Fair-Balance Paradox, Star-tree Paradox, and Bayesian Phylogenetics

Ziheng Yang

Department of Biology, Galton Laboratory, University College London, London, United Kingdom

The star-tree paradox refers to the conjecture that the posterior probabilities for the three unrooted trees for four species (or the three rooted trees for three species if the molecular clock is assumed) do not approach $\frac{1}{3}$ when the data are generated using the star tree and when the amount of data approaches infinity. It reflects the more general phenomenon of high and presumably spurious posterior probabilities for trees or clades produced by the Bayesian method of phylogenetic reconstruction, and it is perceived to be a manifestation of the deeper problem of the extreme sensitivity of Bayesian model selection to the prior on parameters. Analysis of the star-tree paradox has been hampered by the intractability of the integrals involved. In this article, I use Laplacian expansion to approximate the posterior probabilities for the three rooted trees for three species using binary characters evolving at a constant rate. The approximation enables calculation of posterior tree probabilities for arbitrarily large data sets. Both theoretical analysis of the analogous fair-coin and fair-balance problems and computer simulation for the tree problem confirmed the existence of the star-tree paradox. When the data size $n \to \infty$, the posterior tree probabilities do not converge to $\frac{1}{3}$ each, but they vary among data sets according to a statistical distribution. This distribution is characterized. Two strategies for resolving the star-tree paradox are explored: (1) a nonzero prior probability for the degenerate star tree and (2) an increasingly informative prior forcing the internal branch length toward zero. Both appear to be effective in resolving the paradox, but the latter is simpler to implement. The posterior tree probabilities are found to be very sensitive to the prior.

Introduction

Thanks to the implementation of efficient Markov chain Monte Carlo (MCMC) algorithms in the computer program MrBayes (Huelsenbeck and Ronquist 2001), the Bayesian method of phylogeny reconstruction (Rannala and Yang 1996; Yang and Rannala 1997; Mau and Newton 1997; Li, Pearl, and Doss 2000) has gained popularity and is now widely used in analysis of molecular data sets. One concern raised about the method is that it often produces extremely high posterior probabilities for trees or clades (Suzuki, Glazko, and Nei 2002; Cummings et al. 2003; Douady et al. 2003; Erixon et al. 2003; Simmons, Pickett, and Miya. 2004). For example, Rannala and Yang (1996) calculated the posterior probability for a tree of five ape species using 11 mitochondrial tRNA genes to be 0.9999. Even though the tree is sensible, the posterior probability is very high given that the human-chimpanzee-gorilla relationship was hard to resolve and that the data set, with 759 bp, is small. Similarly use of MrBayes in real data analysis has produced high posterior probabilities, often mostly 100%. Sometimes different data sets (such as different genes or different taxon samples) produced contradictory phylogenies, each with strong posterior support (e.g., Bourlat et al. 2006). Simulation studies and empirical data analyses have repeatedly found that the posterior tree probabilities tend to be much higher than bootstrap support values (e.g., Cummings et al. 2003; Douady et al. 2003; Erixon et al. 2003; Simmons, Pickett, and Miya 2004). This discrepancy in itself may not suggest anything inappropriate about posterior probabilities, because the interpretation of bootstrap support values is uncertain (e.g., Berry and Gascuel 1996; Yang and Rannala 2005). Nevertheless, there is widespread concern that posterior probabilities for trees or clades calculated from many data sets may be too high.

Key words: Lindley's paradox, fair-balance paradox, star-tree paradox, prior, clade probabilities.

E-mail: z.yang@ucl.ac.uk.

Mol. Biol. Evol. 24(8):1639–1655. 2007
doi:10.1093/molbev/msm081
Advance Access publication May 7, 2007

In a simulation study, Suzuki, Glazko, and Nei (2002) generated data sets under the star tree for four species and analyzed them using MrBayes, which considers binary trees only. They found that the posterior probability for the inferred binary tree was often too high. The study used a wrong and simplistic model in the analysis, so that the problem was due in part to model violation. However, extreme posterior probabilities were observed in similar simulations without model violation (Cummings et al. 2003; Lewis, Holder, and Holsinger 2005; Yang and Rannala 2005). The failure of the posterior probabilities for the three binary trees to converge to $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ in large data sets simulated under the star tree is somewhat counterintuitive and is called the *star-tree paradox* (Lewis, Holder, and Holsinger 2005). The concern is not so much that the posterior tree probabilities differ from $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ as that they are sometimes either very small or very large when in fact no information is available to resolve the tree one way or another.

The posterior probability for a tree is the probability that the tree is true given the data, the prior, and the likelihood (substitution) model. There are thus three possible reasons for high tree probabilities: (1) errors, including numerical problems in the MCMC algorithm, which cause the posterior probabilities to be calculated incorrectly; (2) misspecification of the substitution model; and (3) misspecification and sensitivity of the prior. The first two reasons may be responsible for high posterior probabilities in some studies. In particular, use of simplistic and unrealistic models is known to inflate posterior probabilities for trees (e.g., Buckley 2002; Lemmon and Moriarty 2004; Huelsenbeck and Rannala 2004). However, high posterior probabilities have also been observed when the first two reasons clearly do not apply (Yang and Rannala 2005). This article deals with the third reason and studies the effect of prior specification on Bayesian phylogenetic inference.

The nature of the problem may be better understood by considering the analogous fair-coin problem (Lewis, Holder, and Holsinger 2005; Yang and Rannala 2005). Suppose a coin is fair with the probability of heads to be $\theta_0 = \frac{1}{2}$. We flip the coin n times and observe y heads. We then calculate the posterior probabilities (P_- and P_+) for two models that the coin is either negatively or

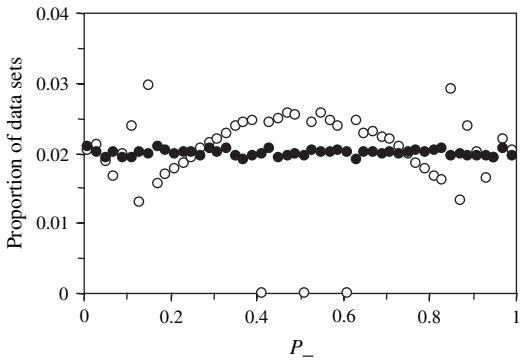

FIG. 1.—The histogram of P_-, the posterior probability that the coin has negative bias (with the probability of heads $\theta < \frac{1}{2}$) in a coin-tossing experiment. A fair coin is tossed $n = 10^3$(○) or $n = 10^6$(●) times. The number of heads y in n tosses is used to calculate P_-, assuming a uniform prior $\theta \sim U(0, 1)$, and the proportion of replicate data sets in which P_- falls into bins of 2% width is calculated to form the histogram. The number of simulated replicates is 10^5. The fluctuation for $n = 10^3$ is mainly due to the discrete nature of the data; for example, in no data sets is P_- in the 0.50–0.52 bin because $P_- = 0.5$ if $y = 500$ and $P_- = 0.525$ if $y = 499$. When $n = 10^6$, the fluctuation disappears and P_- has nearly a $U(0, 1)$ distribution, by which the proportion in each bin is 0.02.

positively biased: H_-: $\theta < \frac{1}{2}$ and H_+: $\theta > \frac{1}{2}$. (It is inconsequential whether the true value $\theta = \frac{1}{2}$ is included in none, one, or both of the two models since a point value has zero probability in a continuous distribution.) We assign equal prior probabilities for H_- and H_+ and uniform priors for θ in each model. When n is large, we may expect P_- and P_+ to approach $\frac{1}{2}$, but they do not. Instead P_- varies considerably among data sets (all generated under $\theta_0 = \frac{1}{2}$) even when $n \to \infty$. This is referred to as the *fair-coin paradox* (Lewis, Holder, and Holsinger 2005). Indeed, the limiting distribution of P_- when $n \to \infty$ is the uniform $U(0, 1)$ (Yang and Rannala 2005, equation 5). Figure 1 shows the histograms of P_- when $n = 10^3$ and 10^6. Intuitively, even though the proportion of heads y/n becomes closer and closer to $\frac{1}{2}$ when n increases, the number of heads y fluctuates around $n/2$ more and more wildly among data sets. Note that the variance of y/n is $1/(4n)$, and the variance of y is $n/4$. The posterior probability P_- depends on the number as well as the proportion of heads.

One has to consider how a sensible Bayesian analysis should behave in this problem. In a significance test, the P value has a uniform distribution $U(0, 1)$ if the null hypothesis is true and the test is exact. The true null hypothesis is falsely rejected 5% of the time if the test is conducted at the 5% significance level. This is the case even with infinitely large data sets, if a fixed significance level is used. However, Bayesian statistics is a more ''optimistic'' and ''aggressive'' methodology (Efron 1998). In Bayesian model selection, the posterior probability for the true model, or the model closest to the truth among the compared models, should converge to one when the amount of data approaches infinity. As H_- and H_+ are equally distant from the truth $\theta_0 = \frac{1}{2}$, one may sensibly expect P_- and P_+ to converge to $\frac{1}{2}$ when $n \to \infty$. Of course, P_- should converge to 1 if $\theta_0 < \frac{1}{2}$ (or to 0 if $\theta_0 > \frac{1}{2}$). For the tree problem, the same argument suggests that if the true tree is the star tree, one would like the posterior probabilities for the three binary trees to converge to $\frac{1}{3}$ each when the number of sites $n \to \infty$. Here I take this position, as did Lewis, Holder, and Holsinger (2005) and Yang and Rannala (2005). It has been unclear how posterior tree probabilities behave in very large data sets or when $n \to \infty$, because problems of phylogeny reconstruction are intractable analytically. Numerical calculation of integrals becomes unreliable in large data sets while MCMC algorithms are too slow and too imprecise.

In this article I develop approximate methods to calculate the posterior probabilities (P_1, P_2, P_3) for the three rooted trees for three species, using data of binary characters evolving at a constant rate. This is the simplest tree-reconstruction problem (Yang 2000), chosen here to make the analysis possible. The approximation allows Bayesian calculation in arbitrarily large data sets, without the need for MCMC algorithms. I conduct large-scale simulations, which confirm the existence of the star-tree paradox; when the data size n increases, the posterior tree probabilities do not converge to $\frac{1}{3}$ each, but continue to vary among data sets according to a statistical distribution. This distribution is characterized. I then explore the sensitivity of Bayesian analysis to the prior and evaluate two strategies suggested to resolve the star-tree paradox. The first assigns a nonzero prior probability for the degenerate star tree (Lewis, Holder, and Holsinger 2005), and the second uses a prior to force the internal branch lengths to approach zero when $n \to \infty$ (Yang and Rannala 2005). The behavior of posterior tree probabilities in large data sets is predicted by drawing an analogy with the fair-coin problem, and the predictions are confirmed numerically by computer simulation.

A synopsis is provided in the next section, which summarizes the major results of this study. The biologist reader may read this section, as well as the Discussion, and skip the Mathematical Analysis section.

Biological Synopsis

The Fair-coin and Fair-balance Problems

The fair-coin problem, as described above, has the same behavior as the fair-balance problem discussed by Yang and Rannala (2005), and in this study their results are treated interchangeably. Here the results are summarized for the fair-coin problem. We assign a beta prior on the probability of heads: $\theta \sim \text{beta}(\alpha, \alpha)$, with mean $\frac{1}{2}$ and variance $1/(8\alpha + 4)$. This is the $U(0, 1)$ prior when $\alpha = 1$ but can be highly concentrated around $\frac{1}{2}$ if α is large. As long as α is fixed, the posterior probability P_- for the model of negative bias approaches the uniform distribution $U(0, 1)$ when the number of coin tosses $n \to \infty$.

Two strategies (priors) are considered to resolve the fair-coin paradox. In the first, α in the beta prior increases with n so that the prior variance of θ approaches 0, forcing θ to be more and more highly concentrated around $\frac{1}{2}$. We require that P_- approach $\frac{1}{2}$ if the coin is fair, and 1 if the coin has a negative bias (or 0 if the coin has a positive bias). These requirements mean that the prior variance for θ should approach 0 faster than $1/n$ and more slowly than $1/n^2$. In the second, a nonzero prior probability is assigned to the degenerate model of no bias H_0: $\theta = \frac{1}{2}$. Then the

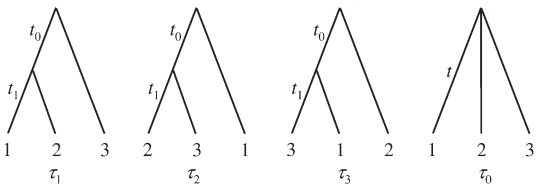

FIG. 2.—The three rooted trees for three species: $\tau_1 = ((12)3)$, $\tau_2 = ((23)1)$, and $\tau_3 = ((31)2)$. Branch lengths t_0 and t_1 are measured by the expected number of character changes per site. The star tree $\tau_0 = (123)$ is also shown with its branch length t.

posterior probability for H_0 approaches 1 when $n \to \infty$, and the method behaves as desired.

The Star-tree Problem

Defining the Problem

The three binary rooted trees for three species are shown in figure 2. The data are three sequences of binary characters, which are assumed to be evolving at a constant rate (that is, under the molecular clock) (Yang 2000). The data can be summarized as counts n_0, n_1, n_2, n_3 of site patterns *xxx*, *xxy*, *yxx*, and *xyx*, where *x* and *y* are any two distinct characters, while the total number of sites is $n = \sum_{i=0}^{3} n_i$. Each binary tree has two branch length parameters t_0 and t_1, measured by the expected number of changes per site. Intuitively, we can see the three variable patterns *xxy*, *yxx*, and *xyx* ''support'' the three binary trees τ_1, τ_2, and τ_3, respectively. Indeed a likelihood analysis will choose tree τ_1 as the maximum-likelihood tree if n_1 is greater than both n_2 and n_3. Let p_0, p_1, p_2, p_3 be the expected site pattern probabilities, with $\sum_{i=0}^{3} p_i = 1$. Then tree τ_1 can be represented by $p_0 > p_1 > p_2 = p_3$, with two free parameters, whereas the star tree is $p_0 > p_1 = p_2 = p_3$ (Yang 2000). In a Bayesian analysis, we assign equal probabilities $(\frac{1}{3})$ to the three binary trees, and exponential priors with means μ_0 and μ_1 on the two branch lengths t_0 and t_1 in each binary tree (fig. 2).

