# Bayesian support is larger than bootstrap support in phylogenetic inference: a mathematical argument

Tom Britton [*] Bodil Svennblad [†] Per Erixon [‡]
Bengt Oxelman [§]

June 20, 2007

**Abstract**

In phylogenetic inference the support of an estimated phylogenetic tree topology and its interior branches are usually measured either with non-parametric bootstrap support values (BS) or with Bayesian posterior probabilities (BPP). Extensive empirical evidence indicate that BPP values are systematically larger than BS when measured on the same data set, but there are no theoretical results supporting such a systematic difference. In the present note we give a heuristic mathematical argument supporting the empirically observed phenomenon. A simulation study is performed to investigate the heuristic arguments and The heuristic arguments are supported in a simulation study evaluating different steps in the argument.

*Keywords:* Bayesian posterior probability, bootstrap support, marginal likelihood, phylogenetic inference, profile likelihood.

## 1 Introduction

The present paper is concerned with phylogenetic inference based on aligned DNA-sequences from a set of species of interest. The aim is to draw conclusions about the phylogenetic tree specifying the evolution of the species. Beside

[*]Department of Mathematics, Stockholm University, SE-106 91 Stockholm, Sweden. E-mail: tomb@math.su.se
[†]Department of Mathematics, Uppsala University.
[‡]Department of Systematic Botany, Uppsala University.
[§]Department of Systematic Botany, Uppsala University.

coming up with an estimated phylogenetic tree it is also of interest to express the support, or belief, in the estimated tree. Here we study and compare two common ways to perform such an analysis.

The first method is to estimate the phylogenetic tree using maximum likelihood methods for an evolutionary model, and to estimate the support for the estimated tree by the corresponding non-parametric bootstrap support. The second method estimates the phylogenetic tree with the tree-topology having largest Bayesian posterior probability, where an evolutionary model and a prior parameter distribution is used in a Bayesian setting.

The analysis is performed assuming that we have the true evolutionary model and that the sequences are perfectly aligned without any gaps – it is beyond the scope of the present paper to study what effect deviations from the evolutionary model and misalignments have on the two support values. Further, we restrict our attention to support values for the whole phylogenetic tree and not for interior branches, although our findings extend also to this case. Under these assumptions we compare bootstrap support values (BS) with Bayesian posterior probabilities (BPP) for the same sequence data and the same evolutionary model.

Extensive empirical evidence suggest that, generally BPP>BS (e.g. Wilcox *et al.*, 2002, Alfaro *et al.*, 2003, Douady *et al.*, 2003, and Erixon *et al.*, 2003). However, there are to our knowledge no theoretical arguments supporting this systematic difference. In fact, a paper by Efron *et al.* (1996) has been used as an argument to why the two support measures should be approximately equal (e.g. Larget and Simon, 1999, Cummings *et al.*, 2003, Simmons *et al.*, 2004). Svennblad *et al.* (2006) explain why this argument is misleading and why the two support measures need not be equal, but they present no argument for the empirically observed systematic difference. In the present paper we give a mathematical argument, admittedly not fully rigorous, to why BPP>BS. The argument uses the fact that BS is strongly related to the relative *profile* likelihood of the most likely phylogeny, and that BPP approximately equals the relative *marginal likelihood* of the most likely phylogeny. When we have long

sequences the two relative likelihoods mentioned are approximated by normal distributions and the conclusion is drawn by showing that the maximal relative marginal likelihood is larger than the corresponding profile likelihood for the normal distribution.

In Section 2 we present the evolutionary model and type of data considered. Sections 3 and 4 present how bootstrap supports and Bayesian posterior probabilities, respectively, are obtained. In Section 5 we give the mathematical argument to why BPP>BSA and in Section 6 a simulation study is performed to investigate and illustrate the support for the arguments.

# 2 Model and data

Consider an evolutionary model in which sites are assumed to evolve independently and identically (we will use the simple Jukes & Cantor (J-C) model, Jukes and Cantor, 1969, in our examples). Suppose we have $k$ aligned DNA sequences, all of length $n$. From the data we want to estimate the underlying unrooted tree topology which we denote by $\tau$. For $k \geq 4$ species, or terminals, there are $(2k-5)!!$ different topologies to choose between. The smallest non-trivial case is $k = 4$ when there are $3!! = 3 \cdot 1 = 3$ different possible tree topologies $\tau_1$, $\tau_2$ and $\tau_3$.

At any given site, there are $4^k$ possible nucleotide "patterns", which we label somehow from 1 to $4^k$ (if there are $k = 4$ terminals we could give the pattern $AAAA$ label 1, $AAAC$ label 2, ..., $TTTT$ label 256). Because we assume that different sites are independent and identically distributed, an alternative way to summarise the sequence data is by $\mathbf{n} = (n_1, \ldots, n_{4^k})$, where $n_i$ denotes the number of sites in the data that have pattern $i$. We call data on this form *pattern* data. In statistical terms, the pattern data of the form $\mathbf{n} = (n_1, \ldots, n_{4^k})$ is a *sufficient statistic* for the parameters of the model. From now on we therefore consider this as our data.

