Tail Paradox, Partial Identifiability, and Influential Priors in Bayesian Branch Length Inference

Bruce Rannala,[1,2] Tianqi Zhu,[3,4] and Ziheng Yang[*,1,5,6]

[1]Center for Computational and Evolutionary Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing, China
[2]Genome Center and Department of Evolution and Ecology, University of California–Davis
[3]School of Mathematical Sciences, Peking University, Beijing, China
[4]National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Science, Beijing, China
[5]School of Life Sciences, Sun Yat-sen University, Guangzhou, China
[6]Department of Biology, University College London, London, United Kingdom

***Corresponding author:** E-mail: z.yang@ucl.ac.uk.

Associate editor: Jeffrey Thorne

Abstract

Recent studies have observed that Bayesian analyses of sequence data sets using the program MrBayes sometimes generate extremely large branch lengths, with posterior credibility intervals for the tree length (sum of branch lengths) excluding the maximum likelihood estimates. Suggested explanations for this phenomenon include the existence of multiple local peaks in the posterior, lack of convergence of the chain in the tail of the posterior, mixing problems, and misspecified priors on branch lengths. Here, we analyze the behavior of Bayesian Markov chain Monte Carlo algorithms when the chain is in the tail of the posterior distribution and note that all these phenomena can occur. In Bayesian phylogenetics, the likelihood function approaches a constant instead of zero when the branch lengths increase to infinity. The flat tail of the likelihood can cause poor mixing and undue influence of the prior. We suggest that the main cause of the extreme branch length estimates produced in many Bayesian analyses is the poor choice of a default prior on branch lengths in current Bayesian phylogenetic programs. The default prior in MrBayes assigns independent and identical distributions to branch lengths, imposing strong (and unreasonable) assumptions about the tree length. The problem is exacerbated by the strong correlation between the branch lengths and parameters in models of variable rates among sites or among site partitions. To resolve the problem, we suggest two multivariate priors for the branch lengths (called compound Dirichlet priors) that are fairly diffuse and demonstrate their utility in the special case of branch length estimation on a star phylogeny. Our analysis highlights the need for careful thought in the specification of high-dimensional priors in Bayesian analyses.

Key words: Bayesian phylogenetics, Markov chain Monte Carlo, branch lengths, identifiability, tail paradox, compound dirichlet distribution.

Introduction

Bayesian phylogenetics has mostly focused on the inference of tree topologies (reviewed in Yang 2006), but branch lengths in the tree may also be important, for example in estimating evolutionary rates and species divergence times. Several recent studies (e.g., Brown et al. 2010; Marshall 2010) have reported that phylogenetic analyses of real data sets using the popular Bayesian program MrBayes (Ronquist and Huelsenbeck 2003) sometimes produced very large branch length estimates, with the total tree length (sum of branch lengths) much larger than the maximum likelihood estimate (MLE) for the same data under the same model. In extreme cases, the credibility interval constructed from the Bayesian analysis did not even contain the MLE. Brown et al. (2010) and Marshall (2010) conducted systematic investigations of this phenomenon. Marshall (2010) focused mainly on the convergence and mixing properties of the Markov chain Monte Carlo (MCMC) algorithms. Brown et al. (2010) proposed three hypotheses that may explain the observed phenomenon, which form the focus of discussion in this paper. First, there may exist multiple local peaks in the posterior so that the MCMC algorithm has difficulty traversing the space of parameters. Second, the posterior may be single moded but large regions of the posterior have roughly equal posterior density, so that the MCMC does not mix well.

Third, an overly informative prior favors unreasonably large branch lengths. We have here modified Brown et al.'s hypothesis 3; those authors also considered a high likelihood as a possible factor (in addition to an informative prior) that could favor unreasonably large branch lengths. However, unless the model is grossly mispecified, the likelihood will represent information in the data and cannot support biologically unreasonable regions of the parameter space.

Under hypotheses 1 and 2, numerical problems of the MCMC algorithm cause an incorrect posterior density to be generated, while under hypothesis 3, the posterior is correctly sampled but is dominated by an unreasonable prior. It is an interesting question how often unreasonably large tree lengths are caused by multiple local peaks in the posterior (hypothesis 1). We note that multiple peaks are common in the space of trees and may cause serious computational problems for MCMC algorithms (Mossel and Vigoda 2005), although they appear to be rare on the likelihood surface when the tree topology is fixed (Rogers and Swofford 1999). Brown et al. (2010) concluded that multiple modes do not

appear to exist for the six data sets they analyzed and that hypothesis 1 was probably not the major cause of branch length overestimation. Our analysis below shows that if the likelihood and the prior are in conflict, multiple peaks may occur in the posterior even on a fixed tree topology (hypothesis 1).

At any rate, the problem of unreasonable branch lengths manifests itself even when the tree topology is fixed (Brown et al. 2010; Marshall 2010), and the present study concerns estimation of branch lengths (and thus tree length) on a fixed tree.

We argue that the probable cause of the extremely large branch lengths (or tree length) observed in previous studies is the poor choice of a default prior on branch lengths implemented in the current version of MrBayes, which assigns independent and identical distributions for branch lengths (hypothesis 3). This has the consequence of strongly favoring unreasonably large tree lengths. Indeed, the poor choice of prior can cause all the problems described by Brown et al. (2010): multiple local peaks due to conflicts between the prior and likelihood (hypothesis 1), mixing problems in the flat tail (hypothesis 2) and unreasonably large branch lengths in the posterior (hypothesis 3). The problem is exacerbated by the strong correlations between branch lengths and parameters in models of variable rates among sites and between parameters of the rate-variation model. We suggest that the solution to the problem is a careful specification of the joint prior for the branch lengths and for the rate-variation parameters. We propose two new joint priors for branch lengths that are less "informative" than priors that are currently used, and we explore the impact of the prior for branch lengths on the posterior of tree length using simulated and real data sets. Before presenting our new priors, we discuss the behavior of MCMC algorithms in the far tails of the posterior and illustrate that the rate of convergence of the MCMC algorithm depends on whether the posterior has a light or heavy tail.

Theory

Light- and Heavy-Tailed Posteriors and MCMC Convergence

Brown et al. (2010) postulated that "the existence of a local maximum in the posterior density at long tree lengths entraps the MCMC chain, keeping it from sampling the parameter space in proportion to the posterior density" (hypothesis 1) and that "large regions of parameter space with roughly equal posterior density reduce the efficiency of the MCMC search, such that it does not sample the parameter space in proportion to the posterior density" (hypothesis 2). The authors further suggested that if hypothesis 2 is true "the MCMC chain is wandering around a large region of roughly equal posterior density." These speculations prompted us to study the behavior of MCMC when the current parameter value is in the tail of the posterior distribution.