Star-tree Paradox

Posterior probabilities for the three binary trees (P_1, P_2, P_3) were calculated from data sets simulated under the star tree, with $n = 3 \times 10^3$, 3×10^6, or 3×10^9 sites in the sequence. It is found that (P_1, P_2, P_3) does not converge to $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ with the increase of n, confirming the star-tree paradox. Instead (P_1, P_2, P_3) vary among data sets, according to a distribution $f(P_1, P_2, P_3)$, which is independent of the branch length t in the star tree and of the prior means μ_0 and μ_1 (see fig. 7 below). There are four modes in the distribution, such that in most data sets, either the three probabilities are all close to $\frac{1}{3}$, or one of them is close to 1 and the other two are close to 0. Suppose we consider very high and very low posterior probabilities for binary trees as ''errors'' since the true tree is the star tree. In 4.2% (or 0.8%) of data sets, at least one of the three posterior probabilities is > 0.95 (or > 0.99%), and in 17.3% (or 2.6%) of data sets, at least one of the three posterior probabilities is < 0.05 (or < 0.01). Those ''error'' rates appear too high, given that the data sets are arbitrarily large and are supposed to represent infinite data sets.

Two Strategies to Resolve the Star-tree Paradox

Further analysis of the tree problem is through an analogy with the fair-coin problem. Note that the fair-coin and fair-balance problems are analytically tractable, but the tree problem is not. My analysis of the tree problem is thus numerical verification by computer simulation, in which only a finite number of replicate data sets can be generated and each data set can only be of finite size. To see the analogy, it is more convenient to consider the site pattern probabilities as parameters in each binary tree instead of branch lengths t_0 and t_1. In the fair-coin problem, the data have a binomial distribution or multinomial distribution with two cells (corresponding to heads and tails). The two models of negative and positive bias assume that one cell probability is greater than the other, yet the truth (the fair-coin model) is that they are equal. In the star-tree problem, the data have a multinomial distribution with four cells (corresponding to the four site patterns). We compare three binary-tree models, which assume that one of three cell probabilities (for the three variable site patterns) is greater than the other two and that these other two are equal. The truth (the star tree) is that all three cell probabilities are equal. In other words, the three binary trees are represented by τ_1: $p_1 > p_2 = p_3$, τ_2: $p_2 > p_3 = p_1$ and τ_3: $p_3 > p_1 = p_2$, while the true star tree is τ_0: $p_1 = p_2 = p_3$. (The probability p_0 for the constant pattern may be considered an unimportant nuisance parameter, shared by all four trees.) Both the proportions of heads and tails in the fair-coin problem and the proportions of the site patterns in the tree problem converge to their expected probabilities, with variances proportional to $1/n$.

We apply the same two strategies as discussed above for the fair-coin problem to resolve the star-tree paradox. The first uses a prior on parameters in the model to force the binary tree to converge to the star tree, or to force the three cell probabilities p_1, p_2, p_3 to approach equality ($p_1 = p_2 = p_3$), when $n \to \infty$. From the analysis of the fair-coin problem, the prior should force $E(p_1 - p_2)^2$ to approach 0 faster than $1/n$ but more slowly than $1/n^2$. This means, as seen by translating the prior on cell probabilities into a prior on branch lengths t_0 and t_1, that the mean μ_0 in the exponential prior for the internal branch length t_0 should approach 0 faster than $1/\sqrt{n}$ but more slowly than $1/n$. This prediction is only partially confirmed. Simulations confirm that to resolve the star-tree paradox—that if, for (P_1, P_2, P_3) to converge to $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ if the star tree is the true tree — μ_0 should approach 0 faster than $1/\sqrt{n}$. Numerical problems (see later) have prevented confirmation that μ_0 should approach 0 more slowly than $1/n$ for P_1 to converge to 1 if tree τ_1 is the true tree.

The second strategy assigns a nonzero prior probability π_0 for the degenerate star tree ($p_1 = p_2 = p_3$). Simulations confirm that when $n \to \infty$, the posterior probability for the star tree approaches 1, and this prior indeed resolves the star-tree paradox. This result is expected from previous theoretical work. Indeed Dawid (1999) has studied the asymptotics of Bayesian model selection when the data size $n \to \infty$. If all models considered in the Bayesian analysis are wrong, the probability for the model closest to the truth, as measured by the Kullback-Leibler divergence, approaches 1. If one model is correct and all others are wrong,

the probability for the true model approaches 1. If several models are true, the probability for the true model with the fewest parameters approaches 1. The case where several models of the same dimension are true is not well specified. Dawid's proof assumes that the parameters are unbounded while here the star tree is at the boundary of the parameter space of the binary trees. However, the qualitative conclusions appear applicable to the tree problem. Here the data are generated under the star tree, so that all four trees are correct, but the star tree has one fewer parameter, and its posterior probability approaches 1.

Discussion

Does the Star-tree Paradox Exist?

Kolaczkowski and Thornton (2006), referred to hereinafter as KT06, recently argued that the star-tree paradox does not exist. The authors performed three analyses, each of which appears to be invalid or misinterpreted.

First, KT06 simulated data sets with up to $n = 10^7$ sites using a star tree of four species, with all four branch lengths equal. The data were analyzed using MrBayes to calculate posterior probabilities (P_1, P_2, P_3) for the three binary unrooted trees without assuming the molecular clock. All five branch lengths in each binary tree are assigned the uniform prior $U(0, 10)$. The variance in the posterior probability for a binary tree, say P_1, was initially small, but increased with the increase of n to a stable value of about 0.06 when $n \geq 10^3$ (KT06, fig. 1b). The standard deviation (SD) of $\sim$0.24 ($=\sqrt{0.06}$) is about the same as that obtained in this article for rooted trees of three species (0.2498; see figure 8a below). It is likely that these two values are indeed identical and that the three-species problem of figure 2, studied here, and the four-species problem with equal branch lengths in the star tree, studied by KT06, produce the same limiting distribution $f(P_1, P_2, P_3)$. It is also likely that the distribution in the four-species case is similarly independent of the branch length used in the star tree and the upper bound in the uniform prior for branch lengths in the binary trees. It would be interesting to know whether this invariance holds also when the four branches in the star tree have different lengths. At any rate, the failure of P_1 to converge to $\frac{1}{3}$ confirms the star-tree paradox. KT06 appeared to have mistaken a stable variance for zero variance when they claimed that their results disproved the star-tree paradox, and they were incorrect to conclude that ''With infinite data, posterior probabilities give equal support for all resolved trees, and the rate of false inferences falls to zero.'' KT06 emphatically criticized the speculation of Lewis, Holder, and Holsinger (2005) that ''Bayesian analyses become increasingly unpredictable'' with the increase of data size when the true tree is the star tree. Technically, this speculation is confirmed rather than refuted by the result of KT06 (and by the results of this study), as the variance of P_1 continues to increase with n, even though the amount of increase approaches zero (KT06, fig. 1b). Clearly, the variance cannot increase without limit, the absolute maximum being 2/9 (with the SD to be $\sqrt{2}/3 = 0.4714$), achieved if the posterior probabilities (P_1, P_2, P_3) take only three sets of values, each with probability 1/3: (1, 0, 0), (0, 1, 0), and (0, 0, 1).

Second, KT06 examined the so-called type-I error rate in finite data sets of 5,000 sites, and find that when the true tree is the star tree, the posterior probability for a binary tree is $> 95\%$ (or $> 99\%$) in less than 5% (or 1%) of data sets. The same pattern holds also for rooted trees in this study, although the posterior probability for a binary tree is $< 5\%$ (or $< 1\%$) in more than 5% (or 1%) of data sets, as mentioned above. It is debatable whether such ''error'' rates are acceptable if they persist in arbitrarily large data sets. While it is appropriate to study so-called Frequentist properties of a Bayesian method, KT06 confused Bayesian posterior probabilities with Frequentist P values when they claimed that ''posterior probabilities never produce strong support for incorrectly resolved phylogenies more often *than they should*.'' Bayesian statistics in general does not provide a guaranty of its performance under Frequentist criteria. KY06 also claimed that the ''type-I'' error rate decreased when n increased from 10^3 to 10^7 (KY06, fig. 2b). This result is inconsistent with the present study and appears to contradict their finding of an increasing and asymptotically stable variance in P_1. The result may be due to numerical problems in the MCMC algorithms in the analysis of KT06.

Third, KT06 used MrBayes to analyze a data set consisting of the expected probabilities of the site patterns calculated under the star tree. This ''infinite'' data set gave $\frac{1}{3}$ as the posterior probability for each binary tree. However, analysis of this average site is not meaningful, as it ignores the variation among data sets and the fact that the number of sites as well as the proportions of site patterns influences Bayesian analysis. In the fair-coin problem, the data set consisting of $\frac{1}{2}$ heads and $\frac{1}{2}$ tails would produce $P_- = P_+ = \frac{1}{2}$, but this average coin toss tells us nothing about the behavior of the Bayesian method when $n \to \infty$ (see fig. 1).

The position of KT06 toward the star-tree paradox is marred by errors in the analysis. The paradox concerns the performance of the Bayesian method in large or infinite data sets, so that finite data sets are not the real issue. Nevertheless the ''error'' rates in finite data sets are higher than KY06 suggested, because the method produced very small posterior probabilities too often (see above). KT06 expected the ''error'' rate to reduce to zero when the data size $n \to \infty$, with the posterior tree probabilities approaching $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. This is the behavior of a sensible Bayesian analysis assumed by Lewis, Holder, and Holsinger (2005) and Yang and Rannala (2005), although KT06 failed to realize that the Bayesian method does not behave in this way.

Priors and Bayesian Phylogenetics

It is a curious fact that to resolve the fair-coin paradox, the prior probability π_0 on the degenerate model of fair coin (H_0: $\theta = \frac{1}{2}$) can be constant and independent of the data size, while the prior on parameter θ (the probability of heads) has to be increasingly concentrated around $\theta = \frac{1}{2}$, depending on the data size n. The difference appears to be due to the fact that any point mass has probability zero in a non-degenerate continuous distribution. Nevertheless, both may be viewed as priors on parameter θ in models of negative and positive biases (H_- and H_+) without considering H_0 in the analysis. The degenerate-model prior is equivalent to assigning

a mixture distribution on θ, with a component at $\frac{1}{2}$ in proportion π_0 and another component from a continuous distribution in proportion $1 - \pi_0$. Similarly the star-tree prior π_0 is equivalent to a mixture-distribution for internal branch lengths in binary trees (with the star tree excluded from the Bayesian analysis), with a component at zero in proportion π_0 and a component from the continuous exponential distribution in proportion $1 - \pi_0$. Implementation of the data size-dependent prior is simpler as it requires only a change to the prior mean for internal branch lengths (Yang and Rannala 2005). The star-tree prior is more complex because bifurcating and multifurcating trees have different numbers of branch length parameters so that algorithms such as reversible jump MCMC (Green 1995) are needed to deal with models of different dimensions (Lewis, Holder, and Holsinger 2005).

Both the star-tree prior and the data size-dependent prior may be criticized. Whether truly simultaneous speciation events ever occur in nature is debatable, and if they do not, assigning a prior probability to a model known to be false runs into a conceptual difficulty. Similarly, the use of data size (although not the data themselves) for prior specification may appear non-Bayesian. The prior is supposed to reflect information concerning the parameter before the data are analyzed and should ideally be independent of the data. Nevertheless, this ideal is often hard to achieve in ''objective'' Bayesian statistics when little information is available about the parameter. Both Jeffreys's prior (Jeffreys 1961) and the reference prior (Bernardo 1979) depend on the likelihood function or the experimental design. One may ask why one's prior ignorance concerning a parameter should depend on how one conducts the experiment to find out about the parameter. An extreme case is Bernardo's (1980) use of the data (not just data size) to specify the prior, although the idea did not appear to be warmly received in the ensuing discussions. Data size-dependent priors were discussed by Bartlett (1957), Davison (2003, pp. 586–587), and Cox (2006, pp. 42–43, 106–107), as a possible way of resolving Lindley's paradox (see below). One may argue that if data sets of such large sizes are needed to resolve the tree, the internal branch must be very short, so that it may be sensible to assume increasingly shorter internal branches in the prior in larger data sets. Yang and Rannala (2005) also discussed the use of empirical estimates of internal branch lengths from real data sets to specify the prior, and pointed out that almost all of the possible phylogenetic trees are wrong, and that most internal branch lengths in wrong trees are estimated to be zero.

The biologist reader should be aware that there have been longstanding fundamental disagreements among statisticians concerning principles of statistical inference. In particular, model selection is a difficult area for both Bayesian and Frequentist statistics, and it is also an area where the two approaches can draw very different conclusions from the same data. A brief overview of this controversy is provided in Yang (2006, §5.1.3). As phylogeny reconstruction unfortunately falls into this class of difficult statistical inference problems (e.g., Yang et al. 1995), biologists may have to think about what constitutes a sensible behavior in a Bayesian phylogenetic analysis. Six decades ago, Egon S. Pearson (1947) wrote that ''Hitherto the user has been accustomed to accept the function of probability theory laid down by the mathematicians; but it would be good if he could take a larger share in formulating himself what are the practical requirements that the theory should satisfy in application.'' This advice may be useful even today.

Almost all controversies surrounding Bayesian inference concern the prior, which is also the focus of this study. For the tree problem, one may take the position that the prior implemented in current computer programs is appropriate and then accept whatever properties Bayesian inference under the prior possesses. The expectation for the posterior tree probabilities to approach $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ when $n \to \infty$ is then seen as false intuition and no paradox remains. This position may be natural to some Bayesian statisticians. Another position is to judge the method by its statistical properties. In the ''objective'' Bayesian method, it is a common practice to specify the prior such that the resulting Bayesian inference is deemed reasonable (e.g., Jeffreys 1961). I have taken that position in this study, motivated by the observation that the posterior tree probabilities are often too extreme. Two a priori criteria are set up: (1) if the true tree is the star tree, the probabilities for the three binary trees should approach $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ when $n \to \infty$, and (2) if a binary tree is the true tree, its posterior probability should approach 1. Two strategies of prior specification are then found to meet those criteria.