If the tree topology $\tau$ and its branch lengths $\mathbf{b}^{(\tau)}$ are known it is, at least in

principle, possible to compute the probability

$$p_i = p_i(\tau, \mathbf{b}^{(\tau)}) = P(\text{pattern } i \text{ occurs at a given site}) \qquad i = 1, \ldots, 4^k.$$

Because sites are assumed independent and identically distributed it follows that the probability to observe a specific data pattern $\mathbf{n} = (n_1, \ldots, n_{4^k})$ follows the multinomial distribution with parameters $n$ (=the sequence length) and $\mathbf{p}(\tau, \mathbf{b}^{(\tau)}) = (p_1(\tau, \mathbf{b}^{(\tau)}), \ldots, p_{4^k}(\tau, \mathbf{b}^{(\tau)}))$:

$$P(n_1, \ldots, n_{4^k}|\tau, \mathbf{b}^{(\tau)}) = P(n_1, \ldots, n_{4^k}|\mathbf{p}(\tau, \mathbf{b}^{(\tau)})) = \binom{n}{n_1 \ \ldots \ n_{4^k}} p_1^{n_1} \cdot \ldots \cdot p_{4^k}^{n_{4^k}}.$$

We stress that this formula is true for both likelihood based inference and Bayesian inference presented below. The Bayesian analysis differs only in that $\tau, \mathbf{b}^{(\tau)})$ are treated as random variables, but conditional on their values the outcome probabilities are identical for the two methods of analysis. The central limit theorem (e.g. Ross, 2006) implies that if $n$ (the sequence length) is large, the multinomial distribution is well approximated by the normal distribution having mean vector $n\mathbf{p}(\tau, \mathbf{b}^{(\tau)})$ and variance elements $np_i(\tau, \mathbf{b}^{(\tau)})(1-p_i(\tau, \mathbf{b}^{(\tau)}))$ and off-diagonal covariance elements $-np_i(\tau, \mathbf{b}^{(\tau)})p_j(\tau, \mathbf{b}^{(\tau)})$.

# 3 Likelihood inference

Suppose that we have observed a pattern data $\mathbf{n}$ and want to make some conclusions about our parameters (topology, branch lengths and model parameters) by using the likelihood. The likelihood is simply the probability defined above, but treating it as a function of the parameters:

$$L(\tau, \mathbf{b}^{(\tau)}) = P(n_1, \ldots, n_{4^k}|\mathbf{p}(\tau, \mathbf{b}^{(\tau)})) = \binom{n}{n_1 \ \ldots \ n_{4^k}} p_1^{n_1} \cdot \ldots \cdot p_{4^k}^{n_{4^k}}. \qquad (1)$$

By definition, the ML-estimate of $(\tau, \mathbf{b}^{(\tau)})$ is obtained by maximising the likelihood $L(\tau, \mathbf{b}^{(\tau)})$ with respect to $\tau$, $\mathbf{b}^{(\tau)}$ and model parameters if there are any. This is done by maximising $L(\tau_i, \mathbf{b}^{(\tau_i)})$ with respect to $\mathbf{b}^{(\tau_i)}$ and possible model parameters for each topology $\tau_i$ and comparing which of the topologies $\{\tau_i\}$ had the largest maximized likelihood. This is equivalent to comparing

which topology has the relatively largest maximised likelihood. Define $\hat{\mathbf{b}}_{ML}^{(\tau_i)}$ to be the set of branch lengths maximising the likelihood for topology $\tau_i$: $\hat{\mathbf{b}}_{ML}^{(\tau_i)} = \text{argmax}_{\mathbf{b}^{(\tau_i)}} L(\tau_i, \mathbf{b}^{(\tau_i)})$. The ML-estimate for the topology is then given by

$$\hat{\tau}_{ML} = \text{argmax}_i \frac{L(\tau_i, \hat{\mathbf{b}}_{ML}^{(\tau_i)})}{\sum_j L(\tau_j, \hat{\mathbf{b}}_{ML}^{(\tau_j)})}. \tag{2}$$

On the right hand side we have inserted the denominator, which is constant with respect to $i$ and is hence irrelevant, for latter use. The terms in the numerator and denominator above are called *profile likelihoods*. The ML-estimate for topology is hence the topology having largest profile likelihood. To actually compute $\hat{\tau}_{ML}$ and its corresponding branch lengths $\mathbf{b}_{ML}^{(\hat{\tau}_{ML})}$ numerically is a non-trivial numerical task, but it is outside the scope of the present paper and is not discussed further.

