*Syst. Biol.* 54(3):455–470, 2005

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150590945313

# Branch-Length Prior Influences Bayesian Posterior Probability of Phylogeny

ZIHENG YANG[1] AND BRUCE RANNALA[2]

[1]*Department of Biology, University College London, Darwin Building, Gower Street, London WC1E 6BT, United Kingdom; E-mail: z.yang@ucl.ac.uk*
[2]*Genome Center and Section of Evolution and Ecology, University of California Davis, One Shields Avenue, Davis, California 95616, USA*

*Abstract*.— The Bayesian method for estimating species phylogenies from molecular sequence data provides an attractive alternative to maximum likelihood with nonparametric bootstrap due to the easy interpretation of posterior probabilities for trees and to availability of efficient computational algorithms. However, for many data sets it produces extremely high posterior probabilities, sometimes for apparently incorrect clades. Here we use both computer simulation and empirical data analysis to examine the effect of the prior model for internal branch lengths. We found that posterior probabilities for trees and clades are sensitive to the prior for internal branch lengths, and priors assuming long internal branches cause high posterior probabilities for trees. In particular, uniform priors with high upper bounds bias Bayesian clade probabilities in favor of extreme values. We discuss possible remedies to the problem, including empirical and full Bayesian methods and subjective procedures suggested in Bayesian hypothesis testing. Our results also suggest that the bootstrap proportion and Bayesian posterior probability are different measures of accuracy, and that the bootstrap proportion, if interpreted as the probability that the clade is true, can be either too liberal or too conservative. [Fair-balance paradox; Lindley's paradox; model selection; molecular phylogenetics; posterior probabilities; prior; star tree paradox.]

For both reconstruction of phylogenetic relationships and use of phylogenies to understand molecular evolution, it is essential to quantify the statistical uncertainty in inferred phylogenies. Yet the phylogeny differs from a conventional statistical parameter and this difference poses obstacles to straightforward application of statistical estimation theory (Yang et al., 1995). Although maximum likelihood (ML) (Felsenstein, 1981) appears to be efficient for obtaining point estimates of phylogenies, determining statistical confidence has proven more difficult (Goldman et al., 2000).

A recent advance in molecular phylogenetics has been the development of the Bayesian approach (Rannala and Yang, 1996; Mau and Newton, 1997; Li et al., 2000), which circumvents some of the controversies surrounding the nonparametric bootstrap, the most commonly used procedure for assessing phylogenetic uncertainty (Felsenstein, 1985). Implementations of efficient Markov chain Monte Carlo (MCMC) algorithms (Larget and Simon, 1999; Huelsenbeck and Ronquist, 2001) have made the method very popular, and it is now widely used to infer species relationships such as the radiation of mammalian orders (Murphy et al., 2001) or the origin of land plants (Karol et al., 2001). However, posterior probabilities for trees or clades produced by the Bayesian method have often appeared surprisingly high (e.g., Suzuki et al., 2002), as was noted in the very first Bayesian phylogenetic analysis (Rannala and Yang, 1996). Several recent studies comparing posterior probabilities and bootstrap proportions using computer simulation suggest that bootstrap proportions tend to be too conservative, whereas posterior probabilities are too liberal (Suzuki et al., 2002; Cummings et al., 2003; Erixon et al., 2003; Simmons et al., 2004). However, most of those studies are hard to interpret as they did not simulate the trees and branch lengths under the same prior as was used in the Bayesian analysis, and thus theoretical expectations for the results are unavailable.

Here we examine the problem of spuriously high posterior probabilities by studying the simplest case of phylogeny reconstruction, namely estimation of the rooted tree for three species using binary characters evolving at a constant rate (Yang, 2000). The analysis of this simple case does not require the use of MCMC algorithms, and thus computational problems such as lack of convergence and poor mixing of the Markov chain are avoided. To establish the relevance of our analysis of the simple case to real data analysis, we corroborate our results by analyzing an empirical data set concerning the origin of land plants.

## SIMULATION EXPERIMENT

### *Bayesian Estimation of Rooted Tree for Three Species*

Here we describe our simulation study of the simple case of three species. Analysis of the real data set concerning land plant divergences is described later. Let the three binary rooted trees for species 1, 2, 3 be $T_1 = ((1, 2), 3)$, $T_2 = ((2, 3), 1)$, and $T_3 = ((3, 1), 2)$ (see Fig. 1). Each tree has an internal branch length $t_0$ and an external branch length $t_1$, measured by the expected number of changes per site. The data consist of three sequences of binary characters, evolving according to a continuous-time Markov process with equal substitution rates between the two characters. The molecular clock (rate constancy over time) is also assumed. This is the binary equivalent of the constant-rate Jukes and Cantor (1969) model for nucleotide substitution. For example, we could imagine the two states as representing purines and pyrimidines in a DNA sequence. The sequence data are summarized as the counts of four site patterns: $n_0$ for the constant pattern $xxx$, and $n_1$, $n_2$, and $n_3$ for the variable patterns $xxy$, $yxx$, and $xyx$, where $x$ and $y$ are two different states. Let $\mathbf{n} = \{n_0, n_1, n_2, n_3\}$.

The Bayesian approach to phylogeny estimation places prior distributions on trees and their branch lengths. The prior can represent either objective information or personal beliefs about the parameters before the data are collected and analyzed. We leave it open whether the prior should be interpreted in an objective or subjective Bayesian framework. We assume a uniform prior probability (1/3) for the three trees, and, given the tree topology, exponential priors for $t_0$ and $t_1$: $f(t_0 \mid \mu_0) =$

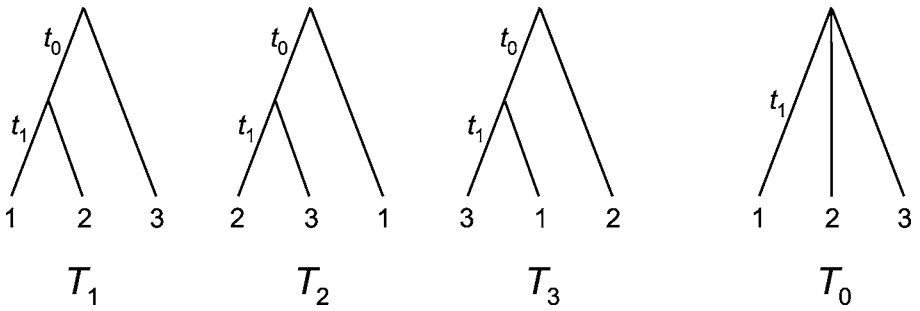


FIGURE 1. The three rooted trees for three species: $T_1 = ((1, 2), 3)$, $T_2 = ((2, 3), 1)$, and $T_3 = ((3, 1), 2)$. Branch lengths $t_0$ and $t_1$ are measured by the expected number of character changes per site. The star tree $T_0 = (1, 2, 3)$ is also shown with its branch length $t_1$.

$\exp(-t_0/\mu_0)/\mu_0$ and $f(t_1 \mid \mu_1) = \exp(-t_1/\mu_1)/\mu_1$, where $\mu_0$ and $\mu_1$ are the means. We also explore a few other priors for $t_0$ and $t_1$, such as uniform and gamma distributions, as described later.

From Yang (2000) (see also Newton, 1996), the likelihood is the multinomial probability of observing data given the tree and branch lengths:

$$\begin{aligned} f(\mathbf{n} \mid T_1, t_0, t_1) &= C p_0^{n_0} p_1^{n_1} p_2^{n_2+n_3}, \\ f(\mathbf{n} \mid T_2, t_0, t_1) &= C p_0^{n_0} p_1^{n_2} p_2^{n_3+n_1}, \\ f(\mathbf{n} \mid T_3, t_0, t_1) &= C p_0^{n_0} p_1^{n_3} p_2^{n_1+n_2}, \end{aligned} \tag{1}$$

where $C = n!/(n_0!\, n_1!\, n_2!\, n_3!)$ is a proportionality constant, and $p_0$, $p_1$, $p_2$, $p_3$ are probabilities of observing the four site patterns, respectively, under tree $T_1$:

$$\begin{aligned} p_0(t_0, t_1) &= \frac{1}{4} + \frac{1}{4}e^{-4t_1} + \frac{1}{2}e^{-4(t_0+t_1)}, \\ p_1(t_0, t_1) &= \frac{1}{4} + \frac{1}{4}e^{-4t_1} - \frac{1}{2}e^{-4(t_0+t_1)}, \\ p_2(t_0, t_1) &= \frac{1}{4} - \frac{1}{4}e^{-4t_1} = p_3(t_0, t_1). \end{aligned} \tag{2}$$

The posterior probability for tree $T_i$, $i = 1, 2, 3$, is

$$\begin{aligned} P_i &= f(T_i \mid \mathbf{n}, \mu_0, \mu_1) \\ &= \frac{\frac{1}{3}\int_0^\infty \int_0^\infty f(t_0 \mid \mu_0) f(t_1 \mid \mu_1) f(\mathbf{n} \mid T_i, t_0, t_1) dt_0 dt_1}{f(\mathbf{n} \mid \mu_0, \mu_1)}, \end{aligned} \tag{3}$$

where

$$\begin{aligned} &f(\mathbf{n} \mid \mu_0, \mu_1) \\ &\quad = \frac{1}{3}\sum_{j=1}^{3}\left[\int_0^\infty \int_0^\infty f(t_0 \mid \mu_0) f(t_1 \mid \mu_1) f(\mathbf{n} \mid T_j, t_0, t_1) \right. \\ &\quad \left. \times\, dt_0 dt_1\right] \end{aligned} \tag{4}$$

is the marginal probability of the data. The integrals are calculated numerically using Mathematica.

### *Computer Simulation*

We simulated data sets to examine the properties of Bayesian posterior probabilities for trees. Except where stated otherwise, we conducted Bayesian simulation, sampling values of parameters from the prior. Each data set is generated by sampling branch lengths $t_0$ and $t_1$ from their prior distributions, calculating the probabilities of the four site patterns $p_0$, $p_1$, $p_2$, $p_3$ according to Equation 2, and then generating counts of site patterns ($n_0$, $n_1$, $n_2$, $n_3$) by sampling from the multinomial distribution $M(n, p_0, p_1, p_2, p_3)$. The procedure is repeated to generate multiple data sets. We use $T_1$ as the correct tree in the simulation, but interpret the results as if the data are simulated from a random tree chosen from $T_1$, $T_2$, and $T_3$ with equal probability.

## SIMULATION RESULTS

### *Effect of Branch Length Prior in Simulated Data*

We simulated data sets using the trees of Figure 1 to examine the effect of the prior for the internal branch length on Bayesian inference of tree topology. Each simulated data set is analyzed using the Bayesian method to calculate the posterior probabilities for the three trees (Equation 3): $P_1$ for the correct tree, and $P_2$ and $P_3$ for the two wrong trees. We contrast *the simulation model*, the model used to generate the data, and *the analysis model*, the model used to analyze the data. The term *model* refers to the full model, including both the prior (for tree topology and branch lengths) and the likelihood (substitution) model. When the simulation and analysis models match, we say that the analysis model is correct. The only possible mismatch between the simulation and analysis models considered here is the prior for internal branch lengths; the correct prior for topology and the correct substitution model are assumed in the analysis.