Based on the perception that the posterior probabilities for trees or clades are often too high, some authors (e.g., Suzuki, Glazko, and Nei 2002; Simmons, Pickett, and Miya 2004) argued that the Bayesian posterior probabilities for trees or clades are not trustable, and alternative methods such as the bootstrap should be used to assess the reliability of estimated trees. Similarly Douady et al. (2003) suggested the bootstrapped Bayesian analysis, in which the Bayesian method is used to analyze bootstrap pseudo-data sets. The method then involves prohibitive computation and is also a strange mix of Bayesian and Frequentist methodologies. Instead, we consider the default priors implemented in current computer programs to be inappropriate and attempt to specify better priors to produce more reasonable posterior tree probabilities. Lewis, Holder, and Holsinger (2005) also emphasized the fact that a number of realistic evolutionary models have been implemented in the MrBayes program, making the method an attractive option for analyzing ever-increasing genetic data sets.

Mathematical Analysis

Bayesian Analysis of the Fair-balance Problem

The Fair-balance Problem

In the fair-coin problem, one may also assign an informative prior on the probability of heads: $\theta \sim \text{beta}(\alpha, \alpha)$. The beta distribution with $\alpha > 1$ has a mode at $\frac{1}{2}$, so that the coin is more likely to be nearly even than seriously biased. For large α, beta(α, α) can be approximated by a normal distribution with mean $\frac{1}{2}$ and variance $1/(8\alpha + 4)$. The likelihood, given by the binomial probability of the number of heads $y \sim \text{bi}(n, \theta)$, is approximated by the normal density $y/n \sim N(\frac{1}{2}, 1/(4n))$. The posterior $\theta|y \sim \text{beta}(y + \alpha, n - y + \alpha)$ can be approximated by the normal distribution with

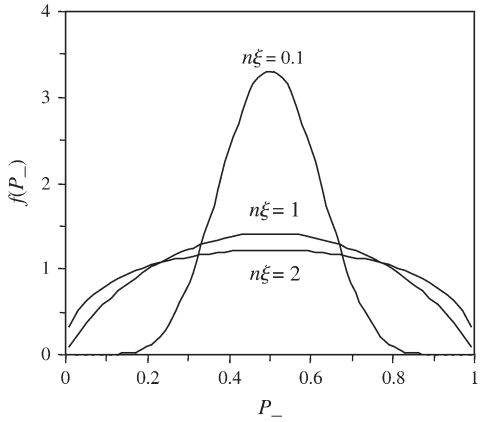

FIG. 3.—The density $f(P_-)$ for the posterior model probability P_- in the fair-balance problem when the prior is $\theta \sim N(0, \xi\sigma^2)$ with $\xi = c/n$. The true $\theta = 0$. The plots correspond to $c = n\xi = 0.1$, 0.5, and 2, calculated using equation (3).

mean $\frac{y+\alpha}{n+2\alpha}$ and variance $\frac{(y+\alpha)(n-y+\alpha)}{(n+2\alpha)^2(n+2\alpha+1)}$. (Note that the variance is approximately proportional to $1/n$ if n is large and α is fixed.) Thus we redefine $\theta - \frac{1}{2}$ as the parameter, use the normal distribution to approximate the prior, the likelihood, and the posterior; and restate the problem as the following *fair-balance problem*. Suppose the data consist of n independent observations $y_1, y_2, \ldots, y_n$, with $y_i \sim N(\theta, \sigma^2)$, where θ is unknown and σ^2 is known. The y_is may be measurement errors on a balance. Let $\bar{y}$ be the sample mean. The two models are then H_-: $\theta < 0$ and H_+: $\theta > 0$. In the prior we assign equal probabilities $(\frac{1}{2})$ for each model, and $\theta \sim N(0, \xi\sigma^2)$, truncated to the appropriate range in each model.

The posterior of θ is then given by $\theta|\bar{y} \sim N\left(\frac{\bar{y}}{1+1/(n\xi)}, \frac{\sigma^2/n}{1+1/(n\xi)}\right)$, from which one can get the posterior probability for model H_- as

$$P_- = \Pr(H_-|\bar{y}) = \int_{-\infty}^{0} f(\theta|\bar{y})\, d\theta = \Phi\left(-\frac{\sqrt{n}\bar{y}/\sigma}{\sqrt{1+1/(n\xi)}}\right), \quad (1)$$

where $\Phi(\cdot)$ is the cumulative distribution function (c.d.f.) of the standard normal distribution (Yang and Rannala 2005, equation 6). Note that $\sqrt{n}\bar{y}/\sigma$ is a random variable from the standard normal distribution.

Suppose the true parameter is θ_0. As $\bar{y}$ varies among data sets as $\bar{y} \sim N(\theta_0, \sigma^2/n)$, P_- has the density

$$f(P_-|\theta_0) = \sqrt{1+\frac{1}{n\xi}} \times \exp\left\{\frac{1}{2}[\Phi^{-1}(P_-)]^2 - \frac{1}{2}\left[\Phi^{-1}(P_-) \times \sqrt{1+\frac{1}{n\xi}} + \frac{\sqrt{n}\theta_0}{\sigma}\right]^2\right\}, 0<P_-<1, \quad (2)$$

where $\Phi^{-1}(\cdot)$ is the inverse c.d.f. of the standard normal distribution (Yang and Rannala 2005, equation 12). This is a function of P_-, $n\xi$, and $\sqrt{n}\theta_0/\sigma$.

If the balance is fair and the true parameter $\theta_0 = 0$, equation (2) becomes

$$f(P_-) = \sqrt{1+\frac{1}{n\xi}} \times \exp\left\{-\frac{1}{2n\xi}\left[\Phi^{-1}(P_-)\right]^2\right\}, 0<P_-<1 \quad (3)$$

(Yang and Rannala 2005, equation 13). This is a function of P_- and $n\xi$. If ξ is a constant, we have $n\xi \to \infty$ and $f(P_-) \to 1$ when $n \to \infty$, so that P_- converges to the uniform distribution $U(0, 1)$ (see fig. 1). This is called the *fair-balance paradox* (Yang and Rannala 2005). We would like P_- to approach $\frac{1}{2}$ when $n \to \infty$, but it fails to do so.

The Data Size-Dependent Prior

One of the ideas suggested in the discussions of Lindley's paradox (Lindley 1957, see below) is to let the prior be increasingly informative with the increase of the data size. Consider $\xi = c/n^\gamma$ as a prior for θ, as a possible way for resolving the fair-balance paradox. From equation (3), it is clear that if $0 < \gamma < 1$, P_- still converges to $U(0, 1)$, even though this prior forces θ to be closer and closer to 0 with the increase of n, converging to a point mass at $\theta = 0$ in the limit. If $\gamma = 1$ so that $\xi = c/n$, $f(P_-)$ peaks at $P_- = \frac{1}{2}$, but the distribution does not degenerate to a point mass at $\frac{1}{2}$ (equation 3). Figure 3 shows a few densities when $c = n\xi = 0.1$, 1, and 2. Note that in this case the prior $\theta \sim N(0, c\sigma^2/n)$ and the likelihood $\bar{y} \sim N(\theta, \sigma^2/n)$ have the same ''precision'' about θ. When $\gamma > 1$, $f(P_-) \to 0$ for all values of P_- except $P_- = \frac{1}{2}$; that is, P_- converges to a point mass at $\frac{1}{2}$. Thus to avoid the fair-balance paradox, we should have $\gamma > 1$ in $\xi = c/n^\gamma$; the variance in the prior of θ should approach 0 faster than $1/n$.

The case of $\theta_0 \neq 0$ (equation 2) is summarized in table 1. The statement of Yang and Rannala (2005, pp. 468–469) that P_- converges to 1 if $\theta_0 < 0$ (or to 0 if $\theta_0 > 0$) irrespective of γ in $\xi = c/n^\gamma$ is inaccurate. Indeed the behavior of $f(P_-)$ depends on γ. To ensure that $P_- \to 1$ if $\theta_0 < 0$ (and $P_- \to 0$ if $\theta_0 > 0$), we require $\gamma < 2$. Any value of γ in the interval $(1, 2)$ will produce sensible Bayesian inference by the criteria used here, and a smaller γ corresponds to a more powerful analysis, as it produces higher posterior probabilities for the true model if the coin is biased. Figure 4a shows that the posterior probability P_- calculated from a data set may be very sensitive to the prior or the value of γ. Furthermore, while $f(P_-)$ converges to a point mass at 1, $\frac{1}{2}$, and 0 if the true $\theta_0 < 0$, $= 0$, and > 0, respectively, the rate of convergence depends on γ. Curves a & b in figure 5 show the density when $\theta_0 = 0$ and 0.1 at $n = 1{,}000$ when $\gamma = \sqrt{2}$ is used.

The Degenerate-Model Prior

Another strategy is to assign a nonzero probability to the degenerate model H_0: $\theta = 0$ (Lewis, Holder, and Holsinger 2005). The three models H_0, H_-, and H_+ then have the prior probabilities π_0, $(1 - \pi_0)/2$, and $(1 - \pi_0)/2$. The unknown parameter in H_- and H_+ is assigned the prior $\theta \sim N(0, \xi\sigma^2)$, truncated to the appropriate range, where ξ is a constant.

The likelihood is given by $\bar{y}|\theta \sim N(\theta, \sigma^2/n)$. The marginal likelihoods are

Table 1
Behavior of Posterior Model Probability P_- in the Fair-balance Problem When the Data Size $n \to \infty$ and the Prior is $\theta \sim N(0, \xi\sigma^2)$ with $\xi = c/n^\gamma$

True θ	Prior	$n\xi$	P_-	Interpretation
$\theta_0 = 0$	$0 \le \gamma < 1$	$n\xi \to \infty$	$f(P_-) \to 1$	P_- varies among data sets like a random number: $P_- \sim U(0, 1)$
(equation 3)	$\gamma = 1$	$n\xi = c$	P_- has a distribution with mode $\frac{1}{2}$	P_- varies among data sets, often close to $\frac{1}{2}$.
	$\gamma > 1$	$n\xi \to 0$	$P_- \to \frac{1}{2}$	P_- is $\frac{1}{2}$ in every data set.
$\theta_0 > 0$	$0 \le \gamma < 1$	$n\xi \to \infty$	$P_- \to 0$	P_- is 0 in every data set.
(eq. 2)	$\gamma = 1$	$n\xi = c$	$P_- \to 0$	P_- is 0 in every data set.
	$1 < \gamma < 2$	$n\xi \to 0$	$P_- \to 0$	P_- is 0 in every data set.
	$\gamma = 2$	$n\xi \to 0$	$P_- \to \Phi(-\sqrt{c}\theta_0/\sigma)$	P_- converges to a constant, which is neither 0 nor 1, in every data set.
	$\gamma > 2$	$n\xi \to 0$	$P_- \to \frac{1}{2}$	P_- is $\frac{1}{2}$ in every data set.

NOTE.— $P_- \to a$ means that $f(P_-)$ approaches 0 for all values of P_- except $P_- = a$. The case of $\theta_0 < 0$ is similar to that of $\theta_0 > 0$.

$$
\begin{aligned}
f(\bar{y}|H_0) &= \frac{1}{\sqrt{2\pi\sigma^2/n}} \times \exp\left(-\frac{n\bar{y}^2}{2\sigma^2}\right),\\
f(\bar{y}|H_-) &= \int_{-\infty}^{0} \left[\frac{2}{\sqrt{2\pi\xi\sigma^2}} e^{-\frac{1}{2\xi\sigma^2}\theta^2} \times \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left(-\frac{n(\bar{y}-\theta)^2}{2\sigma^2}\right)\right] d\theta\\
&= \frac{2}{\sqrt{2\pi\sigma^2(\xi+1/n)}} \exp\left(-\frac{n\bar{y}^2}{2\sigma^2(1+n\xi)}\right) \times \Phi\left(-\frac{n\bar{y}}{\sigma\sqrt{n+1/\xi}}\right),\\
f(\bar{y}|H_+) &= \frac{2}{\sqrt{2\pi\sigma^2(\xi+1/n)}} \exp\left(-\frac{n\bar{y}^2}{2\sigma^2(1+n\xi)}\right) \times \Phi\left(\frac{n\bar{y}}{\sigma\sqrt{n+1/\xi}}\right).
\end{aligned}
\tag{4}
$$

The posterior probabilities for the three models are

$$
\begin{aligned}
P_0 &= P(H_0|\bar{y}) = \frac{1}{D}\pi_0\sqrt{1+n\xi}\exp\left(-\frac{n\bar{y}^2}{2\sigma^2(1+1/(n\xi))}\right),\\
P_- &= P(H_-|\bar{y}) = \frac{1}{D}(1-\pi_0)\Phi\left(-\frac{\sqrt{n}\bar{y}}{\sigma\sqrt{1+1/(n\xi)}}\right),\\
P_+ &= P(H_+|\bar{y}) = \frac{1}{D}(1-\pi_0)\Phi\left(\frac{\sqrt{n}\bar{y}}{\sigma\sqrt{1+1/(n\xi)}}\right),
\end{aligned}
\tag{5}
$$

where $D = \pi_0\sqrt{1+n\xi}\exp\left(-\frac{n\bar{y}^2}{2\sigma^2[1+1/(n\xi)]}\right) + (1-\pi_0)$. Note that P_0, P_-, and P_+ are functions of $\sqrt{n}\bar{y}/\sigma$, π_0, and $n\xi$.

