Choosing among Partition Models in Bayesian Phylogenetics

Yu Fan,[1] Rui Wu,[2] Ming-Hui Chen,[2] Lynn Kuo,[2] and Paul O. Lewis*,[1]

[1]Department of Ecology and Evolutionary Biology, University of Connecticut
[2]Department of Statistics, University of Connecticut

***Corresponding author:** E-mail: paul.lewis@uconn.edu.

Associate editor: Alexei Drummond

Abstract

Bayesian phylogenetic analyses often depend on Bayes factors (BFs) to determine the optimal way to partition the data. The marginal likelihoods used to compute BFs, in turn, are most commonly estimated using the harmonic mean (HM) method, which has been shown to be inaccurate. We describe a new more accurate method for estimating the marginal likelihood of a model and compare it with the HM method on both simulated and empirical data. The new method generalizes our previously described stepping-stone (SS) approach by making use of a reference distribution parameterized using samples from the posterior distribution. This avoids one challenging aspect of the original SS method, namely the need to sample from distributions that are close (in the Kullback–Leibler sense) to the prior. We specifically address the choice of partition models and find that using the HM method can lead to a strong preference for an overpartitioned model. In contrast to the HM method and the original SS method, we show using simulated data that the generalized SS method is strikingly more precise (repeatable BF values of the same data and partition model) and yields BF values that are much more reasonable than those produced by the HM method. Comparisons of HM and generalized SS methods on an empirical data set demonstrate that the generalized SS method tends to choose simpler partition schemes that are more in line with expectation based on inferred patterns of molecular evolution. The generalized SS method shares with thermodynamic integration the need to sample from a series of distributions in addition to the posterior. Such dedicated path-based Markov chain Monte Carlo analyses appear to be a cost of estimating marginal likelihoods accurately.

Key words: phylogenetics, Bayes factor, marginal likelihood, harmonic mean method, stepping-stone method, partitioning.

Introduction

Partitioned analyses are now routine for multi-gene data sets in Bayesian phylogenetics (Nylander et al. 2004; Brandley et al. 2005; Brown and Lemmon 2007; Clarke and Middleton 2008; Brown et al. 2009; Liu et al. 2010). It is widely known that different genes or different codon positions experience different selection pressures. By partitioning, a better fit of model to data can be achieved, and the models used better reflect the molecular evolutionary forces at work. However, more partitions or more complex models mean more parameters are estimated, increasing the variability of estimates given a fixed and finite amount of data. The question is how to choose an economical partition strategy for the data that allows the model to fit the data well but discourages unnecessary partitions that contribute little to goodness-of-fit.

The Bayes factor (BF) has been shown to be a useful criterion for model selection in Bayesian inference:

$$\mathrm{BF}_{01} = \frac{f(\mathbf{y}|M_0)}{f(\mathbf{y}|M_1)}.$$

The BF is the ratio of the marginal likelihood under one model, $f(\mathbf{y}|M_0)$, to the marginal likelihood under an alternative model, $f(\mathbf{y}|M_1)$, for fixed data $\mathbf{y}$. If the ratio is larger than 1.0, model M_0 is favored; if less than 1.0, model M_1 is favored. The language of odds ratios is used in discussions of BFs: For example, BF_{01} represents the BF for model M_0 and against model M_1. The marginal likelihood of model M is a weighted average (expected value) of the likelihood, $f(\mathbf{y}|\boldsymbol{\theta}, M)$, where the weights are provided by the prior, $\pi(\boldsymbol{\theta}|M)$, and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ may be multidimensional and model specific:

$$f(\mathbf{y}|M) = \int_{\boldsymbol{\Theta}} f(\mathbf{y}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)\mathrm{d}\boldsymbol{\theta}.$$

The marginal likelihood is thus a measure of the average fit of model M to data $\mathbf{y}$, which contrasts with the maximized likelihood used by likelihood ratio tests (Wilks 1938), the Akaike information criterion (Akaike 1974), and the Bayesian information criterion (Schwarz 1978), all of which make use of the fit of the model at its best-fitting point in parameter space $\boldsymbol{\Theta}$.

The estimation of marginal likelihoods is a challenging task because no closed-form expression exists for most phylogenetic applications. The solution has been to resort to numerical approximation using Markov chain Monte Carlo (MCMC), and many methods have been proposed, including the harmonic mean (HM) method (Newton and Raftery 1994), bridge sampling (Meng and Wong 1996), path sampling (Gelman and Meng 1998), thermodynamic integration (TI; Lartillot and Philippe 2006), reversible jump MCMC (Huelsenbeck et al. 2004), and the stepping-stone (SS) method (Xie et al. 2010). Among these, the HM of the likelihoods computed from samples taken from the Bayesian posterior probability distribution is the most broadly used

in Bayesian phylogenetics. The popularity of the HM estimator is due to its easy calculation and the fact that currently no other choice is provided by most Bayesian phylogenetics software. In contrast to its popularity in Bayesian phylogenetics, HM has been controversial in the statistics community from the moment it was proposed due to the fact that it is a biased estimator of the marginal likelihood (the estimate is expected to be higher than the true value; Xie et al. 2010) and has a large and unpredictable variance (which can be infinite) (Neal 1994).

Recently, improved means of estimating marginal likelihoods have been introduced into Bayesian phylogenetics. Lartillot and Philippe (2006) described TI and Xie et al. (2010) introduced the SS method, both of which exceed HM greatly in both accuracy and precision. The Savage–Dickey ratio (Verdinelli and Wasserman 1995; Suchard et al. 2001) and reversible jump MCMC (Huelsenbeck et al. 2004) can also be used to accurately estimate the BF directly when models are nested.

The primary question addressed in this paper is: "If the marginal likelihoods of models were more accurately estimated, would less-partitioned models be used in Bayesian phylogenetic analyses?" This is an important question because large Bayesian analyses generally require longer run times, which leads to small effective sample sizes if run times are not adjusted to be proportional to the number of estimated parameters. Reducing the number of data subsets should therefore yield higher effective sample sizes for a given amount of computational effort. Unnecessarily complex models also have more diffuse posterior distributions, so using less-partitioned models is expected to increase confidence in the inferences made. Finally, partitioned analyses can lead to bizarre parameter estimates. For example, it is possible for second codon positions to appear to evolve faster than first or even third codon positions (Marshall 2010), and the estimated proportion of invariable sites can be as high as 0.96 even when all sites are variable (Appendix 1). Such abnormalities do not appear to occur with unpartitioned models. Although eliminating all partitions is an extreme solution that may reduce performance due to poor goodness-of-fit (Brown and Lemmon 2007), using a more accurate marginal likelihood estimator may favor a less-partitioned model that alleviates some of these pathologies without reducing goodness-of-fit appreciably.

Another question addressed is: "Is it possible to further improve the efficiency of SS (Xie et al. 2010) so that results of comparable accuracy can be obtained with less computational effort?" Inspired by the geometric path approach taken in Lefebvre et al. (2010), we show that use of a straightforward reference distribution substantially increases the computational efficiency of SS.

