Assessing Site-Interdependent Phylogenetic Models of Sequence Evolution

Nicolas Rodrigue, Herve Philippe, Nicolas Lartillot

Assessing Site-Interdependent Phylogenetic Models of Sequence Evolution

Nicolas Rodrigue, Hervé Philippe,* and Nicolas Lartillot†*

*Canadian Institute for Advanced Research, Département de Biochimie, Université de Montréal, Montréal, Québec, Canada; and †Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, URM 5506 CNRS-Université de Montpellier 2, Montpellier, France

In recent works, methods have been proposed for applying phylogenetic models that allow for a general interdependence between the amino acid positions of a protein. As of yet, such models have focused on site interdependencies resulting from sequence-structure compatibility constraints, using simplified structural representations in combination with a set of statistical potentials. This structural compatibility criterion is meant as a proxy for sequence fitness, and the methods developed thus far can incorporate different site-interdependent fitness proxies based on other measurements. However, no methods have been proposed for comparing and evaluating the adequacy of alternative fitness proxies in this context, or for more general comparisons with canonical models of protein evolution. In the present work, we apply Bayesian methods of model selection—based on numerical calculations of marginal likelihoods and posterior predictive checks—to evaluate models encompassing the site-interdependent framework. Our application of these methods indicates that considering site-interdependencies, as done here, leads to an improved model fit for all data sets studied. Yet, we find that the use of pairwise contact potentials alone does not suitably account for across-site rate heterogeneity or amino acid exchange propensities; for such complexities, site-independent treatments are still called for. The most favored models combine the use of statistical potentials with a suitably rich site-independent model. Altogether, the methodology employed here should allow for a more rigorous and systematic exploration of different ways of modeling explicit structural constraints, or any other site-interdependent criterion, while best exploiting the richness of previously proposed models.

Introduction

Models of molecular evolution are attempts at describing the process of residue replacement in nucleotide and amino acid sequences. They are the central issue in probabilistic phylogenetics as they have an impact on all stages of an inference. The development of models with greater realism is hoped to lead to improved phylogenetic analyses. This has been the case, for example, with models that allow the global rate of substitution to vary across the positions of an alignment (Yang 1993, 1994). Such rate heterogeneous models are designed to accommodate the different selective pressures operating at each site and have been shown to have a significantly higher statistical fit to most data sets (Yang 1996), often leading to more reasonable phylogenetic estimates (Buckley et al. 2001; Brinkmann et al. 2005). For the study of amino acid sequences, the use of empirical replacement matrices (e.g., Dayhoff et al. 1972, 1978; Jones, Taylor, and Thornton 1992b; Whelan and Goldman 2001) has provided an efficient way of including precompiled information regarding the relative exchangeability of amino acids. A panoply of other models have also been devised, usually aimed at relaxing certain assumptions of standard model formulations, such as the assumption of stationarity (e.g., Galtier and Gouy 1998) or of homogeneity in the substitution pattern across sites (e.g., Lartillot and Philippe 2004; Pagel and Meade 2004).

Although these developments undoubtedly provide more reasonable descriptions of sequence evolution, several oversimplifications persist. Most of the models currently applied continue to operate under the assumption of independence between sites. Yet, this simplification—with practical and computational justifications—is widely regarded as biologically unrealistic; the overall structure adopted by a protein, for example, must involve interactions between various amino acids of a sequence. Indeed, there has been increasing interest in incorporating explicit protein structure constraints into evolutionary models, more recently with ideas borrowed from the literature on statistical potentials (e.g., Bastolla et al. 1999; Babajide et al. 2001; Parisi and Echave 2001; Bastolla et al. 2003).

Statistical potentials are empirically derived scores, generally formulated in terms of pseudo-energy and based on observations from protein structure databases (e.g., Miyazawa and Jernigan 1985; Sippl 1990; Jones, Taylor, and Thornton 1992a). They may be viewed as coarse-grained summaries of observed amino acid interaction patterns. In an evolutionary context, Parisi and Echave (2001), for example, applied a set of statistical potentials in a simulation procedure, which proposes amino acid replacements, and discards sequences resulting in structurally divergent proteins. Fornasari et al. (2002) subsequently used this simulation procedure to construct replacement matrices incorporated into a phylogenetic context, leading to an improved model fit under several contexts (Parisi and Echave 2004, 2005).

The framework adopted in Fornasari et al. (2002) and Parisi and Echave (2004, 2005) is a computationally sensible way of devising a structurally informed model. However, there is also interest in invoking a set of statistical potentials directly within a phylogenetic framework. This is the main motivation behind the models proposed in Robinson et al. (2003) and Rodrigue et al. (2005), which incorporate explicit structural constraints within a Markovian description of the substitution process. The models combine a fitness proxy—a set of potentials considering the overall protein—to common site-independent models, either formulated in mechanistic terms (Robinson et al. 2003) or directly at the level of amino acids (Rodrigue et al. 2005). The potentials are meant to provide an estimate of the compatibility of an amino acid sequence with a given protein structure, so that differences in compatibility, before

Key words: Bayes factor, Markov chain Monte Carlo, thermodynamic integration, posterior predictive distributions, protein structure, statistical potentials.

E-mail: nicolas.rodrigue@umontreal.ca.

Mol. Biol. Evol. 23(9):1762–1775. 2006
doi:10.1093/molbev/msl041
Advance Access publication June 20, 2006

and after inferred amino acid replacement events, influence the probability of an evolutionary scenario.

Formally, these models raise several technical difficulties. For example, they require the use of data augmentation system based on substitution histories (also referred to as transition paths [Jensen and Pedersen 2000; Pedersen and Jensen 2001] or mappings [Nielsen 2002]), which include the timing and nature of each substitution event along each branch (Robinson et al. 2003), eventually leading to the sequences observed in the alignment (Rodrigue et al. 2005). In practice, this means including updates to substitution histories over the tree within Markov chain Monte Carlo (MCMC) procedures, effectively performing a numerical integration over all possible mappings. Other technical complications include computing (the ratio of) stationary probabilities, involving normalizing constants that cannot be computed analytically. Again, this complication is dealt with using MCMC procedures (see Robinson et al. [2003] and Rodrigue et al. [2005] for details). Indeed, these types of elaborate Monte Carlo sampling schemes illustrate some of the practical reasons for assuming independence between sites in the first place. Nevertheless, the methods seem reliable, and reasonably tractable for small single protein data sets (Rodrigue et al. 2005).

The numerical means of applying general site-interdependent models introduces a wide spectrum of possible model configurations; the MCMC procedures allow for a broader class of models than previously proposed methods of incorporating interdependence (e.g., Felsenstein and Churchill 1996; Jensen and Pedersen 2000; Pedersen and Jensen 2001; Arndt et al. 2002; Siepel and Haussler 2004) because the substitution process is effectively defined in the space of sequences. In other words, invoking some sequence fitness criterion could—in theory—accommodate a total interdependence across all sites. An ideal perspective would include full knowledge of the posited fitness landscape of the sequences under study, forming the basis of all evolutionary inferences. In practice, however, it follows that some proxies for sequence fitness may be better suited than others, and that their application may produce different results depending on the specifications of the formally site-independent components of the model. This raises the question of choosing the most relevant combination for a particular data set.

In the Bayesian paradigm, model evaluation strategies can be categorized along 2 broad axes. The first is used to compare the fit of alternative models, often achieved by computing the Bayes factor (Jeffreys 1935; Kass and Raftery 1995). The second, known as posterior predictive checking (Rubin 1984; Gelman et al. 1996), is mainly used as a diagnostic, characterizing discrepancies between features of true data and data simulated under the model of interest. Both strategies have become widely used for the study of phylogenetic models (reviewed in Sullivan and Joyce [2005]).

In the present work, we explore these model evaluation strategies within the site-interdependent framework. From a technical standpoint, posterior predictive checks require nothing more than simulations under the site-interdependent model. The calculation of the relative fit of different models, however, requires more elaborate methods because the models do not allow for a closed form computation of the likelihood. Indeed, in previous studies, the importance of explicit site-interdependent structural considerations was assessed based on the plausibility of associated parameter estimates (Robinson et al. 2003; Rodrigue et al. 2005). Such model assessments remain qualitative; they do not allow for selection between alternative fitness proxies, or even for a quantified comparison against site-independent models. The model comparison framework proposed by Fornasari et al. (2002), on the other hand, can not be applied without waiving one of the original motivations of the models: introducing explicit interdependencies across the positions of a protein.

Here, we propose the use of a numerical technique for the evaluation of Bayes factors, yielding quantitative model comparisons under the fully site-interdependent framework originally proposed by Robinson et al. (2003). Commonly known under the names of thermodynamic integration, path sampling, or Ogata's method, the technique has been used extensively in statistical physics for evaluating (the ratio of) partition functions (for instructive reviews, see Neal [1993] and Gelman [1998]) and more recently for the study of phylogenetic models (Lartillot and Philippe 2004, 2006). We derive an adaptation of the method, which, in combination with previously proposed techniques (Lartillot and Philippe 2006), can provide an overall ranking of models, with or without site-interdependent criteria.