For a smooth unimodal posterior $\pi(x)$, the tail is often the "flattest" part of the distribution in the sense that the slope (gradient) $\nabla\pi(x) = \mathrm{d}\pi(x)/\mathrm{d}x \to 0$ as $x \to \infty$. However, the behavior of the MCMC is not determined by the gradient of the posterior but instead by the gradient of the logarithm of the posterior $\nabla \log \pi(x) = \mathrm{d}\log\pi(x)/\mathrm{d}x$ (Mengersen and Tweedie 1996) and depends on whether the posterior is light-tailed or heavy-tailed. A distribution is light-tailed if there exists $t > 0$ such that

$$\lim_{x\to\infty} \mathrm{e}^{tx}(1 - F(x)) < \infty, \tag{1}$$

where $F(x)$ is the cumulative distribution function. Otherwise, it is heavy-tailed. The exponential, normal, and gamma are examples of light-tailed distributions while heavy-tailed distributions include the Cauchy and inverse gamma. Mengersen and Tweedie (1996) showed that in one dimension, the MCMC algorithm achieves geometric convergence (i.e., the distance between the target distribution π and the distribution that the MCMC reaches in n steps decreases to 0 at least at the rate r^n, where $r < 1$) if and only if

$$\lim_{x\to\infty} \nabla \log \pi(x) = \lim_{x\to\infty} \frac{\mathrm{d}\log\pi(x)}{\mathrm{d}x} < 0, \tag{2}$$

It can be shown that only a light-tailed distribution $\pi(x)$ with $\lim_{x\to\infty} \nabla\pi(x) < 0$ meets this criterion. Geometric convergence is not possible if $\nabla \log \pi(x) \to 0$ as $x \to \infty$. In high dimensions, the posterior π needs to be sufficiently smooth in the tail, besides being light-tailed, to achieve geometric convergence (Roberts and Tweedie 1996).

Below we use specific examples to illustrate the different behaviors of the MCMC depending on the tail of the posterior density.

Case I: Tail paradox for a single-moded light-tailed posterior. Here, we use the normal density as an example of a light-tailed posterior to study the behavior of the MCMC when the chain is in the right tail (table 1a). The posterior density for parameter θ is

$$\pi(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(\theta-\tilde{\theta})^2}, \tag{3}$$

with mean $\tilde{\theta}$ and variance σ^2. The posterior ratio for the proposal $\theta' = \theta + \Delta\theta$ is

$$R = \frac{\pi(\theta')}{\pi(\theta)} = \mathrm{e}^{-\frac{1}{2\sigma^2}(2\Delta\theta(\theta-\tilde{\theta})+(\Delta\theta)^2)} \approx \mathrm{e}^{-\frac{1}{2\sigma^2}2\Delta\theta(\theta-\tilde{\theta})}, \tag{4}$$

since $\theta \gg |\Delta\theta|$. Note that the proposal is accepted with probability $\min(1, R)$. Left moves toward the mode ($\Delta\theta < 0$) are always accepted, whereas right moves away from the mode are accepted with probability R. Far out in the tail θ (i.e., with $\theta = 100\sigma$), R is very different from 1 so that the MCMC moves quickly toward the mode. Near the mode (i.e., when $\theta = \sigma$), R is moderate and the chain wonders. This behavior may appear paradoxical and we refer to it as the "tail paradox": the farther away from the mode, the flatter the tail is but the faster the MCMC moves out of the tail.

Case II: Heavy-tailed posterior density. We use the inverse gamma distribution as an example of a heavy-tailed

Table 1. Acceptance Ratio (R) for Three Different Posterior Distributions.

θ	$\Delta\theta$	$\pi(\theta)$	R_{left}	R_{right}	$\nabla\pi$	$\nabla\log\pi$
(a) Normal posterior $\theta \sim N(0,1)$						
1	1	0.24	1.65	0.22	−0.24	−1
10	1	7.69×10^{-23}	1.3×10^{4}	2.8×10^{-5}	-7.7×10^{-22}	−10
100	1	1.34×10^{-2172}	1.6×10^{43}	2.3×10^{-44}	-1.3×10^{-2170}	−100
1000	1	2.29×10^{-217148}	1.2×10^{434}	3.1×10^{-435}	-2.3×10^{-217145}	−1000
(b) Inverse gamma posterior $\theta \sim invG(3,1)$						
0.5	0.1	1.08	1.481	0.673	−4.33	−4.00
1	0.1	0.18	1.364	0.748	−0.55	−3.00
10	0.1	4.52×10^{-5}	1.040	0.962	-1.76×10^{-5}	−0.39
100	0.1	4.95×10^{-9}	1.004	0.996	-1.98×10^{-10}	−0.0399
1000	0.1	5.00×10^{-13}	1.000	1.000	-2.00×10^{-15}	−0.0040
(c) JC69 distance between human–chimpanzee mitochondrial 12S rRNAs						
0.01	0.001	2.07×10^{-73}	2.39×10^{-5}	1.38×10^{4}	2.084×10^{-69}	1.01×10^{4}
0.1	0.01	6.05	0.09	3.67	1089	180
0.12046768	0.01	33.9	0.68	0.71	0.0	0.0
0.2	0.1	4.78×10^{-6}	1.27×10^{6}	4.86×10^{-19}	−0.0016	−335.6
4	**0.1**	1.56×10^{-365}	1.47	0.81	-4.65×10^{-365}	−2.98
4.204258	0.1	1.17×10^{-365}	1.07	1.06	0.0	***0.0***
4.5	0.1	1.85×10^{-365}	0.78	1.37	$\mathbf{5.24} \times 10^{-365}$	2.83
6	0.1	3.40×10^{-362}	0.59	1.68	$\mathbf{1.79} \times 10^{-361}$	5.25
8	0.1	6.86×10^{-359}	0.79	1.25	1.57×10^{-358}	2.29
9.89311	0.1	$\mathbf{5.26} \times 10^{-358}$	0.995	0.995	0.0	0.0
10	0.1	$\mathbf{5.23} \times 10^{-358}$	1.01	0.98	-5.54×10^{-359}	−0.11
12	0.1	7.40×10^{-359}	1.19	0.84	-1.30×10^{-358}	−1.75
14	0.1	6.47×10^{-361}	1.34	0.74	-1.90×10^{-360}	−2.93
100	1	7.10×10^{-650}	8144	1.22×10^{-4}	-6.40×10^{-649}	−9.01
1000	1	1.60×10^{-4459}	2.00×10^{4}	5.01×10^{-5}	-1.58×10^{-4458}	−9.90

NOTE.—R is the posterior ratio for a move of size $\Delta\theta$ to the left (toward the mode) or to the right (away from the mode) of the current value θ : $R_{left} = \frac{\pi(\theta-\Delta\theta)}{\pi(\theta)}$ and $R_{right} = \frac{\pi(\theta+\Delta\theta)}{\pi(\theta)}$. (a) Standard normal as an example of a light-tailed posterior. The density is $\pi(\theta) = e^{-\theta^2/2}$, so that $\nabla\pi = \frac{d\pi}{d\theta} = -\theta e^{-\theta^2/2} = -\theta\pi$ and $\nabla \log \pi = \frac{d \log \pi}{d\theta} = -\theta$. (b) The inverse gamma as an example of a heavy-tailed posterior. The density is $\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{-\alpha-1}e^{-\beta/\theta}$, so that $\nabla\pi = \pi \cdot \nabla \log \pi$ and $\nabla \log \pi = -\frac{\alpha+1}{\theta} + \frac{\beta}{\theta^2}$. (c) The posterior is given in equation (10), generated by the binomial likelihood under the JC69 model and the gamma prior $G(\alpha,\beta)$, so that $\nabla\pi = \pi \cdot \nabla \log \pi$ and $\nabla \log \pi = \left(\frac{x}{p} - \frac{n-x}{1-p}\right) e^{-4\theta/3} - \beta + \frac{\alpha-1}{\theta}$.

posterior (table 1b)

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{-\alpha-1}e^{-\beta/\theta}, 0 < \theta < \infty, 0 < \alpha, \beta < \infty. \quad (5)$$

The mean is finite only if $\alpha > 1$, and the variance exists only if $\alpha > 2$. The posterior ratio for the proposal $\theta' = \theta + \Delta\theta$ is

$$R = \frac{\pi(\theta')}{\pi(\theta)} = \left(1 + \frac{\Delta\theta}{\theta}\right)^{-\alpha-1} \exp\left\{\beta\frac{\Delta\theta}{\theta(\theta+\Delta\theta)}\right\}$$

$$\approx 1 - \frac{\Delta\theta}{\theta}(\alpha+1)\left[1 - \frac{\beta}{(\alpha+1)\theta}\right], \quad (6)$$

as $\theta \gg \Delta\theta$. If θ is large, R will be close to 1. Thus, the chain will behave like a random walk in the tail and the rate of convergence will be extremely slow.