For this simple case, the ML tree is $T_1$, $T_2$, or $T_3$, depending on whether $n_1$, $n_2$, or $n_3$ is the greatest. More precisely, $T_1$ is the ML tree if and only if $n_1 > \max(n_2, n_3)$ and $n_0 + n_1 > n_2 + n_3$ (Yang, 2000). When $n_1 > \max(n_2, n_3)$ but $n_0 + n_1 \le n_2 + n_3$, the sequences are more divergent

than random sequences, and then none of the binary trees has a higher likelihood than the star tree. The maximum posterior probability tree is similarly determined as long as the prior mean for internal branch lengths $\mu_0 > 0$: that is, if $n_1 > \max(n_2, n_3)$, we have $P_1 > \max(P_2, P_3)$. Changing $\mu_0$ affects the magnitudes of the three posterior probabilities but not their order.

We use the case where the full model is correct—that is, where the analysis model matches the simulation model—to illustrate the interpretation of posterior probabilities for trees. When the data are simulated under the prior and when the full analysis model is correct, the posterior probability for a tree is the probability that the tree is true. Figure 2a ("correct" prior) shows results of such a simulation. Each data set is generated by choosing a tree from $T_1$, $T_2$, and $T_3$ at random and by sampling $t_0$ and $t_1$ from exponential priors with means $\mu_0 = 0.02$ and $\mu_1 = 0.2$, respectively. The sequence length is $n = 500$. For those parameter values, the true tree is recovered by the likelihood or Bayesian methods with probability 0.86. When the data are analyzed (Equation 3), the correct exponential priors with the correct means $\mu_0 = 0.02$ and $\mu_1 = 0.2$ are assumed for $t_0$ and $t_1$, respectively. Each data set produces posterior probabilities $P_1$, $P_2$, $P_3$ for the three trees, and these are collected into 50 bins, at 2% width for each bin. Then in the bin with posterior probability around $P$, the tree should be the true tree with probability $P$. For example, trees in the 94% to 96%

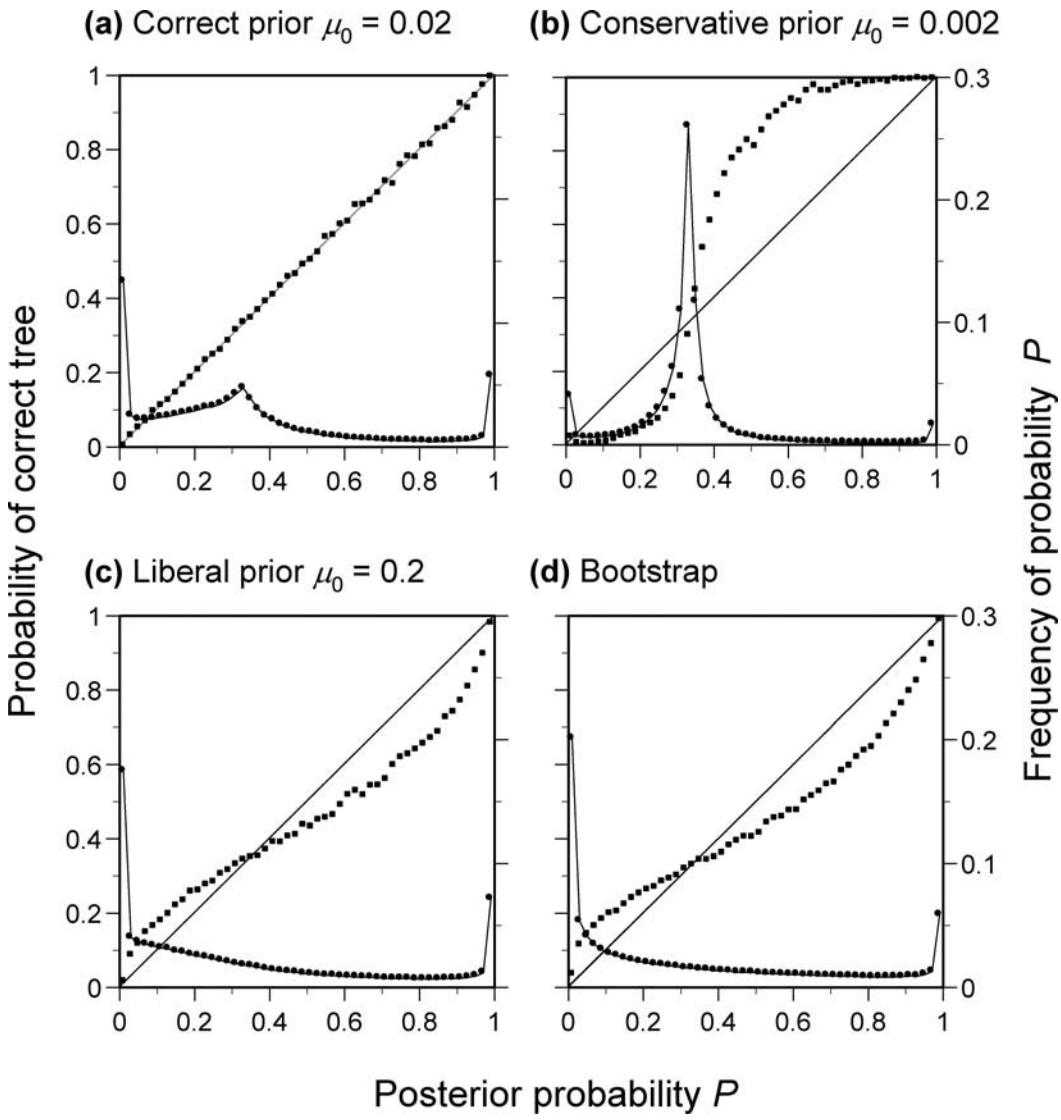


FIGURE 2. The estimated probability that the tree is correct plotted against the posterior probability $P$ for the tree. A total of 100,000 data sets were simulated using branch lengths $t_0$ and $t_1$ drawn from exponential priors with means $\mu_0 = 0.02$ and $\mu_1 = 0.2$. The sequence has $n = 500$ sites. The trees are binned into 50 bins according to their calculated posterior probabilities, and in each bin, the proportion of the true tree is calculated (squares, left $y$-axis). The circles (right $y$-axis) represent the frequency of $P$s in each bin. Each replicate is analyzed using the Bayesian method assuming (a) the correct prior with $\mu_0 = 0.02$, or incorrect priors with (b) $\mu_0 = 0.002$ or (c) $\mu_0 = 0.2$. The correct prior for $t_1$ is always assumed with $\mu_1 = 0.2$. (d) The maximum likelihood method with bootstrap is also used.

bin all have posterior probabilities close to 95%. Among them, about 95% are the true tree while others (about 5%) are either of the two alternative (incorrect) trees (Fig. 2a).

Such a match does not exist when the prior assumed in the analysis model does not match the prior assumed in the simulation. We considered the effect of the prior mean $\mu_0$ for the internal branch length only and used $\mu_1 = 0.2$ as in the simulation model. When the internal branch assumed in the prior is too short ($\mu_0 = 0.002$, "conservative" prior; Fig. 2b), low $P$s (say, $P < 1/3$) overestimate the probability of the correct tree, whereas high $P$s (say, $P > 1/2$ ) underestimate the probability of the correct tree. Thus, the method too often fails to reject or support any tree, and is too conservative. In contrast, when the mean internal branch length assumed in the prior is too large ($\mu_0 = 0.2$, "liberal" prior; Fig. 2c), low $P$s underestimate and high $P$s overestimate the probability of the correct tree, and the method is too liberal. The bootstrap method is also too liberal in these data sets if the bootstrap proportion is interpreted as the probability that the tree is correct.

Figure 2 also shows the distribution of posterior probabilities $P$ over replicate data sets. The three posterior probabilities from each data set are grouped into the 50 bins, and the frequencies in the bins are used to plot the histogram. This procedure ignores the constraint that $P_1 + P_2 + P_3 = 1$ and is not a proper way of representing the distribution of $P_1$, $P_2$ and $P_3$ (which is shown in Fig. 3 below). With the correct prior ($\mu_0 = 0.02$; Fig. 2a), most posterior probabilities are near 0 or 1, although there is a third peak near 1/3. Use of the conservative prior ($\mu_0 = 0.002$; Fig. 2b) shifted the posterior probabilities towards 1/3. Note that in the extreme case where $\mu_0 \to 0$, all three $P$s will approach 1/3. In contrast, the liberal prior ($\mu_0 = 0.2$; Fig. 2c) shifts the density towards the two tails near 0 or 1, and polarizes the probabilities.

The joint density of posterior probabilities, $f(P_1, P_2, P_3)$, is shown in Figure 3, estimated from the same simulated data sets as used in Figure 2. The correct tree is recovered in the data set if and only if $P_1 > \max(P_2, P_3)$. With the correct prior ($\mu_0 = 0.02$; Fig. 3a), there are many data sets in which $P_1$ is near 1 and many data sets in which all three $P$s are near 1/3. Note that data sets in which one of $P_1$, $P_2$, $P_3$ is 0.80 are represented by three line segments in the plot, corresponding to each of the three trees $T_1$, $T_2$, $T_3$ being the ML/Bayes tree. As the full analysis model is correct in this set of simulations (Fig. 3a), exactly 80% of the total density mass on those three line segments is located on the one corresponding to $T_1$. The conservative prior ($\mu_0 = 0.002$; Fig. 3b) shifts the density towards the center of the plot, where all three $P$s are close to 1/3. If $\mu_0 \to 0$ in the analysis model, the density reduces to a point mass at $P_1 = P_2 = P_3 = 1/3$. In contrast, the liberal prior ($\mu_0 = 0.2$; Fig. 3c) shifts the density to the three corners of the plot, so that one of $P_1$, $P_2$, $P_3$ is near 1 whereas the other two are near 0, and high probabilities (say >95%) are produced for wrong trees too often (say, >5% of the time).

Two additional sets of simulations were conducted, using $n = 200$ and 1000 sites, respectively, and using prior means $\mu_0 = 0.1$ and $\mu_1 = 0.2$. The data were analyzed in the same way as in Figures 2 and 3, assuming the correct prior ($\mu_0 = 0.1$, $\mu_1 = 0.2$), a conservative prior ($\mu_0 = 0.01$, $\mu_1 = 0.2$), and a liberal prior ($\mu_0 = 1$, $\mu_1 = 0.2$). The results (not shown) were very similar to those of Figures 2 and 3. In both sets of simulations, use of the correct prior produced a perfect match between the calculated posterior probability of a tree and the probability that the tree is true. Use of the conservative prior caused the posterior probabilities to become less extreme and the method to become too conservative. In contrast, the liberal prior made the posterior probabilities more extreme and the method too liberal. For $n = 1000$, the effect of the liberal prior was noted to be minor, because the posterior probabilities under the correct prior were already very extreme; for that sequence length, the correct tree is recovered in 96% of the simulated replicates. The effect of the conservative prior is always apparent. Furthermore, in both sets of simulations, the bootstrap proportions are noted to be too liberal, as in Figure 2d.