A common measure of support for an estimated tree topology is to use non-parametric bootstrap (Felsenstein, 1985). This is done by repeatedly generating new "pseudo" pattern data $\mathbf{n}^*$ from the original observed pattern $\mathbf{n}$ data by sampling $n$ patterns (i.e. the same length as the original data) *with replacement* from the original data. The resulting pattern data vector will then be an outcome of the multinomial distribution, but now with parameters $n$ and $(p_1^* = n_1/n, \ldots, p_{4^k}^* = n_{4^k}/n)$. For each such "pseudo" pattern data, a maximum likelihood estimate $\hat{\tau}_{ML}^*$ is computed using the method just described. This is repeated many (e.g. 10000) times and the so-called empirical bootstrap support for $\hat{\tau}_{ML}$ is defined as the proportion of bootstrap replicates having the same ML-topology as the original ML-topology:

$$\text{BS}(\hat{\tau}_{ML}) = \text{Bootstrap support for } \hat{\tau}_{ML} = \frac{\#\text{ replicates with } \hat{\tau}_{ML}^* = \hat{\tau}_{ML}}{\#\text{ replicates in total}}.$$

The bootstrap replicates are generated independently, so as more replicates are taken the (empiricial) bootstrap support above converges to the theoretical bootstrap support defined by

$$\text{BS}_{Th}(\hat{\tau}_{ML}) = P(\hat{\tau}_{ML}^* = \hat{\tau}_{ML})$$

It is intuitively clear that a higher bootstrap support should indicate a

stronger belief in the estimated topology since fewer observed site-patterns then talk in favor of other topologies. However, the absolute value of the support has no obvious biological interpretation, and to compare the value of a bootstrap support with other support measures can therefore be misleading.

# 4 Bayesian inference

In Bayesian analysis we use the same evolutionary model but we also need a prior distribution for $\tau$ and $\mathbf{b}^{(\tau)}$. In principle one could specify a prior for the $\mathbf{p}$-vector instead, but since the evolutionary model for the sequences is defined given the topology $\tau$ and branch lengths $\mathbf{b}^{(\tau)}$, it is more natural to define the prior in terms of $\tau$ and $\mathbf{b}^{(\tau)}$. If no prior knowledge about the topology and branch lengths is available a common choice is to have a uniform distribution for the tree topology (all topologies are equally likely prior to the analysis) and to let the branches have independent, exponentially distributed branch lengths (these are the default priors in MrBayes 3.0, Ronquist and Huelsenbeck, 2003). This will induce a prior on the $\mathbf{p}$-vector but it will not be a uniform distribution. In fact, no matter what topology and set of branch lengths are the true ones, certain patterns will always be more likely than others. For example, in the J-C-model a pattern for which all terminals have the same nucleotide (e.g. A) is always more likely than any other pattern, because, for an edge having A at one end the most likely nucleotide at the other end is A, irrespective of the length of the edge.

Given the model and a prior distribution $\pi(\tau, \mathbf{b}^{(\tau)})$ for the topology and branch lengths, Bayesian inference summarizes the knowledge about the parameters in the *posterior distribution* $\pi(\tau, \mathbf{b}^{(\tau)}|\mathbf{n})$. By Bayes' theorem the posterior distribution is proportional to the prior distribution multiplied by the likelihood:

$$\pi(\tau, \mathbf{b}^{(\tau)}|\mathbf{n}) \propto L(\tau, \mathbf{b}^{(\tau)})\pi(\tau, \mathbf{b}^{(\tau)}), \tag{3}$$

where the proportionality factor depends on the vector $\mathbf{n}$ but not on the parameters. As more and more data is collected and/or the prior distribution

$\pi(\tau, \mathbf{b}^{(\tau)})$ is close to uniform/flat, the likelihood plays the dominating role on the right hand side of (3) implying that

$$\pi(\tau, \mathbf{b}^{(\tau)}|\mathbf{n}) \approx const \times L(\tau, \mathbf{b}^{(\tau)}). \tag{4}$$

If we are only interested in the topology parameter $\tau$ our conclusions should be based on the posterior distribution for $\tau$, which is the marginal posterior distribution simply obtained by integrating out the branch lengths: $\pi(\tau|\mathbf{n}) = \int \pi(\tau, \mathbf{b}^{(\tau)}|\mathbf{n}) d\mathbf{b}^{(\tau)}$. We now assume that we have long sequences thus justifying the approximation (4). This assumption together with the fact that a posterior distribution is a proper distribution summing to unity, implies that

$$\pi(\tau|\mathbf{n}) \approx \frac{\int L(\tau, \mathbf{b}^{(\tau)}) d\mathbf{b}^{(\tau)}}{\sum_j \int L(\tau_j, \mathbf{b}^{(\tau_j)}) d\mathbf{b}^{(\tau_j)}}.$$

We would have exact equality above if the integrals contained the prior distributions $\pi(\mathbf{b}^{(\tau)})$ in the numerator and $\pi(\mathbf{b}^{(\tau_j)})$ in the denominator.