While the gradient of the posterior density is nearly zero far out in the tail for both the normal and inverse gamma distributions, the behavior of the MCMC algorithm is very different between the two. In the normal case, virtually all moves away from the mode are rejected whereas in the inverse gamma case, they are nearly always accepted. The behavior near the mode is similar between the two types of posterior distribution.

Case III: Flat posterior density of branch lengths under the uniform prior. When the branch lengths are very large, the probabilities of site patterns are approximately given by the distribution corresponding to random sequences and the likelihood is nearly a nonzero constant. An interesting situation occurs when a uniform prior on branch lengths $U(0, A)$ is used (the MrBayes default value is $A = 100$). If the MCMC is started using very large branch lengths, the posterior ratio of the MCMC proposal will be close to 1, as both the likelihood ratio and prior ratio are approximately 1. Then the chain will wander in the tail of the posterior like a random walk (hypothesis 2) and will not converge until it approaches the mode of the likelihood. Note that even if no mode exists in the prior of branch lengths, there is a mode in the induced prior on tree length, at $nA/2$, where n is the number of branches.

Case IV: Multimodal posterior for branch lengths. If the prior for branch lengths has a mode that is distant from the mode of the likelihood, it is possible for the posterior to have two modes, one near the mode of the likelihood and another near the mode of the prior. Although the likelihood-induced mode can be many orders of magnitude higher than the prior-induced mode, it is possible for the MCMC to be trapped at the smaller mode. For very large branch lengths, the posterior ratio for the proposal $\theta \to \theta'$ is

$$R = \frac{f(\theta')}{f(\theta)}\frac{L(\theta')}{L(\theta)} \approx \frac{f(\theta')}{f(\theta)}\frac{L(\infty)}{L(\infty)} \approx \frac{f(\theta')}{f(\theta)}, \quad (7)$$

so that this part of the posterior is dominated by the prior.

We illustrate this scenario using the case of Bayesian estimation of the distance θ between two sequences under the JC69 model (Jukes and Cantor 1969). The human and chimpanzee 12s rRNA genes from the mitochondrial genome are used (accession numbers D38112 and NC_001643, respectively, Horai et al. 1995). The original data has 953 sites. After alignment gaps are removed, $n = 946$ sites remain, with $x = 11$ differences. Under the JC69 model with one rate for all sites, the likelihood is

$$L(\theta) = p^x(1-p)^{n-x} = \left[\frac{3}{4} - \frac{3}{4}e^{-4\theta/3}\right]^x \left[\frac{1}{4} + \frac{3}{4}e^{-4\theta/3}\right]^{n-x}, \quad (8)$$

where θ is the distance between the two species or the tree length for the tree of two species. Figure 1*a* shows that when $\theta \to \infty$, the likelihood approaches $L(\theta) = (\frac{3}{4})^x(\frac{1}{4})^{n-x} = 3^x/4^n = 5.00705 \times 10^{-565}$, a constant that is greater than 0. The MLE is $\hat{\theta} = 0.01172$, with $L(\hat{\theta}) \to 9.3575 \times 10^{-27}$. Suppose we use the gamma prior

$$f(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}e^{-\beta\theta}\theta^{\alpha-1}, \quad (9)$$

and let $\alpha = 100$ and $\beta = 10$, with the prior mean 10 and variance 1. The posterior is thus

$$\pi(\theta) = f(\theta|x) = \frac{1}{C}f(\theta)L(\theta), \quad (10)$$

with the normalizing constant calculated to be $C = 3.83834 \times 10^{-208}$. This posterior is plotted in figure 1*b*–*d*, showing two modes with a minimum in between. The acceptance (posterior) ratios for proposals at different points of the parameter space as well as the gradient $\nabla\pi$ and $\nabla \log \pi$ are shown in table 1c. Left of the first mode ($\theta \leqslant 0.1$), R_{left} is very small, and far off in the right tail ($\theta \geqslant 100$), R_{right} is very small, so that the chain moves out of both tails very quickly. Between the two modes ($0.2 < \theta \leqslant 8$), the chain may be attracted to either mode, depending on which side of the minimum the current state is. For example, at $\theta = 6$, $R_{\text{left}} = 0.59$ and $R_{\text{right}} = 1.68$, so that the chain is more likely to move to the second lower mode than to the first higher mode. More extreme acceptance ratios can be generated if the two modes are farther apart. Our implementation of the MCMC algorithm using a sliding window to change θ confirms the theoretical analysis: the MCMC can easily stay around the lower peak over as many as 10^8 iterations. While the bump caused by the prior is tiny in magnitude and negligible in terms of probability mass in the posterior, it may have important impact on the convergence properties of the chain in this part of the parameter space. This corresponds to a mixing problem of the type postulated in hypothesis 1 of Brown et al. (2010).

The Prior on Branch Lengths and Its Impact on Tree Length

Let $t = \{t_i\}$ be the vector of $n = 2s - 3$ branch lengths on the unrooted tree of s species. Existing Bayesian phylogenetic programs that do not assume a molecular clock assign a prior density to the branch lengths of the form

$$f(t|\theta) = \prod_{i=1}^{2s-3} f(t_i|\theta), \quad (11)$$

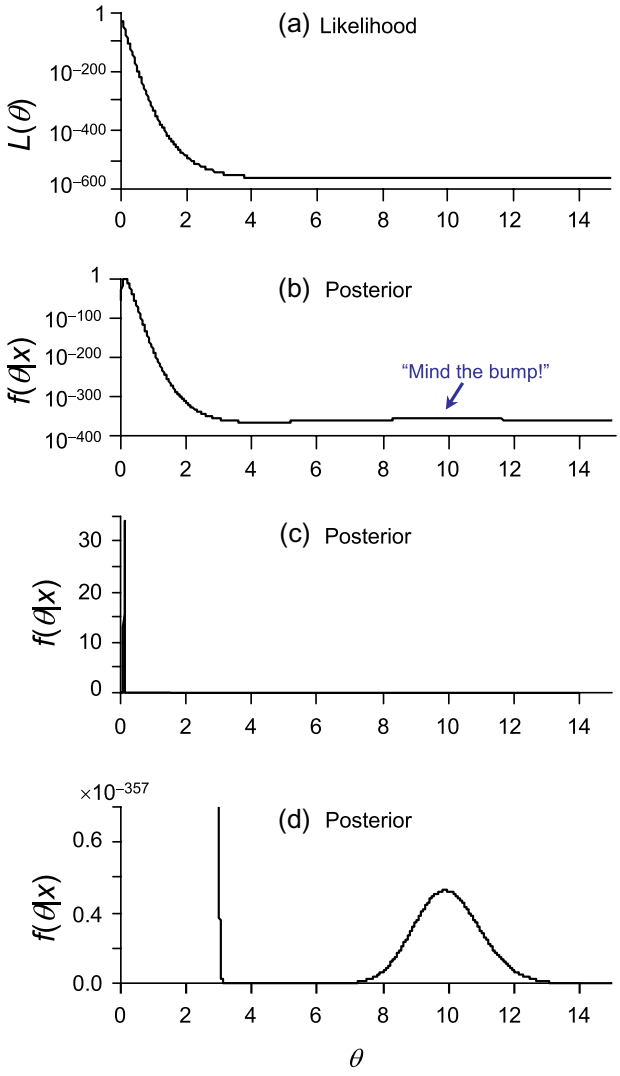

FIG. 1. Bayesian estimation of the sequence distance θ between the human and chimpanzee mitochondrial 12S rRNA genes under a gamma prior $\theta \sim G(100, 10)$. The likelihood (eq. 8) is shown in *a*. The posterior (eq. 10) is shown in *b*, *c*, and *d*, using different scales for the *y* axis. The posterior mean is 0.1217 with the 95% equal-tail credibility interval to be (0.0997, 0.1460). There are two modes in the posterior, at $\theta = 0.120468$ with $\pi(\theta) = 33.9$ and $\theta = 9.89311$ with $\pi(\theta) = 5.26 \times 10^{-358}$. Between the two modes, there is a minimum at $\theta = 4.204258$ with $\pi(\theta) = 1.17 \times 10^{-365}$. See table 1c for more results about this analysis.