We have implemented these model assessment strategies and applied them on real protein data sets, comparing the relevance of 2 sets of statistical potentials (Miyazawa and Jernigan 1985; Bastolla et al. 2001), combined with several different and well known types of models of amino acid sequence evolution. By contrasting different model configurations, we have evaluated the relative contribution of each component to the overall model fit.

Our findings indicate that considering site interdependence due to tertiary structure using statistical potentials always improves the fit of the model for all 3 studied data sets. Yet, the assessment strategies also show that using pairwise contact potentials alone is unsatisfactory. In particular, the statistical potentials we have tested do not suitably account for differences in amino acid exchange propensities or heterogeneous rates of substitution across the sites of an alignment. For such features, the modeling strategies developed under the assumption of independence are still far more appropriate. One pragmatic alternative, which we previously suggested on intuitive grounds, is to layer the use of statistical potentials to complement a suitable site-independent model (Rodrigue et al. 2005). Indeed, for the range of model configurations we have assessed and for all data sets studied, we find that the models receive the highest support when combining an empirical amino acid replacement matrix (Jones, Taylor, and Thornton 1992b), an explicit treatment of variable rates across sites (Yang 1993), and the statistical potentials of Bastolla et al. (2001).

Materials and Methods

Data Sets

We used the following 3 data sets, referred to using a shorthand indicating the protein type, the number of sequences, and their length:

- *FBP20-363*: 20 amino acid sequences of vertebrate fructose bisphosphate aldolase;
- *PPK10-158*: 10 amino acid sequences of bacterial 6-hydroxymethyl-7-8-dihydroxypterin pyrophosphokinase;
- *MYO60-153*: 60 amino acid sequences of mammalian myoglobin.

A complete listing of sequences included in these data sets (with accession numbers) is provided in the supplementary material online, and the alignments are available upon request.

In all the analyses included here, the tree topology is assumed to be known; it was constrained to the maximum likelihood topology, obtained using PhyML (Guindon and Gascuel 2003) under a JTT+F+Γ model.

We apply a simple protein structure representation based on a contact map (see below). The contact map is derived from a reference structure determined by X-ray crystallography for one of the sequences included in the data set (Protein Data Bank accession numbers 1ALD, 1HKA, and 1MBD for *FBP20-363*, *PPK10-158*, and *MYO60-153*, respectively).

Notation

Data sets (D) consist of alignments of P amino acid sequences of length N, assumed related according to a particular phylogenetic tree. The tree is rooted arbitrarily as all models considered here are reversible. We use i to index positions of a sequence $s \in \mathbf{s}$, where $s = (s_i)_{1 \leq i \leq N}$ and $\mathbf{s}$ is the set of all sequences of length N (i.e., $\mathbf{s}$ has 20^N elements). Also, let j specify the nodes, with a node having the same index as the branch leading to it, with the exception of the root node, which has an index 0 ($0 \leq j \leq 2P - 3$). We specify the sequence at node j as s_j (with s_0 being the sequence at the root node) and a particular amino acid state at position i in this sequence as s_{ij}. We write the set of branch lengths as $\lambda = (\lambda_j)_{1 \leq j \leq 2P-3}$ and the set of branch-specific substitution mappings as $\omega = (\omega_j)_{1 \leq j \leq 2P-3}$. The total number of substitutions along a branch is written as z_j ($z_j \geq 0$). We index substitution events as k ($k \leq z_j$) and refer to the time of an event on branch j as t_{jk}. Substitution events alter a single site of the sequence, at position σ_{jk}. When specifying the series of substitution events occurring on a branch j, let s_{jk-1} and s_{jk} represent the sequence states before and after substitution event k. Note that when $k = 1$, we let $s_{jk-1} = s_{j_{up}}$, where j_{up} is the immediate ancestral node of j. Finally, when $k = z_j$, we let $s_{jk} = s_j$.

Statistical Potentials as Sequence Fitness Proxies

Statistical potentials are formulated to associate a pseudo-energy to the different body interactions in a protein tertiary structure. For the potentials used here, interactions are defined between each possible pair of amino acids, with the associated pseudoenergies written as $\varepsilon = (\varepsilon_{lm})_{1 \leq l,\, m \leq 20}$. The protein structure is represented by a contact map $c = (c_{ii'})_{1 \leq i < i' \leq N}$, with elements

$$c_{ii'} = \begin{cases} 1, & \text{if amino acids at sites } i \text{ and } i' \text{ are in contact,} \\ 0, & \text{otherwise, or if } |i - i'| \leq \phi, \end{cases} \tag{1}$$

where ϕ is a threshold, below which contacts due to sequential proximity are ignored. Bastolla et al. (2001) define a contact as 2 amino acids with any heavy atoms (atoms other than hydrogen) within 4.5 Å, whereas Miyazawa and Jernigan (1985) consider side-chain centers within 6.5 Å. Also note that Bastolla et al. (2001) use a threshold of $\phi = 2$, whereas Miyazawa and Jernigan (1985) ignore contacts between immediate neighbors in the sequence ($\phi = 1$).

Given the contact map, the pseudo-energy of a sequence is calculated as:

$$E_s = \sum_{1 \leq i < i' \leq N} c_{ii'} \varepsilon_{s_i s_{i'}}. \tag{2}$$

As explained in Rodrigue et al. (2005), we impose the same protein structure over the tree by applying the same contact map to all sequences considered throughout the inference.

Evolutionary Models

Standard models consider the amino acid states at the positions of an alignment as the realization of a set of independent Markov substitution processes—one for each site—running along the branches of the tree. These processes are described by a rate matrix $Q = [Q_{lm}]$, specifying, in this case, the instantaneous rate of substitution from one amino acid state to another. Rate matrices are typically comprised of 2 sets of parameters: amino acid equilibrium frequencies, $\pi = (\pi_m)_{1 \leq m \leq 20}$, with $\sum_{m=1}^{20} \pi_m = 1$, and exchangeability parameters, $\rho = (\rho_{lm})_{1 \leq l,\, m \leq 20}$ such that

$$Q_{lm} = \frac{1}{Z_Q} \rho_{lm} \pi_m, \, l \neq m, \tag{3}$$

$$Q_{ll} = -\sum_{m \neq l} Q_{lm}, \tag{4}$$

where Z_Q is a normalizing factor such that branch lengths represent the expected number of substitutions per site:

$$Z_Q = 2 \times \sum_{1 \leq l < m \leq 20} \rho_{lm} \pi_l \pi_m. \tag{5}$$

Various combinations of these parameters are possible. In the simplest case, both equilibrium frequencies and exchangeability parameters are fixed to uniform values, rendering all types of substitutions equally probable (referred to as Poisson). More typically, however, equilibrium frequencies and exchangeability parameters are fixed to empirically derived values, such as those of Jones, Taylor, and Thornton (1992b) (written as JTT). Other alternatives might consider equilibrium frequencies as free parameters (designated as +F) or both equilibrium frequencies and exchangeability parameters as free (indicated as GTR).

The independent (20×20) Markov processes operating at each site of the protein can equivalently be considered

as a single Markov process, whose state space is now the set of all sequences of length N. There are 20^N such sequences, and thus, the matrix of this Markov process will be a $20^N \times 20^N$ matrix R:

$$R_{ss'} = \begin{cases} 0, & \text{if } s \text{ and } s' \text{ differ at more than one position,} \\ Q_{lm}, & \text{if } s \text{ and } s' \text{ differ only at site } i,\ s_i = l \text{ and } s'_i = m, \\ -\sum_{s' \neq s} R_{ss'}, & \text{if } s \text{ and } s' \text{ are identical.} \end{cases} \quad (6)$$

With the formulation of equation (6), it is possible to introduce a site-interdependent criterion: the pseudo-energy before and after an amino acid substitution. The new matrix R is then

$$R_{ss'} = \begin{cases} 0, & \text{if } s \text{ and } s' \text{ differ at more than one position,} \\ Q_{lm} e^{\beta(E_s - E_{s'})}, & \text{if } s \text{ and } s' \text{ differ only at site } i,\ s_i = l \text{ and } s'_i = m, \\ -\sum_{s' \neq s} R_{ss'}, & \text{if } s \text{ and } s' \text{ are identical,} \end{cases} \quad (7)$$

where β is a new parameter, effectively weighting the pseudo-energy difference's impact on the rate of substitution. When $\beta = 0$, the model simplifies to the usual site-independent model specified in equation (6). However, when $\beta \neq 0$, the substitution process can no longer be decomposed into a set of N independent processes because the pseudo-energy measure considers the entire amino acid sequence (see the supplementary material online for issues of scaling the rate matrix R as well as the alternative of monitoring branch lengths in terms of the actual number of substitution per site induced by the model, also explained by Rodrigue et al. [2005)]). To indicate the model with statistical potentials ($\beta \neq 0$), we use the suffix +MJ, when using the potentials of Miyazawa and Jernigan (1985), and +BAS, when using those of Bastolla et al. (2001).

Finally, sites of an alignment may have uniform rates or variable (gamma distributed) rates across sites $r = (r_i)_{1 \leq i \leq N}$; (designated as $+\Gamma$). In this case, the rate of substitution from s to s' having a single amino acid difference at site i becomes $R_{ss'} \times r_i$.