To examine which aspects of the prior for internal branch lengths affect posterior tree probabilities, we analyzed a fixed data set in Figure 4. The data are $\mathbf{n} = \{n_0, n_1, n_2, n_3\} = \{300, 80, 65, 55\}$. For tree $T_1$, the maximum likelihood estimates (MLEs) are $\hat{t}_0 = 0.04176$, $\hat{t}_1 = 0.16348$, with log likelihood $\ell_1 = -554.2858$, whereas both trees $T_2$ and $T_3$ reduce to the star tree $T_0$, with estimates $\hat{t}_0 = 0$, $\hat{t}_1 = 0.19054$, and $\ell_2 = \ell_3 = \ell_0 = -556.2283$. The bootstrap proportions for the three trees are (0.887, 0.104, 0.009). The results of Bayesian analysis are shown in Figure 4, using exponential (Fig. 4a), uniform (Fig. 4b), and gamma (Fig. 4c) priors. In Figure 4a, exponential priors are used for $t_0$ and $t_1$, with the means $\mu_0$ varying and $\mu_1 = 0.1$ fixed. When $\mu_0$ increases from 0 to $\infty$, the posterior probabilities ($P_1$, $P_2$, $P_3$) change from (1/3, 1/3, 1/3) to (0.925, 0.052, 0.023). For this data set, the $P$s are most sensitive in the region $0.001 < \mu_0 < 0.1$. The prior mean $\mu_1$ for the external branch length is found to be much less important than is $\mu_0$ (results not shown). In Figure 4b, uniform priors are used for the branch lengths: $t_0 \sim U(0, \mu_0)$ and $t_1 \sim U(0, 1)$. The posterior probabilities for the three trees become more extreme when the upper bound $\mu_0$ increases. In Figure 3c, $t_1$ has an exponential prior with mean $\mu_1 = 0.1$, but $t_0$ has a gamma prior with mean $\mu_0$ and standard deviation $\sigma_0$. The contours represent $P_1$ as a function of $\mu_0$ and $\sigma_0$. The prior mean $\mu_0$ has much greater effect than the standard deviation $\sigma_0$ or variance of the gamma prior.

### *Distribution of Posterior Probabilities in Data Sets Simulated under the Star Phylogeny*

We examine how posterior probabilities $P_1$, $P_2$, $P_3$ change with the increase of the sample size $n$ when the data are simulated under a star phylogeny. The branch lengths are fixed at $t_0 = 0$ and $t_1 = 0.2$ in the simulation, which correspond to site-pattern probabilities $p_0 = 0.58700$ for the constant pattern $xxx$, and $p_1 = p_2 = p_3 = 0.13767$ for the three variable patterns (Equation 2). In the Bayesian analysis, we assumed

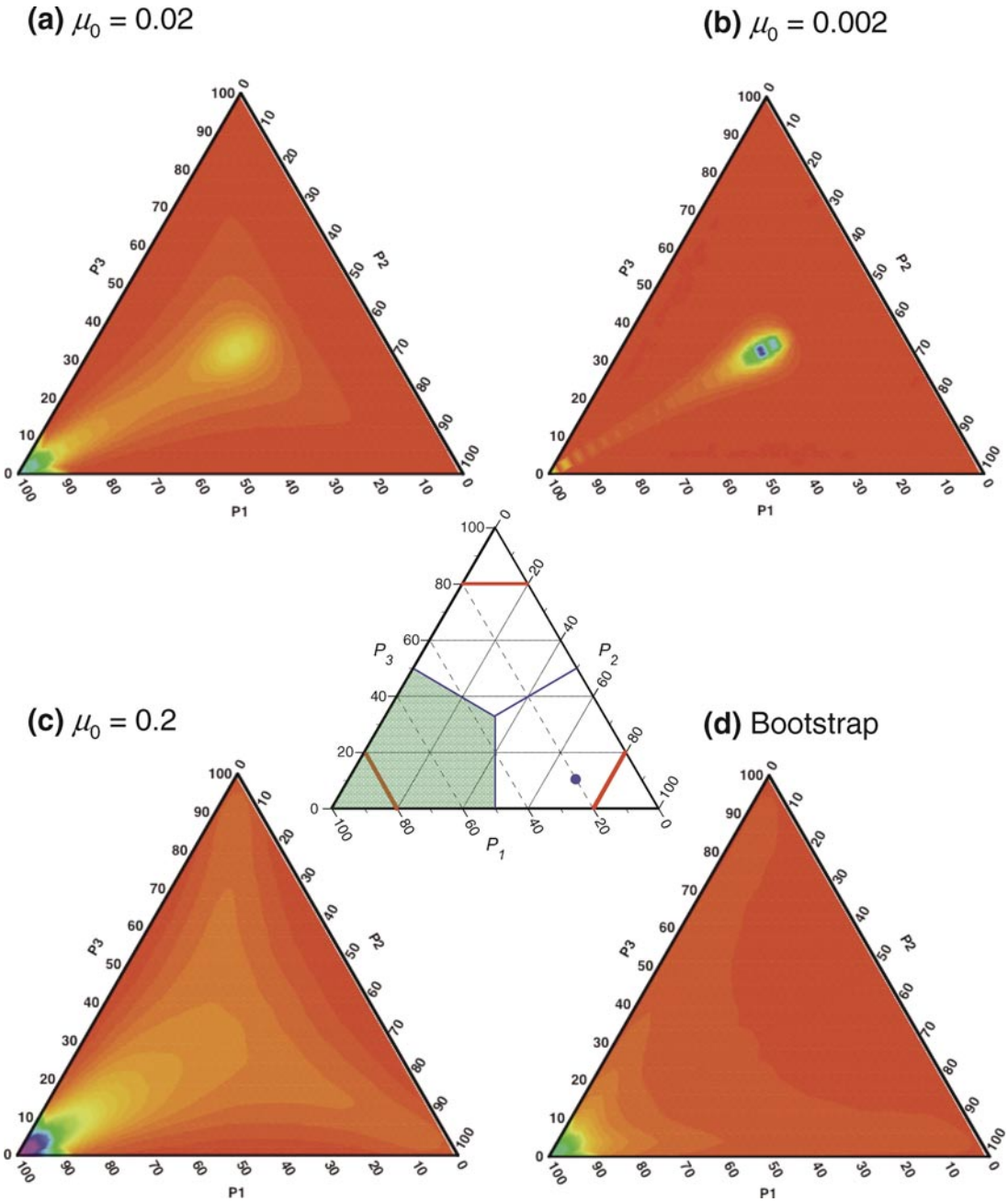


FIGURE 3. Estimated joint probability density, $f(P_1, P_2, P_3)$, of posterior probabilities for the three trees over simulated replicates. See legend to Figure 2 for details of simulation. The data are analyzed by the Bayesian method, assuming (a) the correct prior with $\mu_0 = 0.02$ or incorrect priors with (b) $\mu_0 = 0.002$ and (c) $\mu_0 = 0.2$, and by (d) maximum likelihood with bootstrap. The 2-D density for $P_1$ and $P_2$ (with $P_3 = 1 - P_1 - P_2$) is shown using the color contours, with red, yellow to blue, and purple representing low to high values. The density is estimated using an adaptive kernel smoothing algorithm (Silverman, 1986). Note that the total density mass on the triangle is 1. The inset illustrates the coordinate system of the ternary plot, commonly used to represent proportions of three components $P_1$, $P_2$, and $P_3$, which sum to 1. The point shown has coordinates (0.2, 0.7, 0.1), whereas the center point is (1/3, 1/3, 1/3). Note that the coordinates are represented by lines parallel to the sides of the triangle. The three blue line segments (given by $P_1 = P_2$, $P_2 = P_3$, and $P_3 = P_1$) partition the triangle into three regions, within which trees $T_1$, $T_2$, and $T_3$ is the Bayesian tree, respectively. For example, in the left region (shaded), $P_1 > \max(P_2, P_3)$. The three red line segments represent data sets in which one of $P_1$, $P_2$, $P_3$ is 0.8.

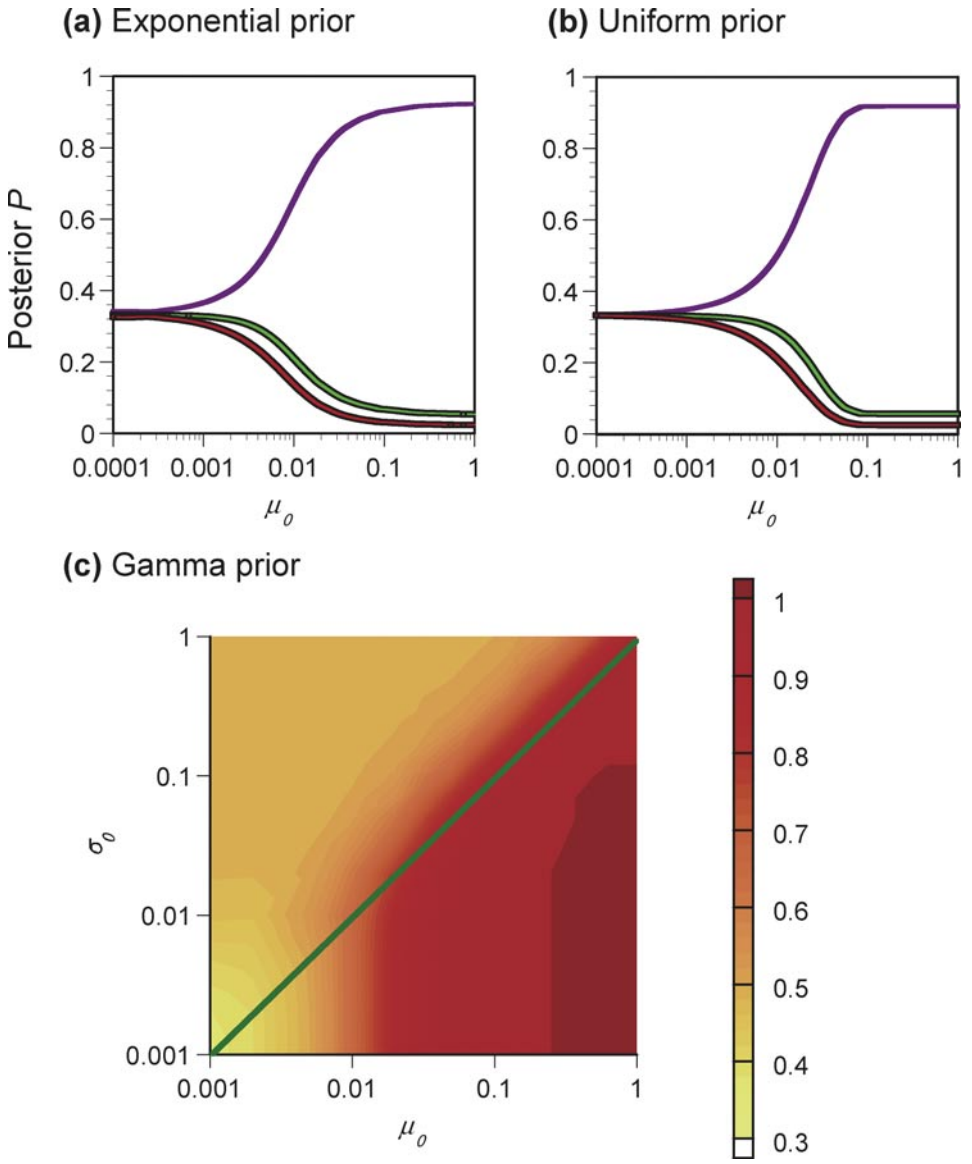


FIGURE 4. Posterior probabilities for the three trees ($P_1$, $P_2$, $P_3$ for the three curves from top to bottom) estimated using different priors for the internal branch length $t_0$. The data are $\mathbf{n} = \{n_0, n_1, n_2, n_3\} = \{300, 80, 65, 55\}$. (a) An exponential prior with mean $\mu_0$ is used for $t_0$, while $t_1$ is exponential with mean $\mu_1 = 0.1$ fixed. (b) A uniform prior $U(0, \mu_0)$ is used for $t_0$, while $t_1 \sim U(0, 1)$. (c) The prior for $t_1$ is exponential with mean $\mu_1 = 0.1$ fixed, and the prior for $t_0$ is gamma with mean $\mu_0$ and standard deviation $\sigma_0$. The exponential distribution is a special case of the gamma with $\mu_0 = \sigma_0$, so the straight line represents the slice examined in (a). The bootstrap proportions for the three trees are (0.887, 0.104, 0.009).