Our Bayesian point estimate for $\tau$ is hence the most probable value in the posterior distribution:

$$\hat{\tau}_B = \text{argmax}_i \pi(\tau_i|\mathbf{n}) \approx \text{argmax}_i \frac{\int L(\tau_i, \mathbf{b}^{(\tau_i)}) d\mathbf{b}^{(\tau_i)}}{\sum_j \int L(\tau_j, \mathbf{b}^{(\tau_j)}) d\mathbf{b}^{(\tau_j)}}.$$

The Bayesian estimator is hence the topology having largest relative *marginal* likelihood.

Further, the Bayesian support for $\hat{\tau}_B$ is simply its posterior probability:

$$\text{BPP}(\hat{\tau}_B) = \pi(\hat{\tau}_B|\mathbf{n}) = \max_i \pi(\tau_i|\mathbf{n}) \approx \max_i \frac{\int L(\tau_i, \mathbf{b}^{(\tau_i)}) d\mathbf{b}^{(\tau_i)}}{\sum_j \int L(\tau_j, \mathbf{b}^{(\tau_j)}) d\mathbf{b}^{(\tau_j)}}. \tag{5}$$

The interpretation of the Bayesian support is clear. Given the Bayesian viewpoint and that the prior distribution and evolutionary model is correct, the Bayesian support for $\hat{\tau}_B$ is the probability that our estimate is correct. The more data we have, the less influential is the choice of prior.

Bayesian methods have been increasingly used recently because of the powerful numerical Markov chain Monte Carlo (MCMC) method (e.g. Gilks *et al.*, 1996) used for obtaining approximations of posterior probabilities even in very complicated settings, and implemented for phylogenetics in MrBayes 3.0 (Ronquist and Huelsenbeck, 2003) for example.

# 5 Comparison of the two estimation methods

Before our main task, to compare the two support measures BS and BPP for an estimated tree topology, we first note that the two methods need not even give the same estimated topology. Recall that $\hat{\tau}_{ML}$ was obtained by maximizing the profile likelihood whereas $\hat{\tau}_B$ was obtained by maximizing the marginal likelihood (assuming long sequences). Since the two methods have different estimators, the actual estimates for a given data set are not necessarily the same (see Svennblad et al, 2006). When this occurs there is of course no reason at all to expect the two support measures to be the same. In practice however, the two estimates most often coincide.

Assume from now on that the two methods gave the same estimated topology, i.e. that $\hat{\tau}_{ML} = \hat{\tau}_B$, from now on denoted $\hat{\tau}$. Recall that the BS for $\hat{\tau}$ equals the probability that the topology having largest relative *profile* likelihood of a bootstrap replicate coincides with $\hat{\tau}$ (which had largest relative *profile* likelihood of the original data set). Recall further that the BPP for $\hat{\tau}$ equals the largest relative *marginal* likelihood. There is hence a clear difference between the two support measures in that they use the profile and marginal likelihoods respectively. From this point of view there is no reason to expect that the two support measures should be approximately equal (see Svennblad et al, 2006, for a further discussion). In what follows we give an argument to why BPP>BS.

## 5.1 A mathematical argument why BPP> BS

Our argument to why BPP> BS for an estimated tree topology $\hat{\tau}$ contains the following steps

$$\mathrm{BS}(\hat{\tau}) \approx \mathrm{BS}_{Th}(\hat{\tau}) \overset{(A)}{\approx} \max_i \frac{L(\tau_i, \hat{\mathbf{b}}_{ML}^{(\tau_i)})}{\sum_j L(\tau_j, \hat{\mathbf{b}}_{ML}^{(\tau_j)})} \overset{(B)}{<} \max_i \frac{\int L(\tau_i, \mathbf{b}^{(\tau_i)}) d\mathbf{b}^{(\tau_i)}}{\sum_j \int L(\tau_j, \mathbf{b}^{(\tau_j)}) d\mathbf{b}^{(\tau_j)}} \approx \mathrm{BPP}(\hat{\tau}). \tag{6}$$

The first approximation relies on that enough bootstrap replicates are taken and was motivated in Section 3, and the last approximation, motivated in Section 4, relies on the sequences being long and/or the prior distribution being close

to uniform so that the prior can be neglected (if the prior is uniform it is an exact equality). There are hence two remaining steps needing motivation. The first part, labelled (A) in equation (6), is that the theoretical bootstrap support approximately equals the largest relative likelihood and the second part, labelled (B) in equation (6), is that the largest relative *profile* likelihood is smaller than the largest relative *marginal* likelihood.

