Syst. Biol. 72(3):639–648, 2023

https://doi.org/10.1093/sysbio/syad007
Advance Access Publication February 28, 2023

LoRaD: Marginal likelihood estimation with haste (but no waste)

Yu-Bo Wang[1], Analisa Milkey[2], Aolan Li[3], Ming-Hui Chen[3], Lynn Kuo[3], and Paul O. Lewis[2,*]

[1]*School of Mathematical and Statistical Sciences, Clemson University, Clemson, SC 29634, USA*
[2]*Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA*
[3]*Department of Statistics, University of Connecticut, Storrs, CT 06269, USA*
**Correspondence to be sent to: Paul O. Lewis, Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Road, Unit 3043, Storrs, CT 06269, USA; E-mail: paul.lewis@uconn.edu*

Received 2 March 2022; reviews returned 15 February 2023; accepted 21 February 2023
Associate Editor: Mark Holder

Abstract.—The Lowest Radial Distance (LoRaD) method is a modification of the recently introduced Partition-Weighted Kernel method for estimating the marginal likelihood of a model, a quantity important for Bayesian model selection. For analyses involving a fixed tree topology, LoRaD improves upon the Steppingstone or Thermodynamic Integration (Path Sampling) approaches now in common use in phylogenetics because it requires sampling only from the posterior distribution, avoiding the need to sample from a series of *ad hoc* power posterior distributions, and yet is more accurate than other fast methods such as the Generalized Harmonic Mean (GHM) method. We show that the method performs well in comparison to the Generalized Steppingstone method on an empirical fixed-topology example from molecular phylogenetics involving 180 parameters. The LoRaD method can also be used to obtain the marginal likelihood in the variable-topology case if at least one tree topology occurs with sufficient frequency in the posterior sample to allow accurate estimation of the marginal likelihood conditional on that topology. [Bayesian; marginal likelihood; phylogenetics.]

The marginal likelihood is a quantity of central importance to Bayesian model selection and hypothesis testing. Defined as the weighted average fit of a model to the data, with weights determined by the prior distribution and fit determined by the likelihood, the marginal likelihood rewards models that fit the data well over the parameter space considered important according to the joint prior distribution, and implicitly punishes models with gratuitous complexity (i.e. models with parameters that do not allow for a substantial increase in the likelihood). The Bayes factor used to compare the fit of one model relative to a single alternative model is a simple ratio of marginal likelihoods. Bayes factors represent a Bayesian alternative to Frequentist model selection criteria such as likelihood ratio tests and AIC/BIC comparisons. As with AIC and BIC, the marginal likelihood may be used to rank models that are not necessarily nested; however, unlike AIC and BIC, the complexity penalty imposed by the marginal likelihood depends on the prior and therefore differs among model parameters.

We present here a new method (LoRaD) that competes favorably with the Thermodynamic Integration (TI; Lartillot and Philippe 2006; Friel and Pettitt 2008) (also known as Path Sampling) and Steppingstone (SS; Fan et al. 2011; Xie et al. 2011; Baele et al. 2015) methods in accuracy for the fixed tree topology case and which can be computed from a sample of points from (only) the posterior distribution. Both TI and SS require sampling from numerous power posterior distributions, which entail considerable extra computational effort (beyond that involved with sampling from the posterior distribution). The samples from these power posterior distributions are normally only used to estimate the marginal likelihood and not to estimate the parameters of the model. An exception is MIGRATE-N, which uses power posteriors for both Metropolis coupling to improve mixing as well as to estimate the marginal likelihood (Beerli and Palczewski 2010). Given that sampling from the posterior distribution is required in order to estimate model parameters, LoRaD is both faster (“haste”) and all samples can be used for both parameter estimation and model assessment (“no waste”).

Interestingly, the LoRaD method is able to accurately estimate the marginal likelihood using only a *subset* of points from the posterior sample. This allows, in some cases, the marginal likelihood of a phylogenetic model in which tree topology is variable to be estimated accurately from only samples involving the most frequently sampled tree topology. LoRaD does not require computation of the likelihood and joint prior for any additional points but does require log transformation of parameters and standardization, which involves computing the inverse of the $p \times p$ variance–covariance matrix, where p is the model dimension (i.e. number of estimated parameters, including edge lengths). Given its importance in Bayesian model selection and the fact that the marginal likelihood may only be approximated numerically in most real-world situations, it is not surprising that many different methods have been proposed for estimating this central quantity (Chib 1995; Meng and Wong 1996; Lartillot and

Philippe 2006; Friel and Pettitt 2008; Fan et al. 2011; Xie et al. 2011; Arima and Tardella 2012; Wang et al. 2018). The marginal likelihood is the normalizing constant that converts a posterior kernel—the product of likelihood and prior probability density—into a posterior probability density, and, ideally, would be estimated from parameter vectors sampled from the posterior distribution. This goal has proved elusive, with early attempts (HM; Newton and Raftery 1994) yielding strongly biased estimates. This led to the development of more accurate methods such as TI and SS.

Fourment et al. (2020) provided a side-by-side comparison of 19 marginal likelihood methods for a simple phylogenetic example (Jukes-Cantor model without rate heterogeneity on a fixed tree topology), finding that the Generalized Steppingstone (GSS) method is arguably the most consistently accurate but also the most computationally intensive of all methods tested. Fourment et al. (2020) also introduced several new methods, such as their Laplus family, that appear to be both extremely fast (i.e. computationally efficient) as well as accurate (boasting accuracy levels rivaling GSS); however, these methods make several assumptions (e.g. independence of edge length parameters) that reduce their generality, and the methods were tested on models involving only edge length parameters. Finally, Arima and Tardella (2012) introduced the Generalized Harmonic Mean (GHM) method and compared it to both the Inflated Density Ratio (IDR) and GSS methods. The GHM and IDR methods are of particular interest because they, like the LoRaD method introduced here, require only samples from the posterior distribution and do not require any additional likelihood evaluations.