Because the model combines statistical potentials with the common site-independent parameterizations, we have referred to this approach as a layering strategy (Rodrigue et al. 2005), aimed at utilizing the most of available information for capturing features of the amino acid replacement process.

Priors

We used the following priors:

- $\lambda \sim$ *Exponential*, with a mean determined by a hyperparameter μ, itself endowed with an exponential prior of mean 1;
- $r \sim$ *Gamma*, with a "shape" hyperparameter α, in turn endowed with an exponential prior of mean 1;
- $\rho \sim$ *Dirichlet*(1, 1, …, 1);
- $\pi \sim$ *Dirichlet*(1, 1, …, 1);
- $\beta \sim$ *Uniform*[$-\beta_{max}$, β_{max}], where, unless stated otherwise, $\beta_{max} = 5$.

MCMC Sampling

Conventional models generally invoke pruning-based likelihood calculations (Felsenstein 1981) and compute a finite-time transition probability matrix $P(\nu) = [P_{lm}(\nu)]$ by rate matrix exponentiation: $P(\nu) = e^{\nu Q}$, giving, in this case, the probability of amino acid l substituting to m, over an evolutionary distance ν. Having computed $P(\nu)$ for all evolutionary distances involved in a tree, the likelihood is calculated by summing transition probabilities for all possible internal node state configurations. Here, given the order of R ($20^N \times 20^N$), an equivalent calculation is not tractable. As an alternative, Robinson et al. (2003) proposed a data augmentation framework based on substitution mappings. For ease of notation, we group all components of the model into a vector θ, i.e., $\theta = \{\pi, \rho, \beta, \lambda, \mu, r, \alpha\}$—when fixing certain elements of the model, the hypothesis vector is obviously reduced in accordance. Given a hypothesis vector $\theta \in \Theta$ under model M, the probability of going from a given sequence to another over branch j, and through a specific substitution history ω_j, can be calculated as

$$p(s_j, \omega_j | s_{j_{up}}, \theta, M) = \left(\prod_{k=1}^{z_j} R_{s_{jk-1} s_{jk}} r_{\sigma_{jk}} e^{-(t_{jk} - t_{jk-1})\Upsilon(s_{jk-1})} \right) \times e^{-(\lambda_j - t_{j z_j})\Upsilon(s_{j z_j})}, \quad (8)$$

where $\Upsilon(s_{jk-1}) = \sum_{i=1}^{N} \sum_{s'_i} R_{s_{jk-1} s'_i} r_i$ represents the "rate away" from sequence s_{jk-1}, with the inner sum being over the 19 sequence states that differ with s_{jk-1} at position i.

The likelihood computations also require the probability of the sequence at the root of the tree:

$$p(s_0 | \theta, M) = \frac{1}{Z_0} e^{-2\beta E_{s_0}} \prod_{i=1}^{N} \pi_{s_{i0}}, \quad (9)$$

with Z_0 being the associated partition function

$$Z_0 = \sum_{\mathbf{s}} e^{-2\beta E_s} \prod_{i=1}^{N} \pi_{s_i}. \quad (10)$$

Assuming lineages evolve independently, we compute the product of equation (8) over all branches, along with the probability in equation (9), yielding the overall likelihood function:

$$p(D, \omega | \theta, M) = p(s_0 | \theta, M) \prod_{j=1}^{2P-3} p(s_j, \omega_j | s_{j_{up}}, \theta, M). \quad (11)$$

The likelihood function (11) is combined with the joint prior probability density $p(\theta \mid M)$ to obtain the posterior probability density $p(\omega, \theta \mid D, M)$ given by Bayes' theorem:

$$p(\omega, \theta | D, M) = \frac{p(D, \omega | \theta, M) p(\theta | M)}{p(D | M)}. \quad (12)$$

Our MCMC procedure consists of using the Metropolis–Hastings algorithm (Metropolis et al. 1953; Hastings 1970) to define a Markov chain with the probability in equation (12) as its stationary distribution; assuming a

current state (ω, θ), an update to a new state (ω', θ') is proposed and accepted with a probability ϑ:

$$\vartheta = \min\left(1, \frac{p(\omega', \theta'|D, M)}{p(\omega, \theta|D, M)} \times \mathcal{H}\right), \quad (13)$$

where $\mathcal{H}$ is the Hastings ratio (of proposal densities).

We used the update operators described by Rodrigue et al. (2005), with minor variations outlined in the supplementary material online, along with more detailed descriptions of MCMC settings used in this work.

Computing Bayes Factors

The denominator of Bayes theorem ($p(D \mid M)$ in eq. 12) is a normalizing constant, such that the posterior probability integrates to 1:

$$p(D|M) = \int_{\Theta}\int_{\Omega} p(D, \omega|\theta, M)p(\theta|M)\mathrm{d}\omega\mathrm{d}\theta. \quad (14)$$

This normalizing constant, also called the integrated or marginal likelihood, is not of concern when working under a fixed model because it cancels out in the Metropolis–Hastings algorithm. When interested in comparing 2 different models, however, the marginal likelihood has an intuitively appealing meaning: expressed as a function of M, it can be viewed as the likelihood of model M given the data D. In the Bayesian framework, the model of highest likelihood would be favored, typically by evaluating the Bayes factor (B_{01}) between M_0 and M_1 (Jeffreys 1935; Kass and Raftery 1995):

$$B_{01} = \frac{p(D|M_1)}{p(D|M_0)}. \quad (15)$$

In principle, a Bayes factor greater (smaller) than 1 is considered as evidence in favor of model M_1 (M_0).

Besides the intuitive interpretation of model likelihoods, comparing models based on equation (15) is attractive for several reasons: the models compared need not be nested, or even rely on the same parametric rationale; the Bayes factor implicitly penalizes for parameter-rich models, thus also allowing for the selection of an appropriate dimensionality. In practice, evaluating Bayes factors for high dimensional models is a computationally demanding task, involving difficult integrals, as shown in equation (14). Here, we rely on the numerical technique of thermodynamic integration (Ogata 1989; Gelman 1998) adapted from the methods described by Lartillot and Philippe (2006). In the present application, the method rests in defining a continuous path connecting a standard site-independent model with the model including the sequence fitness proxy, i.e., the set of statistical potentials. To do so, we make use of the fact that when $\beta = 0$, the site-interdependent model collapses to the usual site-independent model. As shown in the Appendix, for a given value of β, the derivative of the logarithm of the marginal likelihood with respect to β gives:

$$\frac{\partial \ln p(D|\beta)}{\partial \beta} = \left\langle \frac{\partial \ln p(D, \omega|\beta, \theta)}{\partial \beta} \right\rangle, \quad (16)$$

where $\langle\cdot\rangle$ represents an expectation with respect to the posterior distribution (we henceforth omit the dependence on M from the notation, considering it as implicit). Based on a sample $(\theta_h, \omega_h)_{1\le h\le K}$, obtained via the Metropolis–Hastings algorithm, expectations over the posterior probability distribution can be estimated for any value of β using the standard Monte Carlo relation:

$$\left\langle \frac{\partial \ln p(D, \omega|\beta, \theta)}{\partial \beta} \right\rangle \simeq \frac{1}{K}\sum_{h=1}^{K} \frac{\partial \ln p(D, \omega_h|\beta, \theta_h)}{\partial \beta}. \quad (17)$$

Our quasi-static procedure then consists of sampling along a path linking the standard site-independent model, $\beta = 0$, to some arbitrary point, $\beta = x$, by slowly incrementing β by a small value $\delta\beta$ after a set of MCMC cycles. We write such a sample as $(\beta_h, \theta_h, \omega_h)_{0\le h\le K}$, where $\beta_0 = 0$, $\beta_K = x$, and $\forall h$, $0 \le h < K$, $\beta_{h+1} - \beta_h = \delta\beta$. Integrating over the interval $[0, x]$ can then be estimated:

$$\ln\frac{p(D|\beta_K)}{p(D|\beta_0)} = \int_0^x \frac{\partial \ln p(D|\beta)}{\partial \beta}\mathrm{d}\beta \quad (18)$$

$$= \int_0^x \left\langle \frac{\partial \ln p(D, \omega|\beta, \theta)}{\partial \beta} \right\rangle \mathrm{d}\beta \quad (19)$$

$$\simeq x \times \frac{1}{K}\left[\frac{1}{2}\left(\frac{\partial \ln p(D, \omega_0|\beta_0, \theta_0)}{\partial \beta} + \frac{\partial \ln p(D, \omega_K|\beta_K, \theta_K)}{\partial \beta}\right) + \sum_{h=1}^{K-1} \frac{\partial \ln p(D, \omega_h|\beta_h, \theta_h)}{\partial \beta}\right]. \quad (20)$$