When the sample mean $\bar{y}$ varies among data sets according to $N(\theta_0, \sigma^2/n)$, with density

$$
\phi(\bar{y}|\theta_0) = \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left\{-\frac{n}{2\sigma^2}(\bar{y}-\theta_0)^2\right\}, \tag{6}
$$

the posterior model probabilities will have a joint density $f(P_0, P_-, P_+)$. This appears intractable. However, the marginal density of P_0 can be derived as follows. Rewrite P_0 in equation (5) as

$$
P_0 = \left[1 + \frac{1-\pi_0}{\pi_0}\frac{1}{\sqrt{1+n\xi}} \times \exp\left(\frac{n\bar{y}^2}{2\sigma^2(1+1/(n\xi))}\right)\right]^{-1} = \left[1 + ae^{b\bar{y}^2}\right]^{-1}, \tag{7}
$$

with $a = \frac{1-\pi_0}{\pi_0}/\sqrt{1+n\xi}$ and $b = n/[2\sigma^2(1+1/(n\xi))]$. Any P_0 in the interval $(0, 1/(1+a))$ corresponds to two $\bar{y}$:

$$
\bar{y}_i = \pm\sqrt{\frac{1}{b}\log\frac{1/P_0 - 1}{a}}, i = 1, 2. \tag{8}
$$

For each of them, the Jacobi determinant is

$$
\begin{aligned}
|P_0'(\bar{y}_i)| &= \left|\frac{dP_0}{d\bar{y}_i}\right| = \frac{ae^{b\bar{y}_i^2} \times 2b|\bar{y}_i|}{(1+ae^{b\bar{y}_i^2})^2}\\
&= 2P_0(1-P_0) \times \sqrt{b\log\frac{1/P_0-1}{a}}, i = 1, 2.
\end{aligned}
\tag{9}
$$

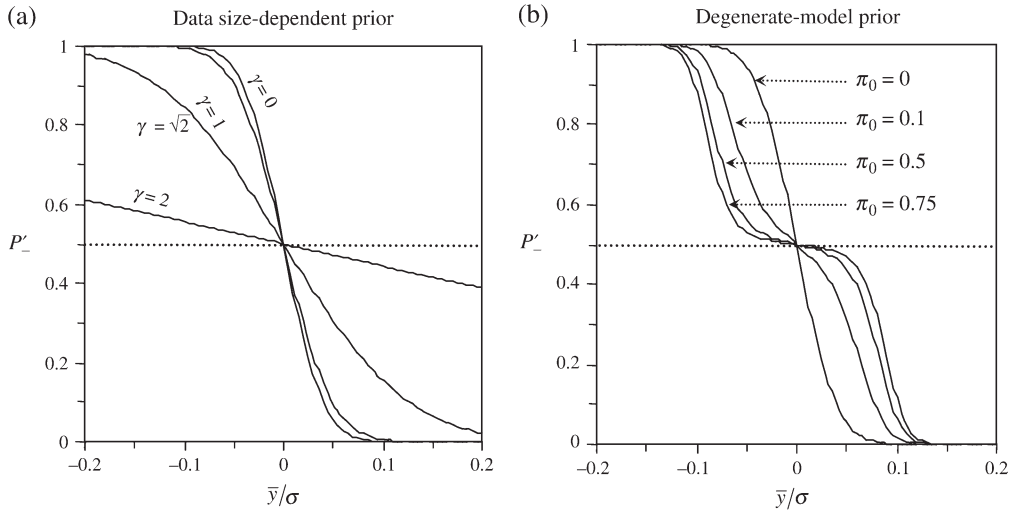

FIG. 4.—The sensitivity of posterior model probability P_- to the prior in the fair-balance problem. (a) The prior is specified as $\theta \sim N(0, \xi\sigma^2)$ with $\xi = c/n^\gamma$, and the posterior P_- is calculated using equation (1). (b) A prior probability π_0 is assigned to the degenerate model H_0: $\theta = 0$, while $\theta \sim N(0, c\sigma^2)$ under models H_- and H_+. The posterior model probabilities P_0 and P_- are calculated using equation (5), and then $P'_- = P_0/2 + P_-$ is used in the plot. In both (a) and (b), $n = 1000$ and $c = 2$ are fixed.

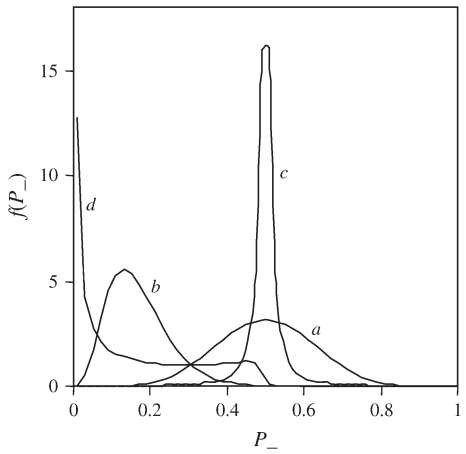

FIG. 5.—The effect of prior on $f(P_-)$, the density of posterior model probability P_- in the fair-balance problem. The sample size is $n = 1000$. The true parameter value is $\theta_0 = 0$ in (a) and (c) and 0.1 in (b) and (d). In (a) and (b), the prior is specified as $\theta \sim N(0, \xi\sigma^2)$ where $\xi = 2/n^\gamma$ with $\gamma = \sqrt{2}$, and $f(P_-)$ is calculated using equation (2). In (c) and (d), the prior probability $\pi_0 = \frac{1}{3}$ is assigned to the degenerate model $\theta = 0$, while $\theta \sim N(0, 2\sigma^2)$ in used under models H_- and H_+. One million data sets are simulated by generating $\bar{y} \sim N(\theta_0, \sigma^2/n)$, and P_0 and P_- are calculated by equation 5. Then $P'_- = P_0/2 + P_-$ is used to construct the histogram and to estimate the density $f(P_-)$.

Thus

$$f(P_0|\theta_0) = \frac{\phi(\bar{y}_1|\theta_0)}{|P_0'(\bar{y}_1)|} + \frac{\phi(\bar{y}_2|\theta_0)}{|P_0'(\bar{y}_2)|} = \frac{\exp\left\{-\frac{n}{2\sigma^2}\left(\sqrt{\frac{1}{b}\log\frac{1/P_0-1}{a}} - \theta_0\right)^2\right\} + \exp\left\{-\frac{n}{2\sigma^2}\left(\sqrt{\frac{1}{b}\log\frac{1/P_0-1}{a}} + \theta_0\right)^2\right\}}{\sqrt{2\pi\sigma^2/n} \times 2bP_0(1-P_0) \times \sqrt{\frac{1}{b}\log\frac{1/P_0-1}{a}}}$$
$$= \frac{\exp\left\{-\frac{1}{2}\left(\sqrt{2(1+1/(n\xi))B} - \frac{\sqrt{n}\theta_0}{\sigma}\right)^2\right\} + \exp\left\{-\frac{1}{2}\left(\sqrt{2(1+1/(n\xi))B} + \frac{\sqrt{n}\theta_0}{\sigma}\right)^2\right\}}{2P_0(1-P_0)\sqrt{\frac{\pi}{1+1/(n\xi)}B}}, \quad (10)$$

where $B = \log\left\{\frac{1-P_0}{P_0}\frac{\pi_0}{1-\pi_0}\sqrt{1+n\xi}\right\}$, and $0 < P_0 < 1/\left[1 + \frac{1-\pi_0}{\pi_0}/\sqrt{1+n\xi}\right]$. Note that $f(P_0|\theta_0)$ depends on P_0, π_0, $n\xi$, and $\sqrt{n}\theta_0/\sigma$.

If $\theta_0 = 0$, equation (10) reduces to

$$f(P_0) = \frac{2\exp\left\{-(1+\frac{1}{n\xi})B\right\}}{2P_0(1-P_0)\sqrt{\frac{\pi}{1+1/(n\xi)}B}} = \frac{\left[\frac{1-P_0}{P_0}\frac{\pi_0}{1-\pi_0}\sqrt{1+n\xi}\right]^{-1-1/(n\xi)}}{P_0(1-P_0)\sqrt{\frac{\pi}{1+1/(n\xi)}\log\left\{\frac{1-P_0}{P_0}\frac{\pi_0}{1-\pi_0}\sqrt{1+n\xi}\right\}}}. \quad (11)$$

This density is specified by π_0 and $n\xi$.

Equations (10) and (11) can be used to confirm that as long as $\pi_0 > 0$, $f(P_0)$ converges to a point mass at 1 when $n \to \infty$ if $\theta_0 = 0$ (so that H_0 is true), and that if $\theta_0 \neq 0$ (so that H_0 is false), $f(P_0)$ will converge to 0, in which case one of P_- and P_+ (the one corresponding to the true model) will converge to 1. In other words, the probability for the correct model always converges to 1 when $n \to \infty$. This is a special case of Dawid's (1999) general proof of the consistency of Bayesian model selection.

Here I consider the prior probability π_0 as a way of resolving the fair-balance paradox and treat P_0 as equal support for H_- and H_+. Thus (P_0, P_-, P_+) calculated from any data set are converted to $(P'_-, P'_+) = (P_0/2 + P_-, P_0/2 + P_+)$. Then if $\theta_0 = 0$, we have $P_0 \to 1$, so that $P'_- \to \frac{1}{2}$ and $P'_+ \to \frac{1}{2}$. Similarly, if $\theta_0 \neq 0$, we have $P_0 \to 0$, so that one of P'_- and P'_+ will approach 1. It is clear that use of the prior probability π_0 resolves the fair-balance paradox.

Nevertheless, the Bayesian analysis may be very sensitive to the value of π_0, and this sensitivity appears to be the nature of the problem. For example, for a data set of size $n = 1{,}000$ with $\bar{y}/\sigma = -0.05$, we have P'_- to be 0.943, 0.683, 0.560 and 0.532, if $\pi_0 = 0$, 1/10, $\frac{1}{3}$, and $\frac{1}{2}$, respectively (fig. 4b). Furthermore, while $f(P'_-)$ converges to a point mass at 1, $\frac{1}{2}$, and 0 if the true $\theta_0 < 0$, $= 0$, and > 0, respectively, the convergence may be at very different rates depending on π_0. Curves c & d in figure 5 show the density for $\theta_0 = 0$ and 0.1, with $n = 1{,}000$ when the prior $\pi_0 = \frac{1}{3}$ is used. This prior produces high posterior probabilities for the true model much more often and may be considered more powerful than the data size-dependent prior with $\gamma = \sqrt{2}$, that is, $\theta \sim N(0, 2\sigma^2/n^{\sqrt{2}})$ (curves a & b in fig. 5).

Lindley's Paradox

If we do not distinguish between models H_- and H_+ and define $P_1 = 1 - P_0 = P_- + P_+$, the problem becomes one of comparing a sharp null hypothesis H_0: $\theta = 0$ with a composite alternative hypothesis H_1: $\theta \neq 0$. This is the case for Lindley's (1957; see also Jeffreys 1939) paradox. If $\sqrt{n}\bar{y}/\sigma$ is fixed but $n \to \infty$, then $P_0 \to 1$ (eq. 7). Lindley's paradox refers to the observation that in a data set, $\sqrt{n}\bar{y}/\sigma$ may differ sufficiently from 0 for H_0 to be rejected by a significance test, while Bayesian analysis of the same data strongly supports H_0 with posterior probability $P_0 \approx 1$. Thus significance test and Bayesian analysis draw opposite conclusions from the same data. Indeed, if large data sets are generated under the null model, such contradictions will occur in ~5% of data sets if the significance test is

conducted at the 5% level. As discussed above, if H_0 is true and n is large, $P_0 \approx 1$ in nearly every data set, but the significance test will still reject the true null hypothesis 5% of the time. This result appears to suggest flaws in the methodology of significance test, as claimed by some Bayesian statisticians (e.g., Good 1982, p. 342; Press 2003, pp. 220–225; Berger 1985, pp. 144–157), rather than in Bayesian analysis, as suggested by, e.g., Bernardo (1980) and Shafer (1982). Furthermore, Davison (2003, pp. 586–587) and Cox (2006, pp. 42–43, 106–107) (see also, Bartlett 1957) suggested the use of $\xi = c/n$, so that $\theta \sim N(0, c\sigma^2/n)$, to resolve Lindley's paradox. By the criteria used here, this prior is not acceptable as it causes $f(P_0)$ to fail to converge to the point mass at 1 when $\theta_0 = 0$ (see equation 11)!

Nevertheless, whatever the true model or the observed data ($\bar{y}$), P_0 can be made arbitrarily close to 1 by the use of a diffuse prior or a large ξ, as $P_0 \to 1$ when $\xi \to \infty$ in equation (7). Bayesian analysis in this case is extremely sensitive to the prior.

Bayesian Tree Estimation in the Three-Species Case

The Tree Problem

There are three (rooted) binary trees for three species (fig. 2): $\tau_1 = ((12)3)$, $\tau_2 = ((23)1)$, and $\tau_3 = ((31)2)$. We consider binary characters, which evolve at a constant rate according to a stationary Markov process. The data are counts n_0, n_1, n_2, n_3 of site patterns *xxx*, *xxy*, *yxx*, and *xyx*. Let $x_i = n_i/n$, $i = 0, 1, 2, 3$, be the proportions of the site patterns. The data may be represented as $\mathbf{n} = \{n_1, n_2, n_3\}$ or $\mathbf{x} = \{x_1, x_2, x_3\}$, with n to be the total number of sites.