where $f(t_i|\theta)$ is a common probability density function (pdf) with the same parameters for all branch lengths. For example, MrBayes uses either a uniform $U(0, \theta)$ prior,

$$f(t_i|\theta) = \frac{1}{\theta}, \text{ for } 0 < t_i < \theta, \; i = 1, 2, \ldots, (2s-3), \tag{12}$$

where θ is the upper bound on branch lengths so that the mean branch length is $\theta/2$, or an exponential prior, with

$$f(t_i|\theta) = \theta e^{-\theta t_i}, i = 1, 2, \ldots, (2s-3), \tag{13}$$

where $1/\theta$ is the average branch length. In both cases, the mean and variance of the branch lengths are completely

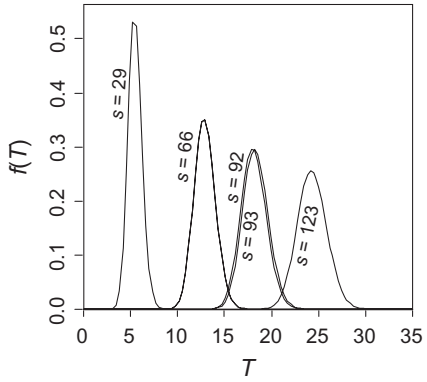

FIG. 2. The induced gamma prior, $T \sim G(2s-3, 10)$, on tree length for the six data sets of Brown et al. (2010), in which the number of sequences is $s = 29, 29, 66, 93, 123$, and 92.

specified by the single parameter θ. The tree length T,

$$T = \sum_{i=1}^{2s-3} t_i, \tag{14}$$

has the expectation

$$\mathbb{E}(T) = \sum_{i=1}^{2s-3} \mathbb{E}(t_i) = (2s-3)\mathbb{E}(t_i). \tag{15}$$

This increases linearly with the number of taxa. The variance is

$$\mathbb{V}(T) = \sum_{i=1}^{2s-3} \mathbb{V}(t_i) = (2s-3)\mathbb{V}(t_i), \tag{16}$$

because of the independent and identical distribution (i.i.d.) assumption.

In the case of the exponential prior for branch lengths, $\mathbb{E}(T) = (2s-3)/\theta$ and $\mathbb{V}(T) = (2s-3)/\theta^2$. Indeed, the induced probability density on T is known to be a gamma distribution with shape parameter $(2s-3)$ and scale parameter θ,

$$f(T|\theta, s) = \frac{T^{2s-4}\theta^{2s-3}e^{-\theta T}}{(2s-4)!}. \tag{17}$$

For typical phylogenetic data sets, $2s - 3 \gg 1$, and the gamma distribution is approximately normal. The shape of this distribution for different numbers of taxa and for $1/\theta = 0.1$ (the default in MrBayes) is shown in figure 2. (These are the same prior densities as shown in fig. 7 of Brown et al. 2010, although the plots for the prior in their fig. 3 do not appear to be correct.)

The induced prior on tree length seems to have been largely overlooked, although it may be strongly informative and contradict the likelihood. Presumably, sufficient data could overcome this influence of the prior. However, it is better to specify a prior on branch lengths that is close to being "uninformative." Below we explore a few possible solutions to this problem. The first is an empirical Bayesian procedure described by Brown et al. (2010), who used independent exponential priors for branch lengths but

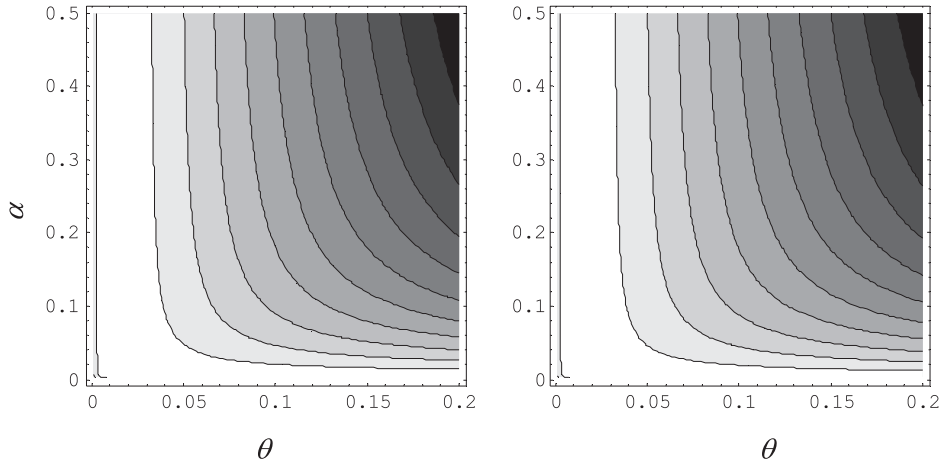

FIG. 3. Log likelihood (equation 44) and log posterior density (eq. 46) plotted against θ and α under the JC69+Γ model. The mitochondrial 12S rRNA genes from the human and the chimpanzee are compared with estimate the sequence distance, using an exponential prior on θ and a uniform prior on α. The same data are analyzed in figure 1.

used the data to estimate the parameter for the exponential, by equating the median of the exponential prior to the estimated average branch length $\bar{b}$, leading to $\theta = 1/(1.4427\bar{b})$. The median appears to be a poor choice, as the prior still favors large tree lengths. The mean is a better choice. Nevertheless, both suffer from the following problems: First, they use the data to derive the prior, which is not fully Bayesian. Second, they are very informative about the tree length. Third, they are very informative about the branch lengths, with a penalty on long branches, which may not be reasonable if there is a mixture of long and short branches on the tree.

