6) as

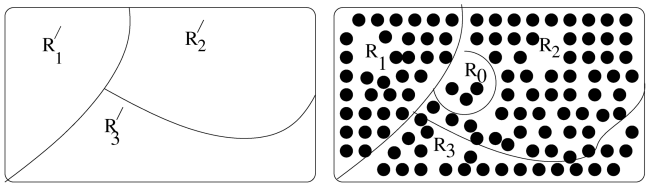

FIGURE 2. The continous $\boldsymbol{p}$-space is k-dimensional where $k = 4^s$ and s is the number of taxa. Here this is illustrated to the left in two dimensions with a hypothetical division of the $\boldsymbol{p}$-space giving subspaces R_i', where $\tau = \tau_i$. To the right the corresponding $\hat{\boldsymbol{p}}$-space is illustrated where each dot represents a possible $\hat{\boldsymbol{p}}$-vector. The regions R_i and R_i' may differ as R_i' is the true division of the $\boldsymbol{p}$ space, whereas R_i is the region where the estimate is equal to τ_i. In the region R_0 there is no unique estimate.

$$\alpha = \sum_{\boldsymbol{n}^* \in R_i} f(n_1^*, n_2^*, \ldots n_k^*). \quad (14)$$

Usually, the $\boldsymbol{n}$-space is very complex and the division of it is therefore practically impossible to determine. The bootstrap support value is then estimated by the fraction of bootstrap replicates giving the same estimate of topology as the original data when drawing bootstrap replicates many times (Fig. 2).

We have seen that theoretically the posterior distribution and the bootstrap distribution are approximately equal. Let the continuous region where $\tau = \tau_i$ be denoted R_i'. To get the posterior distribution of τ_i, the posterior distribution of $\boldsymbol{p}$ is integrated over the region R_i'. If $R_i \simeq R_i'$ and N is large, an integral can be approximated by a sum; i.e.,

$$\int_{\boldsymbol{p} \in R_i'} \pi(\boldsymbol{p}|\boldsymbol{n}) d\,\boldsymbol{p} \simeq \sum_{\boldsymbol{n}^* \in R_i} f(\boldsymbol{n}^*|\boldsymbol{n}) = \alpha. \quad (15)$$

Assuming that the method of inference gives the correct tree, (15) makes the bootstrap support value and the Bayesian posterior probability for a given topology theoretically approximately equal.

APPENDIX 2

In this section we show that there is no unique estimate for topology for the maximum likelihood method when data consists of one column of pattern 8, $XXYZ$, and the Jukes-Cantor model of evolution is used. The three possible topologies are then τ_1, τ_2, and τ_3 (see Fig. 3) with branch lengths $\boldsymbol{t}$, $\boldsymbol{b}$, and $\boldsymbol{v}$, respectively.

Assume the Jukes-Cantor model of evolution, with $\alpha = 0.25$. The likelihood functions for τ_2 and τ_3 will be identical (for τ_3 replace b_i with v_i in the likelihood function for τ_2). Denote the likelihood function for τ_1 with L_1 and for τ_2 (and τ_3) with L_2.

$$\begin{aligned} L_1(\boldsymbol{t}) = \frac{1}{4^4}[1 &+ 3e^{-(t_1+t_2)} - e^{-(t_1+t_3+t_5)} - e^{-(t_1+t_4+t_5)} \\ &- 3e^{-(t_1+t_2+t_3+t_4)} - 2e^{-(t_1+t_2+t_3+t_5)} + 2e^{-(t_1+t_3+t_4+t_5)} \\ &- 2e^{-(t_1+t_2+t_4+t_5)} - e^{-(t_3+t_4)} - e^{-(t_2+t_3+t_5)} \\ &- e^{-(t_2+t_4+t_5)} + 2e^{-(t_2+t_3+t_4+t_5)} + 4e^{-(t_1+t_2+t_3+t_4+t_5)}]. \end{aligned}$$

We want, in the semiclosed five-dimensional continuous space bounded by $0 \le t_1, 0 \le t_2, \ldots, 0 \le t_5$, to find a point $\boldsymbol{t}$ such that $L_1(\boldsymbol{t})$ is maximized. If we would have a closed five-dimensional space, such that $0 \le t_i \le M$, $i = 1, \ldots, 5$ for some positive M, there would be four possibilities for a maximum point $\boldsymbol{t}$ (shown in Fig. 4 for a three-dimensional case):