Under tree τ_1, with branch lengths t_0 and t_1 (fig. 2), the probabilities of observing the four site patterns are

$$\begin{aligned} p_0(t_0, t_1) &= \tfrac{1}{4} + \tfrac{1}{4}e^{-4t_1} + \tfrac{1}{2}e^{-4(t_0+t_1)}, \\ p_1(t_0, t_1) &= \tfrac{1}{4} + \tfrac{1}{4}e^{-4t_1} - \tfrac{1}{2}e^{-4(t_0+t_1)}, \\ p_2(t_0, t_1) &= \tfrac{1}{4} - \tfrac{1}{4}e^{-4t_1}, \\ p_3(t_0, t_1) &= p_2(t_0, t_1) \end{aligned} \tag{12}$$

(Yang 2000). As $0 \le t_0, t_1 \le \infty$, we have $p_0 \ge p_1 \ge p_2 = p_3 \ge 0$ and $p_0 + p_1 + p_2 + p_3 = 1$. The likelihoods under the three trees are

$$\begin{aligned} f(\mathbf{n}|\tau_1, t_0, t_1) &= p_0^{n_0} p_1^{n_1} p_2^{n_2+n_3}, \\ f(\mathbf{n}|\tau_2, t_0, t_1) &= p_0^{n_0} p_1^{n_2} p_2^{n_3+n_1}, \\ f(\mathbf{n}|\tau_3, t_0, t_1) &= p_0^{n_0} p_1^{n_3} p_2^{n_1+n_2}. \end{aligned} \tag{13}$$

We assign prior probability $\frac{1}{3}$ for each binary tree and exponential priors with means μ_0 and μ_1 for t_0 and t_1: $f(t_0) = \exp\{-t_0/\mu_0\}/\mu_0$ and $f(t_1) = \exp\{-t_1/\mu_1\}/\mu_1$. The exponential priors appear more sensible than uniform priors since most branch lengths in real trees are small while very large branch lengths are rare. The marginal likelihood under tree τ_i is

$$\begin{aligned} M_i &= f(\mathbf{n}|\tau_i) \\ &= \int_0^\infty \int_0^\infty f(t_0) f(t_1) f(\mathbf{n}|\tau_i, t_0, t_1)\ dt_0\, dt_1, i = 1, 2, 3. \end{aligned} \tag{14}$$

The posterior tree probability is

$$P_i = \frac{M_i}{M_1 + M_2 + M_3}, i = 1, 2, 3. \tag{15}$$

Thus analysis of each data set requires evaluation of three two-dimensional integrals. (In contrast, the case of four species and no molecular clock requires evaluation of three five-dimensional integrals.) Yang and Rannala (2005) used Mathematica (Wolfram 2003) to calculate the integrals of equation (14) numerically. This is found to be unreliable in large data sets, with $n \ge 5{,}000$, say. A difficulty is that the integrand is nearly a spike at its mode.

Two ideas appear promising. The first is to use the site pattern probabilities as parameters in the binary tree instead of t_0 and t_1 and construct conjugate priors on them. The second is to use large-sample approximations. The latter is explored in this study.

Approximate Calculation of Posterior Probabilities for Trees

We use Laplacian expansion (Copson 1965, pp. 36–47; Bender and Orszag 1999, pp. 261–276) to approximate the integral M_1 for tree τ_1 (eq. 14). The integrals M_2 and M_3 for trees τ_2 and τ_3 are calculated by a permutation of the counts n_1, n_2, n_3. In a typical Bayesian estimation problem under a well-specified model, the likelihood function and the posterior density can be quite accurately approximated using a normal density in large data sets (Lindley 1980; Tierney and Kadane 1986). However, phylogenetic trees are different models (e.g., Yang et al. 1995). For any given data set, the maximum likelihood estimate (MLE) of t_0 is zero in at least one tree, in which case the normal approximation breaks down. Instead the tedious algorithms presented below were derived by trial and error, with intensive testing in comparison with Mathematica.

Rewrite equation (14) as

$$M_1 = \int_0^\infty \int_0^\infty f(t_0, t_1) e^{nh(t_0,t_1)}\ dt_0\, dt_1, \tag{16}$$

where $f(t_0, t_1) = f(t_0)f(t_1)$ is the prior for t_0 and t_1, and

$$\begin{aligned} h(t_0, t_1) &= x_0 \log(p_0) + x_1 \log(p_1) \\ &\quad + (1 - x_0 - x_1)\log(1 - p_0 - p_1). \end{aligned} \tag{17}$$

Here $x_i = n_i/n$ is the observed site pattern frequencies. Define $h_i = \partial h/\partial t_i$, $h_{ij} = \partial^2 h/\partial t_i \partial t_j$, etc., to be the derivatives evaluated at the MLEs $\hat{t}_0$ and $\hat{t}_1$ (see Appendix). Let $\mathbf{H} = \{h_{ij}\}$ be the Hessian matrix. If $\mathbf{H}$ is positive-definite, we let

$$\Sigma = (-n\mathbf{H})^{-1} = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix},$$

with the determinant

$$|\Sigma| = \sigma_0^2 \sigma_1^2 (1 - \rho^2) = n^{-2} |\mathbf{H}|^{-1},$$

where $\rho = \sigma_{01}/(\sigma_0\sigma_1)$. We use the first few terms in the Taylor expansions of f and h as approximations

$$f(t_0,t_1) \simeq f(\hat{t}_0,\hat{t}_1),$$
$$h(t_0,t_1) \simeq h(\hat{t}_0,\hat{t}_1) + h_0(t_0-\hat{t}_0) + h_1(t_1-\hat{t}_1) + \tfrac{1}{2}\sum_{i,j} h_{ij}(t_i-\hat{t}_i)(t_j-\hat{t}_j). \quad (18)$$

The integral of equation (16) is the volume of the solid between the $f\cdot e^{nh}$ surface above the t_0–t_1 plane in the quarterplane $t_0 > 0$, $t_1 > 0$. We consider three cases, depending on whether $\hat{t}_0 > 0$ and whether $\partial h/\partial\hat{t}_0 = 0$ (Yang 2000, Tables 2 and 3). We assume that $x_0 > \frac{1}{4}$.

Case I: $x_1 > (1 - x_0)/3$. We have $\hat{t}_0 = -\frac{1}{4}\log(x_0 - x_1) + \frac{1}{4}\log(2(x_0 + x_1) - 1) > 0$, and $\hat{t}_1 = -\frac{1}{4}\log(2(x_0 + x_1) - 1) > 0$, with $\partial h/\partial\hat{t}_0 = \partial h/\partial\hat{t}_1 = 0$ (Yang 2000, Table 3). The integral can then be approximated by the volume at the neighborhood of the MLE $(\hat{t}_0,\hat{t}_1)$, where the likelihood surface is nearly that of a bivariate normal density function.

$$M_1 \simeq f(\hat{t}_0,\hat{t}_1)e^{nh(\hat{t}_0,\hat{t}_1)} \times \int_0^\infty\int_0^\infty \exp\left\{-\tfrac{1}{2}(\mathbf{t}-\hat{\mathbf{t}})^T(-n\mathbf{H})(\mathbf{t}-\hat{\mathbf{t}})\right\} dt_0\,dt_1 = f(\hat{t}_0,\hat{t}_1)e^{nh(\hat{t}_0,\hat{t}_1)}\cdot 2\pi|\Sigma|^{1/2}. \quad (19)$$

As discussed by Lindley (1980, equation 2), a few more terms may be used in the Taylor expansion of f and h, but this was found to lead to minimal improvement to the approximation. More importantly, the MLE $\hat{t}_0$ is often close to 0, or $\hat{t}_0/\sigma_0$ is small (say, <3), in which case equation (19) is not very reliable. The bivariate normal integral can then be calculated using the algorithm of Drezner and Wesolowsky (1990), which was found to produce good results.

Case II: $x_1 = (1 - x_0)/3$. We have $\hat{t}_0 = 0$, $\hat{t}_1 = -\frac{1}{4}\log((4x_0 - 1)/3) > 0$, with $\partial h/\partial\hat{t}_0 = \partial h/\partial\hat{t}_1 = 0$. The integral is then half that in case I as the volume above the half plane $t_0 < 0$ is missing.

$$M_1 \simeq f(\hat{t}_0,\hat{t}_1)e^{nh(\hat{t}_0,\hat{t}_1)}\cdot\pi|\Sigma|^{1/2}. \quad (20)$$

Case III: $x_1 < (1 - x_0)/3$. We have $\hat{t}_0 = 0$ and $\hat{t}_1 = -\frac{1}{4}\log((4x_0 - 1)/3) > 0$, with $\partial h/\partial\hat{t}_0 < 0$ and $\partial h/\partial\hat{t}_1 = 0$. This situation is complex, and is broken into two cases, depending on whether the Hessian matrix $\mathbf{H}$ is positive definite.

In case IIIa, $\mathbf{H}$ is positive definite. We then use all second-order terms in the Taylor expansion of h.

$$M_1 \simeq \int_0^\infty\int_0^\infty f(t_0,t_1)\exp\{nh(\hat{t}_0,\hat{t}_1) + nh_0t_0 + \tfrac{n}{2}\sum_{i,j} h_{ij}(t_i-\hat{t}_i)(t_j-\hat{t}_j)\}\,dt_0\,dt_1 = f(\hat{t}_0,\hat{t}_1)e^{nh(\hat{t}_0,\hat{t}_1)}\times\int_0^\infty\int_0^\infty e^{nh_0t_0}\times\exp\left\{\tfrac{n}{2}\sum_{ij} h_{ij}(t_i-\hat{t}_i)(t_j-\hat{t}_j)\right\} dt_0\,dt_1. \quad (21)$$

Apply the variable transform $y_0 = t_0/\sigma_0$, $y_1 = (t_1 - \hat{t}_1)/\sigma_1$, and we have

$$M_1 \simeq f(\hat{t}_0,\hat{t}_1)e^{nh(\hat{t}_0,\hat{t}_1)}\int_0^\infty e^{nh_0t_0}\int_{-\hat{t}_1/\sigma_1}^\infty \exp\left\{-\tfrac{1}{2(1-\rho^2)}(y_0^2 - 2\rho y_0y_1 + y_1^2)\right\}\times\sigma_0\sigma_1 dt_1\,dt_0$$
$$= f(\hat{t}_0,\hat{t}_1)e^{nh(\hat{t}_0,\hat{t}_1)}\sigma_0\sigma_1\int_0^\infty e^{nh_0\sigma_0y_0}\left[\int_{-\hat{t}_1/\sigma_1}^\infty \exp\left\{-\tfrac{1}{2(1-\rho^2)}\left[(y_1-\rho y_0)^2 + (y_0^2 - \rho^2y_0^2)\right]\right\} dy_1\right] dy_0$$
$$= f(\hat{t}_0,\hat{t}_1)e^{nh(\hat{t}_0,\hat{t}_1)}\sigma_0\sigma_1\sqrt{2\pi(1-\rho^2)}\int_0^\infty e^{-\frac{1}{2}y_0^2}\Phi\left(\frac{\hat{t}_1/\sigma_1 + \rho y_0}{\sqrt{1-\rho^2}}\right)e^{nh_0\sigma_0y_0}\,dy_0. \quad (22)$$

If $-nh_0\sigma_0 \gg 1$, we may apply Watson's Lemma to approximate the integral in equation (22). Write this as $\int_0^\infty q(y)e^{-cy}\,dy$, where $q(y) = e^{-y^2/2}\Phi(a + by)$, with $a = \hat{t}_1/(\sigma_1\sqrt{1-\rho^2})$, $b = \rho/\sqrt{1-\rho^2}$, and $c = -nh_0\sigma_0$. From the MacLaurin expansion of $q(y)$, we have

$$\int_0^\infty q(y)e^{-cy}\,dy = \int_0^\infty\sum_{k=0}^\infty \frac{q^{(k)}(0)y^k}{k!}e^{-cy}\,dy = \sum_{k=0}^\infty \frac{q^{(k)}(0)}{c^{k+1}}, \quad (23)$$

where $q^k(0)$ is the kth derivative of q, evaluated at $y = 0$. The first few derivatives are as follows:

$$q(0) = \Phi(a),$$
$$q'(0) = b\Phi(a),$$
$$q''(0) = -\Phi(a) - ab^2\phi(a),$$
$$q'''(0) = -3b\phi(a) + b^3(a^2-1)\phi(a), \quad (24)$$

where $\phi(\cdot)$ is the probability density function (p.d.f.) of a standard normal variate. Thus

$$M_1 \simeq f(\hat{t}_0,\hat{t}_1)e^{nh(\hat{t}_0,\hat{t}_1)}\times\sigma_0\sigma_1\sqrt{2\pi(1-\rho^2)} \times\left[\frac{\Phi(a)}{c}\left(1-\frac{1}{c^2}\right) + \frac{b\phi(a)}{c^2}\left(1 - \frac{ab}{c} - \frac{3-b^2(a^2-1)}{c^2}\right)\right]. \quad (25)$$

The algorithm of Hill (1973) is used to calculate $\Phi(\cdot)$.

However, if $c = -nh_0\sigma_0$ is small (< 1), as may be the case if h_0 is nearly zero, equation (25) is unreliable. Then I use the Gauss-Legendre quadrature to calculate the one-dimensional integral of equation (22) numerically, which was found to produce reliable results.

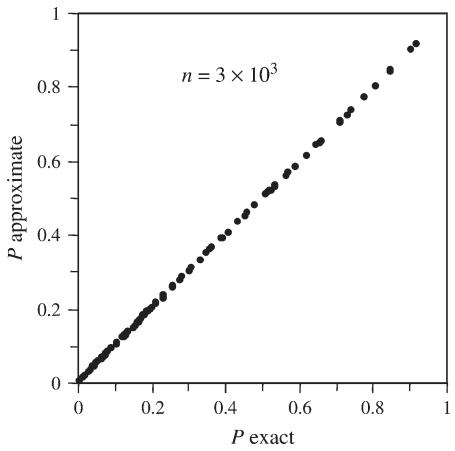

FIG. 6.—The posterior probability P_1 for tree τ_1 calculated using exact (Mathematica) and approximate methods in 100 data sets, simulated using the star tree with $t = 0.2$. The sequence length is $n = 3 \times 10^3$. The prior means are $\mu_0 = 0.1$ and $\mu_0 = 0.2$.

Table 2
The Log Marginal Likelihood log(M_i) and the Posterior Probabilities for the Three Trees Calculated Using Exact (above) and Approximate (below) Methods

n	Log Marginal Probability	Posterior Probabilities
300	−285.42 −286.00 −286.43	0.521 0.291 0.188
	−285.43 −286.01 −286.44	0.521 0.290 0.188
500	−473.78 −474.58 −475.13	0.586 0.263 0.151
	−473.79 −474.59 −475.14	0.587 0.262 0.152
1,000	−944.21 −945.48 −946.22	0.706 0.199 0.095
	−944.22 −945.49 −946.22	0.707 0.198 0.095
3,000	−2,824.62 −2,827.46 −2,828.57	0.928 0.054 0.018
	−2,824.62 −2,827.46 −2,828.58	0.928 0.054 0.018
5,000	−4,704.53 −4,708.86 −4,710.18	0.984 0.013 0.003
	−4,704.53 −4,708.86 −4,710.18	0.984 0.013 0.003
10,000	−9,403.77 −9,411.78 −9,413.41	0.9996 0.0003 0.0001
	−9,403.77 −9,411.79 −9,413.41	0.9996 0.0003 0.0001

The data are $x_1 = 0.11$, $x_2 = 0.10$, and $x_3 = 0.09$ (with $x_0 = 0.70$), while the number of sites is n. The prior is exponential with means $\mu_0 = 0.1$ and $\mu_1 = 0.2$. Equations (19), (20), and (25) are used to calculate log(M_i) for the three trees, respectively.