Equation (20) provides an estimate of the logarithm of the Bayes factor for the model including statistical potentials, with $\beta = x$, over the site-independent model, $\beta = 0$. The value of x is arbitrary. However, with this procedure, we can monitor the Bayes factor anywhere we choose along the dimension of β. Also note that, using the same sample, $\ln p(D \mid \beta_{K'}) - \ln p(D \mid \beta_0)$ can be computed for any value K' ($0 \le K' \le K$). In other words, the curve of the log marginal likelihood along β can be estimated (fig. 1). In practice, because the high-likelihood region is restricted to a very small proportion of the admissible values of β, the integration procedure can be constrained to a small and specific interval; one can consider that outside this specific interval, the marginal likelihood given β is $\sim$0. Thus, exponentiating and integrating this curve yields the overall Bayes factor between the model with statistical potentials (M_1) against the model assuming independence (M_0), with the Monte Carlo estimate derived as

$$B_{01} = \frac{\int p(D|\beta)p(\beta)\mathrm{d}\beta}{p(D|\beta_0)} \quad (21)$$

$$= \int \frac{p(D|\beta)}{p(D|\beta_0)}p(\beta)\mathrm{d}\beta \quad (22)$$

$$\simeq \sum_{h=1}^{K} \frac{p(D|\beta_h)}{p(D|\beta_0)} \times \frac{\delta\beta}{I}, \quad (23)$$

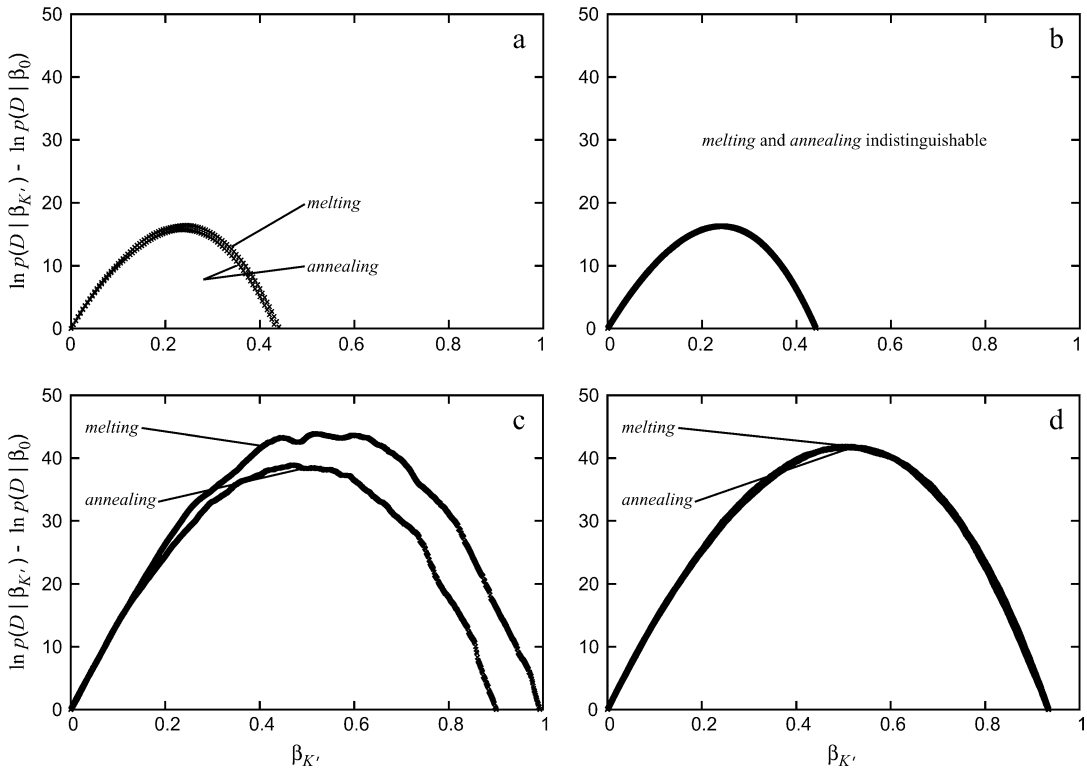

FIG. 1.—Bidirectional integrations along β for JTT+BAS (*a* and *b*) and JTT+F+BAS (*b* and *d*) performed with "fast" (*a* and *c*) and "slow" (*b* and *d*) settings using the *MYO60-153* data set. The trace plots illustrate the empirical tuning of the thermodynamic MCMC sampling, which is more challenging for the model with greater degrees of freedom (bottom).

where I is the interval size of the uniform prior on β and, hence, $\delta\beta/I$ is the density of the prior contained between each successive δβ step of the quasi-static procedure.

The analogy with thermodynamics here is that the inverse of β can be thought of as a "site-interdependence temperature," with $\beta = 0$ effectively "melting" out all structural information. Alternatively, when $\beta > 0$, the models can be said to be "annealed" into site interdependence. From this perspective, plain MCMC runs are in fact sampling the appropriate temperature for the particular sequence fitness proxy.

We also use this analogy in referring to our tuning of the thermodynamic integration, which we explore by applying the procedure in different directions. Specifically, annealing integrations work by first equilibrating a MCMC with $\beta = 0$, followed by a slow and progressive increase to $\beta = x$. If the value of β is increased too quickly, the MCMC run will not have sufficient time to equilibrate, always dragging behind configurations from preceding cycles with each increment of β. Conversely, melting integrations work by equilibrating a MCMC at $\beta = x$ and slowing decreasing to $\beta = 0$. Performing a bidirectional check, i.e., both annealing and melting integrations, forms the basis of our empirical exploration of the MCMC settings needed for refining the estimation procedure (fig. 1).

Obviously, obtaining precise integrations is computationally more challenging when applying the statistical potentials to models with greater degrees of freedom. For example, using *MYO60-153*, figure 1*a* shows that the annealing and melting integrations, applied under JTT+BAS, are very similar for fast runs ($\delta\beta = 0.005$ and $K = 100$) requiring about 2 h of CPU time on a Xeon 2.4 GHz desktop computer. Slower runs ($\delta\beta = 0.0001$ and $K = 5{,}000$) requiring about 2 days of CPU are essentially indistinguishable (fig. 1*b*). When applying the integration under JTT+F+BAS, however, a clear discrepancy is observed between fast (approximately 30 h, $\delta\beta = 0.001$ and $K = 1{,}000$) annealing and melting runs (fig. 1*c*). Nevertheless, by tuning the call frequency of the various Monte Carlo operators, the step size of the quasi-static scheme, and the number of cycles between each increment, the integration settings can be adjusted ($\delta\beta = 0.0005$ and $K = 20{,}000$) to obtain precise Bayes factor estimates within about 15 days (fig. 1*d*).

Our integration scheme along β allows us to compute the Bayes factor between a site-interdependent model and its site-independent counterpart. We also need to compute Bayes factors between site-independent models, which we do using the model-switch integration method described by Lartillot and Philippe (2006). For example, in assessing the model GTR+BAS, we first perform the integration along β, giving the log Bayes factor of GTR+BAS against GTR. Then, applying the model-switch method, we compute the log Bayes factor between GTR and POISSON. With both estimates at hand, we calculate the log Bayes factor of GTR + BAS against POISSON, simply using the additive quality of logarithms:

$$\ln\frac{p(D|\text{GTR+BAS})}{p(D|\,\text{POISSON})} = \ln\frac{p(D|\text{GTR+BAS})}{p(D|\text{GTR})} + \ln\frac{p(D|\text{GTR})}{p(D|\,\text{POISSON})}. \quad (24)$$

Table 1
Natural Logarithm of the Bayes Factor for All Models Studied, with POISSON Used as a Reference

Model	*FBP20-363*	*PPK10-158*	*MYO60-153*
POISSON	0	0	0
POISSON + BAS	10	16	24
POISSON + MJ	6	7	18
POISSON + F	103	34	70
POISSON + F + BAS	158	78	142
POISSON + F + MJ	144	65	129
POISSON + Γ	135	53	138
POISSON + Γ + BAS	138	69	162
POISSON + Γ + MJ	137	58	156
POISSON + F + Γ	238	89	207
POISSON + F + Γ + BAS	296	139	280
POISSON + F + Γ + MJ	285	122	267
JTT	380	144	368
JTT + BAS	391	155	382
JTT + MJ	386	150	379
JTT + F	365	137	389
JTT + F + BAS	397	159	427
JTT + F + MJ	389	145	417
JTT + Γ	*529*	*195*	499
JTT + Γ + BAS	540	206	512
JTT + Γ + MJ	535	200	508
JTT + F + Γ	513	186	*513*
JTT + F + Γ + BAS	**546**	**216**	**551**
JTT + F + Γ + MJ	539	203	537
GTR	310	102	347
GTR + BAS	346	139	394
GTR + MJ	338	121	383
GTR + Γ	434	147	466
GTR + Γ + BAS	471	185	512
GTR + Γ + MJ	462	168	501

NOTE.—The best site-independent models for each data set are emphasized in italics, whereas the best overall models are emphasized in bold.

In this way, it is possible to observe the overall ranking of models for a given data set, by having all Bayes factors against the simplest POISSON model. Note that the error of the integration procedures is cumulative in equation (24); for succinct comparisons of models, we report the mean of the highest and lowest values obtained using bidirectional checks (table 1). For the simpler models, the error can be reduced to less than one natural log unit, whereas the more challenging models can lead to an error approximately ± 4. The actual highest and lowest values obtained are reported in the supplementary material online.