The LoRaD method is a special case of the PWK (Partition-Weighted Kernel) method (Wang et al. 2018). PWK has been formally described elsewhere, but in this paper, we 1) introduce a new LoRaD (Lowest Radial Distance) approach to defining the working parameter space used by PWK, 2) illustrate the method graphically for a simple 2-parameter example, and 3) show that our method can accurately estimate the marginal likelihood in two empirical phylogenetic examples. LoRaD improves on the SS, GSS, and TI methods by allowing the marginal likelihood to be estimated accurately with no further likelihood evaluations other than those needed to generate a posterior sample. It improves on the GHM and IDR methods by requiring only a fraction of the points sampled from the posterior distribution, which allows (e.g.) the variable-topology marginal likelihood to be estimated using samples from only one focal topology, assuming that the focal topology has a marginal posterior probability large enough to be accurately estimated. Another advantage is that LoRaD does not require any change to existing phylogenetic software as long as the log likelihood, log prior, and all parameter values are recorded for the points sampled.

MATERIALS AND METHODS

Notation

Let the posterior probability density for parameter vector θ be denoted by$p(\theta) = q(\theta)/c$, where $q(\theta)$ is the posterior kernel (the prior density multiplied by the likelihood; i.e. the numerator in Bayes' rule) and $c = p(y)$ is the marginal likelihood of interest. The quantities $p(\theta)$ and $q(\theta)$ are both conditioned on data y, but our notation omits this conditioning for simplicity. Likewise, c is a function of y, even though this is not explicit in the notation. The posterior density is simply a scaled version of the posterior kernel. The scaling factor needed (i.e. the marginal likelihood c) equals the volume under the posterior kernel surface; hence, marginal likelihood estimation is equivalent to performing numerical integration to determine the volume under the posterior kernel surface.

Training Sample vs. Estimation Sample

The total sample (comprising T sampled parameter vectors) is first partitioned into two disjoint subsets, the *training sample* and the *estimation sample*, based on a user-specified *training fraction* denoted ψ (Fig. 1a). The training sample thus has size $T_0 = \psi T$ and estimation sample has size $T_1 = (1-\psi)T$. The value $\psi = 0.5$ works well in practice.

The training sample is used to determine the working parameter space (described below), while the estimation sample is used to estimate the marginal likelihood given that pre-specified working parameter space definition. The extent of the working parameter space must be determined without reference to the estimation sample to avoid bias.

The purpose of the transformations described in the next two sections is to remove differences in support and scale among the model parameters so that the posterior density surface is as close as possible (up to a scaling factor) to a multivariate normal density. The goal is to determine the scaling factor (i.e. the marginal likelihood) that makes the unnormalized, log-transformed, and standardized posterior coincide as closely as possible with the multivariate normal reference density.

Log Transformations

Log transformations are performed on parameters that are constrained in their support. All T points in both the training and estimation samples are log-transformed (Fig. 1b), as in Arima and Tardella (2012), such that all parameters have support equal to the entire real line $(-\infty, +\infty)$.

The log transformation used for a strictly-positive continuous parameter X (e.g. edge length, rate ratio) is:

$$Y = \log(X). \tag{1}$$

A logit transformation may be used for a univariate parameter P with support from 0 to 1 (e.g. proportion of invariable sites):

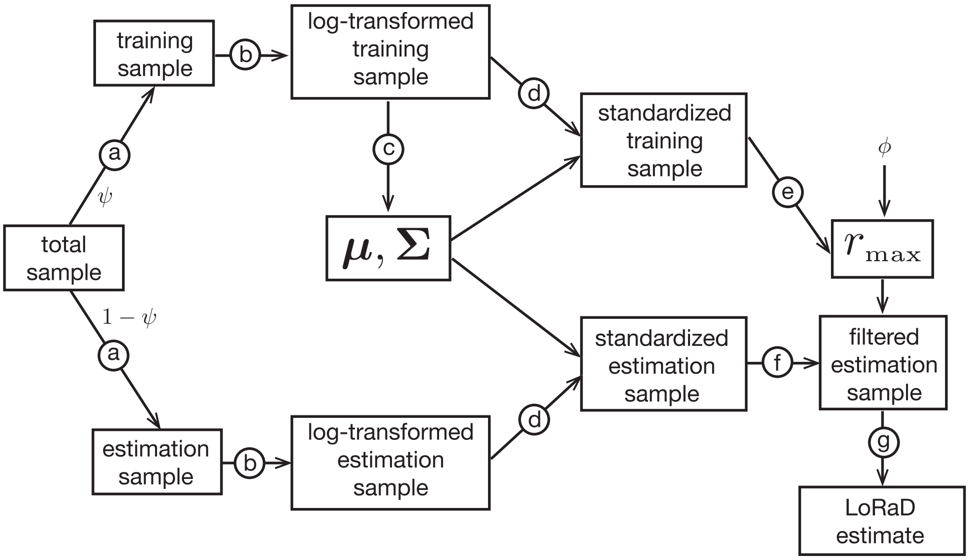
training sample
log-transformed training sample
standardized training sample
total sample
μ, Σ
$r_{\max}$
ϕ
ψ
$1-\psi$
a
b
c
d
e
f
g
standardized estimation sample
filtered estimation sample
estimation sample
log-transformed estimation sample
LoRaD estimate

FIGURE 1. Flow chart showing processing of posterior sample to produce the LoRaD marginal likelihood estimate.

$$Q = \log\left(\frac{P}{1-P}\right). \tag{2}$$

For a vector $U = (U_1, U_2, \cdots, U_M)$ constrained such that $\sum_{m=1}^{M} U_m = 1$ (such as base frequencies, in which $M = 4$, or GTR exchangeabilities, in which $M = 6$), a log-ratio transformation may be used (the log-ratio transformation is a multivariate extension of the logit transformation):

$$V_2 = \log\left(\frac{U_2}{U_1}\right) \tag{3}$$

$$V_3 = \log\left(\frac{U_3}{U_1}\right)$$

$$\vdots \tag{4}$$

$$V_M = \log\left(\frac{U_M}{U_1}\right). \tag{5}$$

Note that the original random variable U has $M-1$ degrees of freedom. The transformation yields variable V, also of dimension $M-1$, with the $M-1$ components equal to the logarithm of the ratio of the corresponding component of U to an arbitrarily chosen reference element (U_1 in this case).