In case IIIb, $x_1 < (1 - x_0)/3$, so that $\hat{t}_0 = 0$ and $\hat{t}_1 > 0$, with $h_0 = \partial h/\partial \hat{t}_0 < 0$ and $h_1 = \partial h/\partial \hat{t}_1 = 0$, but **H** is not positive-definite. This case occurs mainly when the data are very unlikely on the tree and h_0 is very negative. We then use the linear term for t_0 and quadratic term for t_1 in the Taylor expansion of h, as follows.

these two cases about half of the time. Cases II (equation 20) and IIIb (equation 27) are rarely encountered.

The above discussion assumes that the prior on branch lengths are fixed, with μ_0 and μ_1 to be fixed constants. When μ_0 depends on the data size n, some modifications to the above algorithm are necessary.

$$\begin{aligned} M_1 &\simeq \int_0^\infty \int_0^\infty f(t_0, t_1) \exp\left\{ nh(\hat{t}_0, \hat{t}_1) + nh_0 t_0 + nh_{01} t_0 (t_1 - \hat{t}_1) + \tfrac{1}{2} nh_{11} (t_1 - \hat{t}_1)^2 \right\} \mathrm{d}t_0 \, \mathrm{d}t_1 \\ &= f(\hat{t}_0, \hat{t}_1) \mathrm{e}^{nh(\hat{t}_0, \hat{t}_1)} \times \int_0^\infty \mathrm{e}^{\frac{1}{2} nh_{11} (t_1 - \hat{t}_1)^2} \int_0^\infty \mathrm{e}^{(nh_0 + nh_{01}(t_1 - \hat{t}_1)) t_0} \mathrm{d}t_0 \, \mathrm{d}t_1 \\ &= f(\hat{t}_0, \hat{t}_1) \mathrm{e}^{nh(\hat{t}_0, \hat{t}_1)} \times \int_0^\infty \mathrm{e}^{\frac{1}{2} nh_{11} (t_1 - \hat{t}_1)^2} \frac{1}{-nh_0 - nh_{01}(t_1 - \hat{t}_1)} \mathrm{d}t_1 . \end{aligned} \tag{26}$$

Change variables from t_1 to $z = (t_1 - \hat{t}_1)/\sigma_1$, where $\sigma_1 = 1/\sqrt{-nh_{11}}$.

The exact calculation using Mathematica is reliable for small data sets, and unstable for large ones (say, with

$$\begin{aligned} M_1 &\simeq f(\hat{t}_0, \hat{t}_1) \mathrm{e}^{nh(\hat{t}_0, \hat{t}_1)} \times \frac{\sigma_1}{-nh_0} \int_{-\hat{t}_1/\sigma_1}^\infty \left[1 - \tfrac{h_{01}\sigma_1}{h_0} z + \left(\tfrac{h_{01}\sigma_1}{h_0} z \right)^2 - \left(\tfrac{h_{01}\sigma_1}{h_0} z \right)^3 + \ldots \right] \mathrm{e}^{-z^2/2} \mathrm{d}z \\ &\simeq f(\hat{t}_0, \hat{t}_1) \mathrm{e}^{nh(\hat{t}_0, \hat{t}_1)} \times \frac{\sqrt{2\pi}\sigma_1}{-nh_0} \left[1 + \left(\tfrac{h_{01}\sigma_1}{h_0} \right)^2 + 3 \left(\tfrac{h_{01}\sigma_1}{h_0} \right)^4 + 15 \left(\tfrac{h_{01}\sigma_1}{h_0} \right)^6 \right]. \end{aligned} \tag{27}$$

Here $\hat{t}_1/\sigma_1 \gg 1$, and thus the integral from $-\hat{t}_1/\sigma_1$ to ∞ is nearly the same as from $-\infty$ to ∞, while $1/(1 + a) = 1 - a + a^2 - a^3 + \ldots$ when $|a| = h_{01}\sigma_1/h_0 < 1$. The last equality uses the result that if z is a random variable from the standard normal distribution, $E(z^k) = 0$ for odd k or $(k - 1)(k - 3) \cdot 3 \cdot 1$ for even k (e.g., Johnson et al. 1994, p. 89).

Suppose in the data set, $n_1 > n_2 > n_3$. Then $M_1 > M_2 > M_3$. Calculation of M_1 makes use of equation (19) for case I, and calculation of M_3 makes use of equations (22) or (27) for case IIIa. Calculation of M_2 uses each of

$n > 5{,}000$). The approximate calculation is the opposite. It is reliable for large data sets only, say with $n \geq 1{,}000$. Figure 6 shows posterior tree probabilities calculated using the two methods, while table 2 shows the effect of sample size n on the approximation. On a 3.2GHz Pentium IV, analyzing 10^5 data sets took a few seconds using the approximate method and ~15 days using Mathematica. Both methods are much faster than MCMC for this small problem. The approximation allows us to calculate posterior tree probabilities for arbitrarily large data sets.

Simulation of Data

Consider simulation of data sets under tree τ_1 with given branch lengths t_0 and t_1. Simulation under the star tree τ_0 can be done using the same algorithm by fixing $t_0 = 0$. The counts of sites follow a multinomial distribution with four cells: $MN_4(n; p_0, p_1, p_2, p_2)$, with cell probabilities given in equation (12). For large n, the data have approximately a trivariate normal distribution: $\mathbf{n} = (n_1, n_2, n_3) \sim N_3(n\theta_0, n\mathbf{S}_0)$, where

$$n\theta_0 = n\begin{pmatrix} p_1 \\ p_2 \\ p_2 \end{pmatrix}, n\mathbf{S}_0 = n\begin{pmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_2 \\ -p_1p_2 & p_2(1-p_2) & -p_2^2 \\ -p_1p_2 & -p_2^2 & p_2(1-p_2) \end{pmatrix}. \tag{28}$$

We have $|n\mathbf{S}_0| = n^3 p_0 p_1 p_2 p_2$, and

$$(n\mathbf{S}_0)^{-1} = \frac{1}{np_0p_1p_2}\begin{pmatrix} p_2(1-2p_2) & p_1p_2 & p_1p_2 \\ p_1p_2 & p_1(1-p_1-p_2) & p_1p_2 \\ p_1p_2 & p_1p_2 & p_1(1-p_1-p_2) \end{pmatrix}. \tag{29}$$

The normal density is

$$\phi(n_1,n_2,n_3|p_1,p_2) = \frac{1}{\sqrt{(2\pi n)^3 p_0 p_1 p_2^2}} \exp\left\{-\frac{n}{2}(\mathbf{x}-\theta_0)^T \mathbf{S}_0^{-1}(\mathbf{x}-\theta_0)\right\}, \tag{30}$$

where $(\mathbf{x} - \boldsymbol{\theta}_0)^T = (x_1 - p_1, x_2 - p_2, x_3 - p_2)^T$, and T is the transpose.

The Cholesky decomposition of the variance matrix is given as $n\mathbf{S}_0 = \mathbf{L}\mathbf{L}^T$, with

$$\mathbf{L} = \sqrt{n} \cdot \begin{pmatrix} a & 0 & 0 \\ b & d & 0 \\ c & e & f \end{pmatrix}, \tag{31}$$

where

$$\begin{aligned} a &= \sqrt{p_1(1-p_1)}, \\ b &= c = -p_1p_2/\sqrt{p_1(1-p_1)}, \\ d &= \sqrt{p_2(1-p_1-p_2)/(1-p_1)}, \\ e &= -p_2^2/\sqrt{p_2(1-p_1)(1-p_1-p_2)}, \\ f &= \sqrt{p_2(1-p_1-2p_2)/(1-p_1-p_2)}. \end{aligned} \tag{32}$$

Thus to generate a data set, we generate three independent $N(0, 1)$ random variables z_1, z_2, and z_3 to form $\mathbf{z} = (z_1, z_2, z_3)^T$. Then $\mathbf{n} = n\boldsymbol{\theta}_0 + \mathbf{L}\mathbf{z}$ will be the desired counts of site patterns.

Two Strategies to Resolve the Star-tree Paradox

We now consider the two priors for resolving the star-tree paradox, following our discussions of the fair-coin and fair-balance paradoxes above. The first is to let the prior mean for the internal branch length approach zero when the data size increases, and the second is to assign a nonzero probability π_0 for the degenerate star tree.

Data Size-Dependent Prior. This forces the mean μ_0 in the prior for internal branch length t_0 to approach 0, or, equivalently, to force the probabilities of the three variable site patterns p_1, p_2, and p_3 to approach equality ($p_1 = p_2 = p_3$), when $n \to \infty$ (Yang and Rannala 2005). In the fair-coin problem, $1 - \theta$ and θ are the two cell probabilities in the multinomial (binomial) distribution, the models of negative and positive bias are specified as H_-: $1 - \theta > \theta$ and H_-: $1 - \theta < \theta$ while the fair-coin model is H_0: $1 - \theta = \theta$. The distance between H_- (say) and H_0 may be measured by $|1 - \theta - \theta| = |1 - 2\theta|$. It was determined that the prior should force $E(1 - 2\theta)^2$ or the variance of θ to approach 0 faster than $1/n$ but more slowly than $1/n^2$. In the tree problem, the binary tree, say τ_1, is represented by $p_1 > p_2 = p_3$ while the star tree τ_0 is $p_1 = p_2 = p_3$, where p_1, p_2, p_3 are three cell probabilities in a multinomial distribution. The distance between τ_1 and τ_0 can be measured by $|p_1 - p_2|$, and by analogy with the fair-coin problem, we require the prior on branch lengths t_0 and t_1 should force $E(p_1 - p_2)^2$ to approach 0 faster than $1/n$ but more slowly than $1/n^2$.

Let $\mu_0 = c/n^\gamma$ with $\gamma > 0$. The prior for branch lengths t_0 and t_1 is given by the independent exponential distributions

$$f(t_0, t_1) = \frac{1}{\mu_0} e^{-t_0/\mu_0} \times \frac{1}{\mu_1} e^{-t_1/\mu_1}. \tag{33}$$

In place of t_0 and t_1, we use p_0 and $\delta = p_1 - p_2$ as the new parameters in the binary tree; the two sets of parameters are related by equation (12). The prior distribution of p_0 and δ is obtained from equation (33) through a variable transform as

$$\begin{aligned} f(p_0, \delta) = \frac{1}{6\mu_0\mu_1} &\left(\frac{4p_0 - 1 + 4\delta}{3}\right)^{\frac{1}{4\mu_1}-2} \\ &\times \left(1 + \frac{6\delta}{4p_0 - 1 - 2\delta}\right)^{-\frac{1}{4\mu_0}+1}, \\ 0 < \delta < 2p_0 - \frac{1}{2}&, \delta < 1 - p_0. \end{aligned} \tag{34}$$

We have $E(\delta^2) = \int\int \delta^2 f(p_0, \delta)\, d\delta\, dp_0 \simeq \mu_0^2/\left(\frac{1}{8} + \mu_1\right)$. Thus μ_0 should approach 0 faster than $1/\sqrt{n}$ but more slowly than $1/n$; in other words we require $\frac{1}{2} < \gamma < 1$ in $\mu_0 = c/n^\gamma$.

Degenerate-Model Prior π_0. We assign a prior probability $\pi_0 > 0$ for the star tree τ_0, while the three binary trees are assigned prior probabilities $\pi_1 = \pi_2 = \pi_3 = (1 - \pi_0)/3$ (Lewis, Holder, and Holsinger 2005). The branch length t in the star tree is assigned the prior $f(t) = \exp\{-t/\mu_1\}/\mu_1$. The marginal likelihood M_0 under τ_0 is a one-dimensional integral over t, similar to equation (14). This is reliably calculated by approximating the likelihood with a normal density, similarly to the calculation with equation (19). The marginal likelihoods for the three binary trees M_1, M_2, and M_3 are calculated as before. Then $\pi_i M_i$, $i = 0$, 1, 2, 3, are rescaled to sum to one to give the posterior probabilities for all four trees. As the star tree is a special case of the three binary trees with one fewer parameter, all four trees are correct when the data are generated from the