Our interest in the relationship of HM to partitioning was initially aroused by figure 6 of Brown and Lemmon (2007), which plotted twice the natural logarithm of the BF (estimated using HM) for a partitioned model against an unpartitioned model when the unpartitioned model was the true model. In such cases, partitioning is unnecessary and while 2log(BF) is not guaranteed to be less than zero in this case, it is reasonable to expect few, if any, positive values. In contrast to expectation, approximately 31% of Brown and Lemmon's $2\log(\mathrm{BF})$ values were above 0, and more than 5% were above 10. This means that in nearly one third of the simulated data sets analyzed, a clearly overpartitioned model would have been chosen using HM-based BF comparisons. We decided to conduct a study similar to the one represented in figure 6 of Brown and Lemmon (2007) to evaluate the effect of marginal likelihood estimation accuracy on model choice. Our expectation was that few, if any, data sets simulated under an unpartitioned model would be chosen by a BF for a partitioned model against the unpartitioned (true) model when marginal likelihoods of the two models were estimated accurately.

Materials and Methods

Simulated Data

Simulations were similar to those described in Xie et al. (2010). The number of taxa in each simulated data set was decided by drawing from the set of integers from 4 to 20 uniformly (and inclusively), and the number of nucleotide sites was an even number uniformly chosen from 100 to 5,000. For each simulated data set, a tree topology was chosen at random from all possible labeled unrooted binary tree topologies (i.e., the proportional-to-distinguishable model), and internal branch lengths, external branch lengths, base frequencies, and general time reversible (GTR) exchangeabilities were drawn from Gamma(10,0.001), Gamma(1,0.1), Dirichlet(100,100,100,100), and Dirichlet (100,100,100,100,100,100) distributions, respectively. The discrete gamma distribution (ten categories) was used to impart among-site rate heterogeneity, with the gamma shape parameter drawn from a Gamma(2,3) distribution. The 200 data sets used for this example were thus independently and identically distributed. Although, technically, this generating model produced all data sets from a GTR + G distribution, the distributions of parameters were chosen so that many simulation replicates come close to various submodels of the GTR + G model. For example, the Gamma(2,3) distribution used to choose the shape parameter for among-site rate heterogeneity produces values greater than 5 (i.e., effective rate homogeneity) about 50% of the time. Thus, about 50% of data sets could be fit nearly as well by the GTR model as by the (true) GTR + G model. Previously (Xie et al. 2010), we used this simulation scheme to address choice of substitution models in the context of unpartitioned analyses; here, our purpose is to instead explore choice among different possible ways to partition data.

Empirical Data

In addition to simulations, we also evaluated empirical data that have been a focal point for discussions of artifacts associated with partitioning in Bayesian analyses (Marshall et al. 2006). This data set is available at TreeBase (http://www.treebase.org/, study accession number S1679) and comprises 32 taxa and 2,152 nucleotide sites. For our

analyses, we omitted the tRNA gene, leaving data for four protein-coding genes (COI, COII, ATPase8, and ATPase6), and reducing the total number of sites to 2,090.

Generalized SS Method

We describe here a modification of the SS method described by Xie et al. (2010). The modified version is considerably more efficient and does not require sampling from distributions close to the prior (which can be problematic for vague priors). This generalized SS introduces a reference distribution, which in practice is a product of independent probability densities parameterized using samples from the posterior distribution. Although the original SS method does not require samples from the posterior distribution, in practice, the posterior is explored as a means of burning-in the chain and this modified version uses this burn-in period to parameterize its reference distribution. However, if samples from a previous extensive MCMC analysis of the posterior are available, it is advisable to use this previous sample to parameterize the reference distribution. A reference distribution that approximates the posterior closely requires less computational effort to accurately estimate the marginal likelihood. Remarkably, as we will later show, if the reference distribution exactly equals the posterior distribution, the marginal likelihood can be estimated exactly (i.e., in this case, MCMC sampling error does not affect the estimate).

Consider the unnormalized density function q_β, which has normalizing constant c_β yielding the normalized density p_β:

$$q_\beta = [f(\mathbf{y}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)]^{\beta}[\pi_0(\boldsymbol{\theta}|M)]^{1-\beta},$$

$$p_\beta = q_\beta / c_\beta,$$

$$c_\beta = \int_{\Theta} q_\beta \, \mathrm{d}\boldsymbol{\theta},$$

where $\mathbf{y}$ represents the data (e.g., nucleotide sequences), $\boldsymbol{\theta}$ is the vector of model parameters, M is the model under consideration, $f(\mathbf{y}|\boldsymbol{\theta}, M)$ is the likelihood function, $\pi(\boldsymbol{\theta}|M)$ the actual model prior, and $\pi_0(\boldsymbol{\theta}|M)$ is the reference distribution. The density p_β is a form of power posterior that is equivalent to the posterior distribution when $\beta = 1$ but equivalent to the reference distribution when $\beta = 0$. This differs from the original SS method (Xie et al. 2010), where the actual prior distribution is sampled when $\beta = 0$. The power posterior can be difficult to sample if β is near 0.0 and the prior is diffuse (normally the case), so using a reference distribution facilitates sampling from p_β regardless of the value of β. The goal is to estimate the ratio $c_{1.0}/c_{0.0}$, which is equivalent to the marginal likelihood because $c_{0.0} = 1.0$ if the reference distribution is proper (which is assumed throughout). Similar to the original SS method, this ratio can be expressed as a product of K ratios:

$$r = \frac{c_{1.0}}{c_{0.0}} = \prod_{k=1}^{K} \frac{c_{\beta_k}}{c_{\beta_{k-1}}},$$

where $0 = \beta_0 < \cdots < \beta_{k-1} < \beta_k < \cdots < \beta_K = 1$. Each ratio $c_{\beta_k}/c_{\beta_{k-1}}$ is estimated by importance sampling, using $p_{\beta_{k-1}}$ as the importance sampling density. Because $p_{\beta_{k-1}}$ is only slightly different from p_{β_k}, it serves as an excellent importance distribution. One of the K ratios, r_k, can thus be expressed as follows:

$$\begin{aligned}
r_k &= \frac{c_{\beta_k}}{c_{\beta_{k-1}}} \\
&= \frac{\int q_{\beta_k}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{\int q_{\beta_{k-1}}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \\
&= \frac{\int \left(\frac{q_{\beta_k}(\boldsymbol{\theta})}{p_{\beta_{k-1}}(\boldsymbol{\theta})}\right) p_{\beta_{k-1}}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{\int \left(\frac{q_{\beta_{k-1}}(\boldsymbol{\theta})}{p_{\beta_{k-1}}(\boldsymbol{\theta})}\right) p_{\beta_{k-1}}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \\
&= \frac{\int \left(\frac{q_{\beta_k}(\boldsymbol{\theta})}{q_{\beta_{k-1}}(\boldsymbol{\theta})/c_{\beta_{k-1}}}\right) p_{\beta_{k-1}}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{\int \left(\frac{q_{\beta_{k-1}}(\boldsymbol{\theta})}{q_{\beta_{k-1}}(\boldsymbol{\theta})/c_{\beta_{k-1}}}\right) p_{\beta_{k-1}}(\boldsymbol{\theta})d\boldsymbol{\theta}} \\
&= \int \left(\frac{q_{\beta_k}(\boldsymbol{\theta})}{q_{\beta_{k-1}}(\boldsymbol{\theta})}\right) p_{\beta_{k-1}}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\
&= \int \left(\frac{[f(\mathbf{y}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)]^{\beta_k}[\pi_0(\boldsymbol{\theta}|M)]^{1-\beta_k}}{[f(\mathbf{y}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)]^{\beta_{k-1}}[\pi_0(\boldsymbol{\theta}|M)]^{1-\beta_{k-1}}}\right) \\
&\qquad \times p_{\beta_{k-1}}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\
&= E_{p_{\beta_{k-1}}}\left[\left(\frac{f(\mathbf{y}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)}{\pi_0(\boldsymbol{\theta}|M)}\right)^{\beta_k-\beta_{k-1}}\right]. \qquad (1)
\end{aligned}$$

