Syst. Biol. 55(2):195–207, 2006

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150500433722

Computing Bayes Factors Using Thermodynamic Integration

NICOLAS LARTILLOT[1] AND HERVÉ PHILIPPE[2]

[1]*Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier UMR 5506, CNRS-Université de Montpellier 2, 161, rue Ada, 34392 Montpellier Cedex 5, France; E-mail: nicolas.lartillot@lirmm.fr*
[2]*Canadian Institute for Advanced Research, Département de Biochimie, Université de Montréal, Montréal, Québec, Canada*

Abstract.—In the Bayesian paradigm, a common method for comparing two models is to compute the Bayes factor, defined as the ratio of their respective marginal likelihoods. In recent phylogenetic works, the numerical evaluation of marginal likelihoods has often been performed using the harmonic mean estimation procedure. In the present article, we propose to employ another method, based on an analogy with statistical physics, called thermodynamic integration. We describe the method, propose an implementation, and show on two analytical examples that this numerical method yields reliable estimates. In contrast, the harmonic mean estimator leads to a strong overestimation of the marginal likelihood, which is all the more pronounced as the model is higher dimensional. As a result, the harmonic mean estimator systematically favors more parameter-rich models, an artefact that might explain some recent puzzling observations, based on harmonic mean estimates, suggesting that Bayes factors tend to overscore complex models. Finally, we apply our method to the comparison of several alternative models of amino-acid replacement. We confirm our previous observations, indicating that modeling pattern heterogeneity across sites tends to yield better models than standard empirical matrices. [Bayes factor; harmonic mean; mixture model; path sampling; phylogeny; thermodynamic integration.]

Bayesian methods have become popular in molecular phylogenetics over the recent years. The simple and intuitive interpretation of the concept of probabilities underlying the Bayesian paradigm makes it an appealing framework of scientific inference in general (Jaynes, 2003). On the other hand, the Bayesian practice also entails mathematical difficulties, which have prevented its use in most practical fields until recently. Over the last 10 years, the situation has changed, mainly due to the impressive advances in computational power. In addition, general numerical methods based on Markov chains Monte Carlo (MCMC) have been developed, allowing one to conduct Bayesian inferences under a large category of probabilistic models, with few constraints on dimensionality or analytical integrability (Gelman et al., 2004; Holder and Lewis, 2003; Huelsenbeck et al., 2002).

However, this new freedom in model exploration has to be complemented by efficient and reliable methods of model evaluation and selection. More fundamentally, it raises the question of whether devising more complex evolutionary models is indeed relevant in the first place, given the problems that such a project might imply (Rannala, 2002). At first sight, current phylogenetic models offer a good compromise between complexity and tractability. They account for unequal rates of substitution among amino acids through general time-reversible matrices determined empirically or directly inferred from the data (Jones et al., 1992; Whelan and Goldman, 2001). They also allow different positions along the sequence to evolve at different speeds (Yang, 1993, 1994). Both aspects seem to have a significant impact on the quality of the phylogenetic estimates (Brinkmann et al., 2005; Yang, 1996). Yet, this may not be sufficient, as evidenced by all the inconsistencies still observed in many phylogenetic analyses (Gaut and Lewis, 1995; Philippe et al., 2005; Stefanovic, 2004; Sullivan and Swofford, 1997). In an otherwise coherent statistical framework, such as maximum likelihood or the Bayesian method, inconsistencies are a clear indication of model misspecifications, suggesting that some simplifying assumptions common to most current models (e.g., absence of gene conversions or lateral gene transfers, homogeneity of the equilibrium frequencies across sites, stationarity of the substitution process across lineages) may need to be relaxed as well. Hence, in the aim of obtaining more reliable phylogenetic inference, a wider diversity of models than those currently considered has still to be investigated, calling for good methods to perform both parameter estimation and reliable model choice.

Bayesian inference is in general tantamount to exploring the posterior probability distribution over the parameters of interest. Given a model M, with parameter vector $\theta \in \Theta$ (specifying, for instance, the tree topology and branch lengths), and applied on a dataset D, the posterior probability distribution is given by Bayes' theorem:

$$p(\theta \mid D, M) = \frac{p(D \mid \theta, M)p(\theta \mid M)}{p(D \mid M)} \quad (1)$$

where $p(\theta \mid M)$ is the prior distribution, $p(D \mid \theta, M)$ the likelihood function, and

$$p(D \mid M) = \int_{\Theta} p(D \mid \theta, M)p(\theta \mid M)d\theta \quad (2)$$

is the normalization constant, also called the *predictive probability*, or *marginal likelihood*.

As for model fit, the normalization constant, $p(D \mid M)$, is of primary importance. As a function of M, it can literaly read as the likelihood of model M, given the data D. Accordingly, among several models, one is led to choose the one of greatest marginal likelihood. When two models M_0 and M_1 are being compared, one usually defines the *Bayes factor* in favor of M_1 over M_0 as the ratio of their respective marginal likelihoods (Jeffreys, 1935; Kass and

Raftery, 1995):

$$B_{01} = \frac{p(D \mid M_1)}{p(D \mid M_0)}. \quad (3)$$

Values of the Bayes factor greater (smaller) than 1 will be considered as evidence in favor of M_1 (M_0). Other approaches for evaluating model fit in a Bayesian context have been proposed, such as cross-validation (Stone, 1974), posterior predictive approaches (Gelman et al., 1996; Meng, 1994; Rubin, 1984) applied in phylogenetic model comparison (Bollback, 2002), as well as fractional (O'Hagan, 1995), posterior (Aitkin, 1991), or intrinsic (Berger and Pericchi, 1996) Bayes factors. But in the following, we will focus exclusively on the traditional Bayes factor, which is more intuitive in a model-likelihood interpretation perspective.

In practice, posterior expectations can be efficiently estimated by sampling from the posterior distribution, using, for instance, MCMC methods such as the Metropolis-Hastings or the Gibbs sampling algorithms. These methods are now applied extensively in molecular phylogenetics (Huelsenbeck and Ronquist, 2001; Larget and Simon, 1999; Lartillot and Philippe, 2004; Pagel and Meade, 2004; Suchard, 2001). In contrast, the numerical evaluation of the marginal likelihood, and thereby of the Bayes factor, is anything but easy, in particular for high dimensional models, and for large datasets (Han and Carlin, 2000; Kass and Raftery, 1995). Note, in this respect, that the MCMC algorithms used for posterior sampling only involve the ratio of two posterior probabilities (i.e., of the current and the newly proposed parameter value), in which the normalization constant $p(D \mid M)$ cancels out:

$$\frac{p(\theta_1 \mid D, M)}{p(\theta_2 \mid D, M)} = \frac{p(D \mid \theta_1, M)p(\theta_1 \mid M)}{p(D \mid \theta_2, M)p(\theta_2 \mid M)} \quad (4)$$

This implies that these algorithms, however efficient at sampling from the posterior, do not allow one to estimate $p(D \mid M)$ directly.

Among the methods available for evaluating Bayes' factors, many are valid only under very specific conditions. For instance, the Dickey-Savage ratio (Verdinelli and Wasserman, 1995), applied in phylogenetics (Suchard, 2001), assumes nested models. The Laplace estimator (Kass and Raftery, 1995), or the Bayesian Information Criterium (Schwartz, 1978), applied in phylogenetics (Minin et al., 2003; Waddell et al., 2002) are large sample approximations around the maximum likelihood, which cannot always be easily evaluated for complex models. The Laplace estimator (Kass and Raftery, 1995) relies on a normal approximation around the maximum likelihood (ML), which may not be valid for parameter-rich models. The reversible-jump approach (Green, 1995), where a MCMC is devised to jump between models according to the Metropolis-Hastings rule, can in principle be made as general as desired. Yet, in practice, the Metropolis-Hastings moves between models have to be accepted at a sufficient rate for the method to be practical. This requirement can be met quite easily as long as the models being compared are formulated along similar parameterizations; for instance, alternative substitution matrices (Huelsenbeck et al., 2004), or different number of classes for a mixture model (Lartillot and Philippe, 2004). In contrast, the reversible-jump method is not easily applicable when comparing models based on an entirely different parametric rationale.

We are thus left with only a few methods of potentially general applicability, among which (1) the importance sampling estimators, and particularly the harmonic mean estimator (HME) (Newton and Raftery, 1994), and (2) thermodynamic integration, or path sampling (Gelman, 1998; Ogata, 1989). The HME is by far the simplest method, only requiring a sample from the posterior distribution. It has been applied repeatedly, in particular in phylogenetic model comparison (Irestedt et al., 2004; Nylander et al., 2004; Pagel and Meade, 2004). Because its variance may be infinite, a modified, stabilized version has also been proposed (Newton and Raftery, 1994), also used in phylogenetics (Suchard et al., 2003). Thermodynamic integration, on the other hand, is based on a completely different rationale, relies on a more elaborate and computationally more intensive MCMC sampling scheme, but is statistically more well-behaved (Gelman, 1998). Its name stems from an analogy with physics, where the marginal likelihood is equivalent to the so-called partition function and its logarithm to the free energy. In fact, physicists have had to evaluate probabilities formulated in terms of high-dimensional integrals for a long time now (Neal, 2000). Therefore, transposing their well-tried methods into other numerical problems could be a promising approach.

