Computational Methods for Evaluating Phylogenetic Models of Coding Sequence Evolution with Dependence between Codons

Nicolas Rodrigue, Claudia L Kleinman, Hervé Philippe, Nicolas Lartillot

Computational Methods for Evaluating Phylogenetic Models of Coding Sequence Evolution with Dependence between Codons

Nicolas Rodrigue, Claudia L. Kleinman,† Hervé Philippe,† and Nicolas Lartillot†*

*Department of Biology, Center for Advanced Research in Environmental Genomics, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada; and †Département de Biochimie, Centre Robert Cedergren, Université de Montréal, C.P. 6821, Succursale Centre-ville, Montréal, Québec H3C 3J7, Canada

In recent years, molecular evolutionary models formulated as site-interdependent Markovian codon substitution processes have been proposed as means of mechanistically accounting for selective features over long-range evolutionary scales. Under such models, site interdependencies are reflected in the use of a simplified protein tertiary structure representation and predefined statistical potential, which, along with mutational parameters, mediate nonsynonymous rates of substitution; rates of synonymous events are solely mediated by mutational parameters. Although theoretically attractive, the models are computationally challenging, and the methods used to manipulate them still do not allow for quantitative model evaluations in a multiple-sequence context. Here, we describe Markov chain Monte Carlo computational methodologies for sampling parameters from their posterior distribution under site-interdependent codon substitution models within a phylogenetic context and allowing for Bayesian model assessment and ranking. Specifically, the techniques we expound here can form the basis of posterior predictive checking under these models and can be embedded within thermodynamic integration algorithms for computing Bayes factors. We illustrate the methods using two data sets and find that although current forms of site-interdependent models of codon substitution provide an improved fit, they are outperformed by the extended site-independent versions. Altogether, the methodologies described here should enable a quantified contrasting of alternative ways of modeling structural constraints, or other site-interdependent criteria, and establish if such formulations can match (or supplant) site-independent model extensions.

Introduction

Recent years have seen an increasing interest in Markovian models of molecular evolution that incorporate a greater level of realism and that make the explicit distinction between mutational and selective parameterizations (e.g., Robinson et al. 2003; Thorne 2007; Thorne et al. 2007; Yang and Nielsen 2008; Rodrigue, Lartillot, and Philippe 2008). The original proposal of Goldman and Yang (1994) and Muse and Gaut (1994) consisted in assuming that all substitutions arise from point mutations and defining the state space of the Markov process to be the sense codons (i.e., omitting stop codons). An additional feature of such codon substitution models is their distinction between synonymous and nonsynonymous events, which has made them relevant tools for studying selective effects (for recent reviews, see Delport et al. 2008; Anisimova and Kosiol 2009). Huelsenbeck et al. (2006), for instance, proposed the use of a Dirichlet process prior for modeling site-specific modes of selection. Their model considers the rate of nonsynonymous events at a given codon as being mediated by one of several possible components (rate classes); under the Dirichlet process, the number of components is a random variable controlled by a higher level (hyper-)parameterization. Another approach making the distinction between synonymous and nonsynonymous events was presented by Robinson et al. (2003), who described a model and computational methodologies that allow for dependence between codon sites. In their model, Robinson et al. (2003) used an empirical potential originally derived for the protein-fold prediction problem (Jones et al. 1992) to quantify the compatibility of an amino acid sequence with a particular protein tertiary structure; assuming a fixed, coarse-grained protein structure representation, nonsynonymous rates of substitution depend on amino acid sequence pseudo-energy scores returned by the potential before and after the postulated event. For a codon sequence of length N, the Markov generator is defined by a $61^N \times 61^N$ matrix (assuming the universal genetic code), in principle allowing the possibility of a total dependence between all codons (although still based on a point mutation process).

The evolutionary models described by Robinson et al. (2003) and Huelsenbeck et al. (2006) are illustrative of two different modeling stances. The modeling stance taken by Huelsenbeck et al. (2006) may be said to be "phenomenological," in the sense that it aims only at detecting selective effects, in a site heterogeneous manner, without regard to underlying causes. In contrast, the modeling stance taken by Robinson et al. (2003) is "mechanistic," in the sense that it is aimed at explaining underlying selective effects, in this case as pertaining to protein tertiary structure constraints. The mechanistic stance proposed by Robinson et al. (2003) is particularly attractive because it offers the possibility of interrogating real data regarding explicitly defined selective features and because it can be assigned population genetic interpretations (Thorne et al. 2007). However, their modeling approach is computationally challenging.

Specifically, the practical complications of the model presented by Robinson et al. (2003) led these authors to propose the use of a set of Markov chain Monte Carlo (MCMC) techniques based on two different forms of auxiliary variable methods: 1) a data augmentation system, providing a numerical means of integrating over detailed substitution histories (addressing the transient aspects of the model) and 2) an importance sampling argument, providing an approximation of the ratio of two intractable normalizing "constants" (addressing the stationary aspects

Key Words: Markov chain Monte Carlo, data augmentation, auxiliary variables, posterior predictive checking, Bayes factors, protein tertiary structure.

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

Mol. Biol. Evol. 26(7):1663–1676. 2009
doi:10.1093/molbev/msp078
Advance Access publication April 21, 2009

of the model) to draw from so-called doubly intractable distributions. Together, these approaches provided the first proof-of-concept that models with a general dependence between codons could be implemented. However, the resulting sampling devices are elaborate, computationally highly demanding, and do not address the calculation of model fit. By invoking certain modeling simplifications, we have previously sought to explore variants of the approaches presented by Robinson et al. (2003) in the multiple-sequence context (Rodrigue et al. 2005) and to calculate Bayes factors (Rodrigue et al. 2006; Rodrigue et al. 2007), but these exploratory works are based on models that operate only at the level of amino acids, relinquishing the attractive codon-based framework of mutational and selective parameterizations.

Recently, Choi et al. (2007) have presented techniques for evaluating Bayes factors under the stationarity of the original model of Robinson et al. (2003). The main motivation of Choi et al. (2007) was to conduct a meta-analysis on numerous protein-coding genes, but each represented by a single sequence and each considered as a sample from the stationary distribution defined by a mutation selection balance (in which selection is embodied by the potential). Their analysis demonstrates that the use of an empirical potential in this way leads to a significantly improved model fit for most protein-coding genes under study. Focusing on the stationarity of the process, Choi et al. (2007) could dispense with the first of the two auxiliary variable methods (data augmentation), but there remains an interest for approaches that incorporate the phylogenetic component into the overall comparisons as only then can we contrast the performance of parameterizations that only influence the transient aspects of the Markov process (such as models with the Dirichlet process on nonsynonymous rates, or models with implicit versus direct connections to population genetic theory; Thorne et al. 2007). Overall, we still lack a naturally extensible MCMC sampling methodology for quantified phylogenetic explorations of the use of statistical potentials within the context of codon substitution models.

Here, we combine and configure a set of previously proposed numerical techniques in order to implement, assess, and rank codon substitution models including those in the category proposed by Robinson et al. (2003) within a full phylogenetic (multiple-sequence) context. Specifically, we study models that incorporate an empirical potential derived in the context of protein design (Kleinman et al. 2006) with the more well-known codon substitution model formulations inspired by Muse and Gaut (1994). We describe a sampling methodology that exploits recent techniques from the statistical literature derived for approximating posterior distributions under models with intractable normalizing factors in the likelihood function (Murray et al. 2006), as well as data augmentation procedures based on accept/reject simulation (Nielsen 2002) and the uniformization method (Rodrigue, Philippe, and Lartillot 2008). We illustrate empirical explorations of the tuning of the MCMC devices, addressing both forms of auxiliary methods mentioned above, and find that the techniques detailed here allow for tractable applications of the site-interdependent framework for most models of interest. To demonstrate the usefulness of the methods in practice, we conduct simple posterior predictive checks, assessing the ability of different models to reproduce observed features of nonsynonymous rates, and embed the sampling approaches within the thermodynamic integration methods described in Rodrigue et al. (2006), for the calculation of Bayes factors. Using two data sets, we find that although the site-interdependent framework provides an improved fit, it is markedly outperformed by sophisticated site-independent models.