The second is a prior suggested by Suchard et al. (2001). As in MrBayes, the branch lengths have independent exponential priors with the common mean μ,

$$f(t_i|\mu) = \frac{1}{\mu}e^{-t_i/\mu}, i = 1, 2, \ldots, (2s-3). \quad (18)$$

However, an inverse-gamma hyperprior is assigned on μ:

$$f(\mu|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\mu^{-\alpha-1}e^{-\beta/\mu}. \quad (19)$$

Suchard et al. (2001) recommended $\alpha = 2.1$ and $\beta = 1.1$ but such a value of α is extreme and the prior mean branch length of 1.0 seems rather large. We suggest $\alpha = 3$ and $\beta = 0.2$ instead, with the prior mean branch length of 0.1. The joint prior on the branch lengths is thus

$$\begin{aligned} f(t_1,\cdots,t_n) &= \int_0^\infty f(t_1,\ldots,t_n,\mu)d\mu \\ &= \int_0^\infty \left[\prod_{i=1}^n \frac{1}{\mu}e^{-t_i/\mu}\right] f(\mu|\alpha,\beta)d\mu \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty \frac{(\beta+\sum t_i)^{\alpha+n}}{\Gamma(\alpha+n)} \\ &\quad \mu^{-(\alpha+n)-1}e^{-(\beta+\sum t_i)/\mu}d\mu \\ &\quad \times \frac{\Gamma(\alpha+n)}{(\beta+\sum t_i)^{\alpha+n}} \\ &= \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)}\beta^\alpha(\beta+\sum t_i)^{-(\alpha+n)} \quad (20) \end{aligned}$$

Note that

$$\begin{aligned} \mathbb{E}(t_i) &= \mathbb{E}(\mathbb{E}(t_i|\mu)) = \mathbb{E}(\mu) = \beta/(\alpha-1), \quad (21) \\ \mathbb{V}(t_i) &= \mathbb{V}(\mathbb{E}(t_i|\mu)) + \mathbb{E}(\mathbb{V}(t_i|\mu)) = \mathbb{V}(\mu) + \mathbb{E}(\mu^2) \\ &= 2\mathbb{V}(\mu) + (\mathbb{E}(\mu))^2 = \frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)}. \quad (22) \end{aligned}$$

The induced prior on the tree length T can be derived by noting that conditional on μ, T has a gamma distribution

$$f(T|\mu) = f(T|n,1/\mu) = \frac{1}{\mu^n\Gamma(n)}T^{n-1}e^{-T/\mu}, \quad (23)$$

so that

$$\begin{aligned} f(T) &= \int_0^\infty f(T|\mu)f(\mu)d\mu \\ &= \int_0^\infty \frac{1}{\mu^n\Gamma(n)}T^{n-1}e^{-T/\mu}\frac{\beta^\alpha}{\Gamma(\alpha)}\mu^{-\alpha-1}e^{-\beta/\mu}d\mu \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)\Gamma(n)}T^{n-1}\int_0^\infty e^{-(T+\beta)/\mu}\mu^{-\alpha-n-1}d\mu \\ &= \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)\Gamma(n)}\beta^\alpha T^{n-1}(T+\beta)^{-(\alpha+n)}. \quad (24) \end{aligned}$$

The mean and variance of T are

$$\begin{aligned}\mathbb{E}(T) &= \mathbb{E}\{\mathbb{E}(T|\mu)\} = n\mathbb{E}(\mu) = n\beta/(\alpha-1) \quad (25)\\ \mathbb{V}(T) &= \mathbb{V}\{\mathbb{E}(T|\mu)\} + \mathbb{E}\{\mathbb{V}(T|\mu)\}\\ &= \mathbb{V}(n\mu) + \mathbb{E}(n\mu^2)\\ &= \frac{n^2\beta^2}{(\alpha-1)^2(\alpha-2)}\\ &\quad + n\left[\frac{\beta^2}{(\alpha-1)^2} + \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}\right]\\ &= \frac{n\beta^2}{(\alpha-1)^2(\alpha-2)}(n+\alpha-1). \quad (26)\end{aligned}$$

The Compound Dirichlet Priors

We construct a prior of the $2s-3$ branch lengths by first specifying a fairly diffuse prior on the tree length $T = \sum_{i=1}^{2s-3} t_i$ and then partitioning the tree length into the branch lengths according to a Dirichlet distribution. Let T follow a gamma distribution

$$f(T) = \frac{\beta_T^{\alpha_T}}{\Gamma(\alpha_T)} T^{\alpha_T - 1} e^{-\beta_T T}, \quad (27)$$

with mean $\overline{T} = \alpha_T/\beta_T$ and variance α_T/β_T^2. The mean $\overline{T}$ could be assigned a hyperprior and integrated over in the MCMC. Parameter α_T can be chosen to reflect our confidence in the mean $\overline{T}$. Alternatively, both $\overline{T}$ and α_T can be varied to assess the impact of the prior on the posterior estimates. As a default value, we use $\alpha_T = 1$ so that $s_T/\overline{T} = 1/\sqrt{\alpha_T} = 1$, where s_T is the standard deviation of the tree length. We follow the suggestion of Yang and Rannala (2005) to have different prior means for the $s-3$ internal branch lengths and the s external branch lengths. This was suggested as a way of reducing the high posterior probabilities for trees and clades, similar to assigning nonzero probabilities to multifurcating trees (Lewis et al. 2005). Given T, define the ith branch length to be

$$t_i = x_i T, \quad i = 1, 2, \ldots, (2s-3), \quad (28)$$

where $x = \{x_i\}$ has the Dirichlet distribution

$$f(x) = \frac{1}{B(\alpha, c)} \prod_{j=1}^{s} x_j^{\alpha-1} \prod_{k=1}^{s-3} x_k^{\alpha c - 1}, \quad (29)$$

where subscripts j and k denote internal and external branches, respectively, with $x_j < 1$ and $x_k < 1$ for all j, k, and $\sum_j x_j + \sum_k x_k = 1$, and

$$B(\alpha, c) = \frac{\Gamma(\alpha)^s \Gamma(\alpha c)^{s-3}}{\Gamma(s\alpha + (s-3)\alpha c)}. \quad (30)$$

Parameter c is the ratio of the mean internal/external branch lengths. In general, c should be <1. The marginal means and variances of t_j and t_k are

$$\begin{aligned}\mathbb{E}(t_j) &= \mathbb{E}_T[\mathbb{E}_{t_j}(t_j|T)] = \mathbb{E}_T[T\mathbb{E}_{x_j}(x_j)]\\ &= \mathbb{E}_T\left[\frac{T\alpha}{s\alpha + (s-3)\alpha c}\right] = \frac{\overline{T}}{s + (s-3)c}, \quad (31)\\ \mathbb{E}(t_k) &= \mathbb{E}_T[\mathbb{E}_{t_k}(t_k|T) = \mathbb{E}_T[T\mathbb{E}_{x_k}(x_k)]\\ &= \mathbb{E}_T\left[\frac{T\alpha c}{s\alpha + (s-3)\alpha c}\right] = \frac{\overline{T}c}{s + (s-3)c}, \quad (32)\end{aligned}$$

and

$$\begin{aligned}\mathbb{V}(t_j) &= \frac{\alpha(\alpha_0 - \alpha)}{\alpha_0^2(\alpha_0+1)}\left(\frac{\alpha_T + \alpha_T^2}{\beta_T^2}\right) + \frac{\alpha^2\alpha_T}{\alpha_0^2\beta_T^2}, \quad (33)\\ \mathbb{V}(t_k) &= \frac{\alpha c(\alpha_0 - \alpha c)}{\alpha_0^2(\alpha_0+1)}\left(\frac{\alpha_T + \alpha_T^2}{\beta_T^2}\right) + \frac{\alpha^2 c^2\alpha_T}{\alpha_0^2\beta_T^2}, \quad (34)\end{aligned}$$

where $\alpha_0 = s\alpha + (s-3)\alpha c$.