The log/logit/log-ratio transformations (hereafter denoted simply log transformations) require that the original posterior kernel (which has presumably been made available by the Bayesian software that performed the MCMC sampling) be multiplied by the appropriate Jacobian terms so that volumes after transformation are equivalent to volumes before transformation for comparable parameter intervals. The Jacobian terms for the log, logit, and log-ratio transformations are, respectively:

$$\left|\frac{dX}{dY}\right| = e^Y = X \tag{6}$$

$$\left|\frac{dP}{dQ}\right| = \frac{e^Q}{(1+e^Q)^2} = P(1-P) \tag{7}$$

$$\det\left(\frac{d\mathbf{U}}{d\mathbf{V}}\right) = \frac{e^{V_2+\cdots+V_M}}{(1+e^{V_2}+\cdots+e^{V_M})^M} = U_1U_2\cdots U_M. \tag{8}$$

Standardization Transformation

The mean vector and covariance matrix of the (log-transformed) training sample (Fig. 1c) are used to standardize all T sampled points (Fig. 1d). The p -dimensional mean vector of the training sample after standardization is the zero vector, 0_p, and the $p \times p$ variance–covariance matrix of the log-transformed and standardized training sample is equivalent to the identity matrix I_p. The mean vector and covariance matrix of the log-transformed and standardized estimation sample will approximately equal, but will not be identical to, 0_p and I_p, respectively, due to stochastic differences between the training and estimation subsets.

The standardization transformation is

$$\mathbf{Y} = \mathbf{S}^{-0.5}(\mathbf{X} - \overline{\mathbf{X}}), \tag{9}$$

where X is the $p \times T$ matrix of log/logit/log-ratio transformed parameter values, $\overline{\mathbf{X}}$ is the $p \times T$ matrix with each column equal to the p-dimensional mean vector computed from the training sample, and S is the $p \times p$ sample variance–covariance matrix computed from the training sample. The Jacobian determinant for this transformation is $\det(S^{0.5})$.

The notation $\tilde{q}(\widetilde{\theta})$ is used hereafter to denote the posterior kernel of a point $\widetilde{\theta}$ in the log-transformed and standardized parameter space.

Working Parameter Space

The PWK method (of which the LoRaD method is a special case) begins by defining a p-dimensional *working parameter space*, $\tilde{\Theta}$, using the training sample. The working parameter space $\tilde{\Theta}$ is a subset of the p-dimensional full parameter space in which the transformed posterior kernel, $\tilde{q}(\widetilde{\theta})$, is bounded away from zero for every point $\widetilde{\theta} \in \tilde{\Theta}$.

Let $r_i = ||\widetilde{\theta}_i||$ be the Euclidean norm (radius) of point i $(i = 1, \cdots, T_0)$ from the log-transformed and standardized training sample. Transformations preserve volume, so the value c (the marginal likelihood) is not affected. Sort the training sample from smallest to largest r_i: $r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(T_0)}$. Choose the *coverage fraction* ϕ such that points with radius greater than $r_{(t)}$ are discarded, where $t = \phi T_0$. Let $r_{\max} = ||\widetilde{\theta}_{(t)}||$ be the norm of the point $\widetilde{\theta}_{(t)}$ representing the radius that defines the 100ϕ% Lowest Radial Distance (LoRaD) region (Fig. 1e). The working parameter space $\tilde{\Theta}$ is defined as $\{\widetilde{\theta} : ||\widetilde{\theta}|| \leq r_{max}\}$. Note that the working parameter space is not necessarily equivalent to the 100ϕ% HPD region but is expected to be close if the log-transformed, standardized posterior density is reasonably symmetric.

The LoRaD Method

Let Δ be the volume under that portion of the p -dimensional standard normal density $z(\widetilde{\theta})$ corresponding to the working parameter space $\tilde{\Theta}$:

$$\Delta = \int_{\tilde{\Theta}} z(\widetilde{\theta}) d\widetilde{\theta} \tag{10}$$

$$= \int_{\mathbb{R}^p} z(\widetilde{\theta})\ 1_{\widetilde{\theta} \in \tilde{\Theta}}\ d\widetilde{\theta}, \tag{11}$$

where $1_{\widetilde{\theta} \in \tilde{\Theta}}$ is an indicator function that equals 1 if $\widetilde{\theta}$ is in the working parameter space $\tilde{\Theta}$ and 0 otherwise.

The quantity Δ approximates the integral $\int_{\tilde{\Theta}} (\tilde{q}(\widetilde{\theta})/c) d\widetilde{\theta}$. Consider the following modification to (11):

$$\frac{\Delta}{c} = \int_{\mathbb{R}^p} \frac{z(\widetilde{\theta})\ 1_{\widetilde{\theta} \in \tilde{\Theta}}}{\tilde{q}(\widetilde{\theta})} \frac{\tilde{q}(\widetilde{\theta})}{c} d\widetilde{\theta}. \tag{12}$$

This suggests that we can approximate $1/c$ using the expected value of the quantity $z(\widetilde{\theta})\ 1_{\widetilde{\theta} \in \tilde{\Theta}} / \tilde{q}(\widetilde{\theta})$, where the expectation is with respect to the posterior distribution:

$$\frac{1}{c} = \frac{\int_{\mathbb{R}^p} \left(\frac{z(\widetilde{\theta})}{\tilde{q}(\widetilde{\theta})} 1_{\widetilde{\theta} \in \tilde{\Theta}} \right) \frac{\tilde{q}(\widetilde{\theta})}{c} d\widetilde{\theta}}{\Delta} = \frac{E_{\widetilde{\theta}|\mathbf{y}} \left[\frac{z(\widetilde{\theta})}{\tilde{q}(\widetilde{\theta})} 1_{\widetilde{\theta} \in \tilde{\Theta}} \right]}{\Delta}. \tag{13}$$

The LoRaD estimator of c is thus (Fig. 1g)

$$\hat{c} = \frac{\Delta}{\frac{1}{T_1} \sum_{t=1}^{T_1} \frac{z(\widetilde{\theta}_t)}{\tilde{q}(\widetilde{\theta}_t)} \mathbf{1}_{\widetilde{\theta}_t \in \tilde{\Theta}}}, \tag{14}$$

where T_1 is the size of the estimation sample. The denominator of (14) involves a sum of ratios of the multivariate normal reference density to the unnormalized posterior kernel for the filtered estimation sample that contains only points within a distance $r_{\max}$ from the origin (Fig. 1f).