Materials and Methods

Substitution Models

The main motivation of site-interdependent models has been to incorporate explicit protein structure considerations within the phylogenetic context. We follow the nomenclature of Parisi and Echave (2001) and refer to the models as "structurally constrained" (SC) but utilizing the combined contact and solvent accessibility potential developed in Kleinman et al. (2006). The potential, written as $G(s)$ for the pseudo-energy score of the amino acid sequence encoded by the codon sequence $s = (s_i)_{1 \leqslant i \leqslant N}$, is given by

$$G(s) = \sum_{1 \leqslant i < j \leqslant N} \Delta_{ij} \varepsilon_{f(s_i) f(s_j)} + \sum_{1 \leqslant i \leqslant N} \Xi^{w_i}_{f(s_i)} + \sum_{1 \leqslant i \leqslant N} \mu_{f(s_i)}, \qquad (1)$$

where $f(a)$ returns the amino acid state encoded by codon a. The first term in equation (1) is a contact potential: $\Delta_{ij} = 1$ if amino acids encoded at positions i and j are in contact in the three-dimensional structure (see Miyazawa and Jernigan 1985) and 0 otherwise (Δ is the "contact map" and is of dimension $N \times N$) and ε_{lm} is the pseudo-energy associated with observing the amino acids l and m in contact (ε is 20×20, with 210 nonredundant values). The second term is a solvent accessibility potential: $\Xi^{w_i}_l$ is the pseudo-energy associated with observing the amino acid l at position i, where w_i is an index (not an exponent) of the solvent accessibility class defined at site i (the potential uses 14 different solvent accessibility classes, such that Ξ has 14×20 values). The last term accounts for compositional effects, inspired from the random energy approximation (Shakhnovich and Gutin 1993; Sun et al. 1995; Seno et al. 1998): μ_l is the pseudo-energy associated with observing the amino acid l (μ is also called the "chemical potential" of amino acids and has 20 values). We emphasize that the ε, Ξ, and μ parameters are fixed in all models to the values obtained in Kleinman et al. (2006).

The criterion defined in equation (1) is combined to a mutational specification, consisting of two sets of parameters: $\varrho = (\varrho_{lm})_{1 \leqslant l,m \leqslant 4}$ is a set of (symmetrical) nucleotide relative exchangeability parameters, with the (arbitrary) constraint $\sum_{1 \leqslant l < m \leqslant 4} \varrho_{lm} = 1$ and $\varphi = (\varphi_m)_{1 \leqslant m \leqslant 4}$, with $\sum_{m=1}^{4} \varphi_m = 1$, represents a set of global nucleotide equilibrium propensities. The final model also includes a

parameter ω, for now treated as a global factor modulating nonsynonymous rates without regard to the amino acids involved.

Following Robinson et al. (2003), an off-diagonal entry of the Markov generator, corresponding to the instantaneous rate of substitution from one sequence (s) to another (s'), is given by

$$R_{ss'} = \begin{cases} \varrho_{s_{ic}s'_{ic}}\varphi_{s'_{ic}}, & \text{if } \mathscr{A}, \\ \omega\varrho_{s_{ic}s'_{ic}}\varphi_{s'_{ic}}\, e^{\beta(G(s)-G(s'))}, & \text{if } \mathscr{B}, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where

$\mathscr{A}$: s and s' differ only at the cth codon position of the ith site and imply a synonymous change;

$\mathscr{B}$: s and s' differ only at the cth codon position of the ith site and imply a nonsynonymous change;

and s_{i_c} is the nucleotide at the cth codon position of the ith site of sequence s. Diagonal entries are given by the negative sum of off-diagonal entries in a given row. Note that when $\beta = 0$, the model collapses to the type of codon substitution model proposed by Muse and Gaut (1994). Here, given the scaling of the potential, we may fix the value of this parameter at $\beta = 1/2$ (the reason for this is explained below), but we also explore the behavior of the model when β is treated as a free parameter.

The model differs with that of Robinson et al. (2003) in using a set of nucleotide exchangeabilities rather than a single parameter distinguishing transitions and transversions and in using a different pseudo-energy score.

Priors and Model Nomenclature

We used the same priors as described in Rodrigue, Lartillot, and Philippe (2008), along with a uniform prior on β over the range $[-5,5]$ (when its value is not fixed). We refer to the simplest model based on the mutational parameters φ and ϱ only as MG because it is inspired from Muse and Gaut (1994) and write MG-NS to refer to the model with a global nonsynonymous rate factor. When using the SC model based on the statistical potential, we add the suffix -SC for the version with $\beta = 1/2$ and -SC-β for the version treating β as a free parameter, giving MG-SC and MG-SC-β. We also explore combined models, which invoke both a free ω parameter as well as the SC parameterization, and refer to the models as MG-NS-SC and MG-NS-SC-β. We refer to the models with the Dirichlet process prior on ω as MG-NS$^{\mathrm{DP}}$ and those that combine it with the SC parameterization as MG-NS$^{\mathrm{DP}}$-SC and MG-NS$^{\mathrm{DP}}$-SC-β.

Data Augmentation

The model given in equation (2) does not allow for a closed-form calculation of the likelihood function. Rather, we rely on various MCMC computational devices, such as a demarginalization strategy known as "data augmentation." Within the present context, the traditional Bayesian sampling of parameters, collectively denoted as θ, from their distribution conditional on the data, D, and the overall construction of the model, M, is expanded (or augmented) into a two-stage sampling approach: 1) sample a detailed substitution mapping ϕ (including the timing and nature of all events along all branches), conditional on the parameters of the Markov process and the data set under study, and 2) sample a parameter vector θ, conditional on ϕ (and hence also conditional on D). These types of sampling approaches exploit the fact that if a sample can be obtained from the joint distribution $p(\phi,\theta \mid D,M)$, the θ component of this sample is distributed according to the posterior distribution of interest $p(\theta \mid D,M)$. For the traditional substitution models assuming independence between sites, such realizations of the Markov process can be generated directly (e.g., Rodrigue, Philippe, and Lartillot 2008). Although this is not the case for models with dependence, as in equation (2), Robinson et al. (2003) exploit a methodology that consists in using a simple site-independent model to draw a site-independent mapping and accepting or rejecting the mapping with probability ϑ, using the target site-interdependent distribution in the Metropolis–Hastings (MH) (Metropolis et al. 1953; Hastings 1970) rule:

$$\vartheta = \min\left\{1, \frac{p(\phi' \mid \theta,D,M)q(\phi',\phi)}{p(\phi \mid \theta,D,M)q(\phi,\phi')}\right\}, \tag{3}$$

where $q(\phi',\phi)/q(\phi,\phi')$ is the Hastings ratio, computed under the model used to produce the site-independent mappings. Repeatedly proposing and accepting or rejecting mappings in this way form a Markov chain with stationary distribution $p(\phi \mid \theta,D,M)$, and by further alternating with updates on parameter values conditional on the mappings, as we will get to below, the Markov chain formed has the full posterior probability distribution $p(\phi,\theta \mid D,M)$ as its stationary distribution, from which draws can be made at regular intervals to produce large-sample (Monte Carlo) approximations.

Robinson et al. (2003) rely on a nucleotide-level evolutionary model to propose mappings, for one nucleotide position at a time. In neglecting the structure of the genetic code, and other aspects of the site-interdependent model, we can expect the proposal density produced by a nucleotide-level model to be quite distant to the target density (e.g., by producing mappings that include stop codons, which are disallowed under the target model). The mapping proposal system described here is designed to be "as close as possible" to the target site-interdependent density. Our strategy consists in defining a codon-level model for each codon site of the alignment, from all but the contact component of the target model's specifications. Let $G_i(a)$ represent the pseudo-energy associated with observing the amino acid encoded by codon a at site i, but without consideration of the contact component; with the present form of potential, this consists of the solvent and chemical components, that is, $G_i(a) = \Xi^{w_i}_{f(a)} + \mu_{f(a)}$. Then, the Markov generator specifying the instantaneous rate from codon a

to codon b for site i is given by a 61×61 matrix, with entries

$$Q_{ab}^{(i)} = \begin{cases} \varrho_{a_c b_c} \varphi_{b_c}, & \text{if } a \text{ and } b \text{ are synonymous and differ only at } c\text{th codon position,} \\ \omega \varrho_{a_c b_c} \varphi_{b_c} \, \mathrm{e}^{\beta (G_i(a) - G_i(b))}, & \text{if } a \text{ and } b \text{ are synonymous and differ only at } c\text{th codon position,} \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

As before, diagonal entries are given by the negative sum of off-diagonal entries in a given row (see Appendix for details specific to our implementation). Note that from the form of the site-independent components of the potential, only 14 different Markov generators are possible at this point (one for each solvent accessibility class). However, when other site heterogeneous variables are introduced (e.g., the Dirichlet process on ω), working with site-specific matrices as written in equation (4) is both more general and more convenient, with each additional site variable incorporated directly. If we now let $G_\Delta(s)$ be the contact component pseudo-energy of sequence s, that is, $G_\Delta(s) = \sum_{1 \leqslant i < j \leqslant N} \Delta_{ij} \varepsilon_{s_i s_j}$, we can see that the model given in equation (2) can be written as

$$R_{ss'} = \begin{cases} Q_{s_i s'_i}, & \text{if } s \text{ and } s' \text{ differ by one synonymous codon at position } i, \\ Q_{s_i s'_i} \, \mathrm{e}^{\beta (G_\Delta(s) - G_\Delta(s'))} & \text{if } s \text{ and } s' \text{ differ by one nonsynonymous codon at position } i, \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

We instantiate the methodology described by Robinson et al. (2003) by proposing (codon) site-specific mappings under the model given in equation (4) and accepting or rejecting the resulting sequence-wide mapping under the model given in equation (5), which differs with the proposal model only by the factor $\mathrm{e}^{\beta (G_\Delta(s) - G_\Delta(s'))}$ in nonsynonymous rate entries.