We start with (A) in equation (6) which is our weakest point in the argument both on theoretical grounds and empirically in our simulation studies in the next section. Recall that $\hat{\tau}_{ML}$ was defined in (2) as the topology having largest relative profile likelihood, and $\mathrm{BS}_{Th}(\hat{\tau}_{ML}) = P(\hat{\tau}^*_{ML} = \hat{\tau}_{ML})$, i.e. the probability that the topology of the maximized relative likelihood of a bootstrap replicate is the same as the topology of the maximised relative likelihood of the original data. We have no strong argument to why this probability should approximately equal the maximized relative profile likelihood. However, since both the estimate and support values use the *profile* likelihood we find it more plausible that $\mathrm{BS}_{Th}(\hat{\tau}_{ML})$ should resemble this relative profile likelihood rather than the corresponding relative *marginal likelihood*, since neither the estimate nor the BS uses the latter. Our motivation for (A) is thus that if we were to approximate $\mathrm{BS}_{Th}(\hat{\tau}_{ML})$ by the maximized relative *profile* likelihood or the maximized relative *marginal* likelihood, we would recommend the former.

We now motivate inequality (B) in equation (6), which has stronger mathematical grounds as well as strong empirical support from simulations in Section 6.

In equation (1) it was shown that $L(\tau, \mathbf{b}^{(\tau)}) = P(n_1, \ldots, n_{4^k} | \mathbf{p}(\tau, \mathbf{b}^{(\tau)}))$, where $\mathbf{n}$ is multinomially distributed with parameters $n$ (the sequence length) and the vector $\mathbf{p}(\tau, \mathbf{b}^{(\tau)})$. The multinomial distribution can be approximated by the normal distribution when $n$ is large. As a consequence we have that the likelihood $L(\tau, \mathbf{b}^{(\tau)})$ (approximately) equals a high-dimensional normal distribution in terms of $\mathbf{n}$. It can clearly not be normal in the parameters $\tau$ and $\mathbf{b}^{(\tau)}$ since $\tau$ is a discrete nominal parameter. Assume instead that we can reparametrise the vector $(\tau, \mathbf{b}^{(\tau)})$ into a continuous vector $\mathbf{y}$ such that we move between the

topologies $\tau_1,\ \tau_2, \ldots$, as the components of $\mathbf{y}$ vary continuously. Assume further that the different topologies split up the new parameters space into symmetric regions. One way of doing this reparametrisation obeying both assumptions is to let the components of $\mathbf{y}$ express the distance between each *pair* of terminals. This vector has unnecessary high dimension and there are many numeric choices of $\mathbf{y}$ for which there is no matching values for $(\tau, \mathbf{b}^{(\tau)})$. However, each choice of $(\tau, \mathbf{b}^{(\tau)})$ is mapped on to a unique $\mathbf{y}$-vector $\mathbf{y}(\tau, \mathbf{b}^{(\tau)})$, and it is possible to move between topologies as components of $\mathbf{y}$ vary continuously.

Under this assumption the likelihood $L(\mathbf{y})$ still approximately equals a normal distribution in terms of $\mathbf{n}$, where the mean vector and covariance matrix are complicated functions of the vector $\mathbf{y}$. Suppose now that the likelihood $L(\mathbf{y})$ also approximately equals a normal distribution in terms of the $\mathbf{y}$-vector, so $L(\mathbf{y}) = const \times f(\mathbf{y})$ where $f(\cdot)$ is the normal density function of the right dimension. Because of the complicated structure of the vector function $\mathbf{p}(\tau, \mathbf{b}^{(\tau)})$, which is the parameter vector in the multinomial distribution for $\mathbf{n}$, added with the reparametrisation from $(\tau, \mathbf{b}^{(\tau)})$ to $\mathbf{y}$ this assumption is hard or impossible to check. Admittedly, it is a rather strong assumption which we make without proof. A simple univariate comparison, for which the assumption is true, is to let $x$ come from the normal distribution with mean parameter $y$ and standard deviation $\sigma$. Then, treating $x$ as fixed, it follows that $y$ is normally distributed with mean $x$ and standard deviation $\sigma$. With our assumption we have a vector $\mathbf{y}$ of parameters which is normally distributed, and the $\mathbf{y}$-regions corresponding to different topologies splits up the parameter space into symmetric regions.

As before we label the topologies $\tau_1,\ \tau_2, \ldots$, and somewhat incorrectly the corresponding regions in the parameter space for $\mathbf{y}$ are also denoted $\tau_1,\ \tau_2, \ldots$. Let $f_i = \sup_{\mathbf{y} \in \tau_i} f(\mathbf{y})$, so $f_i$ equals the largest likelihood value within topology $\tau_i$, and let $F_i = \int_{\tau_i} f(\mathbf{y}) d\mathbf{y}$, the probability for the parameter being in $\tau_i$ (see Figure 1 for a 1-dimensional illustration).