In this work, we have implemented the HME and the method of thermodynamic integration. We have applied them to the comparison of models of sequence evolution. We show by several means that, whereas thermodynamic integration yields reliable quantitative estimates of Bayes' factors, the HME is unreliable and can even lead to qualitative reversions of the comparisons being conducted. Altogether, considering that some Bayes' factor evaluations performed in a phylogenetic context thus far have relied on the harmonic estimator, we advocate that more caution should be applied, and that thermodynamic integration, or other methods not investigated here, should be used instead. Finally, using thermodynamic integration, we compare several models of amino-acid replacement, among which are site-heterogeneous models that we have proposed previously (Lartillot and Philippe, 2004).

Data and Models

Five datasets were considered in this study. The following nomenclature specifies, for each dataset, the type of protein, the number of taxa (P), and the length of the alignment (N):

- **PGK30-276**: sequences of phosphoglycerate kinase of 30 eubacterial species
- **EF30-627**: sequences of elongation factor 2 from 30 eukaryotes

- **POL39-888**: RNA polymerase Rpb2 of 39 eubacteria
- **DLIG40-430**: DNA ligase of 40 eubacteria
- **UVR30-719**: DNA excision nuclease subunit A of 30 eubacteria

For each dataset, the amino-acid sequences were retrieved from the databases and aligned using ClustalW (Thompson et al., 1994). The alignments were hand-corrected using the MUST package (Philippe, 1993), and regions ambiguously aligned were removed with the help of the GBlocks program (Castresana, 2000).

We assume a uniform prior over topologies, and an exponential distribution on branch lengths, with mean determined by a hyperparameter λ. Rates across sites can be either uniform (UNI model) or distributed according to a Invariant + Gamma (I+Γ) distribution (RAS), in which case both the α parameter and the proportion of invariant sites (p_0) are considered as free parameters. We propose three alternative sets of priors on hyperparameters:

- $P1$: an exponential prior of mean 1 on λ and on α (default prior)
- $P2$: an exponential prior of mean 1 on λ and on $1/\alpha$
- $P3$: a fixed value for λ ($\lambda = 10$), and a flat prior on α, with the restriction that $\alpha < 100$, for the prior to remain proper.

In all cases, we assume a uniform prior on p_0.

For the amino-acid replacement model, we consider five different cases:

- **WAG:** the WAG empirical matrix (Whelan and Goldman, 2001). Stationary probabilities (equilibrium frequencies) will either be set equal to the values reported in the original article (WAG model), or considered as free parameters, with a flat Dirichlet prior (WAG+F model).
- **Poisson:** a Poisson process, which is characterized by its stationary probability vector. We use the same combination of stationary probabilities as for WAG (i.e., Poisson, or Poisson+F).
- **GTR:** the most general time-reversible matrix, which is implemented as described previously (Lartillot and Philippe, 2004).
- **MAX:** each site has its own amino-acid replacement matrix, which is a Poisson process, whose profile, defined by the 20 equilibrium frequencies, is a random variable distributed according to a flat Dirichlet.
- **CAT:** the distribution of amino-acid replacement matrices across sites is modeled by a mixture of a free number of Poisson processes. Each component is defined by a stationary probability vector. The prior is specified by a Dirichlet process (Lartillot and Philippe, 2004).

General MCMC Settings

The methods and implementation for MCMC sampling under these models have been described previously (Lartillot and Philippe, 2004). Briefly, the different components of the parameter vector (topology, branch lengths, site-specific rates, stationary probability vectors, hyperparameters) are updated separately, according to a sequence of calls to all available update mechanisms. One such sequence defines a *cycle*. The number of cycles required for a given chain to reach its stationary equilibrium (burn-in), as well as the total number of cycles and the saving frequency, are first determined empirically. In a second step, the effective size of the sample is determined a posteriori by a time-series variance estimation method based on the empirical autocovariances of the log-likelihood time series (Geyer, 1992). We used a Tukey-Janning lag window (Raftery and Lewis, 1992), with a cutoff at $K/4$, where K is the total number of saved points. Given K and the decorrelation time τ, the effective size of the sample is then estimated as $K_{eff} = K/\tau$.

For each dataset, a first MCMC run under the WAG+F, I+Γ model was conducted, and the consensus of 1,000 trees sampled from the posterior was computed. This consensus was then used for any analysis conducted under a fixed topology.

All source codes, data files, and trees are available from http://systematicbiology.org. Data matrices can also be downloaded from TreeBase (http://www.treebase.org, accession numbers S1388, M2476–M2480).

MARGINAL LIKELIHOOD ESTIMATION

Importance Sampling Estimators

Given an unnormalized density $g(\theta)$, an unbiased estimate of $p(D \mid M)$ is given by the importance sampling formula

$$p(D \mid M) = \frac{E_g\left[\frac{p(D|\theta,M)p(\theta|M)}{g(\theta)}\right]}{E_g\left[\frac{p(\theta|M)}{g(\theta)}\right]}, \tag{5}$$

where $E_g[\cdots]$ is the expectation over g (Kass and Raftery, 1995). Using Monte Carlo procedures, a sample $(\theta_k)_{k=1..K}$ can be drawn from g and used to approximate the expectations $E_g[\cdots]$:

$$p(D \mid M) \simeq \frac{\sum_{k=1}^{K} \frac{p(D|\theta_k,M)p(\theta_k|M)}{g(\theta_k)}}{\sum_{k=1}^{K} \frac{p(\theta_k|M)}{g(\theta_k)}}. \tag{6}$$

The simplest application of this method is to use the prior as the importance sampling distribution ($g(\theta) = p(\theta \mid M)$), in which case Eqs. (5) and (6) lead to the *prior arithmetic mean estimator* (AME):

$$p(D \mid M) = E_{\text{prior}}[p(D \mid \theta, M)] \tag{7}$$

$$\simeq \frac{1}{K}\sum_{k=1}^{K} p(D \mid \theta_k, M). \tag{8}$$

However, a well-known problem with this estimator is that the high-likelihood region can be very small.

Therefore, unless K is very large, the sample drawn from the prior will contain virtually no points from the high-likelihood region, resulting in a very poor estimate of $p(D \mid M)$.

An alternative, proposed by Newton and Raftery (1994), is to draw from the posterior, rather than from the prior ($g(\theta) = p(D \mid \theta, M)$). Intuitively, this should have the advantage of enriching the sample in points from the high-likelihood region. This results in the *posterior harmonic mean estimator* (HME):

$$\frac{1}{p(D \mid M)} = E_{\text{post}}\left[\frac{1}{p(D \mid \theta, M)}\right] \tag{9}$$

$$= \frac{1}{K}\sum_{k=1}^{K}\frac{1}{p(D \mid \theta_k, M)}. \tag{10}$$

The HME converges almost surely to the inverse of the marginal likelihood. However, in many practical situations, its variance is infinite. To circumvent this problem, Newton and Raftery (1994) proposed a third importance sampling scheme, called the *stabilized harmonic mean estimator* (SHME), based on a mixture of the prior and the posterior: $g(\theta) = \delta p(\theta \mid M) + (1 - \delta)p(\theta \mid M)$. Typically, δ is chosen equal to 0.1.

General Principles of Thermodynamic Integration

This method, also called path sampling, is explained in greater details elsewhere (Gelman, 1998; Neal, 2000). Here, we give a slightly less formal introduction to its principles and show how it can be applied to phylogenetic problems.

Let us suppose that we have two unnormalized densities, $q_0(\theta)$ and $q_1(\theta)$, defined on the same parameter space Θ. The corresponding true probability densities are denoted by

$$p_i(\theta) = \frac{1}{Z_i}q_i(\theta),\ i = 0, 1, \tag{11}$$

where

$$Z_i = \int_\Theta q_i(\theta)d\theta,\ i = 0, 1 \tag{12}$$

are the normalization constants. Typically, in a Bayesian context, $q_i(\theta) = p(D \mid \theta, M_i)p(\theta \mid M_i)$, $Z_i = p(D \mid M_i)$, and thus, $p_i(\theta) = p(\theta \mid D, M_i)$.