The first step to proposing a site-independent mapping under equation (4) is to draw the codon states at the internal nodes of the tree from their joint distribution, conditional on the parameters of the Markov process and the observed states in the alignment. We use the method described in Nielsen (2002) for this purpose (also see Bollback 2005) and proceed to sample the series of events along each branch, conditional on the parameters of the Markov process and the codon states at both ends, according to one of two methods: if the codon states at both ends of the branch are identical, or differ by one nucleotide only, we use the accept/reject simulation methods given in Nielsen (2002) to draw the mapping; otherwise, we use the uniformization method described in Rodrigue, Philippe, and Lartillot (2008). We initialize the first mapping of our sampler by calling this procedure over all sites.

Updating Model Parameters

The same types of update operators as used in previous works (see, e.g., Rodrigue et al. 2006) can be applied in the present context to approximate the posterior distribution, based on the site-interdependent MH rule. However, as described in Robinson et al. (2003), for parameters bearing on the stationary distribution of the substitution process, the ratio of two intractable normalizing factors appears in the MH ratio and hence requiring a more elaborate approach. First note that within the data-augmentation framework, the likelihood function can be further decomposed into two factors: one corresponding to the probability of the sequence state at the root node of the tree, which we write as s^{o}, and another corresponding to the probability density of the mapping, conditional on the starting state at the root node. Because the substitution process is reversible, we may take one of the sequences observed at a leaf node to be s^{o} and write this factoring as

$$p(D, \phi \mid \theta, M) = p(s^{\mathrm{o}} \mid \theta, M) p(D^{\emptyset}, \phi \mid s^{\mathrm{o}}, \theta, M), \quad (6)$$

where $D^{\emptyset}$ represents the remaining sequences of the data set once s^{o} has been removed.

The stationary distribution of the full site-interdependent codon model given in equation (2) reads as

$$p(s^{\mathrm{o}} \mid \theta, M) = \frac{1}{Z_\theta} \mathrm{e}^{-2\beta G(s^{\mathrm{o}})} \prod_{i=1}^{N} \left(\prod_{c=1}^{3} \varphi_{s^{\mathrm{o}}_{ic}} \right), \quad (7)$$

where Z_θ is the normalizing factor:

$$Z_\theta = \sum_s \mathrm{e}^{-2\beta G(s)} \prod_{i=1}^{N} \left(\prod_{c=1}^{3} \varphi_{s_{ic}} \right), \quad (8)$$

with the sum being over all 61^N possible sequences. Of course, this sum is not tractable (and hence making the posterior distribution of ultimate interest doubly intractable). When proposing new values for any of the parameters implicated in the stationary distribution, the ratio of two of these terms appears. For simplicity, let $f(s^{\mathrm{o}}, \theta)$ be the unnormalized density:

$$f(s^{\mathrm{o}}, \theta) = \mathrm{e}^{-2\beta G(s^{\mathrm{o}})} \prod_{i=1}^{N} \left(\prod_{c=1}^{3} \varphi_{s^{\mathrm{o}}_{ic}} \right). \quad (9)$$

Expanding the MH rule for the present context, we have

$$\vartheta = \min \left\{ 1, \frac{p(D^{\emptyset}, \phi \mid s^{\mathrm{o}}, \theta', M) p(\theta' \mid M) f(s^{\mathrm{o}}, \theta') q(\theta', \theta) Z_\theta}{p(D^{\emptyset}, \phi \mid s^{\mathrm{o}}, \theta, M) p(\theta \mid M) f(s^{\mathrm{o}}, \theta) q(\theta, \theta') Z_{\theta'}} \right\}, \quad (10)$$

where we have written the complicating factors at the end of the ratio for emphasis.

The importance sampling method proposed by Robinson et al. (2003), which we call the "single-point bridge" (SPB) method, involves an internal Gibbs MCMC (detailed below) to approximate the MH ratio itself based on a large sample of K sequences, written as $(s^{(k)})_{1 \leqslant k \leqslant K}$.

The sample of sequences is drawn from the stationary distribution induced by a third set of parameters (other than θ and θ') written as θ^*, and the approximation in the present context reads as

$$\frac{Z_\theta}{Z_{\theta'}} \simeq \frac{\sum_{k=1}^{K} \mathrm{e}^{-2(\beta-\beta^*)G(s^{(k)})} \prod_{i=1}^{N} \prod_{c=1}^{3} \frac{\varphi_{s_{ic}^{(k)}}}{\varphi^*_{s_{ic}^{(k)}}}}{\sum_{k=1}^{K} \mathrm{e}^{-2(\beta'-\beta^*)G(s^{(k)})} \prod_{i=1}^{N} \prod_{c=1}^{3} \frac{\varphi'_{s_{ic}^{(k)}}}{\varphi^*_{s_{ic}^{(k)}}}}. \quad (11)$$

As discussed in Robinson et al. (2003), the quality of the approximation depends on the choice of θ^*; a parameter vector chosen to be as close as possible to the mid-point of the θ and θ' yields the best approximation. As a simple, yet crude implementation of this approach, we define a new θ^* for each update attempt on parameters involved in the stationary distribution, always at the mid-point between θ and θ', and explore empirically the size of the sample of sequences (see Results and Discussion).

Another strategy to approximating the MH ratio itself is to design a different MH ratio that has the same desired probability distribution (or a good approximation thereof) as its limiting distribution. Here, we propose an adaptation of the "single variable exchange" (SVE) method recently proposed by Murray et al. (2006). In the present context, the method rests on drawing a single auxiliary sequence (written as ς) from the stationary distribution of the substitution process induced by θ'. Then, the MH kernel is expanded to

$$\vartheta = \min\left\{1, \frac{p(D^{\emptyset}, \phi \mid s^{\mathrm{o}}, \theta', M)p(\theta' \mid M)f(s^{\mathrm{o}}, \theta')f(\varsigma, \theta)q(\theta', \theta)Z_\theta Z_{\theta'}}{p(D^{\emptyset}, \phi \mid s^{\mathrm{o}}, \theta, M)p(\theta \mid M)f(s^{\mathrm{o}}, \theta)f(\varsigma, \theta')q(\theta, \theta')Z_{\theta'} Z_\theta}\right\}, \quad (12)$$

where all intractable factors at the end of the ratio cancel. The MH kernel's validity rests on having truly sampled ς from the stationary probability induced by θ', which we cannot do analytically here. Instead, we again make use of an empirically tuned (see Results and Discussion) internal Gibbs MCMC system. This means that our sampler is making draws from an approximation of the desired posterior distribution.

As for updating parameters not involved in the stationary distribution of the substitution process, note that the MH kernel in equation (10) simplifies to

$$\vartheta = \min\left\{1, \frac{p(D^{\emptyset}, \phi \mid s^{\mathrm{o}}, \theta', M)p(\theta' \mid M)q(\theta', \theta)}{p(D^{\emptyset}, \phi \mid s^{\mathrm{o}}, \theta, M)p(\theta \mid M)q(\theta, \theta')}\right\} \quad (13)$$

and hence we need not call any doubly intractable device for such cases.

Internal Gibbs MCMC

As mentioned above, both the SPB and the SVE approaches to updating parameters involved in the stationary distribution of the substitution process rely on an internal Gibbs MCMC to draw one or several sequences, subsequently used in the MH kernel of the main MCMC, sampling from the full posterior distribution. The basic Gibbs update we use is conceptually the same as that described in Robinson et al. (2003); we fix all but one codon site and update the state at that site according to the probability of the 61 possible states. The 61 probabilities required for the Gibbs update can be calculated analytically (up to a multiplicative constant) according to equation (9). By repeating this procedure over all sites, a Markov chain in sequence space is formed, with a limiting distribution corresponding to the stationary distribution of the codon substitution model. In practice, we loop the procedure over all sites, in what we call a "Gibbs sweep" across the sequence, and tune the number of Gibbs sweeps empirically (see Results and Discussion).

General MCMC Settings

We used similar computational settings as in previous works alternating between updates on parameters and mappings (see, e.g., Lartillot 2006; Rodrigue et al. 2006). We use additive, multiplicative, and constrained update operators to propose new parameter values given the current parameter values and the current mapping. Additive operators rest on random draws from a uniform distribution in the interval [0, 1], denoted as U, as well as a tuning parameter, denoted as δ, and propose a new value for a unidimensional parameter by adding $\delta(U - 1/2)$ to the current parameter value; they have Hastings ratio of 1. Multiplicative operators propose a new value by multiplying the current value by $\mathrm{e}^{\delta(U-1/2)}$; they have a Hastings ratio equal to $\mathrm{e}^{\delta(U-1/2)}$. Constrained operators are applicable to profile-like parameters, with the constraint that they sum to 1. For these operators, a pair of entries in the profile is selected at random and their sum is stored; then an additive operator is applied to one of the pair, with back reflection if the value is beyond the stored sum or if it is negative; the other pair is then set under the constraint that the sum of both new values is equal to the original sum. Note that constrained operators can be applied on more than one pair of entries in profile-like vectors, always with a Hastings ratio of 1.