An estimator $\hat{r}_k$ is constructed using samples $\boldsymbol{\theta}_{k-1,i}$ $(i = 1, 2, \ldots, n)$ from $p_{\beta_{k-1}}$:

$$\hat{r}_k = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{f(\mathbf{y}|\boldsymbol{\theta}_{k-1,i}, M)\pi(\boldsymbol{\theta}_{k-1,i}|M)}{\pi_0(\boldsymbol{\theta}_{k-1,i}|M)}\right]^{\beta_k-\beta_{k-1}}.$$

Numerical stability can be improved by factoring out the largest sampled term, $\eta_k = \max_{1\leqslant i\leqslant n}\{f(\mathbf{y}|\theta_{k-1,i}, M)\pi(\boldsymbol{\theta}_{k-1,i}|M)/\pi_0(\boldsymbol{\theta}_{k-1,i}|M)\}$:

$$\begin{aligned}
\hat{r}_k &= \frac{1}{n}(\eta_k)^{\beta_k-\beta_{k-1}} \\
&\quad \times \sum_{i=1}^{n}\left[\frac{f(\mathbf{y}|\boldsymbol{\theta}_{k-1,i}, M)\pi(\boldsymbol{\theta}_{k-1,i}|M)}{\eta_k\pi_0(\boldsymbol{\theta}_{k-1,i}|M)}\right]^{\beta_k-\beta_{k-1}}.
\end{aligned}$$

On the log scale,

$$\begin{aligned}
\log\hat{r}_k &= (\beta_k - \beta_{k-1})\log\eta_k \\
&\quad + \log\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{f(\mathbf{y}|\boldsymbol{\theta}_{k-1,i}, M)\,\pi(\boldsymbol{\theta}_{k-1,i}|M)}{\eta_k\,\pi_0(\boldsymbol{\theta}_{k-1,i}|M)}\right]^{\beta_k-\beta_{k-1}}\right\}.
\end{aligned}$$

Finally, summing $\log \hat{r}_k$ over all K ratios yields the overall estimator:

$$\log \hat{r} = \sum_{k=1}^{K} \log \hat{r}_k$$

$$= \sum_{k=1}^{K} [(\beta_k - \beta_{k-1}) \log \eta_k]$$

$$+ \sum_{k=1}^{K} \log \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[\frac{f(\mathbf{y}|\boldsymbol{\theta}_{k-1,i}, M)\, \pi(\boldsymbol{\theta}_{k-1,i}|M)}{\eta_k\, \pi_0(\boldsymbol{\theta}_{k-1,i}|M)} \right]^{\beta_k - \beta_{k-1}} \right\}. \quad (2)$$

This approach reduces to the original SS method if the reference distribution is equal to the actual prior. However, the reference distribution would normally be chosen to be closer (in the Kullback–Leibler sense) to the posterior than the actual prior, resulting in importance distributions that better approximate the distribution in the numerator of each ratio. In practice, samples from the posterior distribution ($\beta_k = 1$) are used to parameterize the joint reference distribution $\pi_0(\boldsymbol{\theta}|M)$. For each component θ of the model (where a component could be an individual parameter or a block of correlated parameters, such as base frequencies), the marginal posterior sample mean ($\hat{\mu}_\theta$) and variance ($\hat{\sigma}^2_\theta$) are used to parameterize an independent reference distribution for θ. For example, if θ represents the gamma shape parameter used for modeling among-site rate heterogeneity, a Gamma($\hat{\mu}^2_\theta/\hat{\sigma}^2_\theta, \hat{\sigma}^2_\theta/\hat{\mu}_\theta$) distribution would be used as the reference distribution for θ because the mean of a Gamma(a,b) distribution is ab and its variance is ab^2. Relative base frequencies are assigned a Dirichlet(a,c,g,t) distribution. The means ($\hat{\mu}_A$, $\hat{\mu}_C$, $\hat{\mu}_G$, and $\hat{\mu}_T$) and variances ($\hat{\sigma}^2_A$, $\hat{\sigma}^2_C$, $\hat{\sigma}^2_G$, and $\hat{\sigma}^2_T$) of the sampled base frequencies may be used to parameterize the Dirichlet reference distribution as follows (using least squares to estimate m, the sum of all parameters),

$$\hat{m} = \frac{\sum_{i \in \{A,C,G,T\}} \hat{\mu}_i^2 (1 - \hat{\mu}_i)^2}{\sum_{i \in \{A,C,G,T\}} \hat{\sigma}_i^2 \hat{\mu}_i (1 - \hat{\mu}_i)} - 1,$$

$$a = \hat{m}\hat{\mu}_A,$$

$$c = \hat{m}\hat{\mu}_C,$$

$$g = \hat{m}\hat{\mu}_G,$$

$$t = \hat{m}\hat{\mu}_T.$$

Similarly, a Dirichlet(a,b,c,d,e,f) reference distribution can be constructed for the GTR exchangeability parameters using sample means ($\hat{\mu}_{AC}$, $\hat{\mu}_{AG}$, $\hat{\mu}_{AT}$, $\hat{\mu}_{CG}$, $\hat{\mu}_{CT}$, and $\hat{\mu}_{GT}$) and variances ($\hat{\sigma}^2_{AC}$, $\hat{\sigma}^2_{AG}$, $\hat{\sigma}^2_{AT}$, $\hat{\sigma}^2_{CG}$, $\hat{\sigma}^2_{CT}$, and $\hat{\sigma}^2_{GT}$). The joint reference distribution is simply the product of these independent reference distributions.

Different subsets of a partition scheme are often given their own relative substitution rate. These subset relative rates are known as rate multipliers in MrBayes (Ronquist and Huelsenbeck 2003), where they are introduced using the command `prset ratepr = variable`. Subset relative rates, by definition, have mean 1.0, which precludes the use of a Dirichlet prior or reference distribution. We use instead a transformed Dirichlet distribution (which we term here a subset relative rate distribution) to accommodate subset relative rates. Consider the case of a partition that defines n subsets, with proportion p_i of the total sites assigned to subset i. Let $Y \sim \text{Dirichlet}(c_1, c_2, \ldots, c_n)$ and $Y = \{y_i : y_i = x_i p_i\}$. The variable $X = \{x_i\}$ has a subset relative rate distribution with density function

$$f_X(\mathbf{X}) = p_1 p_2 \ldots p_{n-1} \times \left(\frac{(x_1 p_1)^{c_1 - 1} (x_2 p_2)^{c_2 - 1} \ldots (x_n p_n)^{c_n - 1}}{\frac{\Gamma(c_1)\Gamma(c_2)\ldots\Gamma(c_n)}{\Gamma(\sum_{i=1}^{n} c_i)}} \right). \quad (3)$$

To parameterize a subset relative rates reference distribution, we transform sampled relative rate vectors using the subset proportions to form samples that are Dirichlet distributed. The method described above for parameterizing a Dirichlet reference distribution is then used to obtain $c_1, c_2, \ldots, c_n$ for the subset relative rates reference distribution.