We wish to perform a numerical evaluation of the log-ratio

$$\mu = \ln\left(\frac{Z_1}{Z_0}\right) \tag{13}$$

$$= \ln Z_1 - \ln Z_0. \tag{14}$$

To do this, we define a continuous and differentiable path $(q_\beta)_{0\le\beta\le1}$ in the space of unnormalized densities, joining q_0 and q_1. By extension, for any β, $0 \le \beta \le 1$, p_β and Z_β are defined as

$$p_\beta(\theta) = \frac{1}{Z_\beta}q_\beta(\theta), \tag{15}$$

$$Z_\beta = \int_\Theta q_\beta(\theta)d\theta. \tag{16}$$

When β tends to 0 (resp. 1), p_β converges pointwise to p_0 (resp. p_1), and Z_β to Z_0 (resp. Z_1).

Taking the derivative of $\ln Z_\beta$ with respect to β:

$$\frac{\partial \ln Z_\beta}{\partial\beta} = \frac{1}{Z_\beta}\frac{\partial Z_\beta}{\partial\beta} \tag{17}$$

$$= \frac{1}{Z_\beta}\frac{\partial}{\partial\beta}\int_\Theta q_\beta(\theta)d\theta \tag{18}$$

$$= \frac{1}{Z_\beta}\int_\Theta \frac{\partial q_\beta(\theta)}{\partial\beta}d\theta \tag{19}$$

$$= \int_\Theta \frac{1}{q_\beta(\theta)}\frac{\partial q_\beta(\theta)}{\partial\beta}\frac{q_\beta(\theta)}{Z_\beta}d\theta \tag{20}$$

$$= \int_\Theta \frac{\partial \ln q_\beta(\theta)}{\partial\beta}p_\beta(\theta)d\theta \tag{21}$$

$$= E_\beta\left[\frac{\partial \ln q_\beta(\theta)}{\partial\beta}\right], \tag{22}$$

where $E_\beta[\cdots]$ stands for the expectation with respect to p_β. Defining the *potential*

$$U(\theta) = \frac{\partial \ln q_\beta(\theta)}{\partial\beta}, \tag{23}$$

one has thus the first moment identity:

$$\frac{\partial \ln Z_\beta}{\partial\beta} = E_\beta[U]. \tag{24}$$

Integrating over [0, 1] yields the log-ratio one is looking for:

$$\mu = \ln Z_1 - \ln Z_0 \tag{25}$$

$$= \int_0^1 \frac{\partial \ln Z_\beta}{\partial\beta}d\beta \tag{26}$$

$$= \int_0^1 E_\beta[U]d\beta. \tag{27}$$

The key idea of thermodynamic integration is that, for any value of β between 0 and 1, one can run a Markov chain Monte Carlo in which q_β is used as the unnormalized density in the Metropolis-Hastings ratio. By definition, this yields a sample of parameter values drawn from the probability distribution p_β. Expectations over p_β can then be estimated as averages over this sample, which in particular allows one to evaluate $E_\beta[U]$. This computation can be done for a series of values of β regularly spaced between 0 and 1, which implies running

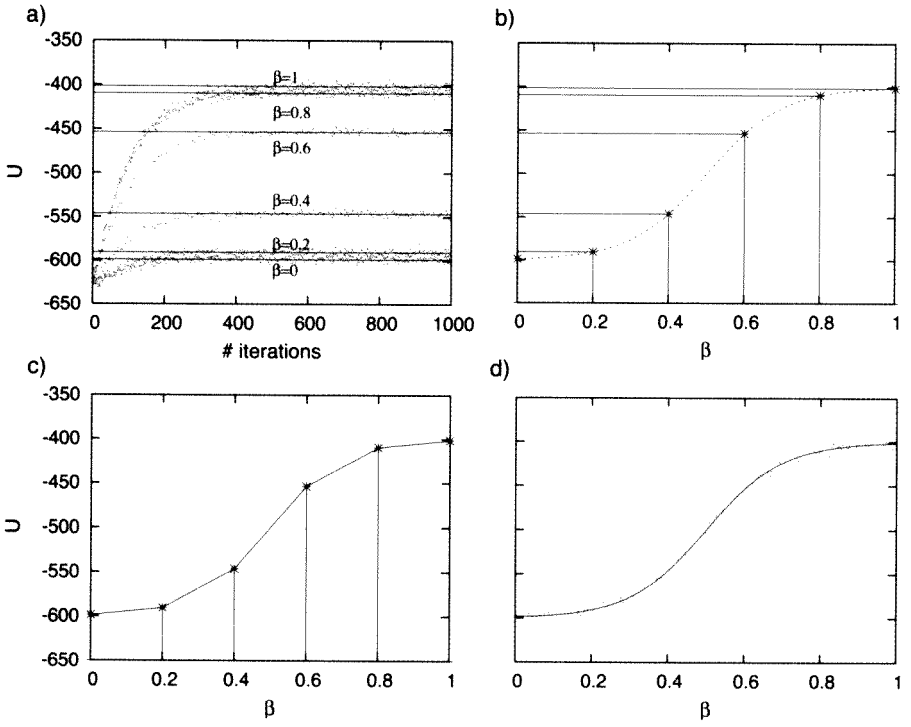

FIGURE 1. Rationale of the thermodynamic integration method. a, A series of independent chains are run under different values of β, and for each of them, the mean posterior expectation of the potential $U = \partial \ln q_\beta / \partial \beta$ is computed (horizontal lines). b, These mean posterior expectations are plotted against β. c, The integral of the curve is estimated by the Simpson procedure. d, Illustration of the quasistatic version, in which β moves continuously from 0 to 1 during MCMC (see text for details).

a Markov chain for each value of β (Fig. 1a, b). These sample expectations are finally used to approximate the integral over [0, 1] (Eq. (25)) using Simpson's triangulation method (Fig. 1c).

Specifically, assuming a discretization step of $\Delta\beta = 1/C$, with C an integer (e.g., $C = 10$), for each $d = 0..C$, we define $\beta_d = d \times \Delta\beta$, and run a Markov chain having p_{β_d} as its stationary distribution. The resulting sample is denoted by

$$\left(\theta_k^d\right)_{k=1..K} \sim p_{\beta_d}. \tag{28}$$

From that, $E_{\beta_d}[U]$ is estimated as

$$\hat{U}_d = \frac{1}{K} \sum_{k=1}^{K} U_{\beta_d}\left(\theta_k^d\right), \tag{29}$$

and by Simpson's triangulation, one gets the *discrete thermodynamic estimate* of $\mu = \ln Z_1 - \ln Z_0$:

$$\hat{\mu}_{ds} = \frac{1}{C}\left(\frac{1}{2}\hat{U}_0 + \sum_{d=1}^{C-1} \hat{U}_d + \frac{1}{2}\hat{U}_C\right). \tag{30}$$

We used this discrete method previously (Lartillot and Philippe, 2004). In the present work, we also introduce a continuous (or *quasistatic*) version, which has the advantage of yielding a greater accuracy. The quasistatic method consists in equilibrating a MCMC under $\beta = 0$, then smoothly increasing the value of β, by adding a constant increment $\delta\beta$ after each series of Q cycles, until $\beta = 1$ is reached (Fig. 1d). During this procedure, points θ_k are saved, for instance, before each update of β. Let us denote $(\beta_k, \theta_k)_{k=0..K}$ the series of points obtained in this way. One has in particular $\beta_0 = 0$, $\beta_K = 1$, and $\forall k\, 0 \leq k < K$, $\beta_{k+1} - \beta_k = \delta\beta$. Then, the *quasistatic* estimate of $\ln Z_1 - \ln Z_0$ is given by:

$$\hat{\mu}_{qs} = \frac{1}{K}\left(\frac{1}{2}U(\theta_0) + \sum_{k=1}^{K-1} U(\theta_k) + \frac{1}{2}U(\theta_K)\right). \tag{31}$$

Equivalently, one can start at $\beta = 1$, equilibrate the MCMC, and then progressively decrease β, while sampling along the path down to p_0. We will make use of this bidirectional method below.

Many different schemes of thermodynamic integration can be devised, depending on the type of path considered. In the present work, we have used two main schemes.