Briefly, each parameter (including each branch length) and hyperparameter is updated multiple times per cycle. Also, under the models using the Dirichlet process on ω, we loop multiplicative update attempts over all components of the current configuration of the process; the configuration of the Dirichlet process itself is resampled by looping a Gibbs update system including five new ω components drawn from the base (hyper-) prior (see Huelsenbeck et al. 2006) over all sites, five times per cycle. Following the series of parameter and hyperparameter updates, a cycle is completed by performing a series of updates on mappings. We established the detailed settings of each cycle empirically (see examples in Results and Discussion), and unless stated otherwise the results presented here are based on 11,000 cycles, removing the first 1,000 cycles as burn-in, and subsampling every 10th cycle, leaving 1,000 draws.

Posterior Predictive Checks

We performed simple posterior predictive checks enabled by the substitution mapping framework (Nielsen 2002; Bollback 2005). The overall framework consists of

contrasting statistics computed on the detailed mappings that constitute the data augmentation, which we call the "posterior mappings" as they are conditional on the parameters and constrained by the observed data, with the same statistics computed on "predictive mappings", which are obtained by an additional simulation step, conditional on the parameters, but unconstrained to the specific codon states observed in the true data. More precisely, using a set of parameters drawn from the posterior distribution, we first sample a sequence from equation (7) using Gibbs MCMC (for site-independent models, the sequence state can be drawn directly), then set the root node to this sequence state, and finally evolve the sequence over the tree according to the Markov process defined by the codon substitution parameters. Repeating this for each draw from the original sample obtained by MCMC allows us to generate posterior predictive distributions of the test statistics of interest. A discrepancy between the distributions of a statistic computed from predictive and posterior mappings indicates a weakness of the model, or, stated reciprocally, a good model should produce predictive mappings that exhibit the same features as posterior mappings.

Given that the SC models mediate nonsynonymous rates of substitution, we explored the properties they induce on statistics that are a function of the nonsynonymous events within the detailed substitution mappings. Many statistics can be envisaged. Here, for illustrative purposes, we compute two simple statistics: the mean number of nonsynonymous events across sites and the variance in the number of nonsynonymous statistics across sites. The distributions of these (or other) statistics, produced from posterior and predictive mappings, can be compared graphically as a first appraisal (Nielsen 2002; Rodrigue et al. 2006; Lartillot et al. 2007, and see Supplementary Material online). We note here, however, that when conducting a posterior predictive check based on statistics that are a function of parameters of the model (as is the case in the present context), one can compute the proportion of draws from the posterior distribution in which the statistic is higher in the predictive mapping than it is in the posterior mapping (displayed graphically in the Supplementary Material online); this proportion is the posterior predictive *P* value, and extreme values (say below 0.05 or above 0.95) are considered as flags of model inadequacy (see, e.g., Gelman et al. 2004). Note that the posterior predictive *P* values are uncalibrated (as they are based on previously inferred parameters) and hence tend to be conservative (e.g., Meng 1994; Gelman et al. 1996; Dahl 2006). Also note that the *P* values obtained with different models should not be viewed as directly comparable to one another nor as a means of quantitative model comparison.

Bayes Factors

We use Bayes factors to perform a formal comparison of models. The same thermodynamic integration methods described in Rodrigue et al. (2006) can be applied here to calculate the (log) Bayes factor comparing a model including the statistical potential with its nonstructural counterpart (i.e., with $\beta = 0$). The procedure consists of first running an MCMC sampler with the value of β progressively incremented from 0 to some sufficiently large value (to encompass all the relevant high-likelihood region), with block series of MH updates performed on other parameters (and hyperparameters, and mappings) between each increment; we denote the values of β over this thermodynamic MCMC run as $(\beta_k)_{0\leqslant k\leqslant K}$, with $\beta_0 = 0$. For any value K' ($0 \leqslant K' \leqslant K$), the difference in log marginal likelihood (marginalized over all but β) between a model fixed at the value $\beta_{K'}$ and a site-independent model fixed at β_0, written as $\ln p(D \mid \beta_{K'}) - \ln p(D \mid \beta_0)$, can be computed, and using this for the entire thermodynamic sample produces a trace of log marginal likelihood differences along β (see supplementary fig. S1 [Supplementary Material online] for a specific example in the the codon context). When using the more rigid SC settings, the value at the $\beta = 1/2$ point along the curve is the log Bayes factor in favor of the structural model. In the case of the MG-NS-SC-β settings (treating β as a free parameter), this is followed by an exponentiation and averaging of the curve over the prior distribution (see eq. 23 in Rodrigue et al. 2006). We used these methods in combination with the site-independent "omega"-switch method described in Rodrigue, Lartillot, and Philippe (2008) to calculate all log-Bayes factors with respect to MG-NS.

Data

We applied the computational framework described above on two data sets taken from Yang et al. (2000). The first, which we refer to as GLOBIN*17-144*, consists of 17 vertebrate nucleotide sequences of the β-globin gene (144 codons). The second, referred to as ADH*23-254*, consists of alcohol dehydrogenase genes taken from 23 species of Drosophila (254 codons). Contact maps and solvent accessibility profiles were derived from the Protein Data Bank files 4HHB and 1A4U for the GLOBIN*17-144* and ADH*23-254* data sets, respectively. The solvent accessibility profiles consider the quaternary structure, though the contact maps do not. For both data sets, we worked under the tree topology used by Yang et al. (2000).

Results and Discussion

Empirical Explorations of Data Augmentation Sampling

We first investigated the properties of the computational devices based on a simplified, rigid codon substitution model with $\beta = 1/2$, $\omega = 1$, and mutational parameters fixed to equal values on their state space ($\varphi_l = 1/4$ and $\varrho_{lm} = 1/6$). Such a model is formally site interdependent, allowing us to explore the data augmentation system with branch lengths (and the hyperparameter governing the prior on branch lengths) being the only parameters updated. In other words, none of the MCMC operators required for sampling from the posterior distribution under such a model involve changes to the limiting distribution of the codon substitution process and hence do not require an internal Gibbs MCMC for doubly intractable operators.

We performed numerous pilot runs to tune the sampler. In a first series of pilot runs, we used the uniformization method (Rodrigue, Philippe, and Lartillot 2008) for proposing mappings. In the codon context, the uniformization is more stable in certain cases than the accept/reject

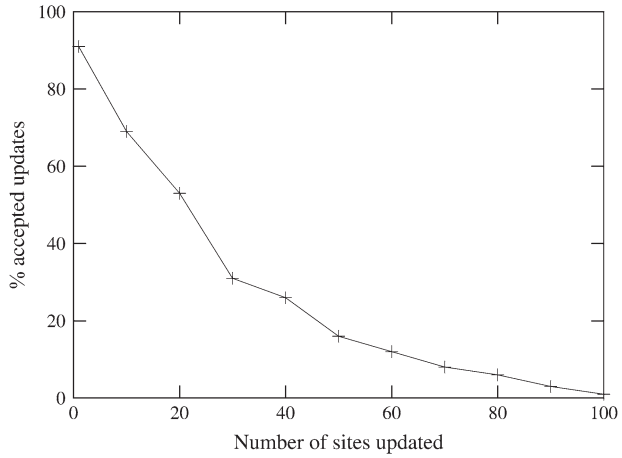

FIG. 1.—Empirical explorations of proportion of accepted updates on substitution mappings as a function of the number of sites with updated mappings in the proposal density. The substitution model used here has $\omega = 1$, $\beta = 1/2$, and all other parameters fixed to uniform values on their state space. The display is for the GLOBIN*17-144* data set.

method, specifically, when the states at the ancestral and descendant nodes differ by two or three nucleotides (Rodrigue, Philippe, and Lartillot 2008). On the other hand, uniformization sampling involves computationally costly matrix powers, which the accept/reject method does not, and in practice, most configurations of internal node states drawn do not correspond to the problematic sampling cases. Using the rigid model described above, for instance, only about 20% of the cases where the ancestral and descendent nodes differ at a codon site do so by more than one nucleotide. When all cases are considered (i.e., when cases with identical ancestral and descendent node states are included), only about 3% of internal node states involve starting and ending states that differ by more than one nucleotide. We observed our sampler to run roughly 120 times faster when using the uniformization method only in these approximately 3% of cases and the accept/reject method in all other cases.

In another example pilot run, we performed 100 multiplicative updates per cycle on branch lengths and 100 additive updates per cycle on the hyperparameter governing the branch length prior; each cycle also included calls to data augmentation updates, with a separate call attempting an update at a single codon, along the entire tree, another call attempting an update at 10 randomly selected codons sites, again over the entire tree, another with 20 sites, and so forth in steps of 10, up to 100 sites. Our objective here was to establish settings that simultaneously update as many site mappings as possible (to "make the most" of calls to the site-interdependent calculations), while obtaining reasonable acceptance rates. Figure 1 displays the percentage of accepted updates as a function of the number of sites in update attempts, working with the GLOBIN*17-144* data set. In this example, the operator attempting an update at a single codon site over the entire tree yielded an acceptance rate over 90%. Such a high acceptance rate is the simple result of having the proposal distribution nearly identical to the target distribution; these differ only in the mapping of a single codon site. However, the operator needs to be called numerous times in order to update the entire sequence-wide mapping. In contrast, an operator attempting an update to many sites has a lower acceptance rate, although even for an operator applied to 50 sites the acceptance rate of 16% is still reasonable. Via these types of experiments, we tuned our sampler to use operators attempting updates to 50, 40, 30, 20, and 10 sites at once, each of these called 10 times per cycle. We note that updating an equivalent number of sites with attempts that alter a single codon site requires about 100-fold more CPU time per cycle. We did not explore a mapping update system that attempts an update along a single branch for a single nucleotide position (Robinson et al. 2003), but presumably such a system would also require a CPU time at least two orders of magnitude greater than the system we use here.