The joint pdf of branch lengths $t = \{t_i\}$ conditional on T is

$$f(t|T) = \frac{1}{B(\alpha, c)} \prod_{j=1}^{s} (t_j/T)^{\alpha-1} \prod_{k=1}^{s-3} (t_k/T)^{\alpha c - 1} \times 1/T^{2s-4}. \quad (35)$$

By transforming the $(2s-3)$ variables $(x_1, x_2, \ldots, x_{2s-4}, T)$ to the variables $(t_1, t_2, \ldots, t_{2s-3})$, we obtain the joint prior density for the $(2s-3)$ branch lengths as

$$\begin{aligned}f(t) &= \frac{\beta_T^{\alpha_T}}{\Gamma(\alpha_T)} e^{-\beta_T \sum_{i=1}^{2s-3} t_i} \left(\sum_{i=1}^{2s-3} t_i\right)^{\alpha_T - 1}\\ &\quad \times \frac{1}{B(\alpha, c)} \prod_{j=1}^{s} t_j^{\alpha-1} \prod_{k=1}^{s-3} t_k^{\alpha c - 1}\\ &\quad \times \left(\sum_{i=1}^{2s-3} t_i\right)^{-\alpha s - \alpha c(s-3) + 1}. \quad (36)\end{aligned}$$

In particular, if we do not distinguish between the internal and external branch lengths (so that $c = 1$) and use a uniform Dirichlet distribution for the proportions $x_1, x_2, \ldots, x_{2s-4}$ (so that $\alpha = 1$), the joint prior density for the $(2s-3)$ branch lengths will be

$$\begin{aligned}f(t) &= \frac{\beta_T^{\alpha_T}}{\Gamma(\alpha_T)} e^{-\beta_T \sum t_i} \left(\sum t_i\right)^{\alpha_T - 1}\\ &\quad \times (2s-4)! \left(\sum t_i\right)^{-2s+4}. \quad (37)\end{aligned}$$

Note that this density is a function of the tree length $T = \sum t_i$ only.

As an alternative, we also consider the inverse gamma prior on the tree length T. This is heavier tailed than the gamma and may be useful if very little information is available about the tree length. Thus, instead of equation (27), we have

$$f(T) = \frac{\beta_T^{\alpha_T}}{\Gamma(\alpha_T)} T^{-\alpha_T - 1} e^{-\beta_T/T}. \quad (38)$$

We assume $\alpha_T > 2$, so that $\mathbb{E}(T) = \beta_T/(\alpha_T - 1)$ and $\mathbb{V}(T) = \beta_T^2/((\alpha_T - 1)^2(\alpha_T - 2))$. Parameters α_T and β_T can be varied to assess the impact of the prior on the posterior estimates. Given T, the x_i are assigned the Dirichlet distribution, as before. The joint distribution of branch lengths under this prior is thus

$$f(t) = \frac{\beta_T^{\alpha_T}}{\Gamma(\alpha_T)} e^{-\beta_T/\sum_{i=1}^{2s-3} t_i} \left(\sum_{i=1}^{2s-3} t_i\right)^{-\alpha_T - 1} \times \frac{1}{B(\alpha, c)} \prod_{j=1}^{s} t_j^{\alpha-1} \prod_{k=1}^{s-3} t_k^{\alpha c - 1} \times \left(\sum_{i=1}^{2s-3} t_i\right)^{-\alpha s - \alpha c (s-3) + 1}, \quad (39)$$

in place of equation (36). If $c = 1$ and $\alpha = 1$, the joint prior density for the branch lengths will be

$$f(t) = \frac{\beta_T^{\alpha_T}}{\Gamma(\alpha_T)} e^{-\beta_T/\sum t_i} \left(\sum t_i\right)^{-\alpha_T - 1} \times (2s-4)! \left(\sum t_i\right)^{-2s+4}, \quad (40)$$

in place of equation (37). We recommend $\alpha_T = 3$ so that $s_T/\overline{T} = \frac{1}{\sqrt{\alpha_T - 2}} = 1$, indicating a fairy diffuse prior. The scale parameter β_T can be chosen to suit different data sets.

For the inverse gamma compound Dirichlet distribution, the marginal means of the branch lengths are given by equations (31) and (32) as before, and the marginal variances are

$$\mathbb{V}(t_j) = \frac{\beta_T^2}{(\alpha_T - 1)^2(\alpha_T - 2)} \left((\alpha_T - 1)\frac{\alpha(\alpha_0 - \alpha)}{\alpha_0^2(\alpha_0 + 1)} + \frac{\alpha^2}{\alpha_0^2}\right), \quad (41)$$

$$\mathbb{V}(t_k) = \frac{\beta_T^2}{(\alpha_T - 1)^2(\alpha_T - 2)} \left((\alpha_T - 1)\frac{c\alpha(\alpha_0 - c\alpha)}{\alpha_0^2(\alpha_0 + 1)} + \frac{c^2\alpha^2}{\alpha_0^2}\right). \quad (42)$$

As an illustration of the differences among those priors on branch lengths, we calculated the mean and standard deviation of the induced prior on tree length for the six data sets analyzed by Brown et al. (2010). We also calculated the probability that the tree length is less than its MLE according to the prior: $\Pr(T < \hat{T})$. When this probability is very small, one may consider the prior to be in conflict with the data (the likelihood). The results are shown in table 2. It is clear that in all these data sets the default MrBayes prior on branch lengths (an exponential with mean 0.1) predicts unreasonably long trees compared with the data. Use of MLEs to specify prior means helps but the procedure is non-Bayesian. The two compound Dirichlet priors provide reasonable branch length distributions for almost all data sets.

We leave it to future work to implement the two new priors of branch lengths in Bayesian phylogenetics programs such as MrBayes. To illustrate the differences among the priors discussed in this paper, we wrote a C program for Bayesian estimation of branch lengths assuming the star phylogeny under the JC69 and JC69+Γ_5 models. We implemented three moves: 1) a multidimensional sliding window to modify the branch lengths, 2) a proportional scaling which multiplies all branch lengths by a random variable that is close to 1, and 3) a sliding window to modify the shape parameter α in the JC69+Γ_5 model.

Partial Identifiability of the Substitution Model and Variable Rates among Sites

Both Brown et al. (2010) and Marshall (2010) observed that unreasonable branch length estimates occurred more often when partition models or models of variable rates among sites (such as the Γ+I model) were used (see also table 3). For such models, the parameters may be only partially identifiable; there may be strong correlations between parameters in the likelihood. This can lead to mixing problems in the MCMC. Here, we present evidence for strong correlations among parameters in empirical data sets. We analyze two data sets (one real and one simulated) to explore the correlation between the shape parameter α of the gamma distribution for variable rates among sites (Yang 1994) and the branch lengths.

First, we consider the estimation of the distance between the human and chimpanzee 12S tRNA genes under the JC69+Γ model. Parameters θ and α are not identifiable: with only two sequences, the likelihood L depends on only the probability p that a site shows a difference, so that L is constant for combinations of θ and α that produce the same p.

$$p = \frac{3}{4} - \frac{3}{4}\left(1 + \frac{4\theta}{3\alpha}\right)^{-\alpha}, \quad (43)$$

$$L(\theta, \alpha) = p^x(1-p)^{n-x} = \left[\frac{3}{4} - \frac{3}{4}\left(1 + \frac{4\theta}{3\alpha}\right)^{-\alpha}\right]^x \left[\frac{1}{4} + \frac{3}{4}\left(1 + \frac{4\theta}{3\alpha}\right)^{-\alpha}\right]^{n-x}. \quad (44)$$

The logarithm of the likelihood is plotted in figure 3*a* as a function of θ and α. An exponential prior with mean 0.1 is assigned on θ and a uniform prior $U(0, 50)$ on α:

$$f(\theta) = \frac{1}{\mu} e^{-\theta/\mu}, \quad f(\alpha) = \frac{1}{50}, 0 < \theta < 50. \quad (45)$$

The posterior is given as

$$f(\theta, \alpha|x) = \frac{1}{C} f(\alpha) f(\theta) L(\theta, \alpha) = \frac{1}{C} \cdot \frac{1}{50} \cdot 10 e^{-10\theta} \cdot L(\theta, \alpha), \quad (46)$$

Table 2. Characterization of the Prior Distribution of Tree Length (T) under Different Branch Length Priors Applied to Five Data sets Analyzed by Brown et al. (2010).