$\mu_0 = 0.1$ and $\mu_1 = 0.2$ in the exponential priors for $t_0$ and $t_1$. The results are shown in Figure 5. In small samples (e.g., $n = 20$), the probabilities are most often close to 1/3, reflecting the paucity of the data. When the sample size increases ($n = 200$ or 1000), however, the probabilities shift to the corners of the plot, with one of the three probabilities close to 1 and the other two close to 0. We encountered problems with numerical integration using Mathematica for large $n$, and it is unclear what the limiting distribution $f(P_1, P_2, P_3)$ is when $n \to \infty$. Note that for those data, the bootstrap proportions are more extreme than the Bayesian probabilities (Fig. 5d).

Similar simulations were conducted by Suzuki et al. (2002) using nucleotide-substitution models without the molecular clock to estimate unrooted trees for four species. The authors observed variable and occasionally very high posterior probabilities for the trees, similar to the pattern for $n = 1000$ in Figure 5. It is important to note that in the simulations of Figure 5 and of Suzuki et al. (2002), the data are generated using fixed branch lengths so that we are examining the frequentist sampling properties of the Bayesian method. Although it is reasonable to use frequentist criteria to evaluate a Bayesian method, there is no theory to guarantee its

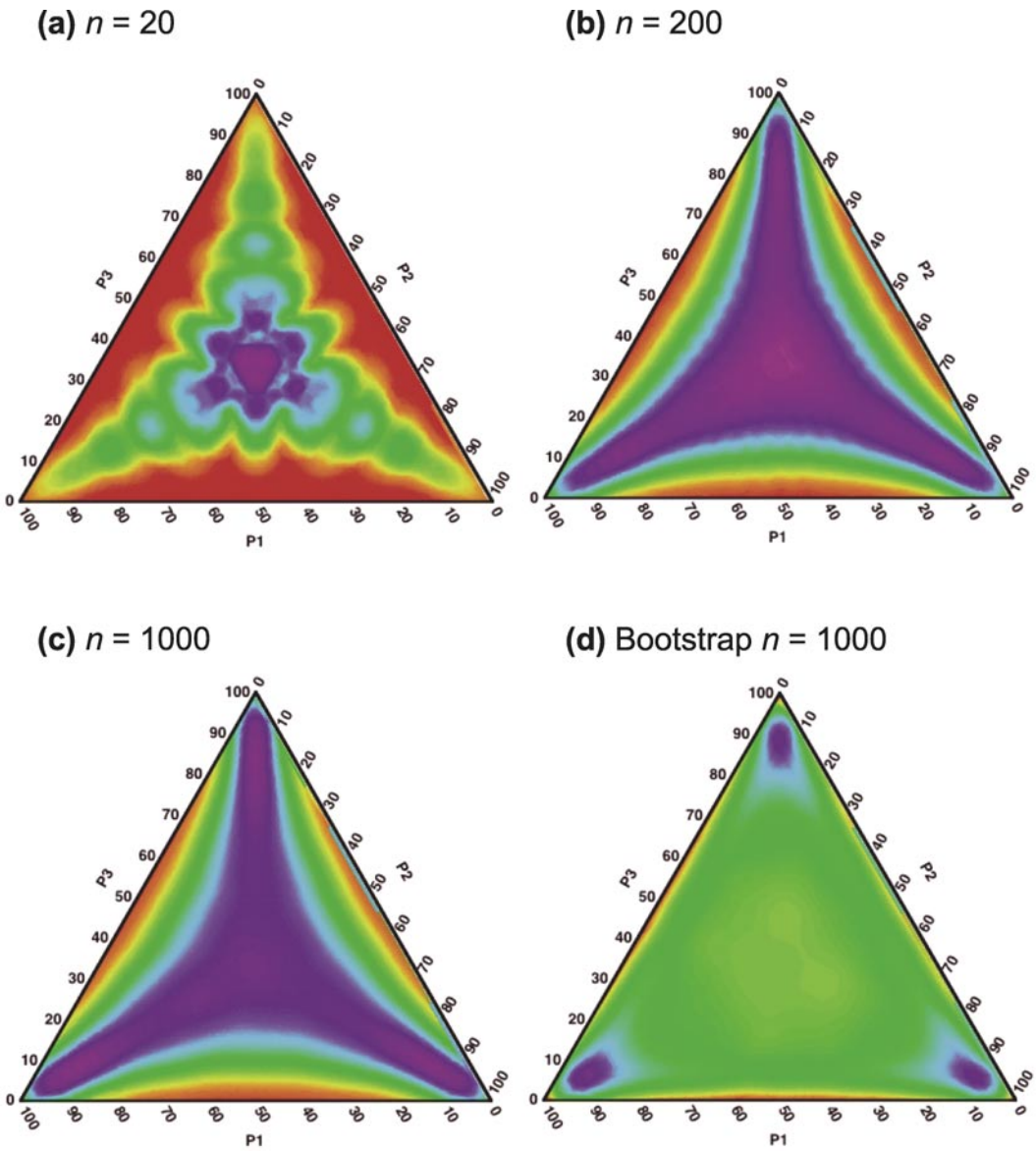


FIGURE 5. Estimated joint probability density, $f(P_1, P_2, P_3)$, of posterior probabilities for the three trees over replicate data sets simulated under the star phylogeny with $t_0 = 0$ and $t_1 = 0.2$. The data are analyzed using the Bayesian method, assuming $\mu_0 = 0.1$ and $\mu_1 = 0.2$ for the means of the exponential priors. Different sample sizes are used: (a) Bayesian probabilities with $n = 20$; (b) Bayesian probabilities with $n = 200$; (c) Bayesian probabilities with $n = 1000$; (d) Bootstrap proportions with $n = 1000$. The density, estimated using an adaptive kernel smoothing algorithm, is shown using the color contours, with red, yellow to blue, and purple representing low to high values. Note that the density is really discrete, especially with $n = 20$.

performance. Suzuki et al. (2002: 16139) incorrectly stated that "Bayesian . . . trees were judged as false-positives when the posterior . . . probability was >95%. . . . Note that the expected false-positive rate (type-I error) is 5% . . . because the confidence level is 95%." These authors have confused posterior probabilities with frequentist $P$-values. Nevertheless, Suzuki et al. (2002) argued that a good method should give about equal probability (1/3) for the three bifurcating trees when the star tree is used to simulate data and when the amount of data is large. In this study, we take this viewpoint as well, as did Lewis et al. (2005). The concern is that if the interior branches are short in the real world, the real data may appear similar to data sets generated under the star tree, and then the posterior probabilities will be highly variable among data sets, sometimes strongly supporting the true tree and other times strongly supporting wrong trees. Lewis et al. (2005) called the phenomenon a *star-tree paradox*.

*Fair-coin and fair-balance paradoxes*.—Lewis et al. (2005) drew an insightful parallel between Bayesian phylogeny reconstruction when the data are simulated under the star tree and a coin-tossing experiment. Suppose a coin is fair with the probability of heads to be exactly

$\theta = 1/2$, but we are required to compare two hypotheses that the coin is either negatively or positively biased: $H_1 : \theta < 1/2$ and $H_2$: $\theta > 1/2$. The truth $\theta = 1/2$ is considered impossible in the analysis. The data are the number of heads $x$ out of $n$ tosses of the coin, with the likelihood given by the binomial probability, $x \mid \theta \sim \text{bino}(n, \theta)$. Lewis et al. (2005) argued that one would like the posterior model probability $P_1 = \Pr(H_1 \mid x)$ to approach 1/2 when $n \to \infty$. Assuming a uniform prior $\theta \sim U(0, 1)$ and prior probability 1/2 for each model, the authors found that $P_1$ instead converged to a uniform distribution. They referred to the phenomenon as the *fair-coin paradox*. Note that the posterior is given by $\theta \mid x \sim \text{beta}(x + 1, n - x + 1)$, which converges to $N(y, y(1 - y)/n)$, where $y = x/n$, when $n \to \infty$. Let $\phi(\cdot)$, $\Phi(\cdot)$, and $\Phi^{-1}(\cdot)$ be the density function, the cumulative density function (CDF) and the inverse CDF (quantile) of the standard normal distribution. We have $P_1 = \Pr(\theta < {}^1/_2 \mid x) \approx \Phi((\frac{1}{2} - y)/\sqrt{y(1 - y)/n})$. Also $\mathrm{d}P_1/\mathrm{d}y = \phi(a) \times 2\sqrt{n} \times (1 + \frac{a^2}{n})^{3/2}$, where $a = \Phi^{-1}(P_1)$. Since $y \sim N(1/2, 1/(4n))$, $P_1$ has the density

$$f(P_1) = f(y(P_1)) \Big/ \left| \frac{\mathrm{d}P_1}{\mathrm{d}y} \right| = \sqrt{\frac{2n}{\pi}} \exp\left\{ -2n \times \frac{a^2}{4(n + a^2)} \right\} \Big/ \left[ \phi(a) \times 2\sqrt{n} \times \left( 1 + \frac{a^2}{n} \right)^{3/2} \right] \to 1, \quad (5)$$

when $n \to \infty$.

A simpler argument can be constructed using the normal distribution. Suppose $n$ measurement errors $x_1$, $x_2, \ldots, x_n$ are observed on a balance, which is fair (calibrated) so that the $x_i$ are independent draws from $N(0, \sigma^2)$ with mean $\theta_0 = 0$ and known variance $\sigma^2$. We are required to test two hypotheses that the balance has either negative or positive bias: $H_1$: $\theta < 0$ and $H_2$: $\theta > 0$. The truth $\theta = 0$ is not allowed in the analysis. We assume a normal prior $\theta \sim N(0, \tau^2)$, with larger $\tau^2$ representing more diffuse priors. Equivalently, $H_1$ and $H_2$ each has prior probability 1/2, and under each model the prior on $\theta$ is $N(0, \tau^2)$, truncated to the range $(-\infty, 0)$ under $H_1$ or $(0, \infty)$ under $H_2$. The likelihood is given by $\bar{x} \mid \theta \sim N(\theta, \sigma^2/n)$, since the sample mean $\bar{x}$ is a sufficient statistic. The posterior of $\theta$ given data $x = \{x_1, x_2, \ldots, x_n\}$ is given by $\theta \mid x \sim N(n\tau^2\bar{x}/(\sigma^2 + n\tau^2), \sigma^2\tau^2/(\sigma^2 + n\tau^2))$. Thus the posterior model probability is

$$P_1 = \Pr(H_1 \mid x) = \int_{-\infty}^{0} f(\theta \mid x)\mathrm{d}\theta = \Phi(z) = \Phi\left( -\frac{\sqrt{n}\bar{x}}{\sigma} \Big/ \sqrt{1 + \frac{\sigma^2}{n\tau^2}} \right), \quad (6)$$

When $n \to \infty$, $z \to -\sqrt{n}\bar{x}/\sigma$, which is a standard normal variable since the sample is from $N(0, \sigma^2)$ and $\bar{x} \sim N(0, \sigma^2/n)$. Thus, $\Phi(z)$ or $P_1 \sim U(0, 1)$ when $n \to \infty$ (see Ripley, 1987: 59). Note that using an increasingly diffuse prior (that is, letting $\tau^2 \to \infty$) has a similar effect as increasing the sample size.