Empirical Explorations of Doubly Intractable Sampling

We next wanted to explore the sampling procedure handling the stationary probability of the codon substitution process within the overall MCMC. In order to focus our exploration on this doubly intractable sampling only, we used data sets consisting of a single sequence, taken from the GLOBIN*17-144* data set, and the MG-SC model. Note that because we consider a single sequence, any expansion to the MG-NS-SC or MG-NS^{DP}-SC models is meaningless because the parameters involved have no bearing on the probability distribution of interest. Likewise, the nucleotide exchangeability parameters are not involved in the stationary probability, and hence only the nucleotide propensity parameters are free in such settings.

The SPB and SVE sampling devices both rest on drawing sequences from equation (7), and hence we ran several pilot runs in order to tune the internal Gibbs MCMC, before embedding it in the main MH MCMC. For example, we performed several pilot runs drawing samples of 100 sequences with a different number of Gibbs sweeps (see Materials and Methods) between draws. Figure 2*a* summarizes the experiment as the autocorrelation function of the sequence pseudo-energy score, from samples with a lag between 1 and 10. The first zero-crossing is observed with a lag of five Gibbs sweeps between draws. Based on these and other similar results, we subsequently devised our implementation to follow a simple procedure: upon starting the overall MCMC, the sequence ς is initialized by performing a random draw from the 61 possible codons at each site, according to the stationary distribution of the site-specific stationary probabilities under the site-specific codon model (lacking the contact component); when calling an operator on a parameters bearing on the stationary distribution of the full site-interdependent process, five Gibbs sweeps across the positions of ς are performed; subsequent calls on parameters bearing on the stationary distribution start from the current ς and again perform five Gibbs sweeps across the sequence.

Using these Gibbs MCMC settings, we explored the effect of the sample size on the SPB approximation given in equation (11). We used constrained moves to propose new values for the nucleotide propensity parameters. Figure 2*b* displays the (log) approximation as a function

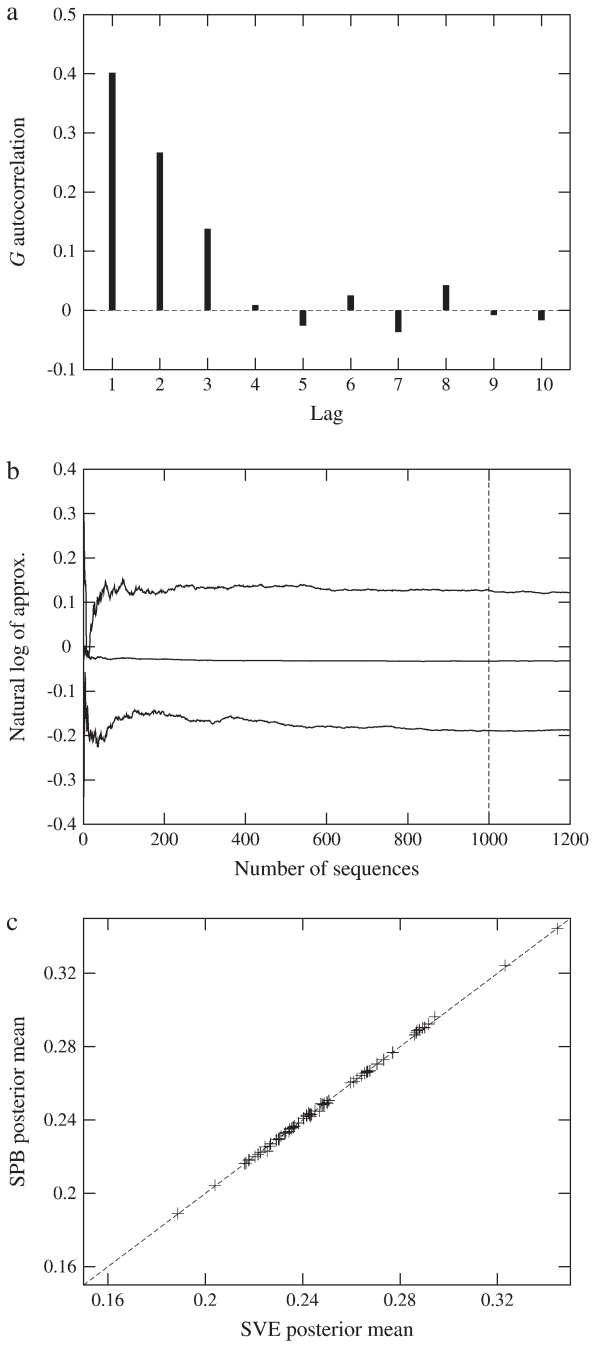

FIG. 2.—Empirical explorations of MCMC sampling behavior using the MG-SC model on the GLOBIN*17-144* data set. In (*a*), we explore the number of Gibbs sweeps needed to decorrelate successive sequences drawn from equation (7). In (*b*), we explore the number of sequences needed to stabilize the importance sampling approximation in equation (11). In (*c*), we compare the posterior means of the nucleotide propensity parameters, inferred separately (one sequence at a time) on each of the 17 sequences of the data set, using the SPB and SVE methods.

of the number of sequences for three different update attempts. These and all other cases we looked at (not shown) were well stabilized with a sample size of 1,000 sequences.

We next applied the SPB method with 1,000 sequences to sample the nucleotide propensity parameters from their posterior distribution (with 100 updates per cycle) using each of the 17 sequences in the GLOBIN*17-144* data set in turn. We also ran the SVE method in these conditions, with the exception that it only requires one sequence (again obtained by updating the auxiliary sequence with five Gibbs sweeps). In practice, we found that both the SPB and SVE methods produce very similar results. Figure 2*c* displays a comparison of the posterior mean of the nucleotide propensity parameters obtained from both methods (for all 17 analyses), which show a strong correlation (0.999). However, we found the SVE sampler to run about 960 times faster than the sampler using SPB method. The SPB method could undoubtedly be made more efficient, for instance using a grid-based system for predefining a small number θ^* parameters (Robinson et al. 2003) as opposed to defining a new θ^* with each update, but because the SVE method is conceptually simpler—without the additional tuning of a grid-based system—and yields the same results, we used it in all subsequent calculations.

We repeated the calculations of drawing nucleotide propensity parameters from their posterior distribution for a few other simple explorations of the SVE methods, using 1, 5, 10, and 50 Gibbs sweeps to update the auxiliary sequence. Surprisingly, even when using a single Gibbs sweep to update the auxiliary sequence, duplicate analyses on the each of the 17 sequences of the GLOBIN*17-144* data set show a good correlation (0.947). When increasing the number of updates per cycle by a factor of 5, a strong correlation (0.998) between duplicate MCMC experiments is obtained again. In other words, even when using an auxiliary sequence that is not exactly decorrelated from one update to the next, the SVE does not appear biased in the present context (although its mixing is mildly degraded). Note that these last sampling settings, with a single Gibbs sweep but five times more update attempts per cycle, in fact have similar computational requirements as previous settings, with five Gibbs sweeps because both amount to performing the same overall number of updates to the auxiliary sequence. We did not observe any differences in the results of the SVE method when using 10 or 50 Gibbs sweeps, indicating such settings to be computationally wasteful.

Combined Data Augmentation and Doubly Intractable Sampling

We finally applied a fully combined MCMC sampler (using both auxiliary variable methods) to draw from the posterior distribution under all models mentioned (see Materials and Methods) for both data sets. The components of the overall sampler are summarized below (see Materials and Methods for details):

- Data augmentation moves: update attempts on substitution mappings, using the MH kernel given in equation (3);
- Doubly intractable moves: update attempts on parameters involved in the stationary distribution, which require (Gibbs) updates to an auxiliary variable sequence used in the MH kernel given in equation (12);

- Plain moves: update attempts on parameters not involved in the stationary distribution, using MH kernel given in equation (13).