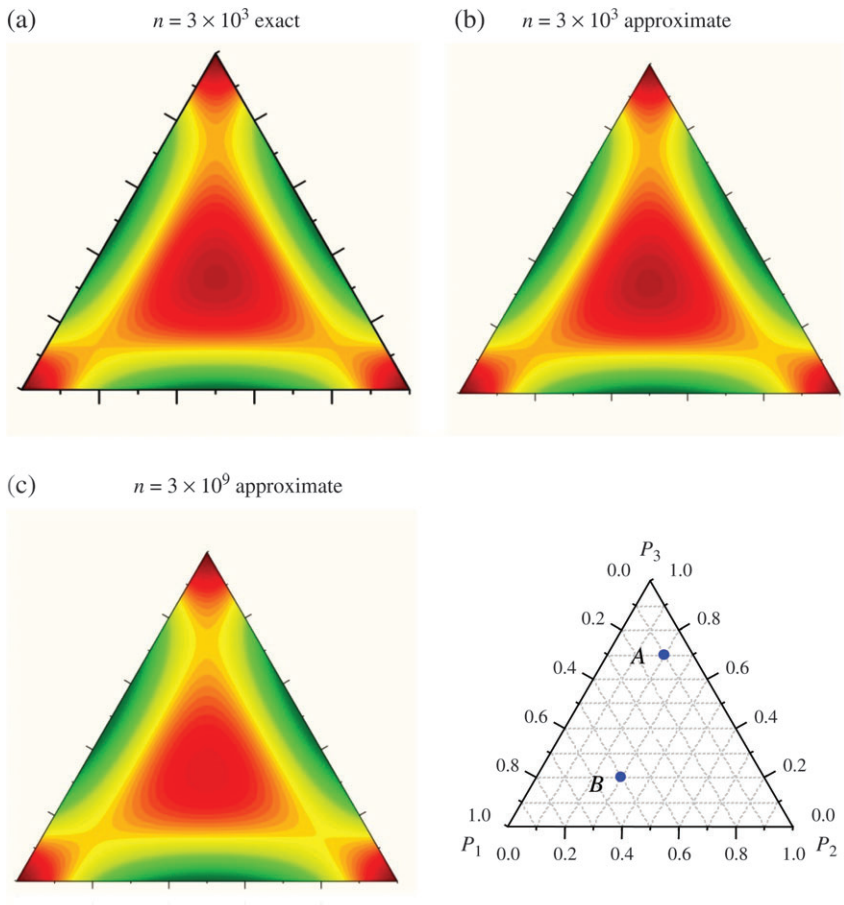

FIG. 7.—Estimated joint density, $f(P_1, P_2, P_3)$, of posterior probabilities for the three trees over replicate data sets. The star tree with branch length $t = 0.2$ is used to generate 10^5 data sets. Each is analyzed to calculate the posterior probabilities P_1, P_2, and P_3 (equation 15), which are then collected to construct a 2-D histogram and to estimate the 2-D density using an adaptive kernel smoothing algorithm (Silverman 1986). The sequence length (and method used to calculate the integrals) is (a) $n = 3 \times 10^3$ sites (exact), (b) $n = 3 \times 10^3$ (approximate), and (c) $n = 3 \times 10^9$ (approximate), where exact calculation is achieved using Mathematica while approximate calculation is based on Laplacian expansion. The density f is shown using the color contours, with green, yellow, to red representing low to high values. The total density mass on the triangle is 1. Note that in the ternary plot, the coordinates (P_1, P_2, P_3) are represented by lines parallel to the sides of the triangle. The two points shown in the key have the coordinates $A(0.1, 0.2, 0.7)$ and $B(0.5, 0.3, 0.2)$, while the center point is $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

star tree. Thus we expect the posterior probability for the star tree τ_0 to converge to 1 as the star-tree model has a lower dimension (Dawid 1999). Here we consider π_0 as a way of resolving the star-tree paradox and divide P_0 among the three binary trees to calculate their posterior probabilities

$$P_i = \frac{\frac{1}{3}\pi_0 M_0 + \frac{1-\pi_0}{3} M_i}{\pi_0 M_0 + \frac{1-\pi_0}{3}(M_1 + M_2 + M_3)}, i = 1, 2, 3. \qquad (35)$$

Thus P_1, P_2, P_3 will converge to the point mass at $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ when $n \to \infty$ if the data are generated under the star tree, and to (1, 0, 0) if the data are generated under the binary tree τ_1.

Simulation Results

The Star-tree Paradox. We use computer simulation to study the variation in posterior tree probabilities (P_1, P_2, P_3) when data sets are generated under the star tree. The branch length is fixed at $t = 0.2$. Each of the 10^5 replicate data sets is analyzed using the Bayesian method to calculate P_1, P_2, P_3, using equal prior probabilities $(\frac{1}{3})$ for the three binary trees and exponential priors for branch lengths with means $\mu_0 = 0.1$ and $\mu_1 = 0.2$ (equation 15). The distribution $f(P_1, P_2, P_3)$ across data sets is estimated by a kernel-density smoothing algorithm (Silverman 1986). Three sequence lengths are used: 3×10^3, 3×10^6, and 3×10^9. For $n = 3 \times 10^3$, both exact calculation using Mathematica and the approximate method by Laplacian expansion are used, while for the two large data sizes, only the approximate method is used.

Figure 7 shows the joint density $f(P_1, P_2, P_3)$ for $n = 3 \times 10^3$ and 3×10^9. Figure 8 shows three univariate densities derived from the same data, for P_1, for $P_{\min} = \min(P_1, P_2, P_3)$ and for $P_{\max} = \max(P_1, P_2, P_3)$. For $n = 3 \times 10^3$, the exact and approximate methods produced results that are indistinguishable, suggesting that the approximation is reliable. The results for $n = 3 \times 10^3$, 3×10^6 (not shown), and 3×10^9 are very similar, indicating that for the parameter values used,

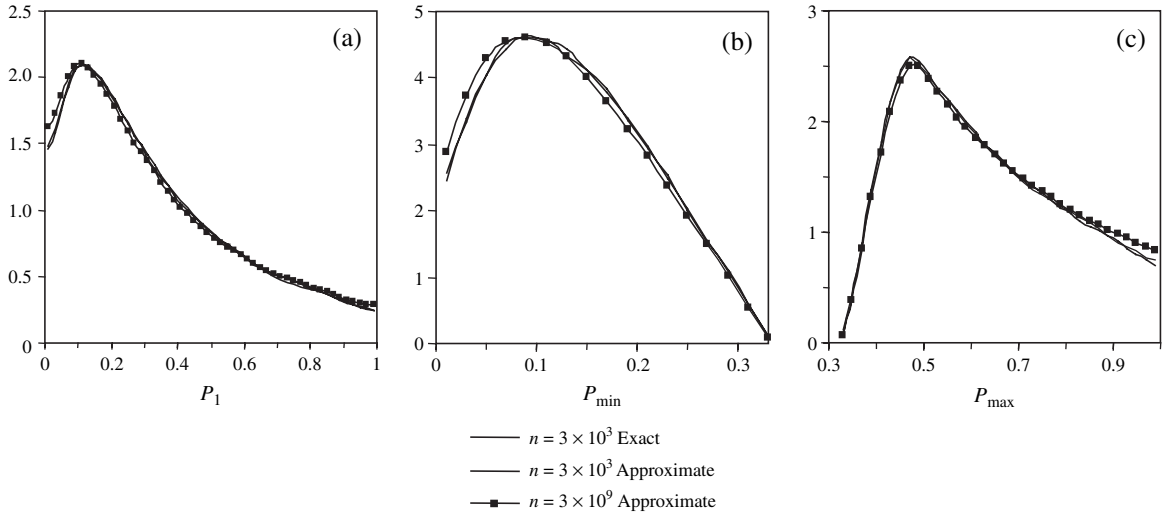

FIG. 8.—The density functions (*a*) for the posterior probability P_1 for any binary tree such as τ_1, (*b*) for the smallest of the three posterior probabilities P_{min}, and (*c*) for the largest of the three probabilities P_{max}. The data of figure 7 are used to estimate the density functions.

$n = 3 \times 10^3$ is close to infinity, although it is noticeable that the posterior probabilities tend to become more extreme (near 0 or 1) in larger data sets (fig. 8a). The SD for P_1 is 0.2440 for $n = 3 \times 10^3$ and 0.2498 for $n = 3 \times 10^6$ and 3×10^9. In general, the means and SDs for P_1, P_{min}, and P_{max} are identical to the fourth decimal place between $n = 3 \times 10^6$ and 3×10^9.

For $n = 3 \times 10^9$, data sets are also simulated using different values of the branch length t in the star tree (such as 0.1, 0.3, 0.4, 0.5, and 1.0), and they are analyzed using different prior means μ_0 and μ_1 (such as $\mu_0 = 0.2, 0.5, 10$ and $\mu_1 = 0.1, 0.3, 0.7$). The number of replicates is also raised to 10^7. As far as can be judged, the distribution $f(P_1, P_2, P_3)$ is independent of t, μ_0 and μ_1. The invariance of $f(P_1, P_2, P_3)$ to parameters t, μ_0 and μ_1 may be generally true as it parallels the fair-balance analysis in which the limiting distribution $f(P_-)$ is uniform, independent of parameter ξ in the prior $\theta \sim N(0, \xi\sigma^2)$. It also indicates that the distribution is unlikely to change when n increases beyond 3×10^9. In all cases examined, every P_i has mean 1/3 and SD 0.2498, and pairwise correlation coefficient −0.5000. The correlation should be exactly $-\frac{1}{2}$, according to the following symmetry argument (Peter Green, pers. comm.). From $1 = P_1 + P_2 + P_3$, we have

$$0 = \sum_{i=1}^{3} \text{var}(P_i) + \sum_{i \neq j} \text{cov}(P_i, P_j)$$
$$= 3\text{var}(P_1) + 6\text{cov}(P_1, P_2), \quad (36)$$

so that $\text{corr}(P_1, P_2) = \text{cov}(P_1, P_2)/\text{var}(P_1) = -\frac{1}{2}$. There are four modes in the distribution, at the center and the three corners of the ternary graph (fig. 7).

We now use the distributions of P_1, P_{min} and P_{max} for $n = 3 \times 10^9$ to examine how often the Bayesian method produces extreme posterior probabilities, assuming that this sequence length represents the limiting case of infinite data (fig. 8). P_{min} has mean 0.1298 and SD 0.0769 while P_{max} has mean 0.6319 and SD 0.1698. In 4.23% of data sets, $P_{max} > 0.95$ (that is, at least one of the three posterior probabilities is > 0.95), and in 0.79% of data sets, $P_{max} > 0.99$. In 17.3% of data sets, $P_{min} < 0.05$ (that is, at least one of the three posterior probabilities is < 0.05), and in 2.6% of data sets, $P_{min} < 0.01$. If we consider any particular binary tree, such as τ_1, we find that the proportion of data sets in which $P_1 < 0.05$ (or 0.01) is 8.1% (or 1.31%), and the proportion of data sets in which $P_1 > 0.95$ (or 0.99) is 1.41% (or 0.26%). Because the true tree is the star tree, we would not want any binary tree to have either a very high or a very low posterior probability. The method appears to produce extreme posterior probabilities, especially very small ones, quite often.

Data Size-dependent Prior. This prior forces the mean μ_0 of internal branch lengths to approach 0 when $n \to \infty$. We let $\mu_0 = 0.1/n^\gamma$ and use different values for γ. When the data are simulated under the star tree, the means of the posterior probabilities for the three binary trees are always $\frac{1}{3}$. Figure 9a shows the SD of P_1 for tree τ_1 when $\gamma = 0$, 0.5, 0.51, 0.707, and 0.8. Our theoretical analysis suggests that γ has to be greater than $\frac{1}{2}$ for P_1 to converge to the point mass $\frac{1}{3}$. If $\gamma = 0$, the SD of P_1 converges to 0.2498 when $n \to \infty$; this is the case of the star-tree paradox discussed above. If $\gamma = 0.5$, the SD stabilizes to 0.064 instead of 0. Thus (P_1, P_2, P_3) have a distribution, which depends on parameters such as branch length t in the star tree, and μ_1 and c in the prior (in $\mu_0 = c/n^\gamma$). This is analogous to the case of $\theta_0 = 0$ and $\gamma = 1$ in table 1 for the fair-balance problem (fig. 3). When $\gamma = 0.51$, slightly larger than $\frac{1}{2}$, the SD decreases monotonically from 0.0608 at $n = 10^3$ to 0.0522 at $n = 3 \times 10^9$. The limit when $n \to \infty$ should be 0, according to the theoretical analysis. If $\gamma = 0.707$ or 0.8, the SD clearly converges to 0 when $n \to \infty$.

Results obtained when the data are simulated under the binary tree τ_1 with $t_0 = 0.01$ and $t_1 = 0.2$ are shown in figure 9b. The theoretical analysis predicts that one has to have $\gamma < 1$ for P_1 to converge to the point mass at 1 when $n \to \infty$. If $\gamma = 0$, 0.5, or 0.707 (all less than 1), the mean

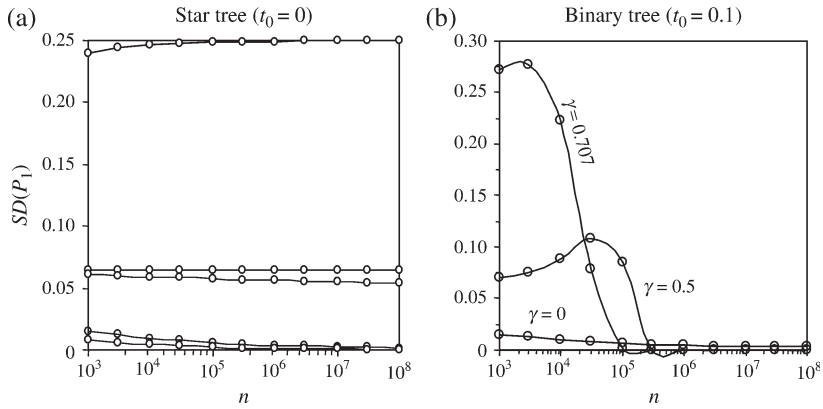

FIG. 9.—(*a*) The SD in the posterior probability P_1 for tree τ_1 is plotted against the data size n when the data are simulated under the star tree with branch lengths $t_0 = 0$ and $t_1 = 0.2$ and analyzed using the prior means $\mu_0 = 0.1/n^\gamma$ and $\mu_1 = 0.2$. The values of γ are 0, 0.5, 0.51, 0.707, and 0.8 from top to bottom. The theoretical expectation is that the SD $\to$ 0 (so that $P_1 \to \frac{1}{3}$) when $n \to \infty$ if and only if $\gamma > 0.5$. (b) The SD in P_1 is plotted against the sample size n when the data are simulated under a binary tree with branch lengths $t_0 = 0.01$ and $t_1 = 0.2$. The same priors are used to analyze the data as in (*a*), with γ to be 0, 0.5, and 0.707. The theoretical expectation is that the SD $\to$ 0 (so that $P_1 \to 1$) if $\gamma<1$ but $P_1 \to \frac{1}{3}$ if $\gamma > 1$; this expectation is not confirmed here as values of γ around 1 caused computational problems.

of P_1 indeed converges to 1 while the SD converges to 0, so that the probability for the true model converges to 1 (fig. 9b). Numerical problems are encountered with larger values of γ, so that the cases in which γ is close to or larger than 1 are not examined. Nevertheless, as long as the star-tree paradox is resolved (with $\gamma > \frac{1}{2}$), small values for γ are preferred to larger ones, as small values lead to higher posterior probabilities for the true tree when the true tree is binary. Three convenient values for γ are 0.667, 0.707, and 0.75. These are the harmonic, geometric, and arithmetic means of $\frac{1}{2}$ and 1, and may represent conservative, moderate, and liberal priors, respectively.