To follow Suzuki et al. (2002) and Lewis et al. (2005), we would like a good method to give equal support for $H_1$ and $H_2$ when $n \to \infty$. The Bayesian method does not achieve this; instead, the posterior model probability $P_1$ converges to $U(0, 1)$. This may be termed *the fair-balance paradox*. When $n \to \infty$, the sample mean $\bar{x}$ will be closer and closer to $\theta_0 = 0$; in 99% of data sets, $\bar{x}$ will be in the narrow interval $(-2.58\sigma/\sqrt{n}, 2.58\sigma/\sqrt{n})$. Also, the confidence interval for $\theta$ from each data set will be narrower and narrower around the true value 0. However, when forced to decide whether $\theta < 0$ or $\theta > 0$, the posterior model probability varies widely among data sets, just like a random variable from $U(0, 1)$, sometimes strongly supporting one of the two hypotheses.

## Analysis of the Land Plant Data of Karol Et Al.

*Sequence alignment*.—To demonstrate the relevance of our analysis of the simple three-species case (Fig. 1) to real data sets used in molecular phylogenetics, we examined the impact of the prior for internal branch lengths on the posterior clade probabilities using the data set of Karol et al. (2001) concerning land plant divergences. The 40 species are identified in Figure 6; see appendix 2 in Karol et al. (2001) for GenBank accession numbers for the sequences. The alignment was retrieved from the *Science* Web site (http://www.sciencemag.org/cgi/content/full/294/5550/2351/DC1/1) and includes four genes concatenated as a supergene. The four genes are *atpB* and *rbcL* from the chloroplast, *nad5* from the mitochondria, and the small subunit rRNA gene (SSU rRNA) from the nuclear genome. We made a few minor corrections to the alignment of Karol et al., leaving 5141 sites in the sequence, compared with 5147 used by Karol et al. The alignment is available at the *Systematic Biology* web site, http://systematicbiology.org.

MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) was used to conduct the Bayesian analysis (see below for our modifications), with a Markov process model of nucleotide substitution used for likelihood calculation. Three of the four genes (*atpB*, *rbcL*, and *nad5*) are protein-coding, with huge differences among the three codon positions in the evolutionary dynamics, such as the evolutionary rate, the base compositions, the transition/transversion rate ratio, and the extent of rate variation among sites. Ideally, such heterogeneity should be taken into account in the analysis (Yang, 1996), and indeed MrBayes provides some models for combined analysis of such heterogeneous data. However, for the posterior probabilities calculated from our analysis to be directly comparable with those of Karol et al. (2001), we followed those authors and ignored the differences among codon positions. Thus, we used the HKY+G model, with five categories in the discrete

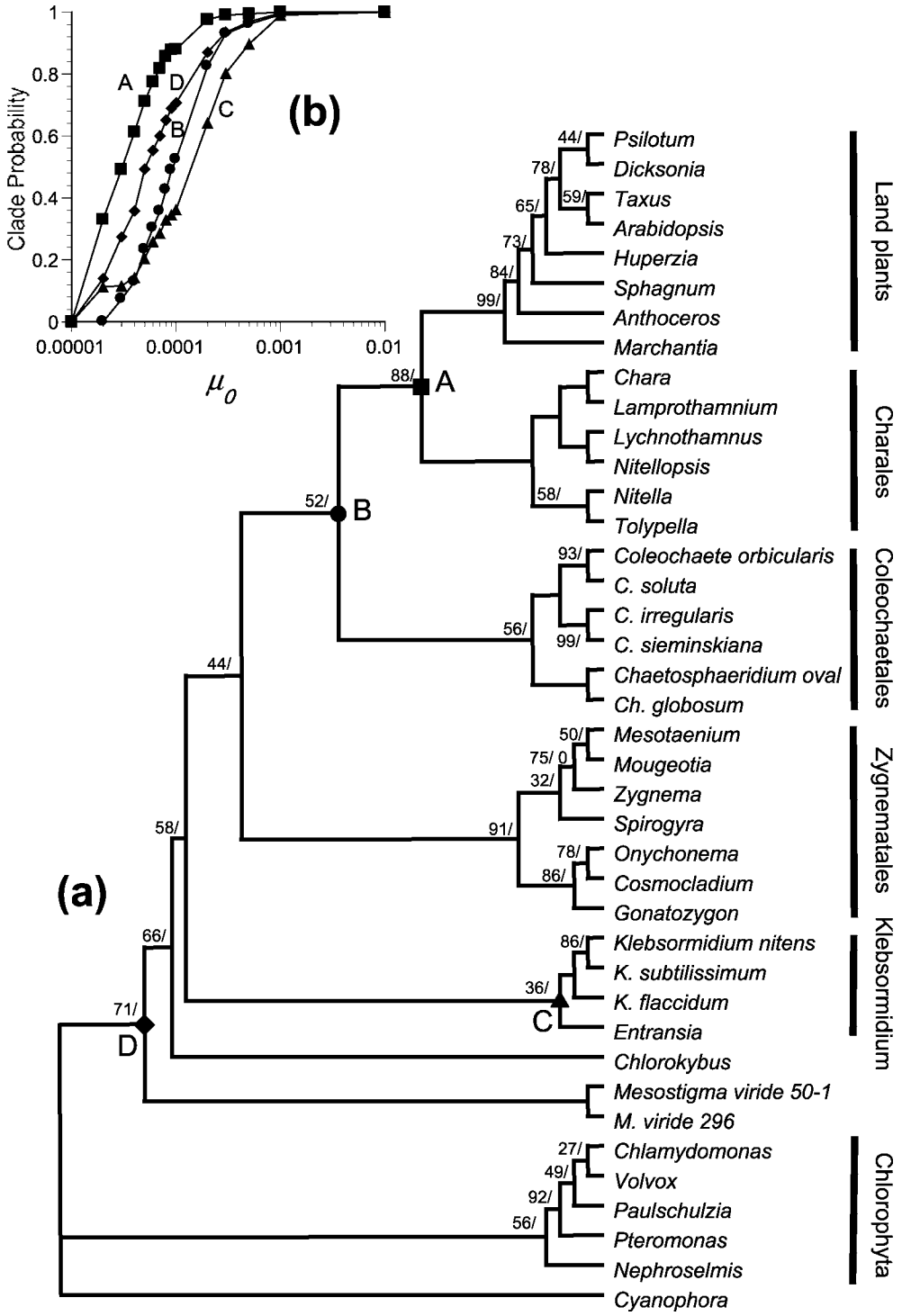


FIGURE 6. The phylogeny and clade probabilities for land plants and their relatives, estimated by Bayesian inference without assuming the molecular clock. Exponential priors with means $\mu_0$ and $\mu_1$ are assumed for internal and external branch lengths, respectively, with $\mu_1 = 0.1$ fixed. (a) Consensus tree published by (Karol et al., 2001), obtained when $\mu_0$ used is not too small (e.g., $>10^{-4}$). The two posterior probabilities ($\times 100$), at the two sides of "/" at each node, are for $\mu_0 = 0.0001/0.1$, with values equal to 100% omitted. (b) The posterior probabilities for four major clades (labeled A, B, C, and D in a) plotted against $\mu_0$. These are (A) the land plants-charales clade, (B) the clade of land plants, charales, and Coleochaetales, (C) the Klebsormidiales clade including Entransia, and (D) charophyta including Mesostigma. See Karol et al. (2001) for the phylogenetic significance of those clades.

gamma model of rates for sites (Yang, 1994a; Hasegawa et al., 1985). To reduce computation, parameters in the substitution model are fixed at their maximum likelihood estimates (MLEs) obtained from a few parsimony trees: $\hat{\kappa} = 3.0$ for the transition/transversion rate ratio; $\hat{\pi}_T = 0.307$, $\hat{\pi}_C = 0.188$, $\hat{\pi}_A = 0.272$, and $\hat{\pi}_G = 0.233$ for nucleotide frequencies; and $\hat{\alpha} = 0.29$ for the gamma shape parameter. Thus, the model accounts for the major features of the more complex model GTR+I+G, which Karol et al. used. Note that the gamma distribution with $\alpha \leq 1$ allows for virtually invariable sites with rates close to 0. We expect that fixing the substitution parameters to their MLEs will have little effect on tree reconstruction (Yang et al., 1995), as those parameters are reliably estimated from the large data set, with standard errors to be about 1% to 2% of the MLEs. The molecular clock is not assumed, and unrooted trees are considered.

*Modifications to MrBayes and MCMC analysis*.—The current version of MrBayes (version 3.0) (Ronquist and Huelsenbeck, 2003) assumes the same prior, either uniform or exponential, for all branch lengths on the unrooted tree. To use exponential priors with different means $\mu_0$ and $\mu_1$ for the internal versus external branch lengths, the source code of the program was modified by Z.Y. The change of the prior only affects the calculation of the prior ratio in MCMC moves that alter branch lengths, and does not affect other parts of the program, such as likelihood calculation or proposals that change other parameters. Thus, the source code was analyzed and appropriate changes were made. Our extensive tests indicated that the modifications were correct. We assume uniform priors for topologies and exponential priors with different means for the internal and external branch lengths. The prior mean of external branch lengths is fixed at $\mu_1 = 0.1$, whereas the prior mean of internal branch lengths $\mu_0$ is varied to see how it affects posterior clade probabilities. We used preliminary runs to determine reasonable settings to ensure convergence of the MCMC algorithm. Results reported in Figure 6 and below were obtained by running two simultaneous chains, using a burn-in of 20,000 iterations, followed by sampling every 10 iterations for a total of 2,000,000 iterations. Each analysis is conducted at least twice, using different random numbers, to confirm consistency between runs.

The original analysis of Karol et al. (2001) using MrBayes (Huelsenbeck and Ronquist, 2001) produced very high posterior probabilities for the inferred clades. The results obtained from applying the modified version of the program, assuming exponential priors with different means $\mu_0$ and $\mu_1$ for the internal and external branch lengths, are summarized in Figure 6. When $\mu_0$ is in the range (0.00004, 0.001), we recover the same phylogeny as reported by Karol et al. (2001), with larger $\mu_0$ producing higher probabilities for the clades. When $\mu_0 \geq 0.01$, posterior probabilities for all but one node on the tree of Karol et al. were calculated to be 100%, but surprisingly one node was not found in the sampled trees (Fig. 6a). Posterior clade probabilities calculated using $\mu_0 = 0.0001$ and 0.1 are listed on the tree of Figure 6a, whereas Figure 6b shows the posterior probabilities for four important clades on the phylogeny as functions of $\mu_0$. Posterior probabilities for other nodes show similar changes with the change of $\mu_0$ (results not shown).