Prior	Mean	SD	$\Pr\{T < \{\hat{T}\}\}$
SimulatedA (Brown and Lemmon 2007): s = 29, $\hat{T} = 0.12$			
exp: G(2s - 3, 10)	(2s - 3)/10 = 5.5	$\sqrt{2s-3}/10$ = 0.7416	5.5×10^{-70}
Brown median $G\,(2s-3,(2s-3)/1.4427\hat{T})$	$1.4427\hat{T} = 0.173$	$1.4427\hat{T}/\sqrt{2s-3} =$ 0.023	0.006
Brown mean $G(2s-3,(2s-3)/\hat{T})$	$\hat{T} = 0.12$	$\hat{T}/\sqrt{2s-3} = 0.016$	0.52
Suchard(2.1, 1.1)	55	175.7	2.9×10^{-54}
Suchard(3.0, 0.2)	5.5	5.60	2.4×10^{-21}
Gamma Dirichlet(1, 1)	1	1	0.11
invGamma Dirichlet(3, 2)	1	1	9.0×10^{-6}
Frogs (Gamble et al. 2008) s = 66, $\hat{T} = 0.64$			
exp: G(2s - 3, 10)	12.9	1.14	3.5×10^{-117}
Brown median $G\,(2s-3,(2s-3)/1.4427\hat{T})$	0.923	0.081	0.000049
Brown mean $G(2s-3,(2s-3)/\hat{T})$	0.64	0.056	0.51
Suchard(2.1, 1.1)	129	409.7	1.1×10^{-54}
Suchard(3.0, 0.2)	12.9	13.0	3.0×10^{-13}
Gamma Dirichlet(1, 1)	1	1	0.47
invGamma Dirichlet(3, 2)	1	1	0.40
Clams (Hedtke et al. 2008) s = 93, $\hat{T} = 1.96$			
exp: G(2s - 3, 10)	18.3	1.35	8.6×10^{-109}
Brown median $G\,(2s-3,(2s-3)/1.4427\hat{T})$	2.83	0.209	1.7×10^{-6}
Brown mean $G(2s-3,(2s-3)/\hat{T})$	1.96	0.145	0.51
Suchard(2.1, 1.1)	183	580.4	3.8×10^{-34}
Suchard(3.0, 0.2)	18.3	18.4	3.1×10^{-6}
Gamma Dirichlet(1, 1)	1	1	0.86
invGamma Dirichlet(3, 2)	1	1	0.92
Lizards (Leache and Mulcahy 2007), s = 123, $\hat{T} = 2.48$			
exp: G(2s - 3, 10)	24.3	1.56	2.3×10^{-148}
Brown median $G\,(2s-3,(2s-3)/1.4427\hat{T})$	3.6	0.23	4.1×10^{-8}
Brown mean $G(2s-3,(2s-3)/\hat{T})$	2.5	0.16	0.51
Suchard(2.1, 1.1)	243	770.2	2.0×10^{-37}
Suchard(3.0, 0.2)	24.3	24.4	1.2×10^{-6}
Gamma Dirichlet(1, 1)	1	1	0.92
invGamma Dirichlet(3, 2)	1	1	0.95
Froglets (Symulaet al. 2008), s = 92, $\hat{T} = 0.55$			
exp: G(2s - 3, 10)	18.1	1.35	1.2×10^{-200}
Brown median $G\,(2s-3,(2s-3)/1.4427\hat{T})$	0.79	0.06	1.9×10^{-6}
Brown mean $G(2s-3,(2s-3)/\hat{T})$	0.55	0.041	0.51
Suchard(2.1, 1.1)	181	574.1	8.2×10^{-85}
Suchard(3.0, 0.2)	18.1	18.2	5.1×10^{-22}
Gamma Dirichlet(1, 1)	1	1	0.42
invGammaDirichlet(3, 2)	1	1	0.30

where the normalizing constant $C = 7.39 \times 10^{-28}$. The logarithm of the posterior density is shown in figure 3*b*.

Second, we analyzed a data set of 200 sequences simulated on a star tree with all 200 branch lengths to be 0.01. The sequence length was 500 sites. We then analyzed the data assuming a star tree and applying both ML and Bayesian methods (using several different priors) under both the JC69 and JC69+Γ_5 models. Under JC69+Γ_5 a uniform prior $U\,(0.05, 50)$ was assigned on the shape parameter α, which is the default prior in MrBayes. The true tree length is 2. The MLE is $\hat{T} = 2.13$ under JC69 and 2.15 under JC69+Γ_5. The posterior means and 95% equal-tail credibility intervals are shown in table 3. With the uniform and exponential priors, which assume i.i.d. distributions for the branch lengths, the MLE is outside the 95% confidence intervals (CIs). In particular, the correlation between the gamma shape parameter α and branch lengths θ under JC69+Γ_5 (see below) has generated extremely long tree lengths under the uniform and exponential priors. With the use of the two new priors suggested in this paper, the MLE is well within the 95% posterior CIs under both JC69 and JC69+Γ_5.

Discussion

In this paper, we have used a combination of analytical theory and simulations to explore the potential causes of the excessively long branch lengths (and tree lengths) observed

Table 3. Posterior Means and 95% Equal-Tail Credibility Intervals of the Tree Length in the Analysis of a Simulated Data set under Different Branch-Length Priors.

Prior		Posterior Mean (95% CI)	Posterior α
$JC69(\mathbf{T} = 2.1337)$			
Uniform	$t_i \sim \mathrm{U}(0, 100)$	**2.540 (2.401, 2.683)**	
Exponential	$t_i \sim \mathrm{E}(1/\mu), \mu = 0.1$	**2.489 (2.353, 2.627)**	
Suchard(2.1, 1.1)	$t_i \mid \mu \sim, Exp\,(1/\mu), \mu \sim invG\,(2.1, 1.1)$	**2.265 (2.136, 2.398)**	
Suchard(3, 0.2)	$t_i \mid \mu \sim Exp\,(1/\mu), \mu \sim invG\,(3, 0.2)$	**2.162 (2.035, 2.293)**	
Gamma Dirichlet	$\mathbf{T} \sim G\,(1, 1)$	**2.132 (2.006, 2.261)**	
Inverse gamma Dirichlet	$\mathbf{T} \sim invG\,(3, 2), t_i \mid \mathbf{T} \sim Dir\,(1, 1, \ldots)$	**2.130 (2.002, 2.261)**	
$JC\,69 + G\,5(\mathbf{T} = 2.1519, \hat{\alpha} = 0.4848)$			
Uniform	$t_i \sim \mathrm{U}(0, 100)$	**11942 (11205, 12661)**	**0.077 (0.076, 0.078)**
Exponential	$t_i \sim Exp\,(1/\mu), \mu = 0.1$	**11.37 (10.02, 12.80)**	**0.223 (0.193, 0.257)**
Suchard(2.1, 1.1)	$t_i \mid \mu \sim Exp\,(1/\mu), \mu \sim invG\,(2.1, 1.1)$	**2.569 (2.315, 2.852)**	**0.531 (0.449, 0.623)**
Suchard(3, 0.2)	$t_i \mid \mu \sim Exp\,(1/\mu), \mu \sim invG\,(3, 0.2)$	**2.247 (2.027, 2.489)**	**0.501 (0.421, 0.591)**
Gamma Dirichlet	$\mathbf{T} \sim G\,(1, 1), t_i \mid \mathbf{T} \sim Dir\,(1, 1, \ldots)$	**2.151 (1.937, 2.384)**	**0.489 (0.409, 0.581)**
Inverse gamma Dirichlet	$\mathbf{T} \sim invG\,(3, 2), t_i \mid \mathbf{T} \sim Dir\,(1, 1, \ldots)$	**2.147 (1.934, 2.381)**	**0.489 (0.409, 0.581)**