To show inequality (B) in equation (6) is then equivalent to showing that $f_{i_{\max}} / \sum_j f_j < F_{i_{\max}} / \sum_j F_j$. We show the result for the 1-dimensional case but it can be extended to higher dimensions. Suppose the normal distribution

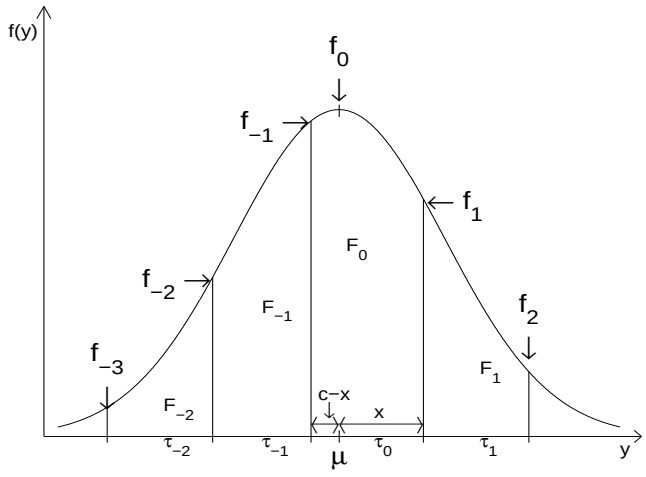


**Figure 1:** In the figure the normal density is plotted and the regions $\tau_i$ are marked. The value $f_i$ equals the maximal density value in the $\tau_i$-interval, and $F_i$ is the corresponding area (the area of the density over $\tau_i$).

has mean $\mu$ and standard deviation $\sigma = 1$ without loss of generality, and let each topology correspond to an interval of length $c$ (by symmetry all topologies should have equal length). Relabel the topologies so that $i_{\max} = 0$ is the interval/topology for which both $f_i / \sum_j f_j$ and $F_i / \sum_j F_j$ are maximized. Let $x$ denote the distance between $\mu$ and the right end-point of interval $\tau_0$ (so $0 \leq x \leq c$) and consequently $c - x$ ($\geq 0$) is the distance between $\mu$ and the left end-point of $\tau_0$ (see Figure 1). We then have $F_{i_{\max}} / \sum_j F_j = F_0/1 = \Phi(x) - \Phi((-(c-x))$, where $\Phi(\cdot)$ is the standard normal distribution. Similarly, the quantity $f_{i_{\max}} / \sum_j f_j$ equals

$$\frac{\varphi(0)}{\varphi(0) + \sum_{k \geq 0} \varphi(x + kc) + \sum_{k \leq 0} \varphi((x-c) + kc)},$$

where $\varphi(y) = \exp(-y^2/2)/\sqrt{2\pi}$ is the standard normal density function. Below we show that $F_{i_{\max}} / f_{i_{\max}} \geq F_j / f_j$ for all $j$. This will prove our statement, since

it then follows that $f_j = F_j \frac{f_j}{F_j} \leq F_j \frac{f_{i_{\max}}}{F_{i_{\max}}}$ implying that

$$\sum_j f_j \leq \frac{f_{i_{\max}}}{F_{i_{\max}}} \sum_j F_j,$$

which is exactly the postulated statement. It remains to show that $F_{i_{\max}}/f_{i_{\max}} \geq F_j/f_j$, for all $j > 0$, which we do for the special case that $x = c$, the proof for $j < 0$ and general $x$ being similar. We have that

$$\begin{aligned} \frac{F_{i_{\max}}}{f_{i_{\max}}} &= \frac{\int_0^c \varphi(x)dx}{\varphi(0)} = \int_0^c \frac{\varphi(x)}{\varphi(0)}dx, \\ \frac{F_j}{f_j} &= \frac{\int_0^c \varphi(x+jc)dx}{\varphi(jc)} = \int_0^c \frac{\varphi(x+jc)}{\varphi(jc)}dx. \end{aligned}$$

From this we see that our statement follows if we can show that $\varphi(x+y)/\varphi(y)$ is a decreasing function of $y$. But $\frac{d}{dy}\varphi(x+y)/\varphi(y) = -x\varphi(x+y)/\varphi(y) < 0$ which completes the "proof" for inequality (B) in equation (6). The inequality is strongly supported in the next section where we have simulated the two sides a number of times.

To summarise, we have motivated all steps in equation (6) which hence gives a mathematical argument to why BPP> BS.