Annealing-Melting Integration

This first scheme involves only one model (M) at a time, the path going from the prior to the unnormalized posterior defined by M. This path is defined as follows:

$$q_\beta(\theta) = p(D \mid \theta, M)^\beta p(\theta \mid M). \tag{32}$$

Note that $q_0(\theta) = p(\theta \mid M)$, and $q_1(\theta) = p(D \mid \theta, M)p(\theta \mid M)$. The corresponding normalization constants are $Z_0 = 1$ and $Z_1 = p(D \mid M)$. The integrand takes a simple form:

$$U(\theta) = \frac{\partial \ln q_\beta(\theta)}{\partial \beta} \tag{33}$$

$$= \ln p(D \mid \theta, M), \tag{34}$$

so that the integration defined above directly leads to an estimate of

$$\ln p(D \mid M) = \ln Z_1 - \ln Z_0 \tag{35}$$

$$= \int_0^1 E_\beta [\ln p(D \mid \theta, M)] d\beta. \tag{36}$$

In a thermodynamic perspective, the inverse of β can be seen as the equivalent of a temperature. Raising the likelihood to a power $\beta < 1$ is equivalent to smoothing out the likelihood surface, which will yield a "looser" Markov chain, more prone to accepting less likely parameter configurations. This is analogous to the behavior of a thermodynamic system, which has a higher probability of visiting high-energy microscopic configurations at higher temperature. Thus, slowly moving from $\beta = 0$ to $\beta = 1$ is equivalent to a quasistatic cooling down, or *annealing*, of the MCMC. Conversely, moving from $\beta = 1$ to $\beta = 0$ amounts to a warming up, or *melting*, of the MCMC. Note that the heated chains of the parallel Metropolis-coupled Markov chains (Altekar et al., 2004) are defined in a similar way (except that in their case, the posterior, rather than only the likelihood, is raised to the power β).

As an illustrating example, the annealing-melting integration method was applied to the PGK dataset, under a simple version of the rate-across-site (RAS) model, assuming a Poisson+F amino-acid replacement matrix, a prior mean branch length of $\lambda = 10$, and γ-distributed rates across sites, with $\alpha = 1$, and no invariant sites. The topology was constrained to the posterior consensus. Figure 2 shows the evolution of $\ln p(D \mid \theta, \text{RAS})$ as a function of β. The increment was $\delta\beta = \pm 0.001$ and was applied every $Q = 10,000$ cycles. The area situated between the curve and the zero axis is equal to the logarithm of the marginal likelihood under RAS and can be estimated using Eq. (31). In the present case, one obtains $\ln p(D \mid \text{RAS}) = -9922.1$ natural log units (nits).

Note that the curve shown in Figure 2 is strictly increasing. This can be shown theoretically: above, we have seen that the first derivative of $\ln Z$ was related to the expectation of U (Equation 24). By a similar argument, the second derivative of $\ln Z$ can be related to the variance of U:

$$\frac{\partial^2 \ln Z_\beta}{\partial \beta^2} = V_\beta[U] + E_\beta\left[\frac{\partial U}{\partial \beta}\right], \tag{37}$$

where $V_\beta[U] = E_\beta[U^2] - E_\beta^2[U]$. In the present case, $U = \ln p(D \mid \theta)$ does not depend on β, so that this simplifies into

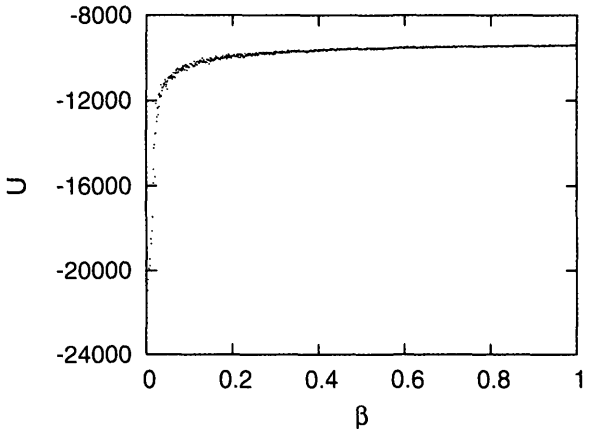

FIGURE 2. Annealing integration under the RAS model for the PGK dataset. The 1,000 values of $\ln p(D \mid \theta, M)$ sampled during the quasistatic integration are plotted against β, shadowing the curve of $E_\beta[\ln p(D \mid \theta, M)]$ as a function of β. The logarithm of the marginal likelihood, as the integral of $E_\beta[\ln p(D \mid \theta, M)]$ over [0, 1], can then be estimated as the integral of the curve.

$$\frac{\partial^2 \ln Z_\beta}{\partial \beta^2} = V_\beta[U] > 0. \tag{38}$$

We will make further use of this identity when evaluating the error on the estimate.

The annealing method was applied on the same dataset, under the Poisson+F model and without rate variation across sites (UNI model), yielding an estimate of $\ln p(D \mid UNI) = -10,309.0$ nits. The logarithm of Bayes' factor between the two models UNI and RAS is then obtained by taking the difference between the two estimated supports:

$$\ln \frac{p(D \mid \text{RAS})}{p(D \mid \text{UNI})} = -9922.1 + 10,309.0 = 386.9, \tag{39}$$

i.e., the PGK dataset gives a support of approximately 386.9 nits in favor of RAS over UNI.

Model-Switch Integration

As is clear in the previous example, the difference between the logarithm of the marginal likelihoods of the two models can be small compared to these two values. This can lead to poor estimates, unless the precision on each marginal likelihood is very high. For this reason, rather than performing two quasistatic moves from the prior to each of the two models' posterior distributions, it might be more convenient to make a single, and shorter, path directly connecting the two models in the space of unnormalized densities. This is the rationale of the model-switch method.

Suppose that we want to compare two models M_0 and M_1 that are defined on the same parameter space Θ. Note that this does not restrict the generality of the procedure, because parameters specific of, say, M_0 can be included

in the common parameter vector, but not be involved in the computation of the likelihood according to M_1. The model-switch scheme involves a path that goes directly from model M_0 to model M_1:

$$q_\beta(\theta) = [p(D \mid \theta, M_0)p(\theta \mid M_0)]^{1-\beta}[p(D \mid \theta, M_1)p(\theta \mid M_1)]^{\beta}. \quad (40)$$

For $\beta = 0$ or 1:

$$q_0(\theta) = p(D \mid \theta, M_0)p(\theta \mid M_0), \quad (41)$$

$$Z_0 = p(D \mid M_0), \quad (42)$$

$$q_1(\theta) = p(D \mid \theta, M_1)p(\theta \mid M_1), \quad (43)$$

$$Z_1 = p(D \mid M_1). \quad (44)$$

Therefore, in the present case, performing the thermodynamic integration leads to computing the logarithm of the Bayes' factor between the two models.

Differentiating p_β with respect to β yields

$$U(\theta) = \frac{\partial \ln q_\beta(\theta)}{\partial \beta} \quad (45)$$

$$= \ln p(D \mid \theta, M_1) + \ln p(\theta \mid M_1) - \ln p(D \mid \theta, M_0) - \ln p(\theta \mid M_0). \quad (46)$$

The Bayes' factor between the two models UNI and RAS was recomputed using this alternative integration scheme, using an increment $\delta\beta = 0.001$, and every $Q = 1{,}000$. Figure 3 shows the collection of values of $U(\theta) = \ln p(D \mid \theta, \text{RAS}) - \ln p(D \mid \theta, \text{UNI})$ collected under a quasistatic transformation from UNI to RAS. In the present case, the two models have the same prior density, so that we only need to compute the difference between their log-likelihoods. In graphical terms, the logarithm of the Bayes factor is equal to the algebraic area situated between the curve and the x-axis, and as before, can be estimated by averaging over the sample (Eq. (31)). In the present case, it yields an estimate of 386.7 nits, very close to our previous estimate obtained with the annealing-melting method.

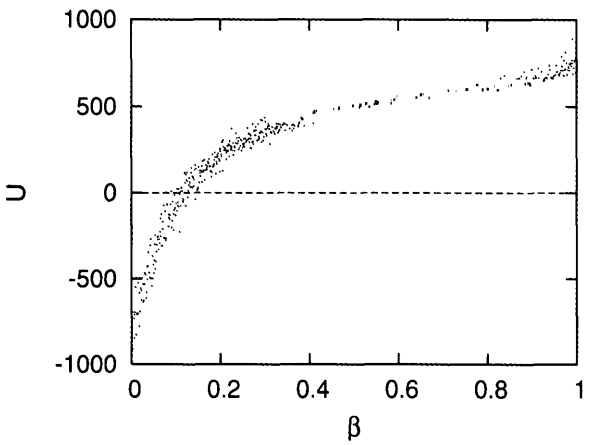

FIGURE 3. Model-switch integration between the uniform-rate (UNI) and the rate-across-site (RAS) models for the PGK dataset. The 1,000 values of $\ln p(D \mid \theta, \text{RAS}) - \ln p(D \mid \theta, \text{UNI})$ sampled during the model-switch integration are plotted against β. The logarithm of the Bayes' factor can be estimated as the algebraic area between the curve and the x-axis.

Note that, as in the case of the annealing-melting scheme, the function being integrated is monotonous (its derivative being equal to $V_\beta[\ln p(D \mid \theta, \text{RAS}) - \ln p(D \mid \theta, \text{UNI})]$).