1. $\boldsymbol{t}$ is in the interior of the closed space,
2. $\boldsymbol{t}$ is in a boundary region of the space,

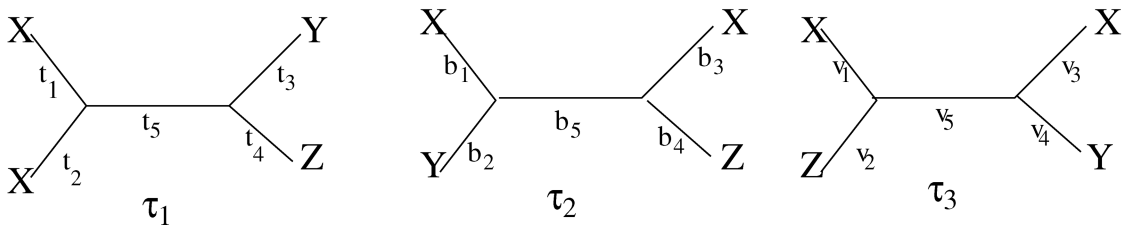

FIGURE 3. The possible topologies τ_1, τ_2, and τ_3 with branches $\boldsymbol{t}$, $\boldsymbol{b}$, and $\boldsymbol{v}$, respectively, when data consist of one site: {XXYZ}.

3. $\boldsymbol{t}$ is on a border between the different boundary regions,
4. $\boldsymbol{t}$ is in a corner of the closed space.

To find the maximum value of L_1 in a closed space, find, to begin with, possible extreme points in the interior of the closed space (case 1). There is no need to further investigate the kind of possible extreme points but evaluate L_1 in those points. Then investigate the boundary regions (case 2) and evaluate L_1 in the possible extreme points of these regions. Continue in the same way with case 3 and finally evaluate L_1 in the corners of the closed space (case 4). The maximum value of L_1 in the closed continuous space is the maximum value of the evaluated points.

If, in the five-dimensional closed space, $\boldsymbol{t}$ is an interior point, the tree will be fully resolved. If $\boldsymbol{t}$ is a boundary point with some $t_i = 0$, the tree will collapse.

If there exists an interior $\boldsymbol{t}$ that is extreme, i.e., maximizes or minimizes L_1 or is a so-called saddlepoint, the gradient $\nabla L_1 = (L'_{1_{t_1}}, \ldots, L'_{1_{t_5}}) = (0, \ldots, 0)$. From $L'_{1_{t_1}} - L'_{1_{t_2}} = 0$ and $L'_{1_{t_3}} - L'_{1_{t_4}} = 0$ it is found that $t_1 = t_2$ and $t_3 = t_4$. From $L'_{1_{t_1}} - L'_{1_{t_5}} = 0$, and by using $t_1 = t_2$ and $t_3 = t_4$, it also follows that $e^{-t_5} = (-3e^{-t_1}(1 + e^{-t_3}))/(2e^{-t_3}) < 0$, which is impossible because $0 < e^{-t_5} < 1$ in the interior of the compact space. Therefore, τ_1 has no interior extreme point (case 1).

Continue with cases 2 and 3. In the closed five-dimensional space the boundary regions to be investigated are obtained by letting combinations of variables be fixed, 0 or M. At least one variable must not be fixed as otherwise a corner is investigated (case 4). When a variable is fixed to 0, or M, the likelihood function looks slightly different and an extreme point is searched for in the usual way. In the five-dimensional space there are 210 boundary regions and borders to be investigated. By investigating all of them it is found that only three of the regions have a possible interior extreme point, no matter how large M is. Evaluating L_1 at those possible extreme points gives $L_1 \leq 4^{-4}$ for all three of them.