Computation of Δ

The quantity Δ represents the cumulative probability of the p-dimensional standard radial error distribution (i.e. the distribution of $R = ||\ X\ ||$ when $X \sim \text{MVNorm}(0_p, I_p)$). Edmundson (1961) provided the cumulative probability of a p-dimensional standard radial error distribution,

$$p(R \leq r_{\max}) = \frac{\gamma(p/2, r_{\max}^2/2)}{\Gamma(p/2)}, \tag{15}$$

where $\gamma(s, x)$ is the lower incomplete gamma function,

$$\gamma(s, x) = \int_0^x t^{s-1} e^{-t} dt, \tag{16}$$

and $\Gamma(s)$ is the (complete) gamma function,

$$\Gamma(s) = \int_0^\infty t^{s-1} e^{-t} dt. \tag{17}$$

Note that Equation (15) lacks the (erroneous) factor 2 in Equation (4), p. 12, of Edmundson (1961).

Applying Chib's Method When Topology Varies

If the model involves estimation of tree topology, and if one particular tree topology τ is frequent enough that its marginal posterior probability can be accurately estimated, then the Chib (1995) method may be used to obtain an estimate of the marginal likelihood for the variable-topology model given a posterior sample filtered to contain only sampled points from the focal topology τ. Chib's method rearranges the Bayes' rule formula to obtain an estimate of the marginal likelihood $p(y)$ given the marginal posterior probability of tree topology τ, $p(\tau \mid y)$, the marginal likelihood conditional on τ, $p(y \mid \tau)$, and the prior probability of tree topology τ, $p(\tau)$:

$$p(\mathbf{y}) = \frac{p(\mathbf{y}|\tau)\ p(\tau)}{p(\tau|\mathbf{y})}. \tag{18}$$

This presumes that the sample size remaining after filtering out sample points with non-focal topologies is sufficiently large to accurately estimate the term $p(y \mid \tau)$.

LoRaD in Practice

Two issues arise in practice when using LoRaD. First, how does one choose the coverage fraction ϕ, which is the fraction of the training sample used in determining the limits of the working parameter space? Second, how does one determine whether the MCMC sample size used was sufficient to obtain an accurate estimate of the marginal likelihood?

Overlapping Batch Statistics (OBS) can provide an estimate of the Monte Carlo standard error (MCSE) of the marginal likelihood estimate, which can be used to assess whether the size of the MCMC sample was large enough to reliably estimate the marginal likelihood. The OBS MCSE is computed as follows:

$$\widehat{MCSE} = \left\{ \left[\frac{B}{T-B} \right] \frac{\sum_{b=1}^{T-B+1} (\hat{\eta}_b - \bar{\eta})^2}{T-B+1} \right\}^{\frac{1}{2}}, \tag{19}$$

where $\hat{\eta}_b$ is an estimate of the marginal likelihood in log scale using the b th batch $(\theta_i, i = b, b+1, \cdots, b+B-1)$, $\hat{\eta}$ is the overall mean,

$$\bar{\eta} = \frac{1}{T-B+1} \sum_{b=1}^{T-B+1} \hat{\eta}_b \tag{20}$$

and the batch size B is suggested to be in the range $10 \leq T/B \quad \leq 20$. Ideally, the estimated MCSE should be one tenth or less of $\hat{\eta} = \log(\hat{c})$.

RESULTS

An Illustrative Example

The following example is non-phylogenetic, but has the advantage that the models involved contain at most only two free parameters (i.e. $p = 2$), which allows for a 3-dimensional graphical depiction of the posterior kernel surface and the corresponding bivariate normal reference density. Assume that the goal is to evaluate which of two models (JC69 or K80) is best for estimating the evolutionary distance between two DNA sequences. Two sequences of length 200 sites were simulated (using the *k80lorad.py* script in the *lorad/k80example* directory of the Supplementary Materials, Dryad DOI:doi:10.5061/dryad.pg4f4qrrw) under the more-complex K80 substitution model with edge length $\nu = 0.2$ and transition/transversion rate ratio $\kappa = 5$. Of the 200 total sites, 142 (71%) were constant, with the 58 (29%) variable sites divided into 36 (18%) showing transition-type substitutions and 22 (11%) showing transversion-type substitutions.

In the K80 model (Kimura 1980) ($p = 2$), the edge length parameter, ν, represents the expected number of substitutions per site, and the parameter κ is the ratio of the instantaneous rate of transition-type substitutions (i.e. $A \leftrightarrow G$ or $C \leftrightarrow T$) to the instantaneous rate of transversion-type substitutions ($A \leftrightarrow C$, $A \leftrightarrow T$, $G \leftrightarrow C$, or $G \leftrightarrow T$). The posterior kernel for this model may be visualized as a fin-like surface above a 2-dimensional plane defined by axes representing each of the two model parameters (Fig. 2a). The joint prior density used for this example is the product of a Gamma(1, 50) = Exponential(1/50) prior density for ν and an identical Gamma(1, 50) prior density for κ.

An MCMC analysis of this posterior distribution of length 1,000,000 iterations, sampling every 100 iterations, yielded 10,000 sampled points. Estimated effective sample sizes were 10213 (ν) and 10684 (κ), indicating good mixing. Posterior means were 0.189 (ν) and 4.50 (κ). The marginal posterior variance for the κ parameter (2.05) was considerably higher than that for the ν parameter (0.00191), indicating that the 400 total nucleotides contain considerably more information about ν than κ, which makes sense given that information needed for estimating κ lies in sites that experienced substitution (29% of sites differed between the two sequences), whereas all sites contribute information relevant to estimating ν. A plot of the posterior kernel surface reflects these differences (Fig. 2a; note the one order of magnitude difference in scale for the two parameter axes).