Error Estimation

We will consider the discrete and the quasistatic procedures separately. For the discrete estimator, two main sources of errors have to be considered: the sampling error and the error induced by the discretization. First, the sampling variance is equal to

$$V[\hat{\mu}_{ds}] = \frac{1}{C^2}\left(\frac{1}{4}V[\hat{U}_0] + \sum_{d=1}^{C-1} V[\hat{U}_d] + \frac{1}{4}V[\hat{U}_C]\right), \quad (47)$$

where

$$V[\hat{U}_d] = \frac{1}{K_{eff}} \sum_{k=1}^{K} (U_\beta(\theta_k) - \hat{U}_\beta)^2 \quad (48)$$

and K_{eff} is the effective sample size. The corresponding standard error is $\sigma_s = \sqrt{V[\hat{U}_d]}$. As for the discretization error, because $E_\beta[U]$ is a monotonous function of β, the worst-case upper (resp. lower) error is given by the area between the piecewise continuous function joining the measured values of $E_\beta[U]$ and the upper (resp. lower) step function built from them. Both areas are equal to:

$$\sigma_d = \frac{|E_1[U] - E_0[U]|}{2C} \simeq \frac{|\hat{U}_1 - \hat{U}_0|}{2C}. \quad (49)$$

The total error can then be estimated as the worst-case discretization error, combined with the 95% confidence interval of the sampling error: $\sigma = \sigma_d + 1.645\sigma_s$.

Concerning the quasistatic estimator, the discretization error is again equal to

$$\sigma_d = \frac{|E_1[U] - E_0[U]|}{2K}. \quad (50)$$

Assuming independence between the successive points of the chain, the sampling variance is given by

$$V[\hat{\mu}_{qs}] = \frac{1}{K^2}\left(\frac{1}{4}V_0[U] + \sum_{k=1}^{K-1} V_{\beta_k}[U] + \frac{1}{4}V_1[U]\right) \quad (51)$$

$$= -\frac{1}{4K^2}(V_0[U] + V_1[U])$$

$$+\frac{1}{K^2}\left(\frac{1}{2}V_0[U]+\sum_{k=1}^{K-1}V_{\beta_k}[U]+\frac{1}{2}V_1[U]\right) \quad (52)$$

$$\simeq -\frac{1}{4K^2}(V_0[U]+V_1[U])+\frac{1}{K}\int_0^1 V_\beta[U]d\beta. \quad (53)$$

Using the second moment identity (Eq. (38)), the second term of Eq. (53) can be reformulated as

$$\frac{1}{K}\int_0^1 V_\beta[U]d\beta = \frac{1}{K}\int_0^1 \frac{\partial^2 \ln Z}{\partial\beta^2} \quad (54)$$

$$= \frac{1}{K}\left(\frac{\partial \ln Z_1}{\partial\beta}-\frac{\partial \ln Z_0}{\partial\beta}\right) \quad (55)$$

$$= \frac{1}{K}(E_1[U]-E_0[U]), \quad (56)$$

so that the sampling variance of the quasistatic estimate is

$$V[\hat{\mu}_{qs}] = -\frac{1}{4K^2}(V_0[U]+V_1[U]) + \frac{1}{K}(E_1[U]-E_0[U]), \quad (57)$$

and the associated standard error is equal to $\sigma_s = \sqrt{V[\mu_{qs}]}$. As in the case of the discretized estimate, the total error is $\sigma = \sigma_d + 1.645\sigma_s$.

All this is valid only if the points are truly independent draws from their respective p_β distributions. If this is not the case, then a factor $\tau = K/K_{eff}$ (i.e., the decorrelation time) has to be put in front of the right-hand side of Eq. (57), to account for the effective sample size. Here, the situation is slightly more complicated due to the fact that, as β moves from 0 to 1, the decorrelation time of the chain might change. In general, we did not observe large variations of the decorrelation time for different valus of β. In practice, we compute the decorrelation time at $\beta = 0$ (τ_0) and $\beta = 1$ (τ_1) and take the larger of the two.

In addition, because β changes continuously during sampling, the chain is never exactly at equilibrium, and this will cause a "thermic lag" of the MCMC: when sampling a value of θ at the current value of β, one is in effect sampling from $p_{\beta'}$, with β' slightly smaller than β. Because $U(\theta)$ is an increasing function of β, one expects this lag to result on average in an underestimation of μ when going from 0 to 1. In contrast, performing a quasistatic move from 1 to 0 will lead to an overestimation of μ. These under- and overestimations obtained by performing a bidirectional estimation are interesting, because they allow us to bracket the true value. Specifically, each direction yields a confidence interval of the form $[\mu-\sigma, \mu+\sigma]$. In the present article, we will always display these two intervals together, but they could as well be merged into a definitive confidence interval (i.e., the smallest interval of **R** containing them). This will account for worst case errors due to thermic lag and discretization, as well as the 95% level confidence related to the sampling error. In principle, there is less than 5% of chances that the true value lies outside. In practice, however, when the discretization error or the thermic lag dominate the sampling error, the true risk is much lower.

In summary, we propose the following overall procedure, allowing to estimate all sources of errors for the quasistatic method:

1. Equilibrate MCMC at $\beta = 0$, and obtain a first sample of K_1 points, on which to estimate $E_0[U]$, $V_0[U]$ and τ_0;
2. Perform the quasistatic sampling, as explained above, moving β progressively from 0 to 1;
3. Once $\beta = 1$, perform an additional series of K_1 steps, to evaluate $E_1[U]$, $V_1[U]$ and τ_1.

An estimate of μ is obtained at step 2 (which we call $\hat{\mu}_-$ to mean that it potentially underestimates μ because of the thermic lag). The corresponding discretization and sampling errors can be computed from steps 1 and 3, and combined into σ_-. Doing the same sampling procedure from 1 to 0 yields another estimate $\hat{\mu}_+$, with an error of σ_+. Finally, the two estimates and their respective errors can be combined together, as explained above.

To illustrate the interplay between the different sources of errors, we applied both the discrete and the quasistatic estimators to the evaluation of the Bayes' factor between the two models UNI and RAS. First, the discrete method was applied, using $C = 10$ intervals across [0, 1]. This implies running 11 chains, each of which was run for a total of 110,000 cycles, including a burn-in of 10,000 cycles, and saving 1 point every 100 cycles. The estimated decorrelation time of the chains varied between 1 and 2.6 saved points (or equivalently, between 100 and 260 cycles). The logarithm of the Bayes' factor was estimated at 379.3 nits, with a total error of 44.2 nits. Not surprisingly, the discretization error is dominant in this case ($\sigma_d = 43.0$), whereas the sampling error is small ($\sigma_s = 0.74$).

Next, we applied the bidirectional model-switch quasistatic method, under several values of Q and $\delta\beta$ (Table 1). In each case, the two separate confidence

TABLE 1. Bayes' factor between the RAS and the UNI models for the PGK dataset: precision of the model-switch estimate as a function of $\delta\beta$ and Q. For each condition, a bidirectional model-switch integration is performed and the total confidence interval is evaluated as indicated in the text. The discretization (σ_d) and sampling (σ_s) errors are reported, as well as the estimated decorrelation time of the chain (in each case, the largest value among the two directions is indicated).

$\delta\beta$	Q	UNI to RAS	RAS to UNI	σ_d	σ_s	τ
0.01	10	[306.9;368.3]	[392.7;472.6]	8.5	19.2	22
0.01	100	[372.3;405.5]	[369.6;406.0]	8.5	5.9	2.1
0.001	10	[372.4;389.6]	[379.7;400.0]	0.9	5.6	19
0.001	100	[378.8;388.1]	[383.1;391.0]	0.9	2.3	3.1
0.001	1,000	[383.7;389.7]	[381.6;387.9]	0.9	1.4	1.1
0.001	10,000	[382.6;388.9]	[381.1;387.9]	0.9	1.5	1.0

intervals are indicated, together with the estimated discretization and sampling errors, and the decorrelation times. The discretization error is much smaller than with the discrete method: it is of the same order of magnitude as the sampling error for $\delta\beta = 0.01$, and negligible (less than a natural unit of logarithm) under $\delta\beta = 0.001$. The sampling error also decreases with $\delta\beta$, as expected. As for the thermic lag, it manifests itself by the fact that the two intervals are shifted with respect to each other (except when $Q > 1,000$, but then, the observed shift is within sampling error). In all cases, the two intervals are overlapping, except when $\delta\beta = 0.001$ and $Q = 10$. Note, however, that in the latter case, the combined interval encompasses all other intervals obtained under more stringent settings. The quasistatic method can thus work under two regimes: either the thermic lag is negligible, in which case the two estimates obtained by the bidirectional method are congruent within sampling error, or it is dominant, and then, what one obtains is modulo sampling error, a bracketing of the true value.