Simulated Data Analysis

Each of the 200 simulated data sets was subjected to six separate MCMC analyses for the purpose of estimating marginal likelihoods: 1) an analysis of unpartitioned data in which HM was used to estimate the marginal likelihood, 2) an analysis in which data were partitioned into two equal-sized subsets and HM was used to estimate the marginal likelihood, 3) an analysis in which original SS was used to estimate the marginal likelihood for the unpartitioned data, 4) an analysis in which original SS was used to estimate the marginal likelihood for the bipartitioned data, 5) an analysis in which generalized SS was used to estimate the marginal likelihood for the unpartitioned data, and 6) an analysis in which generalized SS was used to estimate the marginal likelihood for the bipartitioned data. Separate analyses were required for HM, original SS, and generalized SS because both SS methods require special MCMC analyses to be performed in which the target distribution varies from the posterior to either the actual prior or the reference distribution over the course of the run. HM analyses were allotted approximately the same amount of computational effort as SS analyses. For bipartitioned analyses, the first subset was always the first $n/2$ sites and the second subset always the last $n/2$ sites in a data set of size n.

The GTR + G model was used for all analyses, and the tree topology was fixed to the true tree topology used to generate the data. In the case of partitioned models, all parameters were unlinked except the branch lengths. Prior distributions ($\pi(\boldsymbol{\theta}|M)$) were as follows: exponential(10) for all branch lengths, Dirichlet(1,1,1,1) for base frequencies, Dirichlet(1,1,1,1,1,1) for the GTR exchangeabilities, and exponential(0.01) for the discrete gamma shape parameter.

The subset relative rates were fixed to 1.0 for all subsets in these analyses.

For HM analyses, 22,000 MCMC cycles were employed, and all parameters were updated once per cycle. A slice sampler (Neal 2003) was used to update branch lengths and the discrete gamma shape parameter. A Metropolis–Hastings Dirichlet proposal (Metropolis et al. 1953; Hastings 1970) was used to update base frequencies and GTR exchangeabilities. The Markov chain was sampled every ten cycles, providing 2,200 samples in which the first 200 samples were discarded as burn-in.

For the generalized SS analyses, 2,000 MCMC cycles were devoted to each of the 11 β-power posteriors ($K = 10$), again sampling every ten cycles. The first step served the dual purpose of serving as a burn-in period and providing samples from the posterior distribution for parameterizing the reference distribution. The 11 β values were equally spaced along the path from 1.0 to 0.0 (using a reference distribution that approximates the posterior obviates the need to place more sampling effort near $\beta = 0$).

For the original SS analyses, all the settings were the same as the generalized SS analyses except that: 1) the first step served as a burn-in period but no reference distribution was parameterized and 2) the 11 β values were chosen according to evenly spaced quantiles of the Beta(0.3,1) distribution, placing most values of β near 0 as recommended by Xie et al. (2010).

All analyses were repeated with a different pseudorandom number generator seed so HM, original SS, and generalized SS could be compared on the basis of repeatability.

Empirical Data Analysis

Four partition schemes were compared: "None" (unpartitioned data, data from all four genes concatenated), "Gene" (4 data subsets, each corresponding to a gene), "Codon" (3 data subsets, each corresponding to a codon position), and "Both" (12 data subsets, with each of the three codon positions in each of the four genes given its own partition subset). A GTR + G model was applied to each subset regardless of the number of subsets (1, 3, 4, or 12) in the partition model. Branch lengths were linked across subsets and the tree topology was fixed at the maximum likelihood topology found by heuristic (Tree-Bisection-Reconnection, or TBR, branch swapping) search in PAUP* 4b10 (Swofford 2002) assuming the "None" partition model and the GTR + G substitution model. Thus, the simplest model ("None") has 70 free parameters ($2 \times 32 - 3 = 61$ branch lengths, 5 GTR exchangeabilities, 3 relative base frequencies, and 1 discrete gamma shape parameter), whereas the most complex model ("Both") tested has 180 free parameters ($2 \times 32 - 3 = 61$ branch lengths, $5 \times 12 = 60$ GTR exchangeabilities, $3 \times 12 = 36$ relative base frequencies, $1 \times 12 = 12$ discrete gamma shape parameters, and $12 - 1 = 11$ subset relative rates). The marginal likelihood of each partition model was estimated using two methods: HM and generalized SS. The software Phycas (Lewis et al. 2008) was used.

For HM analyses, a single Markov chain was allowed to burn-in for 500 cycles, where one cycle involved updating all parameters at least once (base frequencies, GTR exchangeabilities, and subset relative rates were updated ten times per cycle). In addition, an update affecting all branch lengths (tree rescaling) was attempted once per cycle. These updates were effected either by slice sampling (branch lengths and discrete gamma shape parameters; Neal 2003) or Metropolis–Hastings proposals (base frequencies, GTR exchangeabilities, subset relative rates, and tree rescaling; Metropolis et al. 1953; Hastings 1970). Following the burn-in period, the chain was allowed to run for 25,000 additional cycles and was sampled once per cycle.

Generalized SS analyses were identical to HM analyses with the exception that after 500 burn-in cycles, 1,000 MCMC cycles were devoted to each of the 25 β-power posteriors ($K = 24$). The first step provided samples from the posterior distribution for parameterizing the reference distribution. The 25 β values were equally spaced along the path from 1.0 to 0.0. The amount of computation was intentionally made identical for HM and the generalized SS analyses so that comparisons of the performance of HM versus the generalized SS would be fair.

Results

Simulations

The simulation experiment compared BF estimated using the HM, original SS, and generalized SS methods. Two hundred data sets of varying sizes (both number of taxa and number of sites) were simulated using a GTR + G model, but the data sets varied in the parameter values used in the generating model. The 200 points in the plots in figure 1 represent the quantity 2log(BF) calculated for each data set. Each BF value measures the marginal likelihood of the partitioned model (two equal-sized subsets) divided by the marginal likelihood of the unpartitioned model. Because the true model was unpartitioned, the expectation is that 2log(BF) would be negative for all data sets, indicating that the unpartitioned model fits the data better on average than the (arbitrarily and unnecessarily) partitioned model. For HM analyses (fig. 1*a*), 43 (21.5%) points are greater than zero. This result is qualitatively similar to that reported by Brown and Lemmon (2007) in their figure 6, even though those authors used a more complicated generating model. In contrast, only one 2log(BF) value was above zero when BF values were estimated using the original and generalized SS methods (fig. 1*b* and *c*), and the variance of generalized SS estimates is clearly smaller than the variance of estimates made using the original SS method. The single 2log(BF) greater than zero was from one of the smallest data sets simulated (only 130 sites and 4 taxa). Using both SS methods, increasing the number of sites generally resulted in a stronger preference for the unpartitioned model, whereas no such trend was evident for HM analyses.