Both parameters were log-transformed and standardized using the sample mean and variance–covariance

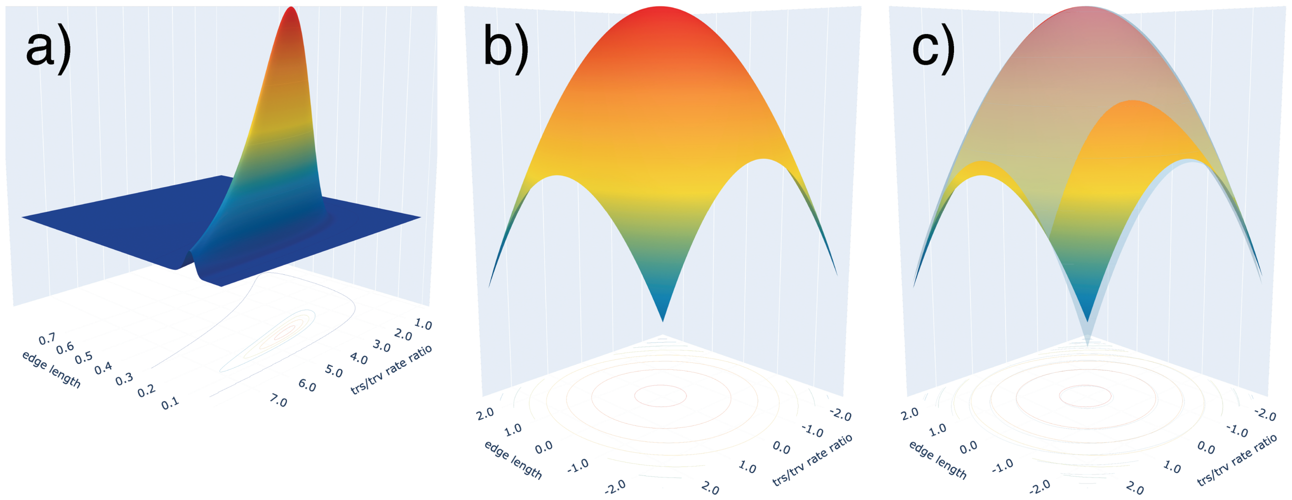

FIGURE 2. Posterior kernel surfaces for the K80 example: a) untransformed; b) log-transformed and standardized; c) transformed, standardized log-posterior surface (heat colors) with the approximating multivariate standard normal density surface (monochrome) superimposed.

matrix of the log-transformed training sample. This yielded a nearly symmetric posterior kernel surface on the log scale (Fig. 2b). The transformed posterior kernel, if normalized, is closely approximated by a bivariate standard normal density (Fig. 2c). This similarity makes the standard normal density an excellent reference function for the PWK method. The $T_0 = 5,000$ log-transformed and standardized training sample points were sorted, retaining the fraction $\phi = 0.1$ having the lowest radial distance (i.e. the 500 points closest to the origin). While the sample mean vector of the training sample was necessarily equal to the 2-dimensional zero vector due to the standardization transformation, the mean vector of the estimation sample was slightly different (−0.00062, −0.02655) due to the fact that it was standardized using the mean and covariance matrix from the training sample.

The log marginal likelihood estimated using the LoRaD method for this example is −460.82239. For comparison, an independent Generalized Steppingstone analysis was carried out using 5 ratios (“stones”), 100,000 iterations/ratio, sampling every 100, and reference distributions set to the Gamma prior distributions parameterized from the marginal posterior means and variances. The log marginal likelihood estimated via the Steppingstone method was −460.86154.

The JC69 model (Jukes and Cantor 1969) is nested within the K80 model because JC69 assumes $\kappa = 1$. Given that the true value of κ used when simulating the data was $\kappa = 5$, we expect that the marginal likelihood for the JC69 model will be lower than that for the K80 model because the JC69 model cannot reach the region of parameter space in which the K80 model achieves its highest likelihoods.

The log marginal likelihood estimated using the LoRaD method is −467.33247 (Generalized Steppingstone estimate using 5 stones is −467.35384). The marginal likelihood for the JC model is indeed lower by more than 6 log units than the marginal likelihood for the K80 model.

Example: Fixed Topology

As an empirical phylogenetic example, we revisited the 32-taxon cicada dataset used to test the Generalized Steppingstone method (Fan et al. 2011). The data are from Marshall et al. (2006) (treebase.org study ID 1679). We removed the tRNA and the equivalent of a single codon (so that the data for no gene contained partial codons), leaving four protein-coding genes: COI (774 sites), COII (702 sites), ATPase8 (462 sites), and ATPase6 (149 sites). We partitioned the data four ways: unpartitioned (all 2087 sites concatenated); by gene (partitioned at gene boundaries); by codon (the partition comprised three subsets corresponding to the first, second, and third codon positions from all four genes); and by both (partitioned by gene and codon yielding 12 subsets). For every partitioning scheme, a separate GTR+G model was applied to each subset (with state frequencies, GTR exchangeabilities, and rate variance unlinked across subsets). Parameters that were linked across all subsets included the tree length and edge length proportions. The tree topology was fixed for all marginal likelihood analyses to the maximum likelihood tree obtained using PAUP* v. 4a166 (Swofford 2003) for the unpartitioned data using a GTR+G model.

Priors used were as follows:

State frequencies	$\text{Dirichlet}(1,1,1,1)$
GTR exchangeabilities	$\text{Dirichlet}(1,1,1,1,1,1)$
Rate variance	$\text{Gamma}(1,1)$
Edge proportions	$\text{Dirichlet}(1,1,\cdots,1)$
Tree length	$\text{Gamma}(1,10)$
Subset relative rates	$\text{Dirichlet}(1,1,\cdots,1)$

The number of Dirichlet parameters for the subset relative rate prior varied depending on the partition model (3 if partitioning by codon, 4 if by gene, and 12 if by both gene and codon).

MCMC analyses and estimation of the marginal likelihood for each partitioning scheme were carried out by software available in directory *lorad* of the Supplementary Materials.

MCMC analyses were run for 10,000,000 iterations following a burn-in of length 100,000 iterations. The burn-in iterations were used to tune Metropolis-Hastings proposals that were then fixed for the sampling iterations. The single chain was sampled every 100 iterations, yielding a sample size of 100,000 from the posterior distribution. For each of the four partition models, the MCSE was estimated for 11 coverage fraction (ϕ) values (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, and 0.99) for the first replicate, and the value of ϕ associated with the smallest MCSE was used to estimate the marginal likelihood using the LoRaD method for all 20 replicates. The ϕ values used were: 0.1 (unpartitioned), 0.2 (partitioned by codon), 0.8 (partitioned by gene), and 0.6 (partitioned by both gene and codon).