## DISCUSSION

### *Bayesian Posterior Probabilities Versus Bootstrap Proportions*

Bayesian posterior probabilities are conceptually straightforward to interpret. The posterior probability for a tree or clade is the probability that the tree or clade is true, given the data and the model (including the prior and the likelihood model). Using the simple case of phylogeny reconstruction for three species, we illustrated this interpretation (Fig. 2a). In contrast, the bootstrap proportion has been much harder to interpret. At least three interpretations have been offered in the literature (see, e.g., Berry and Gascuel, 1996). The first is that it means *repeatability*. A clade with bootstrap proportion $P$ in the original data set is expected to be in the estimated tree with probability $P$ if many new data sets are generated from the same data-generating process and if the same tree reconstruction method is used to analyze the data sets (Felsenstein, 1985). The rationale for resampling the original data by bootstrap is that the distribution of the bootstrap samples around the observed data set is a good approximation of the unknown distribution of observed data from the data-generating process (Efron, 1979; Efron et al., 1996). The simulations of Hillis and Bull (1993) suggest that the bootstrap proportion varies so much among replicate data sets that it is useless as a measure of repeatability. A second interpretation is the frequentist *type-I error rate*, using the star tree as the null hypothesis (Felsenstein and Kishino, 1993) or a confidence interval (Felsenstein and Kishino, 1993; Zharkikh and Li, 1995). If we generate many data samples under the star tree in which the concerned clade (with bootstrap proportion $P$ from the original data set) is absent, then the clade will be in the estimated tree with probability $< 1 - P$. Efron et al. (1996) argued that this interpretation is only approximate, and suggested a more complex, two-step bootstrap procedure for transforming bootstrap proportions into standard frequentist confidence intervals. A third interpretation is phylogenetic *accuracy*: a clade with bootstrap proportion $P$ is in the true tree with probability $P$. This interpretation equates bootstrap proportion with Bayesian posterior probability and appears to be the one that most empirical phylogeneticists use or would like to use (e.g., Hillis and Bull, 1993; Murphy et al., 2001). All studies comparing the two approaches appear to be using this interpretation, as otherwise the two measures are incomparable.

The fact that the posterior probabilities change drastically with the prior for internal branch lengths (e.g., Fig. 4) suggests that the posterior probability and bootstrap proportion are two fundamentally different measures of phylogenetic uncertainty. This result appears to contradict previous claims that the two should be theoretically close (Efron et al., 1996; Newton, 1996),

and to agree with recent simulations demonstrating their differences (Suzuki et al., 2002; Alfaro et al., 2003; Cummings et al., 2003; Douady et al., 2003; Erixon et al., 2003). We observe that bootstrap proportions are more similar to posterior probabilities under some priors than under others. For example, for the data set of Figure 4a, the bootstrap proportions are close to the posterior probabilities under the exponential prior with mean $\mu_0 = 0.06$ but are more or less extreme than the posterior probabilities when $\mu_0 <$ or $>0.06$, respectively. In the data sets of Figure 2 simulated under the prior, the bootstrap proportions are on average comparable to posterior probabilities under $\mu_0 = 0.2$ but are more different from the posterior probabilities under $\mu_0 = 0.02$ or 0.002. Thus, the bootstrap will be too conservative or too liberal, that is, the bootstrap proportions will be too moderate or too extreme relative to the posterior probabilities under the correct prior, depending on whether the prior mean $\mu_0$ used to generate the replicate data sets is very small or very large. It is clear that the bootstrap proportion, if interpreted as the probability that the clade is correct, is not always conservative, as suggested previously (Hillis and Bull, 1993).

We asked the question whether certain priors can produce posterior probabilities that are close to the bootstrap proportions in all replicate data sets. For the data sets of Figures 2 and 3, we used an iterative algorithm to adjust the means $\mu_0$ and $\mu_1$ in the exponential priors to minimize the difference between the posterior probabilities and the bootstrap proportions in each data set. We found that $\mu_0$ and $\mu_1$ "estimated" in this way vary considerably among data sets, suggesting that it is in general impossible for bootstrap proportions to match posterior probabilities under fixed priors. Similarly, we observed in the simple case of three species that the effect of overall sequence divergence is quite different on the two measures. For example, adding constant sites ($n_0$) to the data tends to polarize the posterior probabilities while having little impact on bootstrap proportions. For example, Table 1 lists posterior probabilities calculated for the data set of Figure 4, but with different numbers of constant sites added. The bootstrap proportions are (0.887, 0.104, 0.009), estimated using 100,000 bootstrap pseudosamples. At this level of accuracy, we cannot detect any difference in bootstrap proportions among the data sets. The lack of effect of $n_0$ on the bootstrap is understandable in this case, because in each data set, the maximum likelihood tree is determined by the counts of sites for the three variable patterns and largely independent of $n_0$. In contrast, the posterior probabilities become more extreme when constant sites are added to the data. Intuitively, adding constant sites reduces the overall sequence divergence, and as a result, every observed change becomes less likely and is counted as stronger evidence in calculation of posterior probabilities. Those results also suggest that the posterior probability might be more sensitive to the substitution model, especially concerning rate variation among sites, than bootstrap proportions.

TABLE 1. Posterior probabilities in data sets with different numbers of constant sites. The bootstrap proportions are (0.887, 0.104, 0.009), for all the four data sets. The posterior probabilities are calculated under the exponential priors with means $\mu_0 = 0.02$ and $\mu_1 = 0.2$. The second data set is analyzed in Figure 4.

| Data ($n_0$, $n_1$, $n_2$, $n_3$) | Posterior probabilities ($P_1$, $P_2$,$P_3$) |
|---|---|
| D1: (100, 80, 65, 55) | (0.446, 0.303, 0.251) |
| D2: (300, 80, 65, 55) | (0.786, 0.136, 0.078) |
| D3: (500, 80, 65, 55) | (0.852, 0.098, 0.051) |
| D4: (1000, 80, 65, 55) | (0.890, 0.074, 0.036) |

### *Factors Inflating Posterior Probabilities for Trees*

Bayesian posterior probability for a tree or clade is the probability that the tree or clade is true given the data and model (prior and substitution model). Thus, there can be only three possible reasons for spuriously high clade probabilities: (i) program errors and computational problems, (ii) misspecification of the likelihood (substitution) model, and (iii) misspecification and sensitivity of the prior. Lack of convergence and poor mixing in the MCMC algorithm can cause the chain to stay in an artificially small subset of the parameter space, leading to spuriously high support for the trees visited in the chain. This problem may be a serious concern in Bayesian analysis of large data sets, but in principle may be resolved by running longer chains and designing more efficient algorithms. Model misspecification, that is, use of an overly simple substitution model, is also known to cause spuriously high posterior probabilities (Buckley, 2002; Huelsenbeck and Rannala, 2004; Lemmon and Moriarty, 2004; Suzuki et al., 2002). The problem can in theory be resolved by implementing more-realistic substitution models or taking a model-averaging approach (Huelsenbeck et al., 2004). In this study, we examined the effect of the prior on internal branch lengths and demonstrated that the posterior probabilities are sensitive to the prior specification. We note that high posterior probabilities were observed in simulated data sets where the substitution model is correct (this study) and in analyses that did not use MCMC algorithms (Rannala and Yang, 1996). In those cases, the first two factors do not apply. The sensitivity of Bayesian inference to prior specification is more fundamental and difficult to deal with (see below). The uniform prior with a large upper bound such as 10 or 100 is often advocated as a "non-informative" or "diffuse" prior for branch lengths. However, such a prior causes inflated clade probabilities and is one of the worst in this regard. Exponential priors with small means appear preferable.

### *The Effect of Prior on Bayesian Model Comparison and Phylogeny Estimation*

Phylogeny reconstruction can be viewed as a problem of model selection rather than parameter estimation (Yang et al., 1995). Different trees have different likelihood functions with different branch length parameters, and are equivalent to non-nested models. In contrast to Bayesian parameter estimation under a well-specified model, where the posterior will be

dominated by the likelihood when more and more data are available, Bayesian hypothesis testing or model selection is a difficult area, and weak prior information for model parameters is known to cause problems (e.g., Bernardo, 1980; DeGroot, 1982; Berger, 1985: 144–157). Below, we briefly review the literature on Bayesian model selection in presence of weak prior information, partly because phylogeny estimation appears to be affected by similar difficulties but mainly because some of the suggested remedies appear useful to phylogeny estimation.

An extreme well-known case is Lindley's paradox, in which Bayesian analysis and traditional hypothesis testing approaches reach drastically different conclusions (Lindley, 1957; see also Jeffreys, 1939). Consider test of a simple null hypothesis $H_0$: $\theta = 0$ against the composite alternative hypothesis $H_1$: $\theta \neq 0$ using a random sample $x_1, x_2, \ldots, x_n$ from $N(\theta, \sigma^2)$ with $\sigma^2$ known. The usual test is based on the sufficient statistic $\bar{x}$ having a normal distribution $N(0, \sigma^2/n)$ under $H_0$ and calculates the $P$-value as $\Phi(-\sqrt{n} \mid \bar{x} \mid /\sigma)$. In the Bayesian analysis, suppose the prior is $\Pr(H_0) = \Pr(H_1) = 1/2$, and $\theta \sim N(0, \tau^2)$ under $H_1$. The likelihood is given by $\bar{x} \sim N(0, \sigma^2/n)$ under $H_0$ and by $\bar{x} \mid \theta \sim N(\theta, \sigma^2/n)$ under $H_1$. Then the ratio of posterior model probabilities, which is also the Bayes factor since the prior model probability is uniform, is equal to the ratio of the marginal likelihoods

$$\begin{aligned} B &= \frac{\Pr(H_0 \mid x)}{\Pr(H_1 \mid x)} \\ &= \frac{\frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left\{-\frac{n}{2\sigma^2}\bar{x}^2\right\}}{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{-\frac{\theta^2}{2\tau^2}\right\} \times \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left\{-\frac{n}{2\sigma^2}(\bar{x}-\theta)^2\right\} \mathrm{d}\theta} \\ &= \sqrt{1 + \frac{n\tau^2}{\sigma^2}} \times \exp\left\{-\frac{n\bar{x}^2}{2\sigma^2\left(1 + \frac{\sigma^2}{n\tau^2}\right)}\right\} \end{aligned} \quad (7)$$

Now suppose that $\sqrt{n} \mid \bar{x} \mid /\sigma = z_{\alpha/2}$, so that we reject $H_0$ at the significance level $\alpha$, but as $n \to \infty$, we see that $B \to \infty$ and $\Pr(H_0 \mid x) \to 1$. Hence the paradox: while the significance test rejects $H_0$ decisively at $\alpha = 10^{-10}$, say, the Bayesian method strongly supports $H_0$ with posterior model probability $\Pr(H_0 \mid x)$ approaching 1. We can also fix the sample size $n$ but increase $\tau^2$, making the prior more and more diffuse; again $\Pr(H_0 \mid x) \to 1$ as $\tau^2 \to \infty$. In both cases, the prior distribution under $H_1$ becomes more and more spread out relative to the likelihood, which is concentrated in a small region close to but different from 0.