We also investigated repeatability of HM, original SS, and generalized SS by analyzing each of the 200 data sets twice using different pseudorandom number seeds. Ideally, the same estimate of 2log(BF) should result from independent analyses. Plotting the 2log(BF) values obtained from seed 1

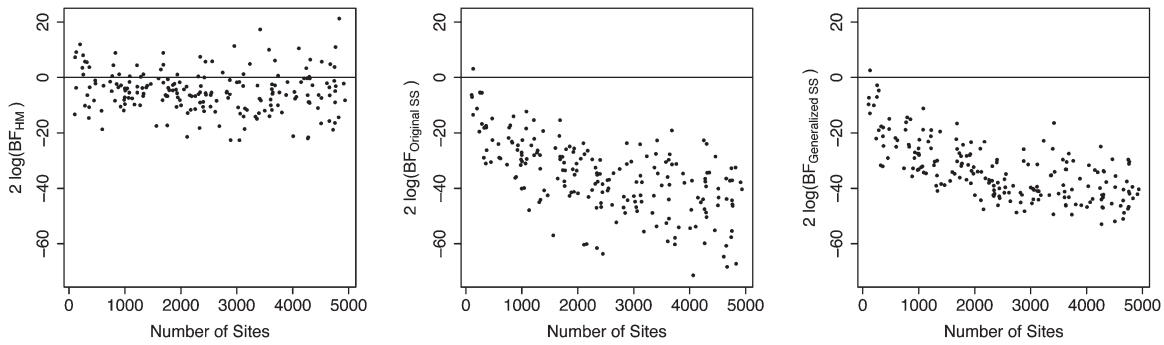

FIG. 1. Plots relating the number of sites to twice the natural logarithm of the BF (2log(BF)) in favor of the partitioned model (with two equal-size subsets) over the unpartitioned model for 200 data sets simulated under a diversity of unpartitioned GTR + G models (see text for details). (*a*) Left: 2log(BF) estimated using the HM method. (*b*) Middle: 2log(BF) estimated using the original SS method. (*c*) Right: 2log(BF) estimated using the generalized SS method.

against the 2log(BF) values from seed 2 should therefore result in all values being very close to the 45° diagonal line indicating perfect identity. It is clear that HM (fig. 2*a*) is far less repeatable than the generalized SS (fig. 2*c*). Principal component analyses reveal that 99.9% of the variance is explained by the first principal component for the generalized SS estimates, whereas only 70.2% of the variance is explained by the first principal component for HM estimates. The original SS (fig. 2*b*) is intermediate in repeatability.

Empirical Example

Although we have shown that HM behaves poorly when used for choosing a partition model on the basis of simulated data, it can be argued that our simulations present a scenario (each site evolving under exactly the same model) that is perhaps never found in the real world. Also, biologists use their experience in choosing a partitioning scheme so that partition placement is not arbitrary, as it was in our simulated data example. The value of the simulation experiment lies in the fact that we know the true model and can judge whether methods are behaving as they should under ideal circumstances. The poor performance of the HM method under such straightforward conditions does not bode well for its use in much more complex real data situations.

A natural question at this point might be: "Does using generalized SS instead of HM make any difference in actual practice?" To answer this question, we reanalyzed a four-gene New Zealand *Kikihia* (cicada) data set used by Marshall et al. (2006) to illustrate problems with branch lengths that can arise in partitioned Bayesian analyses. Table 1 and figure 3 show the results of applying both HM and generalized SS to these data under four possible partition models. Both HM and generalized SS agree that partitioning by codon is a good idea but disagree on whether partitioning by gene is beneficial. Given the choice between partitioning by gene (four subsets) and not partitioning (one subset), HM prefers the partitioned model, whereas generalized SS chooses the unpartitioned model. Likewise, given the choice between partitioning by codon (3 subsets) and partitioning by both gene and codon (12 subsets), HM chooses to partition by both gene and codon, whereas generalized SS chooses the simpler model that partitions by codon only.

Figure 3 also clearly shows the bias in HM: For each partition model, the HM estimate of the marginal likelihood is considerably greater than the generalized SS estimate. The greater variability of HM estimates is also evident.

Discussion

Generalized SS Method

The SS method described here is an important generalization of the original method described by Xie et al. (2010).

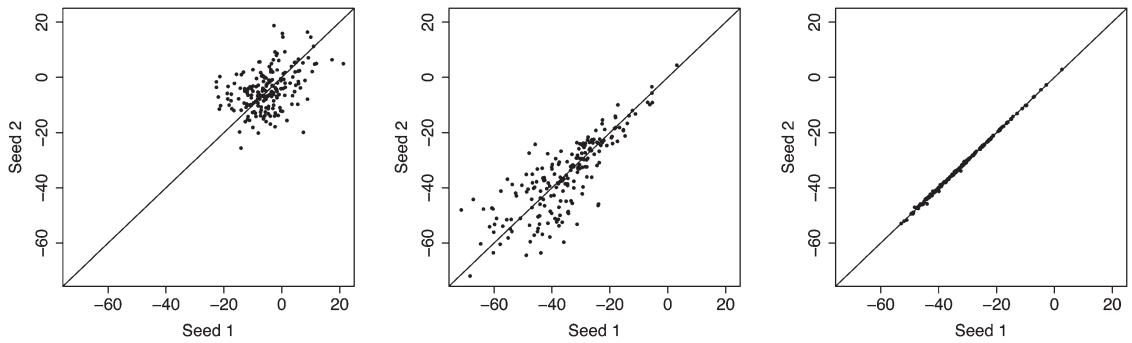

FIG. 2. Scatterplots showing twice the natural logarithm of the BF (2log(BF)) estimated using two independent analyses started with different pseudorandom number seeds. (*a*) Left: 2log(BF) estimated using the HM method. (*b*) Middle: 2log(BF) estimated using the original SS method. (*c*) Right: 2log(BF) estimated using the generalized SS method.

Table 1. Mean Log Marginal Likelihoods and Standard Deviations Based on 20 Independent Replicates from Analysis of the Four-gene New Zealand Cicada Data Set.