Marginal likelihoods for each partitioning scheme were also estimated using the Generalized Steppingstone (GSS) method (Fan et al. 2011). The GSS analyses involved 20 steppingstones, where each steppingstone is one ratio of normalizing constants estimated using a sample of size 10,000 (1,000,000 iterations saving every 100 following a burn-in of 10,000 iterations) from a single power posterior distribution. The 20 β values representing powers for the power posterior distributions were evenly spaced (i.e. $\beta \in \{0.00, 0.05, 0.10, 0.15, \cdots, 0.95\}$). The reference distribution was equivalent to the prior with first and second moments of each component matching the marginal posterior for that component in the posterior sample used for LoRaD. Thus, for this set of analyses, GSS was able to take advantage of the computational effort expended already to obtain the LoRaD estimate in determining its reference distribution.

Finally, the marginal likelihood was estimated using the Generalized Harmonic Mean (GHM) method (Arima and Tardella 2012). GHM, like LoRaD, requires a sample of points from only the posterior distribution, and thus

TABLE 1. Comparison of steppingstone (SS), LoRaD, and GHM for four different partition schemes for the cicada data.

Partition scheme	p	GSS	LoRaD	GHM
Unpartitioned	70	−10334.96	−10335.85	−10333.09
		(0.62)	(0.06)	(0.35)
By codon	90	−9826.69	−9827.86	−9823.55
		(0.82)	(0.18)	(0.68)
		(0.52)	(0.08)	(1.69)
By gene and codon	180	−9884.76	−9887.36	−9873.84
		(0.73)	(0.51)	(1.26)

Note: Values in parentheses represent the standard deviation of the log marginal likelihood across 20 replicates.

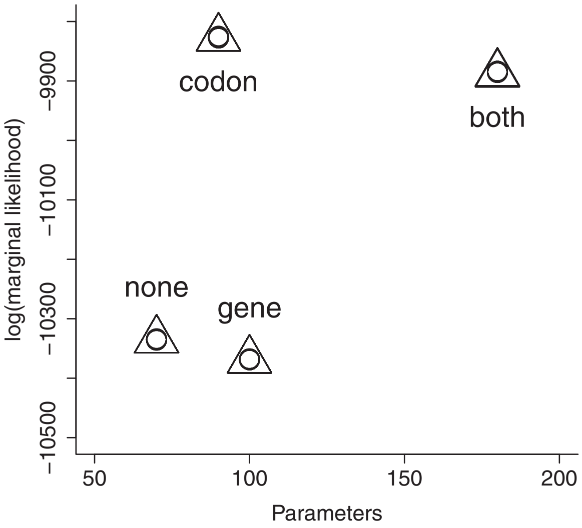

FIGURE 3. Scatter plot of 20 replicate Generalized Steppingstone (GSS; circles) and 20 replicate LoRaD (triangles) log marginal likelihoods for each of four different partition schemes for the cicada data.

comparison with GHM is important to see whether the additional transformation and standardization steps used in LoRaD improve accuracy over the much simpler GHM approach. We did not compare LoRaD to the IDR method because Arima and Tardella (2012) found that GHM was as good or better than IDR in practice.

LoRaD produced estimates that were, on average, slightly lower than GSS, but the marginal likelihoods estimated from the LoRaD and Steppingstone methods differed by less than 0.03% for all partitioning schemes (Table 1 and Fig. 3). Despite using half the computational effort of GSS, LoRaD estimates are more precise than GSS as judged by the standard deviation of log marginal likelihood estimates across MCMC replicates. GHM produced estimates that were considerably higher than either LoRaD or GSS for all partitioning schemes, and GHM was more variable than either of the other two methods across replicate MCMC analyses.

Computational Efficiency of LoRaD vs. GSS

Using the best partition model ("by codon") from the cicada example, we explored the performance of GSS vs. LoRaD for varying MCMC run lengths to examine which method accurately estimates the marginal likelihood with the fewest MCMC iterations. The number of burn-in iterations for each point was 10% of the number of sampling iterations. For GSS, the number of sampling iterations was divided by 20, with a fraction 1/20 going toward estimating the reference distribution and 19/20 spent on sampling 19 different ratios (i.e. 19 "steppingstones"). The LoRaD method does not require sampling from power posteriors, so all sampling involved the posterior distribution. For LoRaD, both training fraction and coverage were 0.5 (except where noted below). In both GSS and LoRaD, only every 10th iteration was sampled.

It is clear that LoRaD outperforms GSS for this example if the number of MCMC iterations is less than 500,000 (5.7 on log_{10} scale) (Fig. 4). Both LoRaD and GSS struggled for 20,000 and 50,000 iterations (4.3 and 4.7, respectively, on log_{10} scale). For these two run lengths, GSS estimates deviated from the true value substantially, and LoRaD could not be computed without increasing the training fraction to 0.8 and the coverage to 0.99 because no points from the estimation sample fell within the working parameter space defined by the training sample.

Example: Variable Topology

Holder et al. (2014) described a new reference distribution for tree topologies that improved the efficiency of Generalized Steppingstone for variable-topology Bayesian analyses. While it is not possible to know the true marginal likelihood, they came as close as possible by using GSS to estimate the fixed-topology marginal likelihood for all 105 tree topologies for data comprising 5 taxa from the green algal genus *Protosiphon* and a *Chlamypodium vacuolatum* outgroup and 1376 sites from the protein-coding chloroplast gene *rbc*L (Lewis and Trainor 2012). The total marginal likelihood can be estimated using these 105 conditional marginal likelihoods as follows:

$$p(y) = \sum_{i=1}^{105} p(y|\tau_i)\, p(\tau_i) \tag{21}$$

$$= \frac{1}{105} \sum_{i=1}^{105} p(y|\tau_i), \tag{22}$$

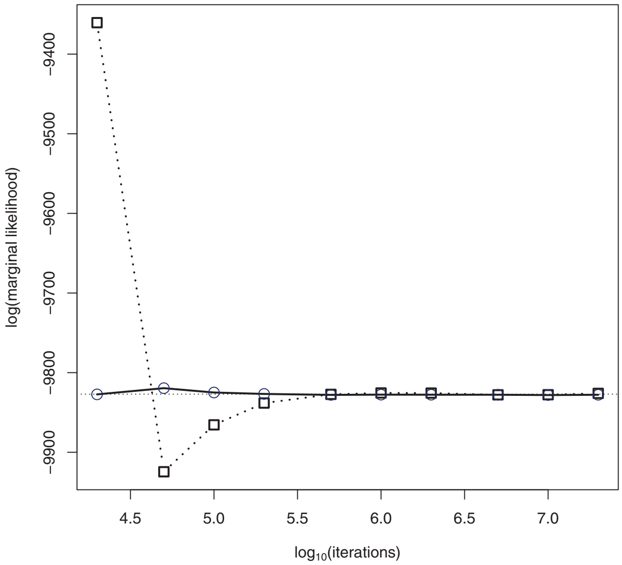

FIGURE 4. Plot showing estimated log marginal likelihood as a function of log_{10} (number of sampling iterations) for LoRaD (solid line with circular points) and GSS (dotted line with square points). The thin dotted horizontal line is at −9827.

where τ_i is the i th distinct tree topology of 105 total. The second line follows from the fact that a discrete uniform prior was assumed for the 105 possible unrooted topologies for 6 taxa. Holder et al. (2014) compared this "brute force" estimate with the GSS estimate (from a single variable-topology analysis using the new topology reference distribution) for 24 models. We estimated the total marginal likelihood using the LoRaD method described here for a subset of 16 of these models and compared the result with the brute force marginal likelihoods reported by Holder et al. (2014) (Table 2). First, an MCMC analysis was performed in which topology was allowed to vary. Second, the resulting posterior sample was filtered so that only sample points associated with the focal tree topology (i.e. the maximum a posteriori, or MAP, tree topology) were retained. Third, LoRaD was used to estimate the marginal likelihood conditional on the focal tree topology, and the Chib method described here was used to estimate the total marginal likelihood using the sampled frequency as the estimate of the marginal posterior probability of the focal tree.

These analyses differed in some details from those used for the cicada data in order to match the marginal likelihoods reported by Holder et al. (2014). For example, an Exponential(10) prior (mean 0.1) was placed on each individual edge length rather than the Gamma-Dirichlet (Rannala et al. 2011; Zhang et al. 2012) prior used for the cicada dataset. Also, rate heterogeneity for the *Protosiphon* analyses was parameterized using the shape parameter from the discrete Gamma among-site rate heterogeneity model; for the cicada data, the parameter used was the rate variance (inverse of the shape parameter).

The following table lists the priors used for the analyses reported in Table 2.

Tree topology	Discrete Uniform $(1,105)$
State frequencies	Dirichlet $(1,1,1,1)$
GTR exchangeabilities	Dirichlet $(1,1,1,1,1,1)$
Gamma shape	Gamma$(1,1)$
Edge length	Exponential(10)
Subset relative rates	Dirichlet$(1,1,1)$

A 6-taxon unrooted tree has 9 edge length parameters. Edge lengths and tree topology were linked across subsets for partitioned analyses; all other parameters were unlinked across partition subsets. For example, the codon-partitioned GTR+I+G model has 41 free (estimated) substitution model parameters (not including tree topology): 9 edge lengths, 2 subset relative rates, 3 sets of 5 exchangeabilities, 3 sets of 3 state frequencies, 3 gamma shapes, and 3 proportion of invariable sites parameters.

All MCMC analyses, regardless of the number of parameters, involved a 10,000 iteration burn-in period in which proposals were tuned followed by 30,000,000 iterations, sampling every 1000, to yield 30,000 sampled parameter values. Approximately 70% of these 30,000 samples involved the focal topology. The LoRaD and brute force GSS estimates differ by less than 0.03% even for the most complex model tested (partitioned GTR+I+G model).

DISCUSSION

Marginal likelihood estimation has become increasingly important in systematics studies in the past decade as methods have improved and software implementing these methods and promoting their use in manuals and tutorials has become more widely available. One feature of state-of-the-art methods for marginal likelihood estimation such as Path Sampling (Lartillot and Philippe 2006) and Steppingstone (Fan et al. 2011; Xie et al. 2011) is the requirement for sampling from multiple power posterior distributions, which involve, like the posterior distribution, the product of the likelihood and joint prior but differ in being intermediate between the posterior and a reference distribution, which equals the joint prior in the case of Steppingstone (Xie et al. 2011). For both the Steppingstone and Generalized Steppingstone (Fan et al. 2011) methods, the posterior distribution is not sampled at all, which means that the computational effort expended in estimating the marginal likelihood cannot be used for estimating model parameters or marginal probabilities of clades or entire tree topologies.

The LoRaD method proposed here provides an accurate means of estimating the marginal likelihood that requires only a sample from the posterior distribution. Thus, no computation is wasted because the same sample used to estimate model parameters may also be used

TABLE 2. Comparison of "brute force" Steppingstone estimate (GSS) and LoRaD for 16 models for the *Protosiphon* data.

Model	Parameters	GSS	LoRaD	S.D.	δ
JC	9	−2776.52	−2776.52	0.04	0.00
JC + I	10	−2744.59	−2744.58	0.01	−0.01
JC + G	10	−2747.44	−2747.44	0.02	0.00
JC + I+ G	11	−2743.56	−2743.57	0.02	0.01
GTR	17	−2714.20	−2714.30	0.02	0.10
GTR + I	18	−2681.00	−2681.22	0.04	0.22
GTR + G	18	−2682.73	−2682.95	0.04	0.22
GTR + I + G	19	−2680.29	−2680.56	0.06	0.27
*JC	11	−2681.79	−2681.79	0.01	0.00
*JC + I	14	−2668.38	−2668.40	0.02	0.02
*JC + G	14	−2668.99	−2668.99	0.02	0.00
*JC + I + G	17	−2667.19	−2667.19	0.08	0.00
*GTR	35	−2551.10	−2551.47	0.15	0.37
*GTR + I	38	−2535.57	−2536.14	0.24	0.57
*GTR + G	38	−2536.75	−2537.17	0.06	0.42
*GTR + I + G	41	−2534.66	−2535.31	0.20	0.65

Notes: Asterisks (*) denote models in which data was partitioned by codon position. The number of parameters listed does not include the tree topology, which was also estimated for each of these models. The δ column shows the difference between the GSS estimate of the log marginal likelihood and the mean of 30 LoRaD estimated log marginal likelihoods from independent MCMC analyses. Standard deviations of log marginal-likelihoods are based on the same 30 replicates.

to estimate the marginal likelihood. While other accurate methods (e.g. bridge sampling; Meng and Wong (1996)) operate on a single posterior sample, these methods all require additional likelihood calculations above and beyond those used to obtain the sample. The LoRaD method does require some computation, but avoids the need to make additional expensive phylogenetic likelihood evaluations, thus obviating the need for modifications to existing Bayesian phylogenetic software. Finally, considerable computational effort is saved by sampling from just one distribution, not a series of distributions along a path from posterior to prior.