The last possibility for a maximum point is at the corners of the closed space (case 4); i.e., where t_i is equal to 0 or M. For 14 of the 32 possible corners, $L_1 = 0$, for the others, $L_1 < 4^{-3}$. As long as we consider finite M, the maximum point for τ_1 is in the corner (0, 0, M, M, M) for which $L_1 = [1 - 3e^{-2M} + 2e^{-3M}]/4^3$. As M increases, the value of L_1 in that corner tends to 4^{-3}, but there are other corners with the same limit value. Hence the maximum value of the likelihood will not be a unique point in the space of branch lengths. The fact that this can happen was pointed out by Steel (1994).

Now, consider topologies τ_2 (and τ_3). The likelihood L_2 is

$$\begin{aligned} L_2(\boldsymbol{b}) = \frac{1}{4^4}[&1 - e^{-(b_3+b_4)} + 3e^{-(b_1+b_3+b_5)} - e^{-(b_1+b_4+b_5)} \\ &- 2e^{-(b_1+b_3+b_4+b_5)} - e^{-(b_2+b_3+b_5)} - e^{-(b_2+b_4+b_5)} \\ &- e^{-(b_1+b_2)} - 2e^{-(b_1+b_2+b_3+b_5)} + e^{-(b_1+b_2+b_3+b_4)} \\ &+ 2e^{-(b_1+b_2+b_4+b_5)} + 2e^{-(b_2+b_3+b_4+b_5)}]. \end{aligned}$$

Analyzing L_2 in the same way as L_1 gives a unique maximum at the corner b = (0, M, 0, M, 0), where $L_2 = (1 - 2e^{-M} + e^{-2M})/4^3$, giving $\max L_2(b) = 4^{-3}$ as $M \to \infty$, which is the same as the maximum value of L_1. Hence, pattern $XXYZ$ is not separately informative for the maximum likelihood method.

In Figure 5 the maximized likelihoods for τ_1 and τ_2 are shown as a function of M. As can be seen, the maximized likelihoods are indistinguishable for large but finite M.

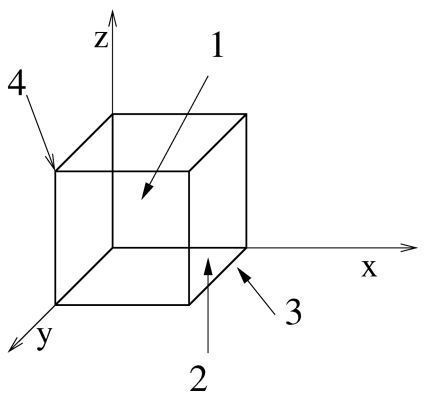

FIGURE 4. The possible locations of a maximum point {x, y, z} in a closed three-dimensional space: 1 an interior point, 2 a point on a boundary region, 3 a point on the border between different boundary regions, and 4 a corner, where each variable is either 0 or M.

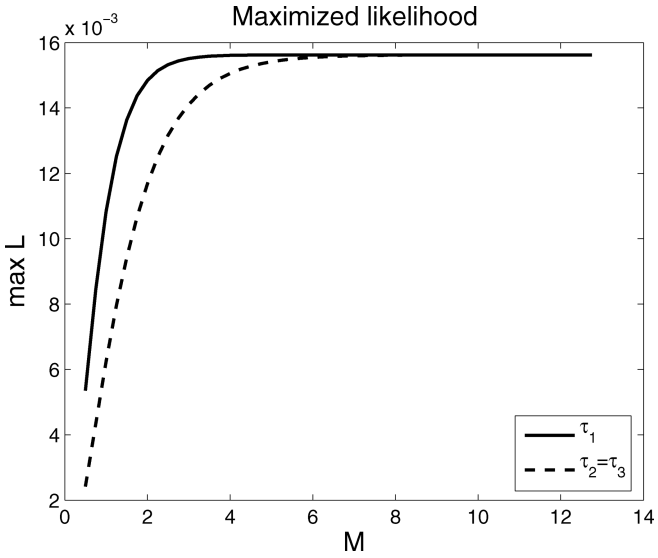

FIGURE 5. The maximized likelihood for topology τ_1 and τ_2 (which has the same likelihood function as τ_3) for pattern $XXYZ$. The likelihood for τ_1 is maximized for branch lengths (0, 0, M, M, M) whereas the likelihood for τ_2 is maximized for branch lengths (0, M, 0, M, 0).