Partition Model	HM	Generalized SS
Unpartitioned	**−10246.78 (1.60)**	**−10336.83 (0.19)**
By gene	**−10215.31 (2.51)**	**−10361.76 (0.78)**
By codon	**−9692.18 (3.05)**	**−9823.35 (0.82)**
By gene and codon	**−9634.64 (3.52)**	**−9875.39 (0.31)**

When the reference distribution equals the actual prior, the generalized method equals the original method; however, choosing a reference distribution that approximates the posterior rather than the prior results in a much more stable and efficient estimator. It is more stable because the series of power posterior distributions being explored are all much more similar to one another, and sampling does not become problematic when the power β is close to zero (if anything, sampling becomes more straightforward at this end of the path). Because it is a generalization of the previous method, we prefer to retain the name SS for the new method. To avoid potential confusion, when the original SS method is applied researchers should explicitly state that the reference distribution equals the actual prior. Because of its improved performance, the generalized version described here (where the reference distribution approximates the posterior distribution) should be the default form of the SS method. To see how the generalized SS method can be more efficient than the original SS method, consider the case in which the reference distribution exactly equals the posterior distribution. In this special case, the overall ratio r can be determined exactly with only a single sampled point! Substituting $\pi_0(\boldsymbol{\theta}) = f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/f(\mathbf{y})$ into equation (2) and assuming only $n = 1$ point was sampled for each value of k (and assuming $K = 1$, if desired),

$$\log \hat{r} = \sum_{k=1}^{K} (\beta_k - \beta_{k-1}) \left\{ \cancel{\log \eta_k} + \log \left(\frac{\cancel{f(\mathbf{y}|\boldsymbol{\theta}_{k-1})\pi(\boldsymbol{\theta}_{k-1})}}{\cancel{\frac{f(\mathbf{y}|\boldsymbol{\theta}_{k-1})\pi(\boldsymbol{\theta}_{k-1})}{f(\mathbf{y})}}} \right) - \cancel{\log \eta_k} \right\} = \log f(\mathbf{y}).$$

(Dependence on the model M suppressed for notational clarity.) Although this result has no application in practice (because the exact value of $f(\mathbf{y})$ must be known in order to compute the reference distribution density), it illustrates the importance of choosing a reference distribution that is a good approximation of the posterior distribution. If considerable effort has already gone into approximating the posterior, it behooves the investigator to use that information in constructing the reference distribution.

Despite the efficiency and stability improvements, the original SS still has a place in Bayesian model selection, particularly in models involving latent variables. For example, assume that each site is assigned a rate category and the number and composition of these rate categories is determined by a Dirichlet process (DP) prior (e.g., see Huelsenbeck and Suchard 2007). The DP prior governs not only model parameters (the rates of rate categories) but also latent variables (the assignments of each site to a rate category), complicating the definition of the reference distribution. In such cases, a hybrid SS approach is possible in which all model parameters unrelated to the DP prior are included in the parameterized reference distribution, with elements such as the DP prior being given a reference distribution equivalent to their actual prior.

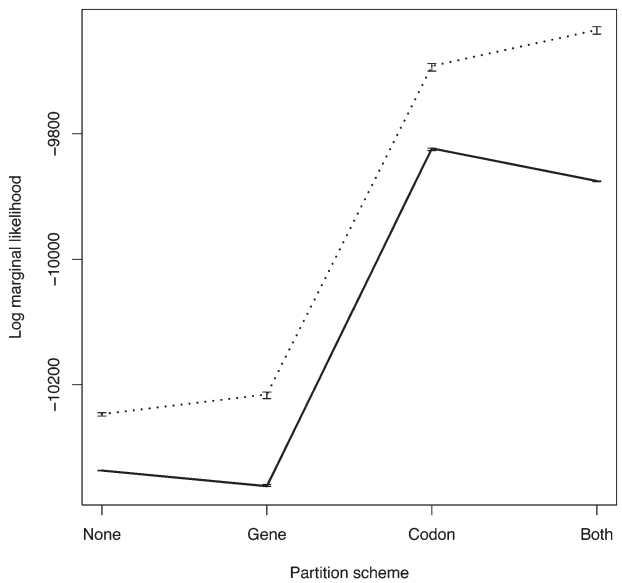

FIG. 3. Results of applying the HM and generalized SS methods to the empirical New Zealand cicada data set for four different partitioning schemes: unpartitioned (None), partitioned by gene (Gene, 4 subsets), partitioned by codon (Codon, 3 subsets), and partitioned by both gene and codon (Both, 12 subsets). Error bars represent standard deviations based on 20 independent replicates. The dotted line connects mean log marginal likelihoods estimated using the HM method, and the solid line connects mean log marginal likelihoods estimated using the generalized SS method.

BFs and Data Set Size

Intuitively, support for the true model over an overparameterized competing model will grow with the size of data sets (more taxa and longer sequences), and in our simulation experiment, this trend is easy to see when generalized SS is used (fig. 1*c*) but not when HM is used (fig. 1*a*) to estimate marginal likelihoods. To support our intuition, we devised the following example using normal distributions, which has the advantage that results are exact (see Appendix 2 for a detailed derivation). Suppose $x_1, \ldots, x_n$ are drawn from a normal distribution with mean 0 and variance σ^2. This is analogous to simulating an unpartitioned nucleotide sequence data set from a given tree topology τ with a branch length set υ and a known nucleotide substitution model. These data may be analyzed using two models. Model M_0 treats $x_1, \ldots, x_{n/2}$ as if drawn from one normal distribution, $N(\mu_1, \sigma^2)$, and $x_{n/2+1}, \ldots, x_n$ as if drawn from a second, potentially distinct normal distribution, $N(\mu_2, \sigma^2)$. The means of the two normal distributions are allowed to be different but variance σ^2 is shared, which is analogous to a Bayesian

phylogenetic analysis in which one or more substitution model parameters are estimated for each partition subset but some (e.g., branch lengths and tree topology) are linked across subsets. Model M_1 treats $x_1, \ldots, x_n$ as if drawn from a single normal distribution $N(\mu_3, \sigma^2)$, which is the analogue of an unpartitioned Bayesian phylogenetic analysis. Both models assume that the variance σ^2 is identical for all n observations.

To obtain a closed-form expression, conjugate priors are used for both mean μ and variance σ^2, and the mean μ is dependent on the variance σ^2. That is, priors are

$$\mu|\sigma^2 \sim N(0, \sigma^2),$$
$$\sigma^2 \sim IG(a, b).$$

After centering the observations $\left(\text{i.e., } \sum_{i=1}^{\frac{n}{2}} x_i = 0 \text{ and } \sum_{j=\frac{n}{2}+1}^{n} x_j = 0 \text{ for model } M_0, \text{ and } \sum_{i=1}^{n} x_i = 0 \text{ for model } M_1\right)$, the BF is

$$\mathrm{BF}_{01} = \frac{f(\mathbf{y}|M_0)}{f(\mathbf{y}|M_1)} = \frac{\sqrt{n+1}}{\frac{n}{2}+1}, \quad (4)$$

which demonstrates that BF_{01} is a monotonically decreasing function of data size, hence, the unpartitioned model is always the expected winner, even for data sets containing as few as $n = 2$ observations. A plot of $2\log(\mathrm{BF}_{01})$ against n (see supplementary fig. S1, Supplementary Material online) is similar to the trend in generalized SS-based 2log(BF) values for the simulated data (fig. 1*c*).

Performance of Generalized SS on Simulated and Empirical Data

Our simulation experiment demonstrated that use of the HM method to compute BFs can potentially be very misleading when using BFs to make decisions about which partition model is best. In more than 1/5 of the data sets analyzed, the HM method would lead one to choose a partitioned model when an unpartitioned model was the true model. In contrast, the generalized SS method described here would have recommended partitioning in only 1 of the 200 data sets. Repeated independent analyses showed that generalized SS is much more repeatable than the HM method.