We note that Lindley's paradox is controversial, even among Bayesian statisticians. Some view it as revealing logical flaws in traditional hypothesis testing (e.g., Good, 1982: 342; Berger, 1985: 144–157; Press, 2003: 220–225), whereas others consider the Bayesian approach to be misleading and suggested fixes (e.g., Bernardo, 1980; Shafer, 1982). However, all appear to agree that the extreme sensitivity of the posterior model probabilities to the prior means that an objective Bayesian analysis is impossible. As remarked by O'Hagan and Forster (2004: 78), "Lindley's paradox arises in a fundamental way whenever we wish to compare different models for the data, and where we wish to express weak prior information about parameters in one or more of the models." For such difficulties to arise, the compared models can have one or more parameters, or one model can be sharp (with no parameters), and the prior can be proper and informative as increasing the size of data while keeping the prior fixed has the same effect. To appreciate the generality of the problem, we contrast Bayesian parameter estimation with model selection using uniform priors for parameters. First, consider estimation of parameter $\theta$ in a well-specified model, with the prior $f(\theta) = 1/(2c)$, $-c < \theta < c$. The posterior is

$$f(\theta \mid x) = \frac{f(\theta) f(x \mid \theta)}{\int_{-c}^{c} f(\theta) f(x \mid \theta)\, \mathrm{d}\theta} = \frac{f(x \mid \theta)}{\int_{-c}^{c} f(x \mid \theta)\, \mathrm{d}\theta}. \quad (8)$$

When the sample is large, the likelihood $f(x \mid \theta)$ is concentrated in a small region (inside the prior interval as long as the prior is diffuse enough to contain the true $\theta$), outside which $f(x \mid \theta)$ is vanishingly small. Then the integral in the denominator is insensitive to $c$, and so is the posterior. In contrast, consider comparison between two models $H_1$ involving parameter $\theta_1$ with prior $f_1(\theta_1) = 1/(2c_1)$, $-c_1 < \theta < c_1$, and $H_2$ involving parameter $\theta_2$ with prior $f_2(\theta_2) = 1/(2c_2)$, $-c_2 < \theta < c_2$. The Bayes factor is

$$\begin{aligned} \frac{\Pr(H_1 \mid x)}{\Pr(H_2 \mid x)} &= \frac{\int_{-c_1}^{c_1} f_1(\theta_1) f_1(x \mid \theta_1) \mathrm{d}\theta_1}{\int_{-c_2}^{c_2} f_2(\theta_2) f_2(x \mid \theta_2) \mathrm{d}\theta_2} \\ &= \frac{c_2}{c_1} \times \frac{\int_{-c_1}^{c_1} f_1(x \mid \theta_1)\, \mathrm{d}\theta_1}{\int_{-c_2}^{c_2} f_2(x \mid \theta_2) \mathrm{d}\theta_2}. \end{aligned} \quad (9)$$

When the data are informative and the likelihood $f_i(x \mid \theta_i)$ under model $i$, $i = 1, 2$, is highly concentrated, the two integrals are more or less independent of $c_i$. However, the Bayes factor or posterior model probability depends on $c_2/c_1$, and that sensitivity will not disappear with the increase of data. The difficulty is that when the prior information is weak, one may not be able to decide whether $U(-10, 10)$ or $U(-100, 100)$ is a more appropriate prior, even though the Bayes factor differs by 10-fold between the two.

We note some differences between Lindley's paradox and the phylogeny estimation problem. First, in tree estimation, the mean of the prior for internal branch lengths is important, whereas in Lindley's paradox, it is the variance. In both cases, increasing the sample size has the same effect of exacerbating the problem. Second, if we view tree estimation as a problem of hypothesis testing, with the binary trees being the alternative hypothesis (hypotheses) and the star tree being the null hypothesis, the pattern in phylogeny estimation is opposite to

that in Lindley's paradox. In the former, posterior probabilities are high for the binary tree, which is considered the "alternative" hypothesis, whereas in the latter, the effect of increasing amounts of data or a progressively diffuse prior is a strong support for the null hypothesis. Lewis et al. (2005) argued that the phylogeny problem, especially the star-tree paradox, is more similar to the fair-coin paradox they constructed (or the equivalent fair-balance paradox discussed in this study). However, this formulation has difficulties as well. First, a simple hypothesis test considers one alternative hypothesis, but we have many binary trees and such composite hypothesis tests can have complex properties. Second, a conventional hypothesis test makes assumptions about parameters in a general model, whereas different trees are equivalent to different models with different parameter spaces. Third, rejection of the star tree is not an appropriate measure of the statistical support for the ML/Bayes tree; one can construct data sets in which all three binary trees are significantly better than the star tree but it is ridiculous to claim that all three binary trees are significantly supported by the same data (Tateno et al., 1994; Yang, 1994b). Here we consider Lindley's paradox, the fair-coin or fair-balance paradoxes, and phylogeny estimation as three distinct manifestations of the deeper problem of sensitivity of Bayesian model selection to the prior for model parameters.

If we use an extreme prior $\mu_0 = 0$, all binary trees will have the same small probability. Thus, high clade probabilities in any data set can be made small by assuming a very small $\mu_0$ in the analysis model, and therefore there must always be a region of $\mu_0$ over which the posterior probabilities are sensitive to changes in the prior. However, in large data sets, this sensitive region may include only very small values of $\mu_0$. In our analysis of the land plant data set, the sensitive region is $(10^{-5}, 10^{-3})$ (Fig. 6b). Such values may seem unrealistically small if we consider estimated internal branch lengths in published trees. The question arises as to whether the prior for the internal branch lengths is relevant for the high posterior probabilities reported in many real data sets. We suggest that the answer is "Yes." In Bayesian model comparison, parameters in different models with different definitions are usually assigned different priors. In phylogeny reconstruction, branch lengths in different phylogenies have different biological meanings, and one can envisage assigning different priors for them. For example, a biologist's information or belief about the internal branch length in the tree ((human, chimpanzee), gorilla) may well be different than about the internal branch length in the tree (human, (chimpanzee, gorilla)), in each case the respective tree topology being assumed to be true. The branch in the latter tree may be expected to be shorter if the tree is considered less likely to be true than the former tree. Estimates of internal branch lengths in wrong or poor trees are typically small and often 0. If we specify the prior to represent our prior knowledge of branch lengths in all binary trees, the majority of which are wrong or poor trees, a very small $\mu_0$ is necessary. This argument also suggests that $\mu_0$ should be smaller in larger trees with more species.

*Possible Remedies to Deal with the Sensitivity to the Prior*

In a traditional parameter estimation problem, two approaches can be taken when the posterior is sensitive to parameters in the prior such as $\mu_0$. The first is the hierarchical or full Bayesian approach, which assigns a hyperprior for $\mu_0$ and integrates $\mu_0$ out in the MCMC algorithm. Suchard et al. (2001) implemented such an approach. Adding a hyperprior to $\mu_0$ is equivalent to specifying a different prior for $t_0$, in the same way that the gamma prior considered in Figure 4 is an extension of the exponential prior. From our results (Fig. 4), the mean of the prior for $t_0$ appears more important than the variance.

The second approach to dealing with the prior parameter $\mu_0$ is the empirical Bayes approach, which estimates $\mu_0$ from the data and uses the estimate to calculate posterior probabilities. We implemented this approach for the three-species case (Fig. 1). We estimate $\mu_1$ as the single branch length in the star phylogeny (i.e., with $t_0 = 0$): $\hat{\mu}_1 = -\log\{(4n_0/n - 1)/3\}/4$, as an overall measure of sequence divergence. We estimate the prior mean $\mu_0$ by maximizing the marginal likelihood: $L(\mu_0, \hat{\mu}_1) = f(\mathbf{n} \mid \mu_0, \hat{\mu}_1)$ of Equation 4. The estimates $\hat{\mu}_0$ and $\hat{\mu}_1$ are then used to calculate $P_i$'s in Equation 3. Application of this approach to the data analyzed in Figure 4 gives parameter estimates $\hat{\mu}_1 = 0.19054$ and $\hat{\mu}_0 = 0.02746$, with the marginal log likelihood $\ell = -558.3644$. The posterior probabilities are (0.8278, 0.1120, 0.0602) (c.f. Fig. 4a). We note that use of the star tree to estimate $\mu_1$ leads to overestimates of $\mu_1$ and underestimates of $\mu_0$, and may be problematic in large trees. However, use of the marginal likelihood function to estimate $\mu_0$ means that the estimate will be dominated by the ML tree, which typically has longer internal branches than poor trees. An alternative strategy is to estimate parameters such as $\mu_0$ and $\mu_1$ from the data for each possible tree topology, and then choose values representative of the collection of estimates among the trees, for use in Bayesian calculation. This strategy is computationally demanding because of the great number of trees, but will produce very small estimates of $\mu_0$ in real data analysis since most trees are wrong trees with small or zero internal branch lengths. As far as we are aware, the empirical Bayes approach has been used only in estimation of parameters in a well-specified model and not in dealing with sensitivity of Bayesian model comparison to the prior.

We note that Bayesian model comparison is an extremely active research area, with much controversy. A number of modifications have been introduced to deal with the sensitivity of Bayes factors to the prior on model parameters, resulting in a plethora of Bayes factors: such as *intrinsic*, *partial*, *fractional*, and *pseudo*-Bayes factors (see, e.g., O'Hagan and Forster, 2004 pp. 183–191). In discussions of Lindley's paradox, several possible remedies were suggested in the literature. We now discuss their potential use in the phylogeny problem. All such remedies

are to some degree subjective and none are generally accepted. (a) Blame the question (e.g., Hill, 1982). It has been suggested that one cannot reasonably expect a parameter to take a fixed value $\theta = 0$; instead, one should consider the null hypothesis that $\theta$ lies within a narrow interval $(-\delta, \delta)$. (b) Blame the data and avoid using large data sets (Bartlett, 1957). Those two options are unlikely to be relevant or appealing to molecular phylogeneticists. (c) Assume that both hypotheses are wrong and add a pinch of probability $\varepsilon$ for errors, i.e., for possibilities unaccounted for in the model (DeGroot, 1982; Hill, 1982; see also Jeffreys, 1961: 129). How to determine $\varepsilon$ is subjective. In the phylogeny problem, one of the binary trees should be correct, so that this idea does not appear logically sound. Factors such as lineage sorting or horizontal gene transfers may cause different genes to have different tree topologies, but this should best be dealt with by allowing data partitions to have different phylogenies, rather than by considering the composite phylogeny to be a star phylogeny. Nevertheless, one might assume a prior probability $\varepsilon$ for the star tree, or, equivalently, assume a point mass at $t_0 = 0$ in the prior for internal branch lengths $t_0$, with the rest of the density coming from a distribution. This should have an effect similar to that achieved by assuming a small $\mu_0$ in the exponential prior and will reduce high posterior probabilities for trees. Lewis et al. (2005) has implemented this strategy, using a reversible-jump MCMC algorithm to move between trees with different numbers of branch lengths. (d) Use data or at least the size of data to specify the prior (Bernardo, 1980; Davison, 2003). As the prior is supposed to reflect information or beliefs before the data are gathered, this idea is outrageous to many Bayesian statisticians but considered useful by others. In the case of Lindley's paradox, one suggestion is to let the variance $\tau^2$ be proportional to $1/n$, that is, $\theta \sim N(0, c\sigma^2/n)$, so that increasingly informative priors are used for $\theta$ under $H_1$ in larger data sets. Values of $c$ in the range 5 to 20 appear to produce results comparable to the traditional significance test (Davison, 2003: 586–587).