The following protocol summarizes:

- for a particular model setting, run a quasi-static thermodynamic integration, estimating the log marginal likelihood curve along β (applying the Monte Carlo estimate given by eq. 20);
- exponentiate and integrate the resulting curve to estimate the overall Bayes factor between the site-interdependent model and the underlying site-independent model (applying the Monte Carlo estimate given by eq. 23);
- given the marginal likelihood comparisons between site-independent models, estimated using the model-switch scheme described by Lartillot and Philippe (2006), compute all Bayes factor with respect to POISSON (applying relations analogous to eq. 24).

Posterior Predictive Resampling

The sampling techniques used here are particularly well suited to performing posterior predictive checks, as described by Nielsen (2002) (also see Bollback [2005]). A posterior predictive scheme is based on a simulation procedure, which consists of drawing a sequence from the stationary probability written in equation (9) under a given $\theta \in \Theta$, and simulating a substitution mapping on the branches of the tree to generate a replication of the data—in other words, these mappings are unconstrained to any states at the leaves of the tree (Nielsen 2002). The simulation procedure is repeated on each successive parameter values of the initial MCMC sampling performed on the true data.

Given a statistic of interest, posterior predictive checks then consist in comparing the value of the statistic observed on the data with the distribution obtained on the replicates; a discrepancy indicates that the model does not adequately account for the phenomena summarized by the statistic. Here, our statistics are not exactly computed on the data but on mappings sampled from their posterior distribution. We refer to the substitution histories obtained from simulations as predictive mappings, in contrast with what we call the "observed" mappings, which are conditioned on the true data. Note, of course, that these latter mappings are not actually observed but rather constitute the data augmentation step of the MCMC methods.

To explore whether a model can explain the level of rate heterogeneity of a given data set, we compared the variance in number of substitutions across sites, calculated based on the number of substitutions counted at each site in predictive and observed mappings. This particular statistic was used by Nielsen (2002) as an example demonstrating the utility of a mapping-based framework.

Also, in order to observe how well a model captures amino acid exchange propensities, we counted each of the 190 possible types of exchange in mappings to generate what we refer to as the residue exchange distribution. We then computed the Euclidean distance between predictive and observed exchange distributions for each sample point from the posterior distribution.

Results and Discussion

Bayes Factors

We applied the thermodynamic integration procedures to all data sets and for all model combinations described herein. The resulting Bayes factors, computed against the simplest model (POISSON), are reported in table 1.

Overall Fit of Site-Independent Models

The most favored site-independent model is JTT+Γ for *FBP20-363* and *PPK10-158* and JTT+F+Γ for *MYO60-153*. This is somewhat expected. The POISSON-based models are obviously unrealistic because the exchangeability between amino acids is clearly not uniform, hence giving support to JTT-based models. Also, allowing for rate heterogeneity is known to nearly always improve the model fit (Yang 1996; Buckley et al. 2001; Posada and Buckley 2004), as is the case here. The equilibrium frequencies of JTT appear to be suitable for the 2 smaller data

Table 2
Mean Posterior Values of β under All Model Combinations Described in the Text

Model	*FBP20-363*	*PPK10-158*	*MYO60-153*
POISSON + BAS	0.107	0.249	0.268
POISSON + MJ	0.0074	0.0279	0.0423
POISSON + F + BAS	0.402	0.462	0.637
POISSON + F + MJ	0.0658	0.0724	0.1086
POISSON + Γ + BAS	0.0989	0.268	0.239
POISSON + Γ + MJ	0.0058	0.0296	0.0397
POISSON + F + Γ + BAS	0.439	0.564	0.717
POISSON + F + Γ + MJ	0.0811	0.0983	0.1406
JTT + BAS	0.176	0.264	0.240
JTT + MJ	0.0231	0.0368	0.0423
JTT + F + BAS	0.305	0.378	0.501
JTT + F + MJ	0.0449	0.0722	0.0816
JTT + Γ + BAS	0.177	0.277	0.244
JTT + Γ + MJ	0.0234	0.0391	0.0424
JTT + F + Γ + BAS	0.333	0.478	0.575
JTT + F + Γ + MJ	0.0541	0.0724	0.0975
GTR + BAS	0.433	0.511	0.625
GTR + MJ	0.0680	0.0777	0.1148
GTR + Γ + BAS	0.440	0.546	0.679
GTR + Γ + MJ	0.0791	0.0929	0.1228

NOTE.—95% credibility intervals are given in the supplementary material online.

sets, in as much as the dimensionality penalty renders a specific adjustment of these parameters unreliable. For *MYO60-153*, however, such a data set–specific adjustment of equilibrium frequencies seems worthwhile. The GTR matrix is always rejected over the JTT-based models, most likely because the data sets considered are too small to reliably infer the 189 additional free parameters introduced by this model. Note, however, that the GTR-based models are still far better than POISSON-based models.

Overall Fit of Site-Interdependent Models

Models including statistical potentials are always favored over their site-independent counterparts, under all configurations explored here. This being the case for all 3 proteins studied suggests that such an improvement in fit is general. Nevertheless, the improved fit observed when including statistical potentials is mild when compared with the overall fit of rich site-independent models. Specifically, the use of an empirical amino acid replacement matrix and a gamma distributed rates model both outperform the sole use of statistical potentials.

Interplay between Model Configurations

Interestingly, the relative improvement brought about by the potentials is very much a function of the site-independent components of the models. In particular, the amelioration in model fit when applying statistical potentials, as well as the equilibrium value of β under plain MCMC sampling (table 2), is noticeably lower when the π-vector is fixed, which is the case irrespective of the other site-independent settings. This is perhaps best understood by observing the stationary probability distribution written in equation (9). Whereas the stationary distribution is given by π under the standard notation of continuous-time Markov chains, under the site-interdependent models studied here, it is given by a combination of π and the exponentiated pseudo-energy factor. This forces a reinterpretation of the usual meaning given to π: rather than representing the amino acid equilibrium frequencies, these parameters should be viewed as "chemical potentials" associated to each residue, and whose effect is combined to that of the statistical potentials in the final amino acid equilibrium frequencies (Rodrigue et al. 2005). From this perspective—related to "random energy" approximations (Shakhnovich and Gutin 1993; Sun et al. 1995; Seno et al. 1998)—fixing the values of π, to uniform values (in the case of POISSON) or to the JTT values, effectively prevents the model from compensating for the coupling to the exponentiated pseudo-energy factor and thus leads to a low support for the site-interdependent models. Indeed, although the +F settings were rejected in favor of JTT for *FBP20-363* and *PPK10-158* under site-independence, when invoking the statistical potentials, this increased parameterization seems favored.

Also of interest, we find that the relative improvement brought about by the potentials is more important when using POISSON-based models than when using a JTT-based models. This is consistent with the fact that the JTT matrix inherently accounts for protein structure features, by assigning greater exchange propensities between amino acids sharing various physico–chemical properties. In other words, explicitly accounting for site interdependencies due to tertiary structure requirements is more important when using the naive POISSON-based model than when using the more informed JTT-based model.

When invoking the GTR configuration, the potentials give a greater improvement in fit than when applying the JTT settings. Nevertheless, site-interdependent GTR-based models are still poorer for these small data sets than the JTT-based models.

The use of a +Γ model seems to give an essentially additive improvement in model fit, with little, or no interaction with other model configurations. Because the statistical potentials could impact directly on site-specific rates, this result is unexpected; the lack of interaction in itself may be indicative that the potentials do not, in fact, acknowledge significant rate heterogeneity.

Comparison of Statistical Potentials

We find that for these applications, the potentials of Bastolla et al. (2001) and Miyazawa and Jernigan (1985) receive similar support, with +BAS models mildly favored over +MJ. The comparable merit of these potentials is somewhat expected; both work with a similar contact-based protein structure representation. The fact that +MJ models receive lower support than +BAS models may be a consequence of the oversimplified quasi-chemical approximation used in the derivation of the potentials of Miyazawa and Jernigan (1985) or to differences in the contact definition itself.

Sensitivity to the Prior on β

It is common practice, when assessing a new class of models, to evaluate the influence of the prior on the resulting model fit (Kass and Raftery 1995). Here, we focus on the distinguishing feature of our model: the prior on β. Note that the trace plots shown in figure 1 display, up to

an additive constant, the marginal likelihood of the model with β successively fixed to each value along the integration procedure. Treating β as a free parameter requires that we define a proper prior probability distribution, over which these curves are averaged (eq. 23). Because little is known regarding the usage of statistical potentials in this context, we follow the practice of assigning a bounded uniform prior, and testing empirically that the posterior distribution of β is well within these bounds (Robinson et al. 2003).