As appealing as this seems, there are some caveats associated with the LoRaD method. Because a multivariate standard normal distribution is used as the reference, the posterior sample must be log-transformed and standardized so that it is approximately standard normal. This means that LoRaD is not expected to work well with multimodal posterior distributions. The standardization involves estimating and inverting a potentially large $p \times p$ variance–covariance matrix, where p is the total number of model parameters. It also may be the case that a larger sample is needed from the posterior than would ordinarily be used in order to achieve the desired accuracy, although the computational effort expended in obtaining that larger sample improves all inferences and is not effort wasted.

An important caveat for the LoRaD method is that a reasonably large posterior sample is needed from a single tree topology. If MCMC analyses involve fixed tree topology, then this is not a problem, but estimating the tree topology itself is the focus in many Bayesian phylogenetic analyses and those analyses thus allow tree topology to vary from one iteration to the next. In such cases, LoRaD is most useful if one tree topology is well represented in the posterior sample. That is, one topology τ has marginal posterior probability high enough that it provides a sample of sufficient size to estimate $p(y \mid \tau)$ using LoRaD. If, for example, nearly every sampled tree from the posterior involves a distinct topology, then clearly LoRaD will not be able to provide an accurate estimate of the unconditional marginal likelihood $p(y)$ because both $p(y \mid \tau)$ and $p(\tau \mid y)$ in Chib's identity, Equation (18), are poorly estimated. In such cases, more computationally expensive methods (Lartillot and Philippe 2006; Fan et al. 2011; Xie et al. 2011) will be required. Estimation of the Monte Carlo Standard Error (MCSE) can be used to determine whether the number of samples obtained for the most frequently sampled topology τ is sufficient for estimating $p(y \mid \tau)$.

In summary, LoRaD provides a way to get accurate marginal likelihood estimates efficiently in Bayesian phylogenetics when either tree topology is fixed or one tree topology occurs at a frequency high enough in the posterior sample to estimate the marginal likelihood conditional on that topology. Unlike other fast methods, LoRaD is unbiased and assumes only that the posterior is unimodal. A further advantage is that LoRaD does not require existing Bayesian software to be modified as long as both the log joint prior and log-likelihood are provided in the output alongside sampled parameter values.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.pg4f4qrrw.

Funding

This material is based upon work supported by the National Science Foundation under grant no. DEB-1354146.

Acknowledgements

The authors would like to thank the Associate Editor, Mark Holder, and two anonymous reviewers for their constructive comments and suggestions.

References

Arima S., Tardella L. 2012. Improved harmonic mean estimator for phylogenetic model evidence. J. Comput. Biol. 19:418–438.

Baele G., Lemey P., Suchard M.A. 2015. Genealogical working distributions for Bayesian model testing with phylogenetic uncertainty. Syst. Biol. 65:250–264.

Beerli P., Palczewski M. 2010. Unified framework to evaluate panmixia and migration direction among multiple sampling locations. Genetics 185:313–326.

Chib S. 1995. Marginal likelihood from the Gibbs output. J. Am. Stat. Assoc. 90:1313–1321.

Edmundson H.P. 1961. The distribution of radial error and its statistical application in war gaming. Oper. Res. 9:8–21.

Fan Y., Wu R., Chen M.H., Kuo L., Lewis P.O. 2011. Choosing among partition models in Bayesian phylogenetics. Mol. Biol. Evol. 28:523–532.

Fourment M., Magee A.F., Whidden C., Bilge A., Matsen F.A., Minin V.N. 2020. 19 dubious ways to compute the marginal likelihood of a phylogenetic tree topology. Syst. Biol. 69:209–220.

Friel N., Pettitt A.N. 2008. Marginal likelihood estimation via power posteriors. J R Stat Soc B 70:589–607.

Holder M.T., Lewis P.O., Swofford D.L., Bryant D. 2014. Variable tree topology stepping-stone marginal likelihood estimation. In Chen M.H., Kuo L., and Lewis P.O., editors. Bayesian phylogenetics: methods, algorithms, and applications, New York: Chapman & Hall/CRC, p. 95–111.

Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In Munro H.N., editor. Mammalian protein metabolism. New York: Academic Press, p. 21–132.

Kimura M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.

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

Lewis L.A., Trainor F. 2012. Survival of *Protosiphon botryoides* (Chlorophyceae, Chlorophyta) from a Connecticut soil dried for 43 years. Phycologia 51:662–665.

Marshall D., Simon C., Buckley T. 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 W.H. 1996. Simulating ratios of normalising constants via a simple identity: a theoretical exploration. Stat. Sin. 6:831–860.

Newton M.A., Raftery A.E. 1994. Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). J. R. Stat. Soc.B 56:3–26.

Rannala B., Zhu T., Yang Z. 2011. Tail paradox, partial identifiability, and influential priors in Bayesian branch length inference. Mol. Biol. Evol. 29:325–335.

Swofford D.L. 2003. PAUP*. phylogenetic analysis using parsimony (*and Other Methods. Version 4). Sunderland, Massachusetts: Sinauer Associates. http://paup.phylosolutions.com

Wang Y.B., Chen M.H., Kuo L., Lewis P.O. 2018. A new Monte Carlo method for estimating marginal likelihoods. Bayesian Anal. 13:311–333.

Xie W., Lewis P.O., Fan Y., Kuo L., Chen M.H. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst. Biol. 60:150–160.

Zhang C., Rannala B., Yang Z. 2012. Robustness of compound Dirichlet priors for Bayesian inference of branch lengths. Syst. Biol. 61:779–784.