A similar strategy can be applied to the fair-balance problem. We can let the prior become increasingly informative as $n$ increases by specifying the prior variance $\tau^2 = c\sigma^2/n^k$. The posterior model probability is then (cf: Equation 6)

$$P_1 = \Pr(H_1 \mid x) = \Phi\left(-\frac{\sqrt{n}\bar{x}}{\sigma} \Big/ \sqrt{1 + n^{k-1}/c}\right). \quad (10)$$

Note that if $z \sim N(0, 1)$, then $y = \Phi(az + b)$, where $a$ and $b$ are constants, has the density

$$f(y) = f(z(y)) \times \left|\frac{dz}{dy}\right| = \frac{1}{|a|}\exp\left\{\frac{1}{2}[\Phi^{-1}(y)]^2 - \frac{1}{2a^2}[\Phi^{-1}(y) - b]^2\right\}, \quad (11)$$

As $-\sqrt{n}\bar{x}/\sigma \sim N(0, 1)$ the density of the posterior model probability $P_1$ becomes

$$f(P_1) = \sqrt{1 + \frac{n^{k-1}}{c}} \times \exp\left\{-\frac{n^{k-1}}{2c}[\Phi^{-1}(P_1)]^2\right\}. \quad (12)$$

If $k = 1$, the prior variance becomes $\tau^2 = c\sigma^2/n$, which decreases at the rate $1/n$. The density of $P_1$ then peaks at $1/2$, so that $P_1$ is more likely to be around $1/2$ than close to 0 or 1. However, the density is independent of $n$, and will not become more concentrated around $1/2$ with the increase of $n$. When $k > 1$, the distribution converges to the point mass $P_1 = 1/2$ when $n \to \infty$, as we wanted, and at a faster rate for larger $k$.

To choose an appropriate $k$, we would also want $f(P_1)$ to converge to the point mass at 1 (or 0) at a reasonable rate when $n \to \infty$, if the true $\theta < 0$ and $H_1$ is the true model (or if $\theta > 0$ and $H_2$ is the true model). From Equation 11, the density of $P_1$ if the true parameter value is $\theta_0$ is given as

$$f(P_1|\theta_0) = \sqrt{1 + n^{k-1}/c} \times \exp\left\{\frac{1}{2}[\Phi^{-1}(P_1)]^2 - \frac{1}{2}(1 + n^{k-1}/c)\left[\Phi^{-1}(P_1) + \sqrt{n/(1 + n^{k-1}/c)} \times \frac{\theta_0}{\sigma}\right]^2\right\}. \quad (13)$$

It is easy to see that when $n \to \infty$, $f(P_1 \mid \theta_0)$ degenerates to a point mass at 1 (or 0) if $\theta_0 < 0$ (or if $\theta_0 > 0$)

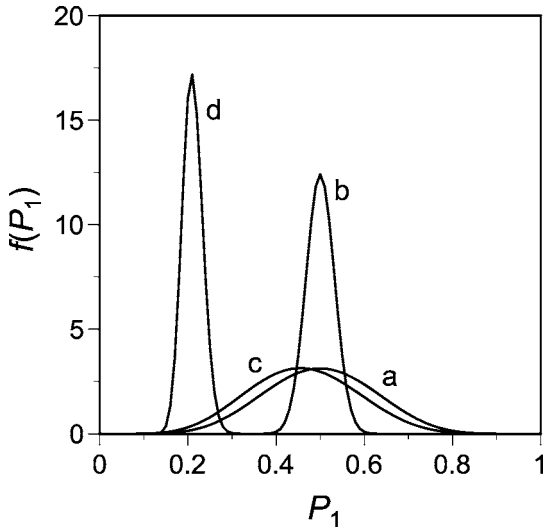


FIGURE 7. The probability density function of the posterior model probability $P_1 = \Pr(H_1 \mid x)$ in the fair-balance problem assuming the prior $\theta \sim N(0, c\sigma^2/n^k)$, with $c = 2$ and $k = \sqrt{2}$. In curves (a) and (b), the true parameter value is $\theta_0 = 0$ and the density is given by Equation (12) while in curves (c) and (d), $\theta_0 = 0.01\sigma$ and the density is given by Equation 13. The sample size is $n = 1000$ for (a) and (c) and $n =$ 1,000,000 for (b) and (d).

irrespective of $k$. However, the convergence is faster if $k$ is smaller. To avoid the fair-balance paradox when $\theta_0 = 0$ and to achieve a fast convergence when $\theta_0 \neq 0$, $k$ should be greater than 1 but not too much greater. Figure 7 plots the densities when $k = \sqrt{2}$ for $\theta_0 = 0$ and $0.01\sigma$ and for two sample sizes $n = 1{,}000$ and 1,000,000, with $c = 2$ fixed.

For the phylogeny problem, we note that the objectives of both strategies (c) and (d), discussed above, can be achieved by applying a small mean in the prior for the internal branch length. The prior mean should be increasingly smaller for longer sequences (greater $n$) and larger trees, and should also reflect the overall information content (e.g., as indicated by overall sequence divergences), in the same way that $\sigma^2$ is used in the priors discussed above. Incorporating those factors in the prior appears hard, and merits further research.

## ACKNOWLEDGMENTS

We thank Paul Lewis, Fredrik Ronquist, and Marc Suchard for many constructive criticisms, and Paul Lewis for making the Lewis, Holder, and Holsinger paper available to us before its publication. This study is supported by a grant from the Biotechnological and Biological Sciences Research Council (UK) to Z.Y. and National Institutes of Health grant HG01988 to B.R.

## REFERENCES

Alfaro, M. E., S. Zoller, and F. Lutzoni. 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol. Biol. Evol. 20:255–266.

Bartlett, M. S. 1957. A comment on D.V. Lindley's paradox. Biometrika 44:533–534.

Berger, J. O. 1985. Statistical decision theory and Bayesian analysis, 2nd edition. Springer-Verlag, New York.

Bernardo, J. M. 1980. A Bayesian analysis of classical hypothesis testing. Pages 605–647 *in* Bayesian statistics (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds.). Valencian University Press, Valencia, Spain.

Berry, V., and O. Gascuel. 1996. On the interpretation of Bootstrap trees: Appropriate threshold of clade selection and induced gain. Mol. Biol. Evol. 13:999–1011.

Buckley, T. R. 2002. Model misspecification and probabilistic tests of topology: Evidence from empirical data sets. Syst. Biol. 51:509–523.

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.

Davison, A. C. 2003. Statistical models. Cambridge University Press, Cambridge, England.

DeGroot, M. H. 1982. Comments on Shafer's paper: Lindley's paradox. J. Am. Stat. Assoc. 77:336–339.

Douady, C. J., F. Delsuc, Y. Boucher, W. F. Doolittle, and E. J. Douzery. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol. Biol. Evol. 20:248–254.

Efron, B. 1979. Bootstrap methods: Another look at the jackknife. Ann. Stat. 7:1–26.

Efron, B., E. Halloran, and S. Holmes. 1996. Bootstrap confidence levels for phylogenetic trees [corrected and republished article originally printed in Proc. Natl. Acad. Sci. U.S.A. 1996, 93:7085–7090]. Proc. Natl. Acad. Sci. USA 93:13429–13434.

Erixon, P., B. Svennblad, T. Britton, and B. Oxelman. 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. Syst. Biol. 52:665–673.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791.

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

Goldman, N., J. P. Anderson, and A. G. Rodrigo. 2000. Likelihood-based tests of topologies in phylogenetics. Syst. Biol. 49:652–670.

Good, I. J. 1982. Lindley's paradox. J. Am. Stat. Assoc. 77:342.

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

Hill, B. M. 1982. Comment on Shafer's paper: Lindley's paradox. J. Am. Stat. Assoc. 77:344–347.

Hillis, D. M., and J. J. Bull. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42:182–192.

Huelsenbeck, J. P., B. Larget, and M. E. Alfaro. 2004. Bayesian phylogenetic model selection using reversible jump markov chain monte carlo. Mol. Biol. Evol. 21:1123–1133.

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.

Jeffreys, H. 1939. Theory of probability. Clarendon Press, Oxford, England.

Jeffreys, H. 1961. Theory of probability, 3rd edition. Oxford University Press, Oxford, England.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21–123 *in* Mammalian protein metabolism (H. N. Munro, ed.) Academic Press, New York.

Karol, K. G., R. M. McCourt, M. T. Cimino, and C. F. Delwiche. 2001. The closest living relatives of land plants. Science 294:2351–2353.

Larget, B., and D. L. Simon. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16:750–759.

Lemmon, A. R., and E. C. Moriarty. 2004. The importance of proper model assumption in Bayesian phylogenetics. Syst. Biol. 53:265–277.

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

Li, S., D. Pearl, and H. Doss. 2000. Phylogenetic tree reconstruction using Markov chain Monte Carlo. J. Am. Stat. Assoc. 95:493–508.

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

Mau, B., and M. A. Newton. 1997. Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. J. Comput. Graph. Stat. 6:122–131.

Murphy, W. J., E. Eizirik, S. J. O'Brien, O. Madsen, M. Scally, C. J. Douady, E. Teeling, O. A. Ryder, M. J. Stanhope, W. W. de Jong, and M. S. Springer. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294:2348–2351.

Newton, M. A. 1996. Bootstrapping phylogenies: Large deviations and dispersion effects. Biometrika 83:315–328.

O'Hagan, A., and J. Forster. 2004. Kendall's advanced theory of statistics: Bayesian inference. Arnold, London.

Press, S. J. 2003. Subjective and objective Bayesian statitics, 2nd edition. John Wiley & Sons, New Jersey.

Rannala, B., and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J. Mol. Evol. 43:304–311.

Ripley, B. 1987. Stochastic simulation. Wiley, New York.

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

Shafer, G. 1982. Lindley's paradox. J. Am. Stat. Assoc. 77:325–334.

Silverman, B. W. 1986. Density estimation for statistics and data analysis. Chapman and Hall, London.

Simmons, M. P., K. M. Pickett, and M. Miya. 2004. How meaningful are Bayesian support values? Mol. Biol. Evol. 21:188–199.

Suchard, M. A., R. E. Weiss, and J. S. Sinsheimer. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. Mol. Biol. Evol. 18:1001–1013.

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

Tateno, Y., N. Takezaki, and M. Nei. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Mol. Biol. Evol. 11:261–277.

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

Yang, Z. 1994b. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. Syst. Biol. 43:329–342.

Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42:587–596.

Yang, Z. 2000. Complexity of the simplest phylogenetic estimation problem. Proc. R. Soc. B Biol. Sci. 267:109–116.

Yang, Z., N. Goldman, and A. E. Friday. 1995. Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. Syst. Biol. 44:384–399.

Zharkikh, A., and W.-H. Li. 1995. Estimation of confidence in phylogeny: The complete-and-partial bootstrap technique. Mol. Phylonenet. Evol. 4:44–63.

*First submitted 17 May 2004; reviews returned 2 October 2004; final acceptance 21 January 2005*
*Associate Editor: Paul Lewis*