It should be noted that the 2 sets of potentials studied here are not scaled equivalently, which leads to different temperature factors at equilibrium (the potentials of Bastolla et al. [2001] lead to higher values of β [table 2]). This means that applying the same uniform prior on β under +BAS and +MJ models amounts to giving favor to the potentials of Bastolla et al. (2001); loosely speaking, the differences in scaling make the space of admissible values for β "appear" larger to +MJ models. To illustrate this problem, we performed a simple exploration of the influence of the size of the interval (*I*) of the uniform prior on β. Using the same sample, the Monte Carlo approximation given by equation (23) can be recomputed for different interval sizes. For example, figure 2 shows the log Bayes factor comparing JTT+F+Γ+BAS and JTT+F+Γ as a function of the interval size *I*. As *I* increases, the density of the prior contained in each increment of the quasi-static procedure decreases, leading to a lower support for JTT+F+Γ+BAS. When *I* reaches an order of magnitude around 10^6, the JTT+F+Γ+MJ model, with prior on $\beta \sim [-5, 5]$, becomes favored over JTT+F+Γ+BAS. Moreover, when *I* reaches an order of magnitude $\sim 10^{17}$, the JTT+F+Γ becomes favored over the site-interdependent model. This illustrates a fundamental theoretical consequence of the Bayesian paradigm: model rankings can change by redefining the space of admissible parameter values (the prior). In the present case, this means that no matter how strong the signal for site interdependence, there exists an interval size *I* for the uniform prior on β such that the site-independent model is favored, an example directly related to the so-called Jeffreys–Lindley paradox (Lindley 1957, 1980; Bartlett 1957).

In practice, the resulting difference in dimensionality penalty does not appear problematic in the present case; the potentials do not differ drastically in scaling, and the maximum marginal likelihood along β was always greater for the potentials of Bastolla et al. (2001) than for those of Miyazawa and Jernigan (1985). For example, for *MYO60-153* under the model JTT+F+Γ+BAS, the maximal point along the marginal likelihood curve gives a Bayes factor of ~553, whereas under JTT+F+Γ+MJ, the maximal point gives ~540.

For this particular comparison, one simple alternative would be to renormalize the potentials to an equivalent scaling. Yet, this solution would still not be applicable when comparing sequence fitness proxies based on fundamentally different rationales. In the longer run, nonuniform priors could be used, particularly, as more data sets are analyzed; Lempers (1971), for example, suggested setting aside some data sets for constructing proper priors to be used in subsequent analyses. Along these lines, we are currently devising other forms of statistical potentials, with each having the same overall temperature scaling (Kleinman et al. 2006).

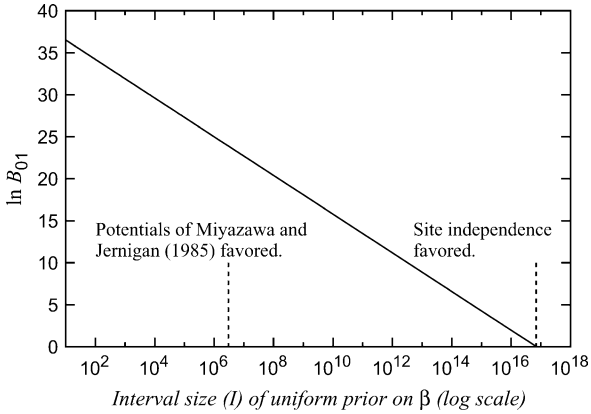

FIG. 2.—Influence of the interval size (*I*) of the uniform prior distribution for β on the calculated Bayes factor. Here, the models being compared are JTT+F+Γ+BAS against JTT+F+Γ, applied to *MYO60-153*. Two thresholds are marked on the graph. The first (leftmost) indicates the point beyond which JTT+F+Γ+MJ (with prior on $\beta \sim [-5, 5]$) is favored over JTT+F+Γ+BAS. The second indicates the point beyond which JTT+F+Γ is favored over JTT+F+Γ+BAS.

Permutations Checks

Overall, the pairwise contact potentials studied here appear inadequate; given the choice between the sole use of statistical potentials and the standard site-independent models, one would opt for the latter. Yet, a signal for site interdependence is clearly detected.

Perhaps the simplest check that can be done when constructing a model accounting for a particular signal is the evaluation of the model's performance when deliberately removing that signal from the data. Following Telford et al. (2005), we explore this through simple permutation tests, whereby we swap the positions of a percentage of random pairs of columns in the alignment. Such permutations have the effect of blurring the structural signal. Indeed, the

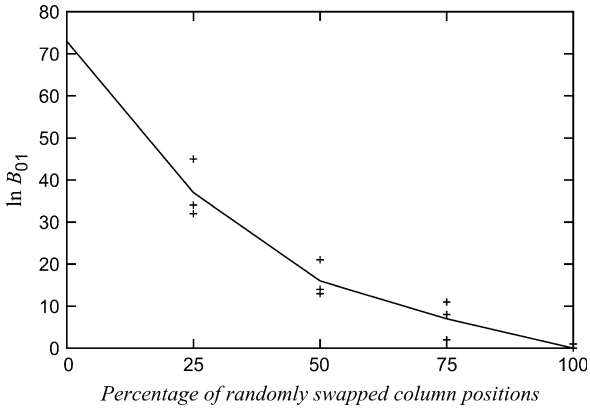

FIG. 3.—Permutation checks randomizing the order of columns in the alignment. The log Bayes factor is estimated between POISSON+F+Γ+BAS and POISSON+F+Γ, for 3 replicates at each randomization level. A line joining the mean values at each randomization level is drawn as a visual aid.

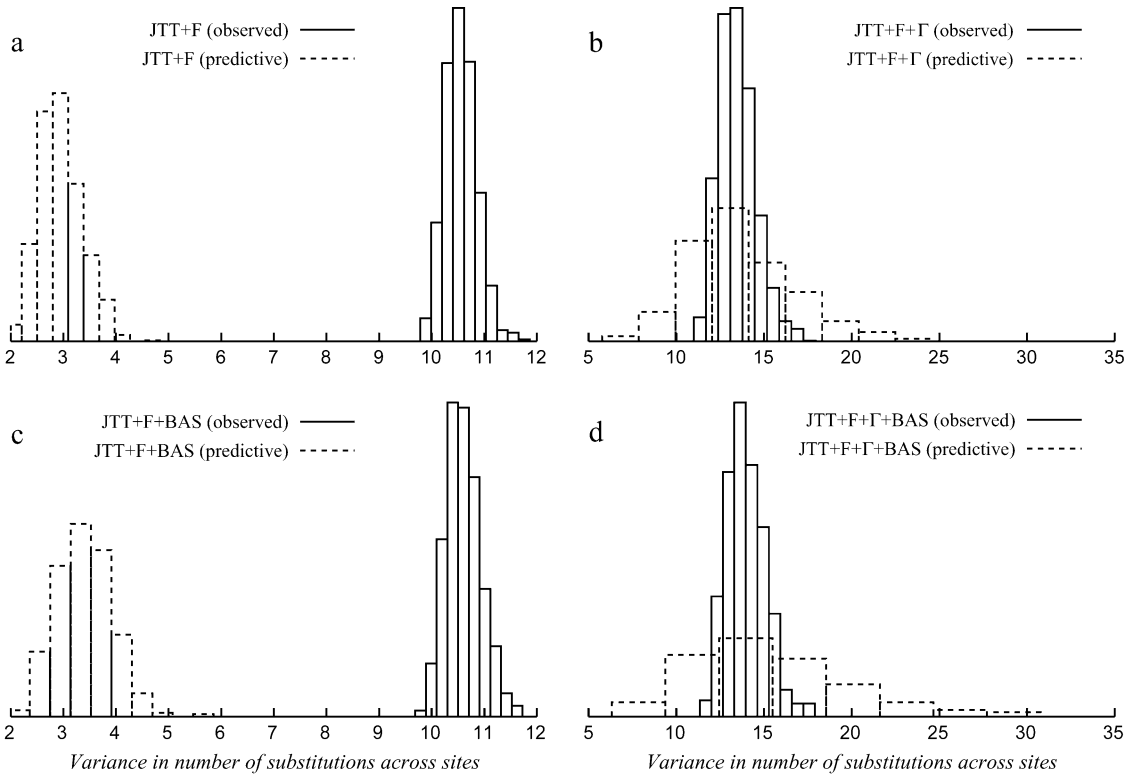

FIG. 4.—Posterior density plots of the variance in the number of substitution across sites obtained in predictive mappings and observed mappings of our sample from the posterior distribution, under the JTT+F (*a*), JTT+F+Γ (*b*), JTT+F+BAS (*c*), and JTT+F+Γ+BAS (*d*) models (using *MYO60-153*).

tests can be viewed as a randomization of the contacts in the contact map. We defined 4 levels of randomization, swapping the position of 25%, 50%, 75%, and 100% of columns. For each randomization, we computed the Bayes factor in favor of the site-interdependent model. Given the computational burden, we performed only 3 replicates for each randomization level.

We performed these permutation checks using the *MYO60-153* data set, comparing the log Bayes factor of POISSON+F+Γ+BAS against POISSON+F+Γ (this is the case giving the greatest improvement in model fit when applying the sequence fitness proxy). As expected, the support for site-interdependent considerations is a decreasing function of the percentage of randomization, essentially dropping to zero for a fully permuted column ordering (fig. 3). Also note that each replicate randomization gives slightly different results; evidently, the interdependencies between different positions of a protein are not all equivalent.

This test plainly illustrates the distinguishing feature of the models in simplistic terms: site-interdependent models give meaning to the order of amino acid columns in the alignment.