Comparing Importance Sampling and Thermodynamic Integration

Technically, the estimation of the marginal likelihood of a model amounts to the numerical evaluation of an integral. Therefore, a simple way of validating an estimation method consists in applying it to cases where the integral can be computed in a closed form. This estimate can then be compared with the analytical value.

A Gaussian Model

We first considered a simple model, parameterized by a vector $\mathbf{x} = (x_1, x_2, \cdots, x_d)$ of dimension d. The prior on $\mathbf{x}$ is a product of independent normals on each x_i, $i = 1 \ldots d$, of mean 0 and variance 1. The likelihood is

$$L(\mathbf{x}) = \prod_{i=1}^{d} e^{-\frac{x_i^2}{2v}}, \tag{58}$$

where v is a hyperparameter. The posterior is then also a product of independent normal distributions, with mean 0 and variance $v/(1+v)$, and the log of the Bayes' factor is $d[\ln(v) - \ln(1+v)]/2$. The prior, the posterior, as well as the posterior's β-heated form, are all Gaussian, and sampling independent values of $\mathbf{x}$ from them is straightforward. The importance sampling estimators (HME, SHME, and AME) and the annealing-melting thermodynamic integration methods can therefore all be applied directly.

As shown in Table 2, for $v = 1$ and $d = 1$, all methods perform reasonably well, with a relative error less than 0.1% for samples of 10^6 points. However, still in the univariate case ($d = 1$), when the variance of the likelihood is small compared to that of the prior ($v = 0.01$), the HME is not reliable, even for large samples. More precisely, it systematically overestimates the marginal likelihood. The stabilized version does better but, in fact, even the primitive AME performs a correct estimation in this case. The thermodynamic method also yields a reliable estimate. Finally, under both high dimension ($d = 100$) and small variance ($v = 0.01$), all three importance sampling methods fail, whereas thermodynamic integration remains well-behaved. Note that the HME and SHME lead to a systematic overestimation, and the AME to an underestimation.

Evaluating the Averaged Likelihood of a Tree

In order to evaluate the reliability of these alternative methods in a phylogenetic context, one would need to find a model for which analytical integration is possible, which is in general not the case. However, there is a very common situation, where an integral (in fact, a sum), is performed analytically: the classical likelihood, evaluated at a given site, is a sum over the 20^{P-3} possible configurations specifying the amino-acid state at each internal node of the tree (for that reason, it is sometimes called the *averaged* likelihood). Denoting such a configuration by $\mathbf{s}$:

$$p(C_i \mid \theta, M) = \sum_{\mathbf{s}} p(\mathbf{s} \mid \theta, M), \tag{59}$$

where C_i stands for the ith column of the alignment. Usually, this summation is done by dynamic programming, using the well-known "pruning" algorithm (Felsenstein, 1981), but we can also perform this summation using the harmonic or thermodynamic methods, and compare the resulting estimates with the true value, obtained by pruning.

To compute the HME, we have to be able to sample values of $\mathbf{s}$, according to $p(\mathbf{s} \mid \theta, M)$, which we can do using the algorithm proposed by Nielsen (Nielsen, 2001). As for the thermodynamic integration, a straightforward generalization of this algorithm (obtained by

TABLE 2. Logarithm of the marginal likelihood of a Gaussian model, evaluated by the harmonic mean estimator (HME), its stabilized version (SHME, using $\delta = 0.1$), the prior arithmetic mean estimator (AME), and the annealing thermointegration procedure. For each setting, the mean and standard error of 10 independent estimations are displayed.

Settings and true value	Sample size	HME	SHME	AME	Thermointegration
$v = 1, d = 1$	10^3	-0.3354 ± 0.0073	-0.3472 ± 0.0064	-0.3450 ± 0.0026	-0.3447 ± 0.0019
-0.346574	10^6	-0.3460 ± 0.0003	-0.3468 ± 0.0002	-0.3465 ± 0.0001	-0.3462 ± 0.0005
$v = 0.01, d = 1$	10^3	-1.2166 ± 0.1020	-2.2363 ± 0.0465	-2.3415 ± 0.0250	-2.2656 ± 0.0835
-2.30756	10^6	-1.5475 ± 0.0787	-2.2720 ± 0.0019	-2.3066 ± 0.0008	-2.3083 ± 0.0017
$v = 0.01, d = 100$	10^3	-68.365 ± 0.9807	-69.470 ± 0.9196	-3035.9 ± 37.323	-230.933 ± 0.5476
-230.756	10^6	-77.619 ± 0.6912	-76.486 ± 0.3166	-2353.2 ± 16.918	-230.781 ± 0.0230

TABLE 3. Logarithm of the integrated likelihood of the mean posterior consensus topology, for the PGK dataset. For each setting, the mean and standard error of 10 independent estimations are displayed.

True value	Sample size	Harmonic mean	Thermointegration
−10,534.4	10^2	$-10, 302.4 \pm 4.5$	$-10, 501.6 \pm 7.4$
	10^3	$-10, 336.5 \pm 2.5$	$-10, 536.1 \pm 1.7$
	10^4	$-10, 348.5 \pm 4.1$	$-10, 534.0 \pm 0.9$
	10^5	$-10, 369.3 \pm 3.0$	$-10, 534.2 \pm 0.7$

replacing all instances of $p(\mathbf{s} \mid \theta, M)$ by $p(\mathbf{s} \mid \theta, M)^\beta$ in the computations) allows one to sample as well from the heated distribution $p(\mathbf{s} \mid \theta, M)^\beta$ for any $\beta \in [0, 1]$. We applied this rationale to the evaluation of the integrated likelihood under the PGK dataset. Specifically, we computed the likelihood of the marginal posterior consensus tree, assuming a simple model with uniform rates across sites, and a Poisson process of amino-acid replacement, with uniform stationary probability vector (Table 3). Here again, whereas the thermodynamic integration method yields estimates close to the true value, the HME is not reliable, even when samples of 10^5 independent points are used.

APPLICATION: MODELS OF AMINO-ACID REPLACEMENT

Finally, we applied the HME and thermodynamic integration to the comparison of alternative models of amino-acid replacement (Tables 4 to 6). Concerning thermodynamic integration, we used both the discretized method, with $C = 10$, and the quasistatic model-switch schemes (with $\delta\beta = 0.01$ and $Q = 100$). For the HME, we ran chains of 1,100,000 cycles, saving one point every 100 cycles. The effective sample sizes ranged from 150 to 2000.

Bayes' factors are known to be sensitive to the choice of the prior. In the present cases, we used by default exponential priors of mean 1 on the two hyperparameters tuning the mean of the prior on the branch lengths (λ) and the variance of the rate distribution (α). However, to measure the impact of the choice of the prior, we also tried two alternative sets of priors on the PGK dataset (see Data and Models). This did not fundamentally change the results (Table 4), indicating that Bayes' factors are robust to the choice of the prior on these parameters.

According to the results obtained by model-switch thermodynamic integration, for all the investigated datasets, the empirical matrix WAG is much better than Poisson (Tables 4 to 6). In addition, considering the stationary probabilities as free parameters (WAG+F) yields a better fit than fixing them to their default values (WAG), a fact that was also observed by the authors of the WAG matrix, by a likelihood ratio test (Whelan and Goldman, 2001). The general reversible model, GTR is in general better than WAG+F, except for the smaller dataset, PGK. In the case of POL, the confidence intervals obtained for the two models, GTR and WAG+F, are overlapping, and it is thus not clear which model has a better fit. Finally, in all cases, MAX is worse than all models but Poisson. In contrast, the fit of CAT is dataset dependent, as it performs better than WAG on the DLIG and EF alignments, but not on POL, nor on PGK. In the case of UVR, there is again a slight overlap between CAT and GTR.

All these Bayes' factors were computed under a fixed topology, constrained according to external criteria (see Data and Models). However, as shown for the PGK and the EF datasets, the ordering of the models is totally identical when averaging over topologies (Tables 4, 5).

In general, the discrete version of thermodynamic integration yields estimates consistent with those computed using the model-switch method (Tables 4, 5), but with a much greater uncertainty. In most cases, the confidence intervals obtained for the alternative models are widely overlapping, which makes it difficult to decide which model is best. In contrast, the estimates obtained

TABLE 4. Logarithm of the marginal likelihood of alternative amino-acid replacement models evaluated on the PGK dataset, by model-switch (MS, the two confidence intervals are reported), discrete (DS) thermointegration, and harmonic mean estimation (HME). The three alternative sets' priors over the hyperparameters ($P1$, $P2$, $P3$, see Data and Models) were considered ($P1$ is the default prior). All evaluations were performed on a predefined topology, except one bidirectional quasistatic run, performed under free topology (FT). Poisson is taken as the reference model. Highest scores are indicated in bold face.