As suggested in Robinson et al. (2003) and Rodrigue et al. (2005), we checked that when fixing $\beta = 0$, the approximated posterior distributions match well with the results obtained under site-independent sampling (Rodrigue, Lartillot, and Philippe, 2008) and that all substitution mapping proposals are accepted because the MH ratio cancels out in such conditions. We found the sampler to be reasonably tractable with models based on a fixed or homogeneous ω, requiring about 5 days of CPU time on a Intel P4 3.2 GHz computer. Combined models using both the structural approaches with the Dirichlet process prior on ω, however, are computationally very demanding. The difficulty with these last models is that the sampling methods for updating the configuration of the Dirichlet process were designed to exploit site-specific log-likelihood (or augmented log-likelihood) calculations (see, e.g., Neal 2000; Lartillot and Philippe 2004; Huelsenbeck et al. 2006), whereas under the site-interdependent settings, all calculations are sequence wide. In practice, our sampler spends over 90% of its time updating the Dirichlet process, requiring well over a month of CPU time for obtaining a sample with these data sets. Such combined models are only of moderate interest, however, as we discuss below.

Posterior Distributions

As in previous studies (Robinson et al. 2003; Rodrigue et al. 2005; Rodrigue et al. 2006), we first focus on the posterior distribution of β (for those models that consider it a free parameter), displayed in figure 3 for both data sets. In all cases, we find the distribution to be well above zero, corresponding to the biologically plausible case with evolution favoring sequences that are compatible with the protein tertiary structure (Robinson et al. 2003; Rodrigue et al. 2005; Rodrigue et al. 2006; Choi et al. 2007). Note that with the potential we use here, we expect the posterior distribution of β to be situated around 1/2; the potential was optimized to maximize a probability similar to equation (7), but lacking the nucleotide propensity factors, and with $\beta = 1/2$ (Kleinman et al. 2006). Under the models with fixed or homogeneous ω, the posterior distribution of β is slightly below 1/2; under the MG-SC-β model the 95% credibility intervals are [0.337, 0.463] and [0.361, 0.479] for the GLOBIN*17-144* and ADH*23-254* data sets, respectively, and under the MG-NS-SC-β model, the intervals are [0.310, 0.455] and [0.325, 0.461]. These intervals do not contain the $\beta = 1/2$, which might reflect that the structural features of the protein are slightly at odds with the average structural features of the database used to derive the potential or that the mutational biases, which were not accounted for in the potential's optimization, play an important role, or that yet other model violations are at play. This last scenario is consistent with the observation that when the Dirichlet process on ω is invoked (i.e., under the MG-NS^{DP}-SC-β model), the posterior distributions are found to be around $\beta = 1/2$, with the 95% credibility intervals at [0.493, 0.569] and [0.395, 0.534] for the GLOBIN*17-144* and ADH*23-254* data sets, respectively.

Also noteworthy are the distributions of ω under the MG-NS, MG-NS-SC, and MG-NS-SC-β models. For the GLOBIN*17-144* data sets, for instance, the posterior mean of ω under the MG-NS model is 0.305, indicating that most sites are under strong purifying selection. The SC configurations can also, in theory, capture purifying selection such that we might expect a level of redundancy in the MG-NS-SC and MG-NS-SC-β models. Indeed, we view the combined models (using both NS and SC settings) as confounded in their approach; we would rather adhere to either the phenomenological modeling stance of capturing selective effects using the ω parameter or the mechanistic modeling stance of explaining selective effects using the SC approach. However, if the structural models indeed capture the most important facets of purifying selection, a combined model could exhibit a very different distribution of ω than when using the NS setting alone. Note that the traditional interpretation of ω as the nonsynonymous/synonymous rate ratio no longer holds for combined models and should rather be viewed as the "residual" nonsynonymous/synonymous rate ratio (not captured by the SC approach). The posterior means of ω of 0.354 and 0.349, respectively, under the MG-NS-SC and MG-NS-SC-β models, are mildly higher than under the MG-NS. Nonetheless, the distributions are situated well below 1, indicating that the SC models leave important aspects of purifying selection unaccounted for. The MG-NS^{DP}, MG-NS^{DP}-SC, and MG-NS^{DP}-SC-β models also exhibit only mild differences in posterior distributions. For example, still referring to the GLOBIN*17-144* data set, the posterior mean number of ω classes under the Dirichlet process is approximately 11 with the MG-NS^{DP} model and approximately 10 with the MG-NS^{DP}-SC and MG-NS^{DP}-SC-β models, indicating that much of the heterogeneity of nonsynonymous rates across sites remains unexplained by the structural models. The same trends were observed with the ADH*23-254* data set (see supplementary tables S1–S6 [Supplementary Material online] for posterior distributions of all substitution parameters under all models for both data sets). Together, we interpret these results as first indications that the overall influence of the structural models is mild.

Posterior Predictive Checks

Using our samples from the posterior distribution, we performed simple posterior predictive checks of model adequacy (see Materials and Methods and supplementary figs S2–S13). The statistic for the first test we performed is simply the mean number of nonsynonymous events across sites. The full distributions (from both posterior and predictive mappings) are displayed in the Supplementary Material online, along with graphical displays of the calculation of P values (also see, e.g., Gelman et al. 2004). The list of P values reported in table 1 summarizes the results. We first note that for the MG model, the P value is 0.995. This reveals that the mean number of nonsynonymous events across sites is greater in predictive mappings than it is in posterior mappings, and the discrepancy is indicative of a problem with the model. For the simple MG model, this result is trivially expected: the model essentially assumes no purifying selection, except against stop codons. Therefore,

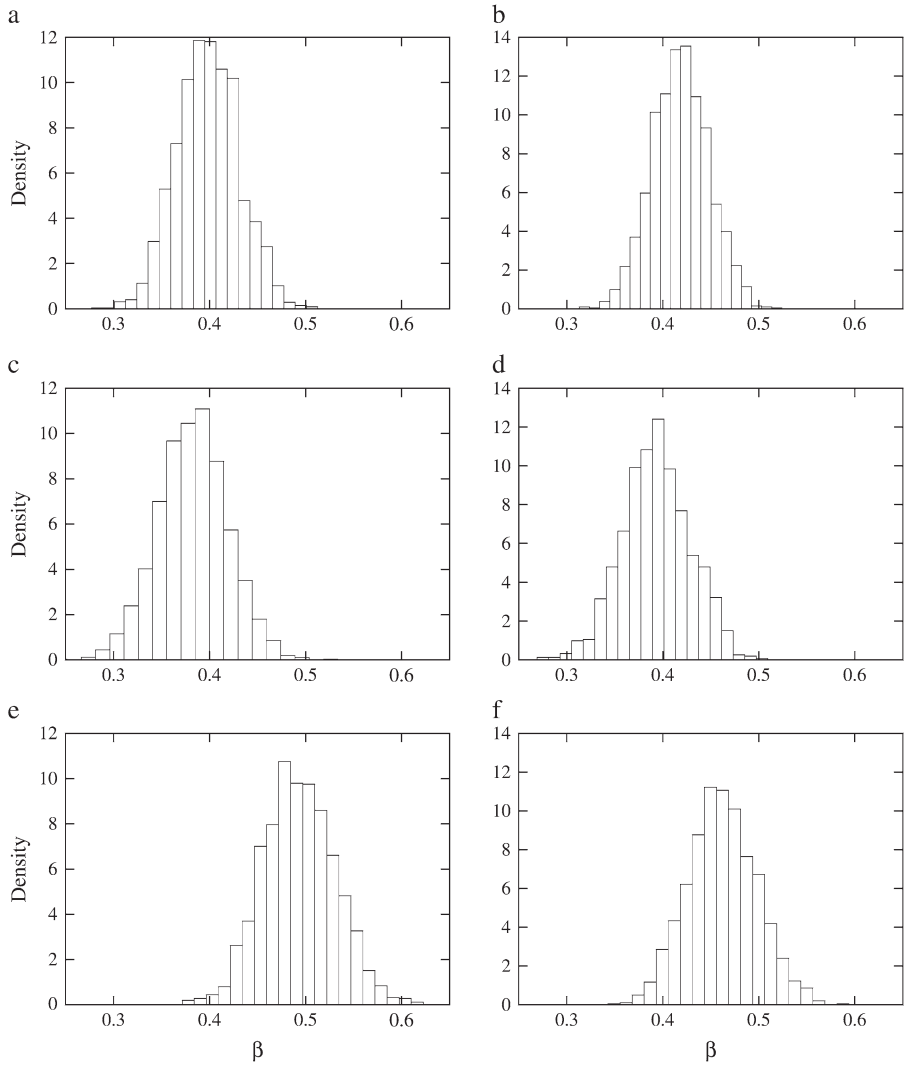

FIG. 3.—Posterior distributions of the β parameter under the MG-SC-β (*a* and *b*), MG-NS-SC-β (*c* and *d*), and MG-NS^{DP}-SC-β (*e* and *f*) models. Left panels are for the GLOBIN*17-144* data set, and right panels are for the ADH*23-254* data set.

it is strongly rejected by protein sequences that are subject to strong purifying selection (Yang 2006; Delport et al. 2008; Anisimova and Kosiol 2009). Specifically here, whereas posterior mappings will be constrained by the data to display a low nonsynonymous/synonymous ratio, predictive mappings under MG will have a ratio of approximately 1.