Posterior Predictive Resampling

Two of the most fundamental patterns of amino acid sequence evolution are 1) the heterogeneity of substitution rates across sites, and 2) the heterogeneity of amino acid exchange propensities. Both of these heterogeneities could be effects induced by structural constraints and, hence, could be accounted for—at least in part—by the sequence fitness proxy. However, accommodating rate-across sites variations (+Γ) and using an empirical amino acid replacement matrix (JTT) also accounts for these heterogeneities. As such, the best model obtained for all 3 data sets (JTT+F+Γ+BAS) seemingly corresponds to a redundant configuration. To further explore this point, we have applied simple posterior predictive checks, as described in Materials and Methods.

Rate Heterogeneity

Under a model assuming uniform rates across sites, and if there is rate variation in the data set considered, the observed rate variance is likely to depart significantly from the predictive rate variance; by the definition of the model, the predictive rate variance will tend to be very low. This is indeed the case, as can be seen from figure 4*a*. The extreme discrepancy between observed and predictive rate variance is in itself enough to reject the uniform rates model (Nielsen 2002). Comparing figures 4*a* and *c* shows that using the potentials of Bastolla et al. (2001) essentially leaves the observed rate variance unchanged, and the predictive rate variance is only slightly higher than the simple model assuming uniform rates—the mean predictive rate variance increases from 2.95 in figure 4*a* to 3.40 in figure 4*c*.

In contrast (fig. 4*b*), under the +Γ model, the observed rate variance is even greater than under the uniform rates model. As can be appreciated graphically, and according

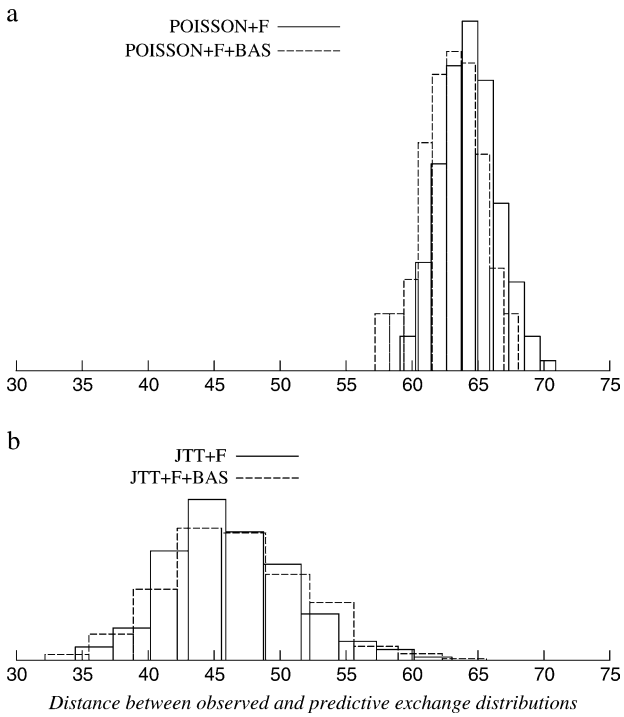

FIG. 5.—Posterior density plots of the Euclidean distance between predictive and observed exchange distribution for *MYO60-153* (see Materials and Methods). In (*a*), the models used are POISSON+F and POISSON+F+BAS. In (*b*), the models are JTT+F and JTT+F+BAS.

to the calculated Bayes factors, an explicit treatment of rate variation (+Γ) gives a better correspondence between model and data, with the predictive distribution centered on the observed (fig. 4*b* and *d*).

Note that predictive distributions tend to have a greater spread than observed distributions. This is a result of predictive distributions comprising 2 levels of uncertainty: the fundamental uncertainty associated with the inferred parameter values of the model (the posterior distribution)—an uncertainty which tends to be greater for higher dimensional models—and the uncertainty associated to the data replication (the simulation procedure). Indeed, this effect is displayed in the more pronounced spread in rate variance under the more complex +BAS model (comparing fig. 4*b* and *d*).

Amino Acid Exchange Propensities

Figure 5 is a comparison of the Euclidean distance between predictive and observed exchange distributions, as explained in Materials and Methods. In principle, a model yielding a lower distance between observed and predictive amino acid exchange distributions would be favored.

In figure 5*a*, the distance between predictive and observed distributions under the POISSON+F is high and is only slightly reduced when applying the potentials of Bastolla et al. (2001)—the mean distance goes from 63.92 under POISSON+F to 62.51 under POISSON+F+BAS. In the case of JTT (fig. 5*b*), the distance between predictive and observed distributions is much lower. This is indicative that a much better adequation is obtained between the types of substitutions of mappings conditioned on the data, with those predicted under the model when using the empirical amino acid exchange propensities of JTT, even when applying the potentials of Bastolla et al. (2001).

Conclusions and Future Directions

The results of the different model assessment strategies converge to the same fundamental conclusion: although an improved model fit is observed when applying the statistical potentials, the improvement does not justify abandoning the successful techniques previously developed for modeling complexities such as across-site rate heterogeneity or variations in amino acid exchange propensities. It would indeed have been surprising to see such a simple 0/1 contact map supplanting all strategies developed under the assumption of independence. For the moment, the best pragmatic alternative seems to be a layering of approach, combining a sequence fitness proxy with an appropriate underlying site-independent configuration.

More generally, it seems clear that the study of site-interdependent models is at an early stage; many alternative model settings can be envisaged, and the utility of each for different applications has yet to be explored. Robinson et al. (2003) presented one particular mechanistic version but also suggested several other possible choices, including codon position–specific nucleotide equilibrium frequencies or a heterogeneous nonsynonymous/synonymous substitution ratio. Working directly at the level of amino acids, we have evaluated some model configurations in the present work. Yet, here again, many other alternatives could be assessed in this context, such as the use of different empirical matrices (e.g., Whelan and Goldman 2001) or considering a mixture of amino acid profiles (Lartillot and Philippe 2004). In addition, it will be of particular interest to conduct broader comparisons of various statistical potentials, based on different functional forms (e.g., Jones, Taylor, and Thornton 1992a; Singh et al. 1996; Gan et al. 2001). From a thermodynamic perspective, a more appropriate use of statistical potentials would include a comparison of the pseudo-energy of the target structure with the pseudo-energy distribution over a set of alternative contact maps for each sequence state considered in the inference. We are currently exploring such alternatives, as well as designing statistical potentials within the overall evolutionary model. In any case, the techniques applied here need not be restricted to a thermodynamic standpoint, but may include any other measurement, formulated as a function of the overall sequence. The relative merit of all these possible models could be explored using the methodology employed here.

Investigating the absolute merits of alternative models, using posterior predictive techniques should also be useful in making explicit the strengths and weaknesses of the different choices. Although we have explored 2 statistics here—for which the site-interdependent models performed poorly—the theoretical implications of the model suggest additional posterior predictive checks. We hope to investigate an expanded set of test statistics in future work, with a particular interest in studying shifts in site-specific evolutionary rates over the tree (i.e., heterotachy), as well as the patterns of concerted evolution across the positions of the alignment.

Many other research directions can also be envisaged using the present framework. The permutation tests, for example, illustrate an interesting perspective: Bayes factors can be computed between alternative contact maps. Here, the permutations effectively scrambled the contact map. In theory, however, one could imagine the opposing scenario of searching for the most supported contact map within an extended set of decoys, suggesting potential applications to structure prediction.

The main limitations presently hindering these broader studies are the computational requirements of the models, which are substantial. Indeed, the total CPU time for the present study is estimated at about 1,000 days on a Xeon 2.4 GHz computer. We are currently developing the methods presented here in order to obtain maximum likelihood parameter estimates, which we believe will be much faster than the current sampling procedures. This would allow for the application of similar models on more and larger data sets, so as to begin to investigate their usefulness in uncovering structural determinants to molecular evolution, as well as assessing their impact on phylogenetic inference and broader domains of molecular biology.

Supplementary Material

A complete listing of sequences included in our data sets is given in the accompanying supplementary material as well as more detailed descriptions of MCMC update operators, tuning strategies, and settings. Other technical issues regarding scaling the rate matrix R are discussed. Finally, highest and lowest values of Bayes factors are also reported, as well as 95% credibility intervals of β under plain MCMC sampling. All the supplementary materials are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We wish to thank Jeff Thorne, Gavin Naylor, Ivana Kostic, Rachel Gouin, and Claudia Kleinman for comments on the manuscript and the Réseau Québécois de calcul de haute performance for computational resources. This work was supported by Génome Québec, the Canadian Institute for Advanced Research, the Canadian Research Chair Program, the Centre National de la Recherche Scientifique (through the ACI-IMPBIO Model-Phylo funding program), and the Robert Cedergren Centre for bioinformatics and genomics.

Appendix

For a given value of β, the derivative of the logarithm of the marginal likelihood is given as:

$$\frac{\partial \ln p(D|\beta)}{\partial \beta} = \frac{1}{p(D|\beta)} \frac{\partial p(D|\beta)}{\partial \beta} \tag{25}$$

$$= \frac{1}{p(D|\beta)} \int_\Theta \int_\Omega \frac{\partial p(D, \omega|\theta, \beta) p(\theta)}{\partial \beta} d\omega d\theta \tag{26}$$

$$= \int_\Theta \int_\Omega \frac{\partial \ln p(D, \omega|\theta, \beta)}{\partial \beta} p(\omega, \theta|D, \beta) d\omega d\theta \tag{27}$$

$$= \left\langle \frac{\partial \ln p(D, \omega|\theta, \beta)}{\partial \beta} \right\rangle, \tag{28}$$

where $\langle \cdot \rangle$ stands for an expectation with respect to the posterior distribution.