Dataset	Poisson+F	WAG	WAG+F	GTR	MAX	CAT
PGK-MS-P1	[94;118]	[952;955]	**[969;993]**	[825;878]	[142;181]	[733;791]
	[91;115]	[952;955]	**[959;983]**	[847;898]	[140;183]	[768;817]
PGK-MS-P2	[90;114]	[951;954]	**[965;990]**	[839;899]	[149;191]	[718;765]
	[93;120]	[953;956]	**[969;995]**	[839;897]	[138;177]	[754;804]
PGK-MS-P3	[81;111]	[965;968]	**[976;1002]**	[852;901]	[136;177]	[732;777]
	[87;115]	[964;968]	**[986;1011]**	[861;915]	[140;178]	[781;826]
PGK-DS	[3;139]	[926;955]	**[828;985]**	[629;928]	[47;284]	[591;870]
PGK-HME	153	925	985	1030	917	1557
PGK-FT	[81;107]	[950;954]	**[965;991]**	[842;892]	[147;185]	[745;797]
	[92;121]	[950;954]	**[962;985]**	[835;889]	[143;182]	[765;821]

TABLE 5. Logarithm of the marginal likelihood of alternative amino-acid replacement models evaluated on the EF dataset, by model-switch (MS) discrete (DS) thermointegration and harmonic mean estimation (HME), under fixed (default) or free (FT) topology. Poisson is taken as the reference model. Highest scores are indicated in bold face.

Dataset	Poisson+F	WAG	WAG+F	GTR	MAX	CAT
EF-MS	[191;227]	[1818;1823]	[1911;1948]	[2024;2109]	[866;938]	**[2133;2220]**
	[186;228]	[1813;1820]	[1904;1940]	[1997;2081]	[868;941]	**[2184;2273]**
EF-DS	[44;264]	[1783;1822]	[1671;1951]	[1532;2178]	[645;1152]	**[1762;2408]**
EF-HME	277	1796	1957	2324	2521	3830
EF-FT	[189;229]	[1797;1805]	[1906;1945]	[1983;2071]	[845;916]	**[2097;2184]**
	[199;247]	[1815;1821]	[1892;1929]	[2001;2087]	[869;940]	**[2152;2241]**

TABLE 6. Logarithm of the marginal likelihood of alternative amino-acid replacement models evaluated on the POL, DLIG, and UVR datasets, by model-switch (MS) thermointegration and harmonic mean estimation (HME). Poisson is taken as the reference model. Highest scores are indicated in bold face.

Dataset	Poisson+F	WAG	WAG+F	GTR	MAX	CAT
POL-MS	[324;376]	[2642;2649]	[2723;2770]	**[2706;2805]**	[691;775]	[2347;2449]
	[321;367]	[2647;2655]	[2754;2787]	**[2689;2812]**	[694;776]	[2423;2519]
POL-HME	451	2680	2818	3130	2707	4477
DLIG-MS	[307;351]	[2241;2248]	[2380;2430]	[2281;2380]	[1419;1526]	**[2688;2803]**
	[321;364]	[2240;2246]	[2399;2438]	[2305;2405]	[1422;1522]	**[2748;2868]**
DLIG-HME	442	2153	2360	2571	3616	4416
UVR-MS	[238;284]	[2640;2652]	[2681;2716]	**[2668;2780]**	[1027;1125]	**[2733;2853]**
	[228;281]	[2636;2647]	[2665;2716]	**[2660;2780]**	[1019;1117]	**[2790;2909]**
UVR-HME	359	2629	2743	3053	3392	4972

by the HME are incongruent with those computed using thermodynamic integration (Tables 4 to 6), except when comparing models having a similar number of parameters (such as WAG+F and Poisson+F). When comparing models of widely differing dimensionality, however, the estimated Bayes' factors are so different that the two methods even differ in their conclusions. For instance, the HME always gives CAT as the best model, whereas thermodynamic integration sometimes favors WAG (PGK) or GTR (POL).

DISCUSSION

Reliability of Marginal Likelihood Estimators

In this work, we have applied two main alternative methods of Bayes' factor evaluation: the harmonic mean estimator (HME) and thermodynamic integration. Our comparative analysis shows a striking discrepancy between them, and comparisons with true values that can be analytically computed, in the case of the normal model (Table 2) or in the context of pruning (Table 3), indicate that this is due to a lack of reliability of the HME. At the same time, these comparisons provide a validation of our implementation of the method of thermodynamic integration.

To see why the HME is misleading, we can rely on the following intuitive reasoning. Supposing, for simplicity, that the likelihood is unimodal, the marginal likelihood is more or less the product of two factors: the likelihood reached in the high-likelihood region (the mode height) and the relative size of this region (the mode width). This latter factor, the mode width, is more precisely defined as the ratio of the size of the region under the posterior mode to the overall size of the parameter space. It acts as an Ockham factor (Jaynes, 2003), as it will be smaller for more ad hoc models, which reach a significant likelihood only under very specific values of the parameters. Note also that, in general, it will tend to be smaller for higher dimensional models.

For an estimator such as the HME to work, it has to retrieve reliable information about both the mode height and the mode width from a posterior sample. Concerning the mode height, the value of the likelihood reached at equilibrium is a good indication. As for the mode width, the only way to extract information about it is by measuring the relative frequency at which points of the sample fall inside and outside the mode. However, obtaining reliable estimates of this frequency requires that a sufficient number of points outside the mode be included in our sample. Yet, in practice, the contrast between the low and the high likelihoods is in general so great that even a posterior sample of astronomical size will be virtually confined within the mode. The estimated frequency at which the low-likelihood region is visited is then 0, which means that, in effect, the HME behaves as if the mode was occupying the entire parameter space (Ockham factor = 1), and therefore, completely underestimates the dimensional penalty.

As a result, the HME overestimates $p(D \mid M)$. Furthermore, this overestimation will be more pronounced in the case of higher dimensional models, for which the Ockham factor is smaller, which implies that the harmonic estimator will be effectively biased in favor of such models. This is, in fact, exactly what we see when comparing models of amino-acid replacement (Tables 4 to 6): whereas the HME yields a more or less correct value of the Bayes' factor between models of equivalent dimensions (i.e., Poisson versus WAG), it completely reverses the conclusions when comparing models of widely differing dimensionality, such as WAG versus CAT. This is particularly striking in the case of MAX, the most parameter-rich model, for which the error is more than fourfold on the logarithm scale.

The fact that the HME systematically overestimates the marginal likelihood may well explain a few odd results obtained recently. First, it was observed that Bayes' factors tend to support higher dimensional models in a too systematic way, to the point that it was concluded that Bayes' factors may not "strike a reasonable balance between model complexity and error in parameter estimates" (Nylander et al., 2004). Second, and more disturbingly, in simulation studies, Bayes' factors seemed to favor models more complex than the actual model used to simulate the data (Pagel and Meade, 2004). Given what we have shown above, these outcomes could also be due, not to a fundamental lack of reliability of the Bayes' factor, but instead to the systematic distortion of the HME in favor of more complex models.

Our analysis stresses the importance of using more robust and well-validated methods for Bayes' factor evaluations. Neither the HME nor its stabilized version fall into this category. We have also tested other estimators based on the importance sampling principle (Geyer, 1994; Meng and Wong, 1996), in particular the estimator proposed by Meng and Wong, which can be shown to be optimal in its category (that includes the HME and the AME). Yet none of them gave reliable results (not shown). More generally, our experience is that importance sampling estimators do not work well on large datasets.

As a general alternative, we propose to employ thermodynamic integration. This method is certainly not straightforward. It is theoretically quite involved, requires additional code-writing for sampling along paths in the space of distributions, and, furthermore, is computationally intensive. According to our experience, as a rule of thumb, thermodynamic integration will require about 10 times more CPU time than a plain posterior sampling under the more demanding among the two models being compared. In general, this means running a chain for several days, up to several weeks for models like CAT, for which mixing is more challenging. On the other hand, it seems to be more reliable. In the present work, it has led to correct estimates in the two cases in which we can compute the corresponding integral in a closed form. In addition, it has better theoretical properties (Gelman, 1998). In particular, its variance is well within control: as can be seen from our error estimation, the variance is at most quadratic in the logarithm of the

likelihood of the dataset, which is itself linear in the size of the alignment. Hence, using more complicated methods, such as thermodynamic integration, seems to be the price to pay for correctly evaluating high-dimensional integrals such as the marginal likelihood. Thus far, no other method of equivalent precision and generality is yet available, although some ideas have been proposed (Chib, 1995; Chib and Jeliazkov, 2001), which we are currently exploring.