NOTE.—Data are simulated under $JC\,69 + \Gamma_5$ with $\alpha = 0.5$ on a star tree of 200 sequences, with all 200 branch lengths to be 0.01. The sequence length is 500. The data are analyzed assuming the star tree by ML and Bayesian methods using different priors under the JC69 and $JC\,69 + \Gamma_5$ models. Under $JC\,69 + \Gamma_5$, an uniform prior $U\,(0.05, 50)$ is assigned on α, which is the default prior used in MrBayes.

in Bayesian MCMC phylogenetic analyses (e.g., Brown et al. 2010; Marshall 2010). Our analyses indicate that the choice of the prior and the initial branch lengths for the MCMC may have a critical influence. The default prior in MrBayes is exponential with mean 0.1. The default initial values of the branch lengths for the MCMC are also 0.1. For many data sets, these values are much greater than the mode of the posterior distribution—thus, the prior places too much weight on extreme branch lengths and the chain is initiated with the branch lengths far out on the tail of the posterior density. Thus, both the rate of convergence may be slow and the posterior may be unduly influenced by the default prior in analyses using MrBayes. Because the tree length depends on the number of sequences sampled, the prior becomes more unrealistic for larger data sets with more species. If the data are not very informative or if there are strong correlations in the posterior, as in the rate variation or partition models, the large prior mean of tree length can exert a strong influence on the posterior causing overestimates of the branch lengths even when the chain has converged.

We explore the convergence properties of the Metropolis–Hastings MCMC algorithm for various forms of the posterior analytically. Existing theories (Mengersen and Tweedie 1996; Roberts and Tweedie 1996) suggest that if the posterior is univariate and light tailed, the MCMC is guaranteed to converge at a geometric rate. In higher dimensions, similar geometric convergence can be achieved if the posterior satisfies certain "smoothness" conditions in addition. If the posterior is heavy tailed, the MCMC will not converge at a geometric rate and may wander in the tails of the posterior for a very long time.

To improve MCMC convergence and to yield a prior that is less informative, we propose a new class of "compound Dirichlet" priors under which the branch lengths follow a Dirichlet distribution given the tree length, while the tree length is assigned a gamma or inverse gamma distribution. Our calculations suggest that our new priors for branch lengths induce reasonable priors on tree length not influenced by the number of sequences sampled. Our analysis of a simulated data set with 200 sequences on a star tree confirms that the current exponential and uniform priors in MrBayes can lead to gross overestimates of tree length and branch lengths and that the problem may become worse when the likelihood model incorporates variable rates among sites. In contrast, the new compound Dirichlet priors produce reasonable posterior estimates.

We note that in general specification of multidimensional priors in Bayesian inference is challenging. In this paper, we have attempted to specify reasonable default priors that may be usable in phylogenetic analysis of all data sets. From the subjective Bayesian viewpoint, one may argue for a careful specification of the prior to reflect one's personal belief for each data set being analyzed. However, such a procedure is never practized. Neither does it seem practicable, except for small data sets with very few species. It is already challenging to specify the $(2s - 3)$-dimensional joint prior for the branch lengths on a fixed tree topology, and the task of doing that for each of the huge number of possible tree topologies is simply too daunting. From the objective Bayesian viewpoint, one would use an uninformative prior so that the posterior reflects information in the data rather than the researcher's personal biases. However, it is now generally acknowledged that no prior is truly uninformative. The reference prior is a default prior that leads to the same inference however the model is parametrized. For the estimation of the distance between two sequences, Ferreira and Suchard (2008) found that the Bayesian estimate under the reference prior had better frequentist properties than the MLE. However, that analysis was based on the assumption that one has information about the substitution parameters, and furthermore, the likelihood function (eq. 4 in Ferreira and Suchard 2008) appears to be incorrectly

defined, with a term of equilibrium base frequency being missing (cf. eq. 1.65 in Yang 2006). Specification of the reference prior requires the likelihood function and the Fisher information matrix, so that it is unclear whether the idea can be made to work for more than two sequences. Another promising approach is to use relaxed-clock models to infer both rooted and unrooted trees (Drummond et al. 2006), with the prior for branch lengths generated by specifying a prior on the divergence times and another prior on the rates for branches based on a model of stochastic rate drift (Thorne et al. 1998). This approach may create a problem of partial unidentifiability since the sequence data provide information about the branch lengths but not about the times and rates individually and as a result there is no convergence to the true parameter values when the amount of sequence data approaches infinity (Yang and Rannala 2006). The approach has also been noted to generate highly counterintuitive priors on divergence times with multiple modes (Heled and Drummond 2011, figure 2). In summary, those recent studies as well as the present one highlight the complexity and importance of specifying the joint prior on branch lengths in Bayesian phylogenetic inference.

Acknowledgments

Part of this research was completed while the authors were guests of Institute of Zoology, Chinese Academy of Sciences, Beijing, supported by the Center for Computational and Evolutionary Biology. Z.Y. gratefully acknowledges the support of K.C. Wong Education Foundation, Hong Kong. Z.Y. was supported by a Biotechnological and Biological Sciences Research Council (BBSRC) grant and a Royal Society Wolfson Merit Award.

References

Brown JM, Hedtke SM, Lemmon AR, Lemmon EM. 2010. When trees grow too long: investigating the causes of highly inaccurate Bayesian branch-length estimates. *Syst Biol.* 59:145–161.

Drummond AJ, Ho SY, Phillips M, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4:e88.

Ferreira M, Suchard M. 2008. Bayesian analysis of elapsed times in continuous-time Markov chains. *Can J Statist.* 36:355–368.

Heled J, Drummond AJ. 2011. Forthcoming. Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Syst Biol.*

Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N. 1995. Recent African origins of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc Natl Acad Sci U S A.* 92:532–536.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism, volume III. New York: Academic Press. p. 21–132.

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

Marshall DC. 2010. Cryptic failure of partitioned Bayesian phylogenetic analyses: lost in the land of long trees. *Syst Biol.* 59:108–117.

Mengersen KL, Tweedie RL. 1996. Rates of convergence of the Hastings and Metropolis algorithms. *Ann Statist.* 24:101–121.

Mossel E, Vigoda E. 2005. Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science* 309:2207–2209.

Roberts GO, Tweedie RL. 1996. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* 83:95–110.

Rogers J, Swofford D. 1999. Multiple local maxima for likelihoods of phylogenetic trees: a simulation study. *Mol Biol Evol.* 16:1079–1085.

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

Suchard MA, Weiss RE, Sinsheimer JS. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. *Mol Biol Evol.* 18:1001–1013.

Symula R, Keogh JS, Cannatella DC. 2008. Ancient phylogeographic divergence in southeastern Australia among populations of the widespread common froglet, Crinia signifera. *Mol Phy Evol.* 47:569–580.

Thorne J, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol.* 15:1647–1657.

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

Yang Z. 2006. Computational molecular evolution. Oxford (UK): Oxford University Press.

Yang Z, Rannala B. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. *Syst Biol.* 54:455–470.

Yang Z, Rannala B. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol Biol Evol.* 23:212–226.