Our analysis of data on New Zealand *Kikihia* cicadas illustrated that using HM can result in overpartitioning in real data as well. The four genes used in this study are quite similar in their pattern of substitution (see supplementary table S1, Supplementary Material online). Based on the tree length estimates, all the genes evolve at a rate within 35% of the average rate. Likewise, patterns of rate heterogeneity (shape parameter α), base frequencies (π_i), and GTR exchangeabilities (r_{ij}) (with the exception of the transition rates r_{AG} and r_{CT}) are similar in magnitude across genes. The similarity of parameter estimates across genes makes it surprising that partitioning by gene would be favored. It therefore makes sense that accurately estimated marginal likelihoods do not support partitioning by gene.

One drawback of SS is that it requires a special MCMC analysis that explores a series of power posteriors. This appears to be a requirement for accurate direct estimates of marginal likelihoods. The HM estimate, on the other hand, can be obtained essentially for free because it requires only samples already needed to approximate the posterior distribution. The extra cost of SS does not appear to be prohibitive, however. If a sample from the posterior is available, it can be used to parameterize the reference distribution and provide good starting values for the power posterior MCMC. Very few additional samples are required if a very accurate reference distribution is available. Even if no previous posterior sample is available, the SS method requires less computational effort than a HM estimate would require to deliver comparable accuracy. One slight advantage of SS over HM is that the last step $(\beta = 0.0)$ requires only draws from ordinary probability distributions and is thus relatively much faster due to the fact that the likelihood need not be computed for proposed values that are ultimately rejected.

In this study, the topology was fixed when estimating the marginal likelihood using the SS method; however, there is often interest in estimating marginal likelihoods that account for uncertainty in the topology. There is nothing to prevent the estimation of marginal likelihoods using the original SS method when the topology is allowed to vary during an MCMC run, but varying the topology complicates generalized SS because of the need to define a reference distribution for topologies that provides a good approximation to the posterior. This important expansion of SS to accommodate topological uncertainty is the subject of ongoing research in our group.

We have demonstrated that the SS method using a reference distribution that approximates the posterior is much more accurate and repeatable for estimating marginal likelihoods than the currently popular HM method. We have shown that using SS can result in the choice of a different (and simpler) partition model than HM method for empirical data. We therefore recommend that SS be used instead of HM when using BFs to decide which partitioning scheme is best. The SS method is implemented in the free, open source software Phycas (Lewis et al. 2008).

Supplementary Material

Supplementary figure S1 and table S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was stimulated by very constructive criticism we received from the reviewers of our original paper describing SS method. In particular, Nicolas Lartillot and Marc Suchard pointed out the limitations of sampling distributions near the prior. We hope that the new approach eliminates their concerns and thank them for their very helpful reviews of our previous work. We thank Nicolas Lartillot again and one

anonymous reviewer for their relevant suggestions on this paper. We also thank the UConn Biotechnology Center's Bioinformatics Facility for use of their Linux cluster. This work was supported by the National Institutes of Health (GM70335 and CA74015 to M.H.C.) and the National Science Foundation (EF0331495 to P.O.L. and DMS0723557 to M.H.C.).

Appendix 1

Case in Which the Proportion of Invariable Sites Exceeds the Proportion of Constant Sites

To demonstrate that invariable sites (*I*) models can misbehave when data are partitioned, we used Seq-Gen 1.3.2 (Rambaut and Grassly 1997) to simulate a partitioned data set in which 5,000 sites were assigned to subset 1 (the "large slow" subset), and 100 sites were assigned to subset 2 (the "small fast" subset). The sites in the small fast subset evolved 50 times faster than those in the large slow subset. The tree model was the maximum likelihood tree (GTR + G model) for the same 32-taxon cicada data set used as the empirical example in this paper. The generating model was Jukes–Cantor (JC) with no among-site rate heterogeneity other than the difference in rate among subsets. We analyzed this single simulated data set with Phycas using a JC + I model for both subsets, with topology fixed to the true tree and branch lengths linked across subsets but unlinking the proportion of invariable sites parameter, p_{invar}. Results are shown in the first row of table 2. Despite the fact that all but one of the sites in the small fast subset were variable, p_{invar} was estimated to be 0.96 for this subset. The model thus considers 96% of the sites in this subset to be incapable of varying when in reality 99% of them are, in fact, indisputably variable. This serves to show that partitioning can force otherwise well-behaved models to explain data in biologically unreasonable ways. In this case, the branch lengths are largely determined by the large slow subset (note the underestimated tree length), which makes it difficult for the model to explain the fast-evolving sites in the small fast subset. To explain these sites, the model finds that it can increase the effective tree length for the second subset by increasing the proportion of invariable sites parameter to absurd levels. The invariable sites model is a mixture model involving two relative rates that, by definition, have expectation 1.0:

$$E[r] = p_{\text{invar}} r_0 + (1 - p_{\text{invar}}) r_1 = 1,$$

$$r_1 = \frac{1 - p_{\text{invar}} r_0}{1 - p_{\text{invar}}} = \frac{1}{1 - p_{\text{invar}}}.$$

The last step results from the fact that $r_0 = 0$ (i.e., invariable sites evolve, by definition, at zero rate). Bumping up p_{invar} to 0.96 allows the model to effectively increase each branch length by a factor of 27, which is very close to the estimated relative rate (25) for this subset in a model (JC + M) that allows subset relative rates to be free parameters (second row of table 2). Note that using a model (JC + I + M, third row of table 2) having both p_{invar} parameters for each subset as well as subset relative rates behaves more sensibly than JC + I. This is because the subset relative rates can account for the difference in rate among subsets, allowing p_{invar} to go back to measuring the proportion of invariable sites. This JC + I + M model is considered best by the HM method, yet still rather seriously overestimates p_{invar} for the first partition. In contrast, the generalized SS method described here shows a slight preference for the (true) JC + M model over the more complex JC + I + M model.

Appendix 2

Derivation of the BF in the Normal Example

Assume data $x_1, \ldots, x_n \sim N(0, \sigma^2)$ and analyze it with two models, partitioned and unpartitioned. For the partitioned model (M_0), suppose $x_1, \ldots, x_{\frac{n}{2}} \sim N(\mu_1, \sigma^2)$ and $x_{\frac{n}{2}+1}, \ldots, x_n \sim N(\mu_2, \sigma^2)$, and set priors as:

$$\mu_1 | \sigma^2 \sim N(0, \sigma^2),$$

$$\mu_2 | \sigma^2 \sim N(0, \sigma^2),$$

$$\sigma^2 \sim IG(a, b).$$

The joint prior is

$$\pi(\mu_1, \mu_2, \sigma^2) = \frac{b^a \sigma^{-2(a+2)}}{2\pi\Gamma(a)} \exp\left(-\frac{\mu_1^2 + \mu_2^2 + 2b}{2\sigma^2}\right).$$

According to the definition of marginal likelihood,

$$f(x_1, \ldots, x_n | M_0) = \iiint f(x_1, \ldots, x_{\frac{n}{2}} | \mu_1, \sigma^2) \times f\left(x_{\frac{n}{2}+1}, \ldots, x_n | \mu_2, \sigma^2\right) \times \pi(\mu_1, \mu_2, \sigma^2) \mathrm{d}\mu_1 \, \mathrm{d}\mu_2 \, \mathrm{d}\sigma^2$$

Table 2. Tree Length, Subset Relative Rates (m_1 and m_2), and Proportion of Invariable Sites Parameter Values ($p_{\text{invar},1}$ and $p_{\text{invar},2}$) for Two Subsets (subscripts 1 and 2). In Total, 549 (11%) of the 5,000 Sites in Subset 1, and 99 (99%) of the 100 Sites in Subset 2 Were Variable.