In contrast with the MG model, under the MG-NS model, which explicitly distinguishes nonsynonymous and synonymous events via the parameter ω, the *P* values obtained are quite low (0.024 and 0.022 for the GLOBIN*17-144* and ADH*23-254* data sets, respectively). This indicates that for this model, the number of nonsynonymous events is lower in predictive mappings than in posterior mappings. One possible explanation for is that the MG-NS model treats all types of nonsynonymous events as equivalent, whereas it is now well understood that most sites typically involve replacements between a small subset of amino acids (e.g., Lartillot and Philippe 2004; Lartillot et al. 2007).

As mentioned above, the SC versions take a more challenging approach, consisting in explicitly modeling selection using the statistical potential, rather than just fitting its effect using ω (without regard to underlying causes). The result in practice, however, is that the MG-SC and MG-SC-β models do not not appear to perform better than the MG model for this test, with *P* values again above the traditional 0.95 threshold. The result is consistent with our analysis of posterior distributions showing that the structural models have only a mild effect. Interestingly, combining the SC and NS settings results in *P* values within the 0.05/0.95 threshold. The phenomenologically motivated MG-NS^{DP} model also performs well for this statistic, with *P* values within the traditional 0.05/0.95 cutoff, as do the combined MG-NS^{DP}-SC and MG-NS^{DP}-SC-β models.

We next performed a posterior predictive check using a second statistic: the variance in the number of nonsynonymous events across sites. The full distributions are again displayed in the Supplementary Material online, and the *P* values are reported in table 2. We first note that

Table 1
Posterior Predictive *P* Values for the Mean Number of Nonsynonymous Events across Sites

Model	GLOBIN*17-144*	ADH*23-254*
MG	0.995	1.000
MG-SC	0.985	1.000
MG-SC-β	1.000	1.000
MG-NS	0.024	0.022
MG-NS-SC	0.082	0.126
MG-NS-SC-β	0.258	0.270
MG-NS^{DP}	0.061	0.118
MG-NS^{DP}-SC	0.434	0.510
MG-NS^{DP}-SC-β	0.448	0.515

Table 2
Posterior Predictive *P* Values for the Variance in the Number of Nonsynonymous Events across Sites

Model	GLOBIN*17-144*	ADH*23-254*
MG	0.000	0.000
MG-SC	0.000	0.008
MG-SC-β	0.000	0.003
MG-NS	0.000	0.000
MG-NS-SC	0.000	0.000
MG-NS-SC-β	0.000	0.000
MG-NS^{DP}	0.356	0.226
MG-NS^{DP}-SC	0.663	0.569
MG-NS^{DP}-SC-β	0.629	0.610

all models with fixed or homogeneous ω have *P* values of zero, or nearly so, indicating that the variance in number of nonsynonymous events across sites is greater in posterior mappings than in predictive mappings. For the MG and MG-NS models, this is again trivially expected: owing to the constraints of the true data, some sites will have more nonsynonymous events than others in order of produce consistent mappings, whereas when simulating predictive mappings, the variance in the number of nonsynonymous events should be low by the definition of the model (which attributes the same nonsynonymous rate to all sites). Once again, in theory, the SC models could induce heterogeneity in the number of nonsynonymous events across sites: structural constraints may vary across sites, leading to an uneven number of nonsynonymous events across sites. Again, in practice, this does not appear to be the case. However, under the MG-NS^{DP} model, the variance in the number of nonsynonymous events is well matched in predictive and posterior mappings and likewise under the combined MG-NS^{DP}-SC and MG-NS^{DP}-SC-β models.

Bayes Factors

Table 3 reports the log-Bayes factors for both data sets. In each case, two values are reported: all calculations were done in duplicates but with different model-switch orientations (see Lartillot and Philippe 2006; Rodrigue et al. 2006; Rodrigue, Lartillot, and Philippe 2008), and we display the lowest and highest values obtained from these procedures. Here, such bidirectional calculations are used to tune the MCMC settings (Rodrigue et al. 2006; Rodrigue, Lartillot, and Philippe 2008). The procedures require about 10 days under the models with homogeneous ω (although crude approximations can be made in a day or two) and over 4 months for models with the Dirichlet process prior, due to the basic MCMC sampling costs described above.

We first note that the most important steps in Bayesian model fit that we observe here are those between the MG and MG-NS models and between the MG-NS and MG-NS^{DP} models. In other words, although the SC and SC-β configurations result in an improved model fit, these configurations in themselves do not provide as good a fit as even the use of a simple global ω parameter. This result is somewhat disappointing (though expected from our posterior predictive analysis) because, as already mentioned, the MG-SC and MG-SC-β models are theoretically attractive in their mechanistic formulation.

We next note that the amelioration brought about by the SC settings is greater when combined with the pure MG model than when combined with the MG-NS model. One possible explanation is that the SC settings mainly capture aspects of purifying (negative) selection, as does the global ω factor (in these cases), such that the combination of the two approaches leads to a level of overlap. This is also consistent with the mild upward shift of the distribution of ω in comparing MG-NS, MG-NS-SC, and MG-NS-SC-β models. In contrast with an overlap effect, the models combining the SC settings with the Dirichlet process on ω produce a synergistic effect; the log-Bayes factor in favor of the MG-NS^{DP}-SC-β model is greater than the sum of log-Bayes factors in favor of MG-NS-SC-β and MG-NS^{DP}. This is likely related to our observation that the posterior mean of β is closer to 1/2 under the MG-NS^{DP}-SC-β model than it is under the MG-NS-SC-β model.

Finally, we note that although the effect is mild, under the SC configurations combined with the Dirichlet process device on ω, setting $\beta = 1/2$ is slightly favored over treating it as a free parameter. In other words, the Bayes factor gives favor to the lower dimensional configuration. Treating β as a free parameter in the present context amounts to purporting having no knowledge about it, beyond its location being between -5 and 5. However, as explained above, we in fact do have knowledge about this parameter's value (i.e., the potentials were originally optimized according to a similar model, with $\beta = 1/2$). This result is also consistent with the posterior distribution inferred for β: the 95% credibility interval straddles the value

Table 3
Natural Logarithm of the Bayes Factor for Models Considered, Computed with MG-NS Used as a Reference

Model	GLOBIN*17-144*	ADH*23-254*
MG	[−92.0, −91.8]	[−319.1, −316.3]
MG-SC	[−22.3, −21.8]	[−220.8, −217.7]
MG-SC-β	[−22.4, −21.7]	[−221.7, −218.6]
MG-NS	—	—
MG-NS-SC	[48.9, 49.3]	[58.1, 58.6]
MG-NS-SC-β	[49.2, 49.6]	[57.9, 58.4]
MG-NS^{DP}	[102.2, 104.2]	[96.8, 100.3]
MG-NS^{DP}-SC	[185.7, 188.4]	[177.7, 181.6]
MG-NS^{DP}-SC-β	[180.9, 183.8]	[173.5, 177.4]

NOTE.—Values given are the upper and lower estimates from bidirectional thermodynamic integrations.

$\beta = 1/2$, in effect inferring the value that we expect from the original scaling of the potential. The mildness of the dimensionality penalty reflected in the Bayes factor can be expected from the fact that these models differ by a single degree of freedom. It may nevertheleess be worthwhile to treat β as a free parameter as a preliminary exploration of possible model violations (as we have done here), at least until the parameters of the potentials themselves can be free within the overall framework.

Conclusions and Future Directions

Overall, our analysis comes to similar conclusions as previous works concerned only with amino acid data (Rodrigue et al. 2006; Rodrigue et al. 2007): models employing simplified structural representations and empirical potentials provide an improved fit but are outperformed by the sophisticated site-independent models of codon substitution. In particular here, the Dirichlet process prior on the nonsynonymous rate factor markedly outperforms the pure site-interdependent SC model. As discussed in Thorne et al. (2007), assigning a meaningful interpretation to ω, let alone with the Dirichlet process, can be difficult, and other directions inspired from population genetic theory should be explored. In the meantime, given that the Dirichlet process modeling proposed by Huelsenbeck et al. (2006) provides an important improvement in model fit and is able to reproduce observed nonsynonymous rate features, it can serve as a phenomenological benchmark, to which biologically motivated alternatives can be compared.

In particular, efforts should now be made to ameliorate the basic form of functions used as proxies for sequence fitness. For instance, structural representations could include other features, such as main-chain dihedral angles (e.g., Betancourt and Skolnik 2004), or consider different contact classes (e.g., based on different distance categories or by distinguishing between the context of contacts—say, solvent exposed contacts as opposed to buried contacts). Other extensions unrelated to protein structure could also be combined, such as a modeling of codon usage preference (Rodrigue, Lartillot, and Philippe 2008; Yang and Nielsen 2008), or context-dependent mutational features (e.g., Baele et al. 2008). We are currently exploring the use of the methods as a means of merging the protein design framework proposed in Kleinman et al. (2006) within the phylogenetic modeling we perform here (along with extensions just mentioned), with the parameters of the potential themselves jointly inferred, along with other parameters of the model. Though such a model will likely require large data sets—of the sort used to construct amino acid replacement matrices (e.g., Whelan and Goldman 2001)—and be computationally demanding, it is not unfeasible and could allow for a more definitive evaluation of the strengths and limitations of alternative functional forms of sequence fitness proxies.