# 6 Simulations

In order to investigate the approximations involved in equation (6) we have performed simulations as follows. We have considered the 4 taxon case implying that there are only three possible topologies. We have generated the topology of the tree uniformly (each tree having probability 1/3). For the branch lengths we have assumed independent exponential branch lengths, as is default in MrBayes 3.0, but varied the mean of the prior branch length. Given the tree, i.e. topology and branch lengths, we have generated a data set of aligned sequences according to the Jukes-Cantor model (Jukes and Cantor, 1969). Then, for this data set we have estimated a tree and its support, using maximum likelihood methods combined with bootstrap support as described in Section 3, and using Bayesian methods as described in Section 4. The analyses were performed using Paup* (Swofford, 2003) and MrBayes 3.0 (Ronquist and Huelsenbeck, 2003)

respectively. Additional to the two support measures we have also computed the maximized relative profile likelihood (mrpl) numerically using Fortran subroutines. Thus, we have obtained the quantities $\mathrm{BS}(\hat{\tau})$, $\max_i \frac{L(\tau_i, \hat{\mathbf{b}}_{ML}^{(\tau_i)})}{\sum_j L(\tau_j, \hat{\mathbf{b}}_{ML}^{(\tau_j)})}$ and $\mathrm{BPP}(\hat{\tau})$ in equation (6). To obtain the theoretical bootstrap support and the maximized relative marginal likelihood is not feasible. So when investigating part (A) in equation (6) in the simulation study we instead investigate if the empirical bootstrap support more resembles mrpl than it resembles the Bayesian support, and for part (B) if the maximized relative profile likelihood is smaller than the Bayesian support. The simulation study hence investigates if

$$\mathrm{BS}(\hat{\tau}) \overset{(A)}{\approx} \max_i \frac{L(\tau_i, \hat{\mathbf{b}}_{ML}^{(\tau_i)})}{\sum_j L(\tau_j, \hat{\mathbf{b}}_{ML}^{(\tau_j)})} \overset{(B)}{<} \mathrm{BPP}(\hat{\tau}). \tag{7}$$

Simulations were performed for two sequence lengths, $n = 100$ sites and $n = 1000$ sites. For the shorter sequence length, the mean of the prior branch length distribution was varied from 0.025 substitutions per site up to 0.2 substitution per site. For each prior distribution 15 trees were generated. For each such tree we simulated sequences (according to the model) and performed the two statistical procedures and also computed the maximized relative profile likelihood. In the Bayesian analysis we used the same prior distribution for the branch lengths as the tree was generated from.

The results from the simulations are given in Table 1 for the shorter sequence length ($n = 100$) and in Table 2 for the longer sequence length ($n = 1000$). To investigate (A) in (7) we have, for each mean prior branch length, listed the fraction of times the bootstrap support was closer to the maximized relative profile likelihood than it was to the Bayes support (which approximately equals the relative maximized marginal likelihood). Before comparing the support measures we have transformed the support values using the log-odds transformation ($\ln(x/(1-x))$). This transformation for example makes a support value of 98% closer to 96% than to 99.9% which agrees with general opinion (the results are very similar without the transformation). To investigate (B) we present the fraction of times the mrpl is smaller than the Bayes support as suggested by (B). In Tables 1 and 2 we only present results from the trees for which the Bayesian

posterior probability was larger than 0.5 (smaller support values are not very interesting and when getting close to 1/3 they are also subject to rounding off errors) and smaller than 0.99 (for higher support values the precision is of the order of the support why the latter cannot be relied upon). This explains why much less than 15 (and 50 respectively) analyses are presented for each prior.

Table 1: Simulation study of (7) for sequence length $n = 100$ sites. Approximation (A) is investigated by the proportion of times the bootstrap support lies closer to the mrpl than it does to the Bayes support, and (B) is investigated by proportion of times the mrpl is smaller than the Bayes support. See text for further details.

| Mean branch length in prior | Frequency BS closer to mrpl | Frequency $mrpl < BPP$ | Frequency $BS < BPP$ |
|---|---|---|---|
| 0.025 | 5/5 (100%) | 5/5 (100%) | 5/5 (100%) |
| 0.05 | 4/4 (100%) | 4/4 (100%) | 4/4 (100%) |
| 0.075 | 4/4 (100%) | 4/4 (100%) | 4/4 (100%) |
| 0.1 | 3/5 (60%) | 3/5 (60%) | 4/5 (80%) |
| 0.2 | 3/5 (60%) | 5/5 (100%) | 3/5 (60%) |

For shorter mean prior branch lengths than presented in the table the support for (A) is still strong, although branch lengths become short such that data sequences of length 100 are not sufficient to estimate the tree topology with any precision. For mean prior branch lengths longer than presented the support for (A) decreases, but then branches are becoming saturated not containing much signal.

For sequences length equal to $n = 1000$ the corresponding results are presented in Table 2. For shorter mean prior branch length not supported in the Table 2 hardly no simulated trees hade Bayesian support between 0.5 and 0.99. For longer mean prior branch length nearly all Bayesian support values were close to 1 thus not very useful when investigating how our approximations perform.