Degenerate-Model Star-tree Prior π_0. Here a nonzero probability π_0 is assigned for the degenerate star tree τ_0, while the three binary trees have prior probabilities $\pi_1 = \pi_2 = \pi_3 = (1-\pi_0)/3$. The posterior probabilities for the three binary trees are calculated using equation (35). We are interested in the behavior of the joint density $f(P_1, P_2, P_3)$ when the data size $n \to \infty$ and when the data are generated under either the star tree or a binary tree.

A few different values are used for π_0: 1/10, 1/4, and $\frac{1}{3}$. In every case, the joint density $f(P_1, P_2, P_3)$ converges to $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ when $n \to \infty$. For example, with $t = 0.2$ in the star tree and $\pi_0 = 0.25$, $\mu_0 = 0.1$, and $\mu_1 = 0.2$ in the prior, the SD of P_1 is calculated to be 0.125, 0.025, and 0.004 for $n = 3 \times 10^3$, 3×10^6, and 3×10^9, respectively. The mean of the distribution is clearly $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, and the convergence of the SD to 0 means that the distribution is becoming degenerate to the point mass at the mean. When $\pi_0 = 0.1$, the SD of P_1 is 0.177, 0.044, and 0.007 for the three values of n, and the rate of convergence is slower than when $\pi_0 = 0.25$.

Furthermore, analysis of data sets simulated under a binary tree with $t_0 > 0$ confirms that when n increases, the posterior probability for the true binary tree approaches 1. In sum, use of the prior π_0 resolves the star-tree paradox, as long as $0 < \pi_0 < 1$. This result is expected from Dawid's (1999) general proof of consistency of Bayesian model selection.

Addendum

Steel and Matsen (2007) recently published a mathematical analysis of the star-tree problem (fig. 2), proving that when the number of sites $n \to \infty$, the posterior probability for any binary tree, say, P_1, does not converge to $\frac{1}{3}$ and will maintain a strictly positive probability of being large (say, > 0.99). The result is consistent with this study, contra Kolaczkowski and Thornton (2006). Note that the limiting distribution $f(P_1, P_2, P_3)$ when $n \to \infty$ remains unknown.

Acknowledgments

I am grateful to Professor Peter Green of University of Bristol for pointing out that the correlation between any two posterior probabilities in the star-tree distribution is exactly $-\frac{1}{2}$. I thank Professor Philip Dawid (UCL) for very useful discussions, and Jim Mallet and Max Telford for comments on the first part of the manuscript. This study is supported by a grant from the Natural Environment Research Council (UK).

Appendix. Derivatives for Laplacian Expansion

Consider tree τ_1. The data can be represented as $x_0 = n_0/n$, $x_1 = n_1/n$, and the likelihood $L = nh$, where

$$h(x_0, x_1|t_0, t_1) = x_0 \log(p_0) + x_1 \log(p_1) + (1 - x_0 - x_1)\log(p_2). \quad (37)$$

where p_0, p_1, p_2 are given in equation (12). Let $e_0 = e^{-4(t_0+t_1)}$ and $e_1 = e^{-4t_1}$ and note that

$$\begin{array}{lll} \frac{\partial p_0}{\partial t_0} = -2e_0, & \frac{\partial p_1}{\partial t_0} = 2e_0, & \frac{\partial p_2}{\partial t_0} = 0, \\ \frac{\partial p_0}{\partial t_1} = -e_1 - 2e_0, & \frac{\partial p_1}{\partial t_1} = -e_1 + 2e_0, & \frac{\partial p_2}{\partial t_1} = e_1. \end{array} \quad (38)$$

Then the gradient $g = \left(\frac{\partial h}{\partial t_0}, \frac{\partial h}{\partial t_1}\right)$ and Hessian matrix $\mathbf{H} = \left\{\frac{\partial^2 h}{\partial t_i \partial t_j}\right\}$ are

$$h_{10}=\frac{\partial h}{\partial t_0}=-2e_0\left(\frac{x_0}{p_0}-\frac{x_1}{p_1}\right),$$
$$h_{01}=\frac{\partial h}{\partial t_1}=-2e_0\left(\frac{x_0}{p_0}-\frac{x_1}{p_1}\right)-e_1\left(\frac{x_0}{p_0}+\frac{x_1}{p_1}-\frac{1-x_0-x_1}{p_2}\right), \quad (39)$$

and

$$\begin{aligned}h_{20}&=\frac{\partial^2 h}{\partial t_0^2}=8e_0\left(\frac{x_0}{p_0}-\frac{x_1}{p_1}\right)-4e_0^2\left(\frac{x_0}{p_0^2}+\frac{x_1}{p_1^2}\right),\\ h_{11}&=\frac{\partial^2 h}{\partial t_0\partial t_1}=8e_0\left(\frac{x_0}{p_0}-\frac{x_1}{p_1}\right)-4e_0^2\left(\frac{x_0}{p_0^2}+\frac{x_1}{p_1^2}\right)\\ &\quad-2e_0e_1\left(\frac{x_0}{p_0^2}-\frac{x_1}{p_1^2}\right),\\ h_{02}&=\frac{\partial^2 h}{\partial t_1^2}=8e_0\left(\frac{x_0}{p_0}-\frac{x_1}{p_1}\right)-4e_0^2\left(\frac{x_0}{p_0^2}+\frac{x_1}{p_1^2}\right)-4e_0e_1\left(\frac{x_0}{p_0^2}-\frac{x_1}{p_1^2}\right)\\ &\quad+4e_1\left(\frac{x_0}{p_0}+\frac{x_1}{p_1}-\frac{1-x_0-x_1}{p_2}\right)\\ &\quad-e_1^2\left(\frac{x_0}{p_0^2}+\frac{x_1}{p_1^2}+\frac{1-x_0-x_1}{p_2^2}\right).\end{aligned} \quad (40)$$

The third derivatives are

$$\begin{aligned}h_{30}&=\frac{\partial^3 h}{\partial t_0^3}=-32e_0\left(\frac{x_0}{p_0}-\frac{x_1}{p_1}\right)+48e_0^2\left(\frac{x_0}{p_0^2}+\frac{x_1}{p_1^2}\right)-16e_0^3\left(\frac{x_0}{p_0^3}+\frac{x_1}{p_1^3}\right),\\ h_{21}&=\frac{\partial^3 h}{\partial t_0^2\partial t_1}=\frac{\partial^3 h}{\partial t_0^3}+8e_0e_1\left(\frac{x_0}{p_0^2}-\frac{x_1}{p_1^2}\right)-8e_0^2e_1\left(\frac{x_0}{p_0^3}+\frac{x_1}{p_1^3}\right),\\ h_{12}&=\frac{\partial^3 h}{\partial t_0\partial t_1^2}=\frac{\partial^3 h}{\partial t_0^3}+24e_0e_1\left(\frac{x_0}{p_0^2}-\frac{x_1}{p_1^2}\right)-16e_1\left(\frac{x_0}{p_0}+\frac{x_1}{p_1}-\frac{1-x_0-x_1}{p_2}\right)\\ &\quad-16e_0^2e_1\left(\frac{x_0}{p_0^3}+\frac{x_1}{p_1^3}\right)-4e_0e_1^2\left(\frac{x_0}{p_0^3}-\frac{x_1}{p_1^3}\right),\\ h_{03}&=\frac{\partial^3 h}{\partial t_1^3}=-32e_0\left(\frac{x_0}{p_0}-\frac{x_1}{p_1}\right)+8e_0\left(\frac{x_0}{p_0^2}(2e_0+e_1)+\frac{x_1}{p_1^2}(2e_0-e_1)\right)\\ &\quad+32e_0^2\left(\frac{x_0}{p_0^2}+\frac{x_1}{p_1^2}\right)-8e_0^2\left(\frac{x_0}{p_0^3}(2e_0+e_1)-\frac{x_1}{p_1^3}(2e_0-e_1)\right)\\ &\quad+32e_0e_1\left(\frac{x_0}{p_0^2}-\frac{x_1}{p_1^2}\right)-8e_0e_1\left(\frac{x_0}{p_0^3}(2e_0+e_1)\right.\\ &\quad\left.+\frac{x_1}{p_1^3}(2e_0-e_1)\right)\\ &\quad-16e_1\left(\frac{x_0}{p_0}+\frac{x_1}{p_1}-\frac{1-x_0-x_1}{p_2}\right)\\ &\quad+4e_1\left(\frac{x_0}{p_0^2}(2e_0+e_1)-\frac{x_1}{p_1^2}(2e_0-e_1)+\frac{1-x_0-x_1}{p_2^2}e_1\right)\\ &\quad+8e_1^2\left(\frac{x_0}{p_0^2}+\frac{x_1}{p_1^2}+\frac{1-x_0-x_1}{p_2^2}\right)\\ &\quad-2e_1^2\left(\frac{x_0}{p_0^3}(2e_0+e_1)-\frac{x_1}{p_1^3}(2e_0-e_1)-\frac{1-x_0-x_1}{p_2^3}e_1\right).\end{aligned} \quad (41)$$

The above formulae are confirmed by using the difference method to approximate the derivatives.

Literature Cited

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

Bender CM, Orszag SA. 1999. Advanced Mathematical Methods for Scientists and Engineers: Asymptotic Methods and Perturbation Theory. New York: Springer-Verlag.

Berger JO. 1985. Statistical Decision Theory and Bayesian Analysis. New York: Springer-Verlag.

Bernardo JM. 1979. Reference posterior distributions for Bayesian inference. J R Stat Soc B. 41:113–147.

Bernardo JM. 1980. A Bayesian analysis of classical hypothesis testing. In: Bernardo JM, DeGroot MH, Lindley DV, Smith AFM, eds. Bayesian Statistics. Valencia, Spain: Valencian University Press. p. 605–647.

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

Bourlat SJ, Juliusdottir T, Lowe CJ, Freeman R, et al. 2006. Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. Nature. 444: 85–88.

Buckley TR. 2002. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. Syst Biol. 51:509–523.

Copson ET. 1965. Asymptotic Expansions. Cambridge, UK: Cambridge University Press.

Cox DR. 2006. Principles of Statistical Inference. Cambridge, UK: Cambridge University Press.

Cummings MP, Handley SA, Myers DS, Reed DL, et al. 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Syst Biol. 52:477–487.

Davison AC. 2003. Statistical Models. Cambridge, UK: Cambridge University Press.

Dawid AP. 1999. The trouble with Bayes factors. Research Report 202. Department of Statistical Science. University College London.

Douady CJ, Delsuc F, Boucher Y, Doolittle WF, et al. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol Biol Evol. 20:248–254.

Drezner Z, Wesolowsky GO. 1990. On the computation of the bivariate normal integral. J Statist Comput Simul. 35:101–107.

Efron B. 1998. R.A. Fisher in the 21st Century. Stat Sci. 13:95–122.

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

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

Green PJ. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika. 82:711–732.

Hill ID. 1973. The normal integral. Appl Stat. 22:424–427.

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

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

Jeffreys H. 1939. Theory of Probability. Oxford, UK: Clarendon Press.

Jeffreys H. 1961. Theory of Probability. Oxford, UK: Oxford University Press.

Johnson NL, Kotz S, Balakrishnan N. 1994. Continuous Univariate Distributions. New York: Wiley.Volume 1

Kolaczkowski B, Thornton JW. 2006. Is there a star tree paradox? Mol Biol Evol. 23:1819–1823.

Lemmon AR, Moriarty EC. 2004. The importance of proper model assumption in Bayesian phylogenetics. Syst Biol. 53:265–277.

Lewis PO, Holder MT, Holsinger KE. 2005. Polytomies and Bayesian phylogenetic inference. Syst Biol. 54:241–253.

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

Lindley DV. 1957. A statistical paradox. Biometrika. 44:187–192.

Lindley DV. 1980. Approximate Bayesian methods. In: Bernardo JM, DeGroot MH, Lindley DV, Smith AFM, eds. Bayesian statistics. Valencia, Spain: Valencian University Press. p. 223–237.

Mau B, Newton MA. 1997. Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. J Computat Graph Stat. 6:122–131.
Pearson ES. 1947. The choice of statistical tests illustrated on the interpretation of data classed in the 2 x 2 table. Biometrika. 34:139–167.
Press SJ. 2003. Subjective and Objective Bayesian Statitics. New Jersey: John Wiley & Sons.
Rannala B, Yang Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol. 43:304–311.
Shafer G. 1982. Lindley's paradox. J Am Statist Assoc. 77:325–334.
Silverman BW. 1986. Density Estimation for Statistics and Data Analysis. London: Chapman and Hall.
Simmons MP, Pickett KM, Miya M. 2004. How meaningful are Bayesian support values? Mol Biol Evol. 21:188–199.
Steel M, Matsen FA. 2007. The Bayesian ''star paradox'' persists for long finite sequences. Mol Biol Evol. 24:1075–1079.
Suzuki Y, Glazko GV, Nei M. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc Natl Acad Sci USA. 99:16138–16143.
Tierney L, Kadane JB. 1986. Accurate approximations for posterior moments and marginal densities. J Am Stat Assoc. 81:82–86.
Wolfram S. 2003. Mathematica 5. Cambridge, UK: Cambridge University Press.
Yang Z. 2000. Complexity of the simplest phylogenetic estimation problem. Proc R Soc B: Biol Sci. 267:109–116.
Yang Z. 2006. Computational Molecular Evolution. Oxford, England: Oxford University Press.
Yang Z, Rannala B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo Method. Mol Biol Evol. 14:717–724.
Yang Z, Rannala B. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. Syst Biol. 54:455–470.
Yang Z, Goldman N, Friday AE. 1995. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst Biol. 44:384–399.

Arndt von Haeseler, Associate Editor

Accepted April 18, 2007