Several other directions could be taken from present work. Exploiting the posterior predictive framework we use here, future studies could consider a broader range of test statistics, such as the proportion of various types of nonsynonymous events (e.g., between specific amino acid pairs), the relative timing of different nonsynonymous events in relation to their location in the protein structure, as well as statistics that are a function of synonymous events in the detailed mappings (and hence potentially motivating further model expansions; e.g., Pond and Muse 2005). Regarding the data augmentation sampling methods, the use of other methods and substitution models to propose mappings (e.g., Hobolth and Stone 2008; Minin and Suchard 2008) could be explored, and we are currently studying mapping proposal methods that do not rely on any substitution model. Regarding the doubly intractable sampling methods, we are currently investigating if other techniques inspired from statistical physics (e.g., Propp and Wilson 1996) could be adapted to perform exact sampling from equation (7). The methods explored here may also serve as a stepping-stone toward the types of Laplace approximations used in Rodrigue et al. (2007) for more economical computations, which would in turn enable quantitative large-scale studies of the factors influencing protein-coding sequence evolution.

Supplementary Material

Supplementary figures S1–S13 and tables S1–S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We wish to thank Stéphane Aris-Brosou, Vladimir Minin, and two anonymous reviewers for comments on the manuscript. This work was supported by the Natural Sciences and Engineering Research Council of Canada, the biT fellowships for excellence (a Canadian Institutes of Health Research strategic training program grant in bioinformatics), the Robert Cedergren Centre for bioinformatics and genomics, and the Canadian Research Chair Program.

Appendix: Site-Specific Codon Substitution Models

Our implementation constructs the entries of site-specific Markov generators $Q^{(i)} = [Q^{(i)}_{ab}]$ from two sets of specifications: a 61-dimensional vector of "stationary probabilities", $\pi^{(i)} = (\pi^{(i)}_a)_{1 \leqslant a \leqslant 61}$, and a set of "transient specification", $\rho^{(i)} = (\rho^{(i)}_{ab})_{1 \leqslant a,b \leqslant 61}$. The entries are given as

$$Q^{(i)}_{ab} = \rho^{(i)}_{ab} \pi^{(i)}_b, \quad a \neq b, \tag{14}$$

$$Q^{(i)}_{aa} = -\sum_{b \neq a} Q^{(i)}_{ab}. \tag{15}$$

The site-specific stationary probabilities are given by

$$\pi^{(i)}_a = \frac{\varphi_{a_1} \varphi_{a_2} \varphi_{a_3} \, \mathrm{e}^{-2\beta G_i(a)}}{\sum_{b=1}^{61} \varphi_{b_1} \varphi_{b_2} \varphi_{b_3} \, \mathrm{e}^{-2\beta G_i(b)}}. \tag{16}$$

The transient specifications are given by

$$\rho_{ab}^{(i)} = \begin{cases} \dfrac{\varrho_{a_c b_c}}{\varphi_{a_{c'}} \varphi_{a_{c''}} \mathrm{e}^{-2\beta G_i(a)}}, & \text{if } a \text{ and } b \text{ are synonymous and differ only at } c\text{th codon position,} \\ \dfrac{\omega \varrho_{a_c b_c}}{\varphi_{a_{c'}} \varphi_{a_{c''}} \mathrm{e}^{-\beta G_i(a)} \mathrm{e}^{-\beta G_i(b)}}, & \text{if } a \text{ and } b \text{ are nonsynonymous and differ only at } c\text{th codon position,} \\ 0, & \text{otherwise,} \end{cases} \quad (17)$$

where c' and c'' are the two nucleotide positions that are not involved in the event. Substituting equation (16) and equation (17) into equation (14) yields the model given in equation (4).

Literature Cited

Anisimova M, Kosiol C. 2009. Investigating protein-coding sequence evolution with probabilistic codon substitution models. Mol Biol Evol. 26:255–271.

Baele G, Van de Peer Y, Vansteelandt S. 2008. A model-based approach to study nearest-neighbor influences reveals complex substitution patterns in non-coding sequences. Syst Biol. 57:675–692.

Betancourt MR, Skolnik J. 2004. Local propensities and statistical potentials of backbone dihedral angles in proteins. J Mol Biol. 342:635–649.

Bollback JP. 2005. Posterior mapping and posterior predictive distributions. In: Nielsen R, editor. Statistical methods in molecular evolution. New York: Springer. p. 439–462.

Choi SC, Hobolth A, Robinson DM, Kishino H, Thorne JL. 2007. Quantifying the impact of protein tertiary structure of molecular evolution. Mol Biol Evol. 24:1769–1782.

Dahl FA. 2006. On the conservativeness of posterior predictive p-values. Stat Probab Lett. 76:1170–1174.

Delport W, Scheffer K, Seoighe C. 2009. Models of coding sequence evolution. Brief Bioinform. 10:97–109.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2004. Bayesian data analysis. Boca Raton (FL): Chapman and Hall.

Gelman A, Meng XL, Stern H. 1996. Posterior predictive assessment of model fitness via realised discrepancies. Stat Sin. 6:733–807.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 11:725–736.

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

Hobolth A, Stone EA. 2008. Markov chain Monte Carlo expectation maximization algorithm for statistical analysis of DNA sequence evolution with neighbour-dependent substitution rates. J Comput Graph Stat. 17:138–164.

Huelsenbeck JP, Jain S, Frost SWD, Pond SLK. 2006. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. Proc Natl Acad Sci USA. 103: 6263–6268.

Jones DT, Taylor WR, Thornton JM. 1992. A new approach to protein fold recognition. Nature. 358:86–89.

Kleinman CL, Rodrigue N, Bonnard C, Philippe H, Lartillot H. 2006. A maximum likelihood framework for protein design. BMC Bioinform. 7:326.

Lartillot N. 2006. Conjugate Gibbs sampling for Bayesian phylogenetic models. J Comput Biol. 13:1701–1722.

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol 7:S4.

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

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

Meng XL. 1994. Posterior predictive p-values. Ann Stat. 22:1142–1160.

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

Minin VN, Suchard MA. 2008. Counting labeled transitions in continuous-time Markov model of evolution. J Math Biol. 56:391–412.

Miyazawa S, Jernigan RL. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules. 18: 534–552.

Murray I, Ghahramani Z, MacKay DJC. 2006. MCMC for doubly-intractable distributions. In: Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence. Available from: http://uai.sis.pitt.edu/papers/06/170_paper.pdf.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitutions, with applications to chloroplast genome. Mol Biol Evol. 11:715–724.

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

Nielsen R. 2002. Mapping mutations on phylogenies. Syst Biol. 51:729–739.

Parisi G, Echave J. 2001. Structural constraints and emergence of sequence patterns in protein evolution. Mol Biol Evol. 18:750–756.

Pond SK, Muse SV. 2005. Site-to-site variation of synonymous substitution rates. Mol Biol Evol. 22:2375–2385.

Propp J, Wilson D. 1996. Exact sampling with couple Markov chains and applications to statistical mechanics. Random Struct Algorithms. 9:223–252.

Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol. 18:1692–1704.

Rodrigue N, Lartillot N, Bryant D, Philippe H. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene. 347:207–217.

Rodrigue N, Lartillot N, Philippe H. 2008. Bayesian comparisons of codon substitution models. Genetics. 180:1579–1591.

Rodrigue N, Philippe, Lartillot N. 2006. Assessing site-interdependent phylogenetic models of sequence evolution. Mol Biol Evol. 23:1762–1775.

Rodrigue N, Philippe H, Lartillot N. 2007. Exploring fast computational strategies for probabilistic phylogenetic analysis. Syst Biol. 56:711–726.

Rodrigue N, Philippe H, Lartillot N. 2008. Uniformization for sampling realizations of Markov processes: applications to Bayesian implementations of codon substitution models. Bioinformatics. 24:56–62.

Seno F, Micheletti C, Martian A. 1998. Variational approach to protein design and extraction of interaction potentials. Phys Rev Lett. 81:2172–2175.

Shakhnovich EI, Gutin AM. 1993. Engineering of stable and fast-folding sequences of model proteins. Proc Natl Acad Sci USA. 90:7195–7199.

Sun S, Bren R, Chan R, Dill K. 1995. Designing amino acid sequences to fold with good hydrophobic cores. Protein Eng. 8:1205–1213.

Thorne JL. 2007. Protein evolution constraints and the model-based techniques to study them. Curr Opin Struct Biol. 17:337–341.

Thorne JL, Choi SC, Yu J, Higgs PG, Kishino H. 2007. Population genetics without intraspecific data. Mol Biol Evol. 24:1667–1677.

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: 691–699.

Yang Z. 2006. Computational molecular evolution. Oxford Series in Ecology and Evolution. Oxford: Oxford University Press.

Yang Z, Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol. 25:568–579.

Yang Z, Nielsen R, Goldman N, Pedersen AMK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 155:431–449.

Jeffrey Thorne, Associate Editor

Accepted April 14, 2009