The logarithm of the likelihood function gives:

$$\ln p(D, \omega|\theta) = \ln p(s_0|\theta) + \sum_{j=1}^{2P-3} \ln p(s_j, \omega_j | s_{j_{up}}, \theta). \tag{29}$$

The derivative of equation (29) therefore involves 2 terms. For the first term, the derivative gives:

$$\frac{\partial \ln p(s_0|\theta)}{\partial \beta} = \frac{\partial}{\partial \beta} \ln \left(e^{-2\beta E_{s_0}} \prod_{i=1}^{N} \pi_{s_{i0}} \right) - \frac{\partial}{\partial \beta} \ln Z_0 \tag{30}$$

$$= -2E_{s_0} + 2\langle E \rangle \tag{31}$$

$$= -2(E_{s_0} - \langle E \rangle), \tag{32}$$

where $\langle \cdot \rangle$ represents an expectation with respect to the stationary probability (as written in eq. 9), which can be estimated based on a sample of sequences $(s^{(h)})_{1 \le h \le K}$ drawn from equation (9) using the Gibbs sampling procedure described by Robinson et al. (2003):

$$\langle E \rangle \simeq \frac{1}{K} \sum_{h=1}^{K} E_{s^{(h)}}. \tag{33}$$

Referring back to equation (29), the second type of term needed is:

$$\frac{\partial \ln p(s_j, \omega_j | s_{j_{up}}, \theta)}{\partial \beta} = \left(\sum_{k=1}^{z_j} \frac{\partial \ln R_{s_{jk-1} s_{jk}} r_{\sigma_{jk}}}{\partial \beta} - \frac{\partial (t_{jk} - t_{jk-1}) \sum_{i=1}^{N} \sum_{s_i'} R_{s_{jk-1} s_i'} r_i}{\partial \beta} \right) - \frac{\partial (\lambda_j - t_{jz_j}) \sum_{i=1}^{N} \sum_{s_i'} R_{s_{jz_j} s_i'} r_i}{\partial \beta}. \tag{34}$$

Equation (34), in turn, requires 2 types of derivatives:

$$\frac{\partial \ln R_{ss'}}{\partial \beta} = E_s - E_{s'} \tag{35}$$

and

$$\frac{\partial R_{ss'}}{\partial \beta} = (E_s - E_{s'}) R_{ss'}. \tag{36}$$

Substituting equations (35) and (36) appropriately into equation (34) and substituting the result of equation (34) back into the derivative of equation (29) completes calculation.

Literature Cited

Arndt PF, Burge CB, Hwa T. 2002. DNA sequence evolution with neighbor-dependent mutation. In: Myers GS, Hannenhalli S,

Istrail S, Pevzner P, Waterman M, editors. Proceedings of the Sixth Annual International Conference on Computational Biology. New York: Association for Computing Machinery. p 32–8.
Babajide A, Farber R, Hofacker IL, Inman J, Lapedes AS, Stadler PF. 2001. Exploring protein sequence space using knowledge-based potentials. J Theor Biol 212:35–46.
Bartlett MS. 1957. A comment on D. V. Lindley's statistical paradox. Biometrika 44:533–4.
Bastolla U, Farwer J, Knapp EW, Vendruscolo M. 2001. How to guarantee optimal stability for most representative structures in the protein data bank. Proteins 44:79–96.
Bastolla U, Porto M, Roman HE, Vendruscolo M. 2003. Connectivity of neutral networks, overdispersion, and structural conservation in protein evolution. J Mol Evol 56:243–54.
Bastolla U, Roman HE, Vendruscolo M. 1999. Neutral evolution of model proteins: diffusion in sequence space and overdispersion. J Theor Biol 200:49–64.
Bollback JP. 2005. Posterior mapping and posterior predictive distributions. In: Nielsen R, editor. Statistical methods in molecular evolution. New York: Springer. p 439–62.
Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol 54:743–57.
Buckley TR, Simon C, Chambers GK. 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. Syst Biol 50:67–86.
Dayhoff MO, Eck RV, Park CM. 1972. A model of evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure. Washington, DC: National Biomedical Research Foundation. p 88–9.
Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure. Washington, DC: National Biomedical Research Foundation. p 345–52.
Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–76.
Felsenstein J, Churchill GA. 1996. A hidden Markov model approach to variation among sites in rate of evolution. Mol Biol Evol 13:93–104.
Fornasari ME, Parisi G, Echave J. 2002. Site-specific amino acid replacement matrices from structurally constrained protein evolution simulations. Mol Biol Evol 19:352–6.
Galtier N, Gouy M. 1998. Inferring pattern and process: maximum likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol Biol Evol 15:871–9.
Gan HH, Tropsha A, Schlick T. 2001. Lattice protein folding with two and four-body statistical potentials. Proteins 43:161–74.
Gelman A. 1998. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. Stat Sci 13:163–85.
Gelman A, Meng XL, Stern H. 1996. Posterior predictive assessment of model fitness via realised discrepancies. Stat Sin 6:733–807.
Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704.
Hastings WK. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.
Jeffreys H. 1935. Some tests of significance, treated by the theory of probability. Proc Camb Philos Soc 31:203–22.
Jensen JL, Pedersen A.-MK. 2000. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. Adv Appl Prob 32:499–517.
Jones DT, Taylor WR, Thornton JM. 1992a. A new approach to protein fold recognition. Nature 358:86–9.
Jones DT, Taylor WR, Thornton JM. 1992b. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–82.
Kass RE, Raftery AE. 1995. Bayes factors and model uncertainty. J Am Stat Assoc 90:773–95.
Kleinman CL, Rodrigue N, Bonnard C, Philippe H, Lartillot N. 2006. A maximum likelihood framework for protein design. BMC Bioinformatics 7:326.
Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095–109.
Lartillot N, Philippe H. 2006. Computing Bayes factors using thermodynamic integration. Syst Biol 55:195–207.
Lempers FB. 1971. Posterior probabilities of alternative linear models. Rotterdam: Rotterdam University Press.
Lindley DV. 1957. A statistical paradox. Biometrika 44:187–92.
Lindley DV. 1980. L. J. Savage—his work on probability and statistics. Ann Stat 8:1–24.
Metropolis S, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equation of state calculation by fast computing machines. J Chem Phys 21:1087–92.
Miyazawa S, Jernigan RL. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18:534–52.
Neal RM. 1993. Probabilistic inference using Markov chain Monte Carlo methods. Technical report CRG-TR-93-1. Toronto: University of Toronto.
Nielsen R. 2002. Mapping mutations on phylogenies. Syst Biol 51:729–39.
Ogata Y. 1989. A Monte Carlo method for high dimensional integration. Num Math 55:137–57.
Pagel M, Meade A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst Biol 53:561–81.
Parisi G, Echave J. 2001. Structural constraints and emergence of sequence patterns in protein evolution. Mol Biol Evol 18:750–6.
Parisi G, Echave J. 2004. The structurally constrained protein evolution model accounts for sequence patterns of the Lβ h superfamily. BMC Evol Biol 4:41.
Parisi G, Echave J. 2005. Generality of the structurally constrained protein evolution model: assessment on representatives from the four main fold classes. Gene 345:45–53.
Pedersen A.-MK, Jensen JL. 2001. A dependent rates model and MCMC based methodology for the maximum likelihood analysis of sequences with overlapping reading frames. Mol Biol Evol 18:763–76.
Posada D, Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst Biol 53:793–808.
Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol 18:1692–704.
Rodrigue N, Lartillot N, Bryant D, Philippe H. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene 347:207–17.
Rubin DB. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann Stat 4:1151–72.
Seno F, Micheletti C, Martian A. 1998. Variational approach to protein design and extraction of interaction potentials. Phys Rev Lett 81:2172–5.
Shakhnovich EI, Gutin AM. 1993. Engineering of stable and fast-folding sequences of model proteins. Proc Natl Acad Sci USA 90:7195–9.

Siepel A, Haussler D. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol 21:468–88.

Singh RK, Tropsha A, Vaisman II. 1996. Delaunay tessellation of proteins. J Comput Biol 2:213–21.

Sippl MJ. 1990. Calculation of conformational ensembles from potentials of mean force; an approach to the knowledge-based prediction of local structure in globular proteins. J Mol Biol 213:859–83.

Sullivan J, Joyce P. 2005. Model selection in phylogenetics. Ann Rev Ecol Evol Syst 36:445–66.

Sun S, Bren R, Chan R, Dill K. 1995. Designing amino acid sequences to fold with good hydrophobic cores. Protein Eng 8:1205–13.

Telford MJ, Wise MJ, Gowri-Shankar Y. 2005. Consideration of RNA secondary structure significantly improves likelihood-based estimates of phylogeny: examples from the bilateria. Mol Biol Evol 22:1129–36.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18: 691–9.

Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10:1396–401.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–14.

Yang Z. 1996. Among site variation and its impact on phylogenetic analyses. Trends Ecol Evol 11:367–70.

Martin Embley, Associate Editor

Accepted June 13, 2006