Our comparison of alternative models of amino-acid replacement confirms and extends what we have presented previously (Lartillot and Philippe, 2004), except in one case: we previously reported that GTR was less fit than WAG+F on the EF dataset, whereas we now find that GTR is in fact better than WAG+F (Table 5). As we can check by the error analysis developed in the present article, this is due to the lack of precision of the discrete version. The quasistatic method thus appears to be much more precise than the discrete version, all the more so as the error can be controlled with great flexibility (Table 1).

The disrete and quasistatic schemes that we have introduced here are not the only possible approaches to thermointegration, however. For instance, an alternative method consists in simulating the joint distribution on (β, θ) (Gelman, 1998). This has the advantage of eliminating both the thermic lag and the discretization error. In the applications presented in this article, the thermic lag and the discretization error are not too problematic, thanks to the monotony of the integrand. However, there are many other potentially interesting paths, not all of which have this monotony property. On the other hand, simulating from the joint distribution on (β, θ) also entails some practical difficulties. In any case, the two approaches need to be compared in practice.

Otherwise, two major conclusions can be drawn from these comparisons. First, in the cases investigated here, the ordering of the models is the same, whether the topology is fixed to that obtained under the standard model (WAG+F, I+Γ), or whether it is averaged away. This confirms that model comparisons seem to be robust to the choice of the topology, as long as this topology is reasonable (Posada and Crandall, 2001). Nevertheless, we do not think that this should be considered as a generalizable rule. In particular, this might not hold anymore if uncertainty is high in the tree, or if each model strongly supports a distinct phylogeny. In such situations, it may be more reasonable to average over topologies, at least if CPU requirements are not limiting.

Second, accounting for pattern heterogeneity across sites by a mixture model results in a better fit in the majority of the cases, although, importantly, some datasets give a greater support for simpler models, like GTR or WAG+F for POL, or WAG for PGK. This could mean that the Dirichlet process requires alignments larger than those investigated here to correctly learn its parameters. Alternatively, it could be due to our approximation consisting in considering only mixtures of Poisson processes, instead of more general mixtures. A different mixture model was proposed recently in which, in contrast to CAT, the stationary probabilities are set equal across the mixture, whereas the relative exchangeability parameters of the matrix are category specific (Pagel and Meade, 2004). Obviously, a combination of the two, i.e., a general mixture of GTR matrices, should be tested.

More generally, many other models of protein evolution can be imagined, allowing for diverse kinds of heterogeneity across sites, or across lineages, which raises a problem of how to choose among all these possibilities. In this article, we have proposed a general method for this purpose. This method can be used as a guide, allowing one to progressively focus on better models of molecular evolution.

ACKNOWLEDGMENTS

We wish to thank Henner Brinkmann for making available the aligned sequences on which this work was based. We are also grateful to Nicolas Rodrigue, David Bryant, Olivier Gascuel, Thomas Lepage, and the two referees for their useful comments on the work and on the manuscript. This work was funded by the Centre National de la Recherche Scientifique, the French funding program ACI IMP-BIO "Model Phylo," and the 60th "Comission Mixte Permanente Franco-Québécoise."

REFERENCES

Aitkin, M. 1991. Posterior Bayes factors. J. R. Stat. Soc. B 53:111–142.

Altekar, G., S. Dwarkadas, J. Huelsenbeck, and F. Ronquist. 2004. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics 20:407–415.

Berger, J., and L. Pericchi. 1996. The intrinsic Bayes factor for model selection and prediction. J. Am. Stat. Assoc. 91:109–122.

Bollback, J. P. 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. 19:1171–1180.

Brinkmann, H., M. van der Giezen, Y. Zhou, G. Poncelin de Raucourt, and H. Philippe. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst. Biol. 54:743–757.

Castresana, J. 2000. Selection of conserved blocks from multiple alignment for their use in phylogenetic analysis. Mol. Biol. Evol. 17:540–552.

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

Chib, S., and I. Jeliazkov. 2001. Marginal likelihood from the Metropolis-Hastings output. J. Am. Stat. Assoc. 96:270–281.

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

Gaut, B. S., and P. O. Lewis. 1995. Success of the maximum likelihood phylogeny inference in the four taxon case. Mol. Biol. Evol. 12:152–162.

Gelman, A. 1998. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. Stat. Sci. 13:163–185.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. Bayesian data analysis. London: Chapman and Hall/CRC.

Gelman, A., X. L. Meng, and H. Stern. 1996. Posterior predicive assessment of model fitness via realised discrepancies. Stat. Sinica 6:733–807.

Geyer, C. J. 1992. Practical Markov chain Monte Carlo. Stat. Sci. 7:473–483.

Geyer, C. J. 1994. Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Technical report 568, school of statistics, University of Minnesota.

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

Han, C., and B. P. Carlin. 2000. MCMC methods for computing Bayes factors: A comparative review. Biometrika 82:711–732.

Holder, M., and P. O. Lewis. 2003. Phylogenetic estimation: Traditional and Bayesian approaches. Nat. Rev. Genet. 4:275–284.

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

Huelsenbeck, J. P., B. Larget, R. E. Miller, and F. Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51:673–688.

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

Irestedt, M., J. Fjeldsa, J. A. Nylander, and P. G. Ericson. 2004. Phylogenetic relationships of typical antbirds (Thamnophilidae) and test of incongruence based on Bayes factors. BMC Evol. Biol. 4:23.

Jaynes, E. 2003. Probability theory. The logic of science. Cambridge University Press, Cambridge, UK.

Jeffreys, H. 1935. Some tests of significance, treated by the theory of probability. Proc. Camb. Phil. Soc. 31:203–222.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. CABIOS 8:275–282.

Kass, R., and A. Raftery. 1995. Bayes factors and model uncertainty. J. Am. Stat. Assoc. 90:773–795.

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

Lartillot, N., and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.

Meng, X. L. 1994. Posterior predictive p-values. Ann. Stat. 22:1142–1160.

Meng, X. L., and W. H. Wong. 1996. Simulating ratios of normalising constants via a simple identity: A theoretical exploration. Stat. Sinica 6:831–860.

Minin, V., Z. Abdo, P. Joyce, and J. Sullivan. 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst. Biol. 52:674–683.

Neal, R. M. 2000. Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Stat. 9:249–265.

Newton, M. A., and A. E. Raftery. 1994. Approximating Bayesian inference with the weigthed likelihood bootstrap. J. R. Stat. Soc. B 56:3–48.

Nielsen, R. 2001. Mapping mutations on phylogenies. Syst. Biol. 51:729–739.

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

Ogata, Y. 1989. A Monte Carlo method for high dimensional integration. Num. Math. 55:137–157.

O'Hagan, A. 1995. Fractional Bayes factors for model comparison. J. R. Stat. Soc. B 57:99–138.

Pagel, M., and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst. Biol. 53:561–581.

Philippe, H. 1993. MUST, a computer package of management utilities for sequences and trees. Nucleic Acid Res. 21:5264–5272.

Philippe, H., N. Lartillot, and H. Brinkmann. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecysozoa, Lophotrochozoa and Protostomia. Mol. Biol. Evol. 22:1246–1253.

Posada, D., and K. Crandall. 2001. Selecting the best-fit model of nucleotide substitution. Syst. Biol. 50:580–601.

Raftery, A. E., and S. M. Lewis. 1992. [Practical Markov chain Monte Carlo]: Comment: One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. Stat. Sci. 7:493–497.

Rannala, B. 2002. Identifiability of parameters in MCMC Bayesian inference of phylogeny. Syst. Biol. 51:754–760.

Rubin, D. B. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann. Stat. 4:1151–1172.

Schwartz, G. 1978. Estimating the dimension of a model. Ann. Stat. 6:461–464.

Stefanovic, S., D. Rice, and J. Palmer. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: Amborella or monocots? BMC Evol. Biol. 4:35.

Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. B 36:111–147.

Suchard, M., C. M. R. Kitchen, J. Sinsheimer, and R. E. Weiss. 2003. Hierarchical phylogenetic models for analyzing multipartite sequence data. Syst. Biol. 52:649–664.

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

Sullivan, J., and D. L. Swofford. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J. Mammal. Evol. 4:77–86.

Thompson, J., D. Higgins, and T. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

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

Waddell, P. J., H. Kishino, and R. Ota. 2002. Very fast algorithms for evaluating the stability of ML and Bayesian phylogenetic trees from sequence data. Genome Inform. 13:82–92.

Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18:691–699.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396–1401.

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

Yang, Z. 1996. Among site variation and its impact on phylogenetic analyses. Trends Ecol. Evol. 11:367–370.

First submitted 4 March 2005; reviews returned 19 May 2005; final acceptance 16 September 2005
Associate Editor: Paul Lewis