From the simulations we see, from the second column of Table 1 and Table 2, that approximation (A) in (7) is empirically supported by simulations (for $n = 100$ in particular when the prior generating the trees does not have too long mean branch lengths). From the third column we can also see that in-

Table 2: Simulation study of (7) for sequence length $n = 1000$ sites. Approximation (A) is investigated by the proportion of times the bootstrap support lies closer to the mrpl than it does to the Bayes support, and (B) is investigated by proportion of times the mrpl is smaller than the Bayes support. See text for further details.

| Mean branch length in prior | Frequency BS closer to mrpl | Frequency $mrpl < BPP$ | Frequency $BS < BPP$ |
|---|---|---|---|
| 0.003 | 6/6 (100%) | 6/6 (100%) | 6/6 (100%) |
| 0.004 | 7/8 (88%) | 7/8 (88%) | 7/8 (88%) |
| 0.005 | 12/12 (100%) | 12/12 (100%) | 12/12 (100%) |
| 0.008 | 8/8 (60%) | 8/8 (60%) | 8/8 (100%) |
| 0.010 | 5/6 (83%) | 5/6 (83%) | 6/6 (100%) |

equality (B) in (7) seems empirically valid irrespective of the mean prior branch length. From the simulations we hence conclude that our mathematical arguments are supported by our simulations. In Tables 1 and 2 we have also listed the frequency with which the Bayes support exceeds the bootstrap support, a frequency which is very high for both sequence lengths and all mean prior branch lengths. This systematic difference, which the present paper tries to give mathematical arguments for, also has strong empirical evidence from many other studies (e.g. Wilcox *et al.*, 2002, Alfaro *et al.*, 2003, Douady *et al.*, 2003, and Erixon *et al.*, 2003).

# 7 Discussion

We have given mathematical arguments indicating why BPP for estimated topologies are larger than corresponding BS values. We admit that several assumptions are made without proofs implying that the results are not rigorous, but our hope is to trigger further research strengthening the arguments. Since both estimation methods, maximum likelihood and Bayesian statistical inference, are consistent, the support values will both tend to 1 as the sequence length $n$ increases keeping everything else fixed. To show that BPP>BS could therefore for be formalized in a result like $P((1-\mathrm{BS}(\hat{\tau}_{ML}))/(1-\mathrm{BPP}(\hat{\tau}_B)) > 1) \to 1$ as $n$ tends to infinity for a wide class of evolutionary models, but a complete

proof for this seems hard to obtain.

## Acknowledgements

Financial support from The Swedish Research Council and The Linneaues Centre for Bioinformatics, Uppsala University, is gratefully acknowledged.

## References

Alfaro, M.E., Zoller, S., and Lutzoni, F. (2003) *Mol. Biol. Evol.,* **20**, 255-266.

Cummings. M.P., Handley, S.A., Myers, D.S., Reed, D.L., Rokas, A., and Winka, K. (2003). Comparing Bootstrap and Posterior Probability Values in the Four-Taxon Case. *Syst. Biol.* **52**, 477-487.

Douady, C.J., Delsuc, Boucher, Y., Doolittle, W.F., and Dourzery, E.J.P. (2003) *Mol. Biol. Evol.,* **20**, 248-254.

Efron, B., Halloran, E., and Holmes, S. (1996) Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci.,***93**, 13429-13434.

Erixon P., Svennblad B., Britton, T. and Oxelman B. (2003): The reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Syst. Biol.,* **52**, 665-674.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution,* **39**, 783-791.

Gilks, W.R., S. Richardson, and D.J.Spiegelhalter. (1996). *Markov chain Monte Carlo in practice*. Chapman & Hall, London.

Jukes T.H. and Cantor C.R. (1969) Evolution of protein molecules. Pp. 21-32 *in Mammalian protein metabolism* (H. N. Munro, ed.). Academic press, New York.

Larget, B., and Simon, D.L. (1999). Markov Chain Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees *Mol. Biol. Evol.* **16**, 750-759.

Ronquist F., and Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics,* **19**, 1572-1574.

Ross, S.M. (2006) *A first course in probability, 7 ed.* Prentice Hall, London.

Simmons, M.P., Pickett, K.M., and Miya, M. (2004). How meaningful are Bayesian support values? *Mol. Biol. Evol.* **21**, 188-199.

Svennblad B., Erixon P., Oxelman B., **Britton T**. (2006): Fundamental differences between maximum likelihood and Bayesian inference in phylogenetics. *Systematic Biology,* **55**, 116-121.

Swofford, D.L. 2003. PAUP* Phylogenetic analysis using parsimony, 4th ver. Sinauer, Sunderland, MA.

Wilcox, T.P., Zwickl, D.J., Heath, T.A., and Hillis, D.M. (2002) *Mol. Phyl. Evol.,* **25**, 361-371.