Model	HM	Generalized SS	Tree Length	m_1	m_2	$p_{\text{invar},1}$	$p_{\text{invar},2}$
JC + I|JC + I	−13788.91	−14025.85	0.15	1.00	1.00	0.27	0.96
JC + M|JC + M	−13442.35	−13642.55	0.24	0.52	25.03	0.00	0.00
JC + I + M|JC + I + M	−13433.07	−13646.04	0.24	0.52	24.90	0.22	0.01
True	—	—	0.22	0.51	25.50	0.00	0.00

NOTE.—I, invariable sites model; M, subset relative rates model; JC, Jukes–Cantor model.

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\frac{n}{2}+1}\frac{b^a}{\Gamma(a)}\frac{\Gamma(\alpha'')}{(\beta'')^{\alpha''}},$$

where $\alpha'' = \frac{n}{2} + a$ and $\beta'' = \frac{1}{2}\left(\sum_{i=1}^{n} x_i^2 + 2b\right) - \frac{1}{n+2}\left[\left(\sum_{i=1}^{\frac{n}{2}} x_i\right)^2 + \left(\sum_{j=\frac{n}{2}+1}^{n} x_j\right)^2\right]$.

For the unpartitioned model (M_1), suppose $x_1, \ldots, x_n \sim N(\mu_3, \sigma^2)$ and set priors as follows:

$$\mu_3|\sigma^2 \sim N(0, \sigma^2),$$

$$\sigma^2 \sim IG(a, b).$$

The marginal likelihood is

$$f(x_1, \ldots, x_n|M_1)$$

$$= \iint f(x_1, \ldots, x_n|\mu_3, \sigma^2)\pi(\mu_3|\sigma^2)\pi(\sigma^2)\mathrm{d}\mu_3\,\mathrm{d}\sigma^2$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n \sqrt{\frac{1}{n+1}}\frac{b^a}{\Gamma(a)}\frac{\Gamma(\alpha')}{(\beta')^{\alpha'}},$$

where $\alpha' = \frac{n}{2} + a$ and $\beta' = \frac{1}{2}\left(\sum_{i=1}^{n} x_i^2 + 2b\right) - \frac{1}{2(n+1)}\left[\left(\sum_{i=1}^{n} x_i\right)^2\right]$.

The BF in favor of the partitioned model is

$$\mathrm{BF}_{01} = \frac{f(x_1, \ldots, x_n|M_0)}{f(x_1, \ldots, x_n|M_1)} = \frac{\sqrt{n+1}}{\frac{n}{2}+1} \times \left\{\frac{\sum_{i=1}^{n} x_i^2 + 2b - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n+1}}{\left(\sum_{i=1}^{n} x_i^2 + 2b\right) - \frac{2}{n+2}\left[\left(\sum_{i=1}^{\frac{n}{2}} x_i\right)^2 + \left(\sum_{j=\frac{n}{2}+1}^{n} x_j\right)^2\right]}\right\}^{a+\frac{n}{2}}.$$

Assuming that the data are centered $\left(\text{i.e., } \sum_{i=1}^{\frac{n}{2}} x_i = 0, \sum_{j=\frac{n}{2}+1}^{n} x_j = 0, \text{and } \sum_{i=1}^{n} x_i = 0\right)$, the BF simplifies to

$$\mathrm{BF}_{01} = \frac{f(x_1, \ldots, x_n|M_0)}{f(x_1, \ldots, x_n|M_1)} = \frac{\sqrt{n+1}}{\frac{n}{2}+1}.$$

References

Akaike H. 1974. A new look at statistical model identification. *IEEE Trans Automat Contr.* 19:716–723.

Brandley M, Schmitz A, Reeder T. 2005. Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. *Syst Biol.* 54:373–390.

Brown JM, Lemmon AR. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. *Syst Biol.* 56:643–655.

Brown MW, Spiegel FW, Silberman JD. 2009. Phylogeny of the "Forgotten" Cellular Slime Mold, Fonticula alba, Reveals a Key Evolutionary Branch within Opisthokonta. *Mol Biol Evol.* 26: 2699–2709.

Clarke JA, Middleton KM. 2008. Mosaicism, modules, and the evolution of birds: results from a Bayesian approach to the study of morphological evolution using discrete character data. *Syst Biol.* 57:185–201.

Gelman A, Meng X-L. 1998. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Stat Sci.* 13:163–185.

Hastings WK. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109.

Huelsenbeck JP, Larget B, Alfaro ME. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Mol Biol Evol.* 21:1123–1133.

Huelsenbeck JP, Suchard MA. 2007. A nonparametric method for accommodating and testing across-site rate variation. *Syst Biol.* 56:975–987.

Lartillot N, Philippe H. 2006. Computing Bayes factors using thermodynamic integration. *Syst Biol.* 55:195–207.

Lefebvre G, Steele R, Vandal AC. 2010. A path sampling identity for computing the Kullback-Leibler and J-divergences. *Comput Stat Data Anal.* 54:1719–1731.

Lewis PO, Holder MT, Swofford DL. 2008. Phycas: software for phylogenetic analysis. Storrs (CT): University of Connecticut. Available from http://www.phycas.org/.

Liu H, Aris-Brosou S, Probert I, de Vargas C. 2010. A time line of the environmental genetics of the haptophytes. *Mol Biol Evol.* 27: 161–176.

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

Marshall DC, Simon C, Buckley TR. 2006. Accurate branch length estimation in partitioned Bayesian analyses requires accommodation of among-partition rate variation and attention to branch length priors. *Syst Biol.* 55:993–1003.

Meng X-L, Wong WH. 1996. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Stat Sin.* 6: 831–860.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equation of state calculations by fast computing machines. *J Chem Phys.* 21:1087–1092.

Neal RM. 1994. Contribution to the discussion of "Approximate Bayesian inference with the weighted likelihood bootstrap" by Michael A. Newton and Adrian E. Raftery. *J R Stat Soc Ser A (Methodological)* 56:41–42.

Neal RM. 2003. Slice sampling. *Ann Stat.* 31:705–741.

Newton MA, Raftery AE. 1994. Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *J R Stat Soc Ser B (Methodological)* 56:3–48.

Nylander J, Ronquist F, Huelsenbeck J, Nieves-Aldrey J. 2004. Bayesian phylogenetic analysis of combined data. *Syst Biol.* 53:47–67.

Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci.* 13:235–238.

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

Schwarz GE. 1978. Estimating the dimension of a model. *Ann Stat.* 6:461–464.

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

Swofford DL. 2002. PAUP*: phylogenetic analysis using parsimony (*and other methods). Sunderland (MA): Sinauer Associates.

Verdinelli I, Wasserman L. 1995. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *J Am Stat Assoc.* 90:614–618.

Wilks SS. 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann Stat.* 9:60–62.

Xie W, Lewis PO, Fan Y, Kuo L, Chen M-H. Forthcoming 2010. Improving Marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst Biol.*