DOI: 10.1534/genetics.108.092254

Bayesian Comparisons of Codon Substitution Models

Nicolas Rodrigue,*,†,[1] Nicolas Lartillot* and Hervé Philippe*

**Canadian Institute for Advanced Research, Département de Biochimie, Université de Montréal, Montréal, Québec H3C 3J7, Canada and* †*Department of Biology, Center for Advanced Research in Environmental Genomics, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada*

Manuscript received June 3, 2008
Accepted for publication September 5, 2008

ABSTRACT

In 1994, Muse and Gaut (MG) and Goldman and Yang (GY) proposed evolutionary models that recognize the coding structure of the nucleotide sequences under study, by defining a Markovian substitution process with a state space consisting of the 61 sense codons (assuming the universal genetic code). Several variations and extensions to their models have since been proposed, but no general and flexible framework for contrasting the relative performance of alternative approaches has yet been applied. Here, we compute Bayes factors to evaluate the relative merit of several MG and GY styles of codon substitution models, including recent extensions acknowledging heterogeneous nonsynonymous rates across sites, as well as selective effects inducing uneven amino acid or codon preferences. Our results on three real data sets support a logical model construction following the MG formulation, allowing for a flexible account of global amino acid or codon preferences, while maintaining distinct parameters governing overall nucleotide propensities. Through posterior predictive checks, we highlight the importance of such a parameterization. Altogether, the framework presented here suggests a broad modeling project in the MG style, stressing the importance of combining and contrasting available model formulations and grounding developments in a sound probabilistic paradigm.

CODON-BASED Markovian substitution models are widely recognized as attractive descriptions of molecular evolution (Yang 2006). There has also been an increasing recognition that some of these modeling approaches can be attributed population genetic interpretations (Thorne *et al.* 2007), and with the emerging juncture of population genetics and molecular evolutionary modeling (Felsenstein 2007), the codon-based framework could have much to contribute to these domains. A panoply of codon substitution models are now available, most of which consist of modifications and extensions of the seminal works of Muse and Gaut (1994) (MG) and Goldman and Yang (1994) (GY).

Suppose an alignment of nucleotide sequences, related according to a given phylogenetic tree. The more traditional level of interpretation surmises each nucleotide column of the alignment as arising from an independent continuous-time Markov process running over the tree and with a state space consisting of the four different nucleotides. In its general reversible form (*e.g.*, Lanave *et al.* 1984), the model is represented using six relative exchangeability parameters (with 5 effective d.f.), for each possible (unordered) pair of nucleotides, and four nucleotide propensities (with 3 effective d.f.), and is often referred to as the *general time reversible* (GTR) model. Taking this model as a starting point in the case of protein-coding sequences, a first step to mechanistically acknowledging the coding nature of the data is to suppose a strong purifying selection against stop codons; the process is reformulated in a state space consisting of nucleotide triplets, omitting triplet states corresponding to stops. In effect, such a model is equivalent to the same GTR type of model applied at the nucleotide level, but with the constraint that the nucleotide sequence must encode some full-length amino acid sequence (one-third the length of the nucleotide sequence). This is the rationale of the codon substitution models in the style of MG. Importantly, it implies that the rates of substitution (entries of the generator of the Markov processes) are proportional to the *target nucleotide* (at the mutating position). From this point, a further model construction step in the MG style is to distinguish between synonymous and nonsynonymous events, for instance utilizing the parameterization presented in the original work or the more compact representation of fixing the synonymous rate factor at one, with a free nonsynonymous rate factor.

In contrast with the MG style of model formulation, the GY models have entries of the Markov generator proportional to the stationary probability of the *target codon*. The contrast between the two formulations can be made very subtle. Indeed, a GY-style model can be specified from the same six nucleotide relative exchangeability and four nucleotide propensity parameters used in the MG-style model above. In such a case, codon

[1]*Corresponding author:* Department of Biology, University of Ottawa, 30 Marie Curie, Ottawa, ON K1N 6N5, Canada.
E-mail: nicolas.rodrigue@uottawa.ca

stationary probabilities are approximated as proportional to the product of the three propensity parameter values associated with the nucleotides at each of the three codon positions (however, such a model entails peculiar properties, as we discuss in later sections). Another option available in the GY style is based on a full 61-dimensional (with 60 effective d.f., assuming the universal genetic code) vector of codon stationary probabilities (F61) (*e.g.*, Huelsenbeck and Dyer 2004; Huelsenbeck *et al.* 2006; Yang 2006). The GY-F61 approach has been suggested as important in giving "more freedom for the model to explain the data by modifying substitution rates using codon frequencies" (Huelsenbeck and Dyer 2004, p. 670). This may be the case, but the GY-F61 model has no natural mechanistic interpretation at the nucleotide level; nucleotide propensities have no direct parameterization in this formulation, and are only implicitly modeled, in a manner confounded with other effects inducing uneven codon stationary probabilities.

Another widely used modeling idea, adopted in both MG and GY formulations, has been to assign a separate set of nucleotide propensity parameters to each of the three codon positions. The distinction with the previously mentioned models is commonly referred to as F1 $\times$ 4 *vs.* F3 $\times$ 4, reflecting the use of a single *vs.* three vectors of dimension 4 (3 d.f.). However, the F3 $\times$ 4 configuration stands only as a phenomenological account of how the coding structure of the data induces a periodic pattern along the nucleotide sequence; there is no mechanistic sense to modeling features induced by selection via an expanded parameterization at the nucleotide level. In other words, this expanded parameterization may capture net resultants of the coding nature of the sequences, but it is not representative of our understanding of the causative factors bearing on the evolutionary process. Differences observed at each of the three codon positions are most likely the result of factors bearing on amino acid or codon preferences, or other high-order features, and should logically be modeled as such.

Recently, Yang and Nielsen (2008) studied models that address these issues, with explicit parameterizations bearing on amino acid or codon preferences within the MG formulation. From their analysis of mammalian genes, they find support for their formulation and hence further validate a modeling strategy aimed at disentangling mutational and selective features (Thorne 2007; Thorne *et al.* 2007; Yang and Nielsen 2008).

However, the ongoing developments of codon substitution models need to be complemented with rigorous evaluation and selection approaches, to contrast the relative merit of different modeling ideas. The Bayes factor (Jeffreys 1935) provides a natural framework for this purpose, with the theoretical advantage of being applicable to the comparison of any pair of models; it does not require models to be nested, and it intrinsically penalizes for higher-dimensional formulations. Bayes factors have already been used for comparing Markovian models of substitution at either the nucleotide level (*e.g.*, Suchard *et al.* 2001) or the amino acid level (*e.g.*, Lartillot and Philippe 2006). As discussed in Lartillot and Philippe (2006), reliable evaluations of Bayes factors for complex evolutionary models require elaborate computational devices. These computational demands are compounded at the codon level of interpretation, where the state space invoked leads to more taxing likelihood calculations than at the nucleotide or amino acid levels. Nonetheless, the application of Monte Carlo procedures running on modern computing machines makes these calculations feasible, thus allowing for broad codon substitution model comparisons.

In this work, we conduct such a Bayesian analysis of model fit and include all of the above-mentioned MG- and GY-style models, comparing the F1 $\times$ 4, F3 $\times$ 4, and F61 (in the GY context) configurations. We also include similar versions to the models of Yang and Nielsen (2008), which allow for a flexible account of either global amino acid preferences or global codon preferences. Each of these model configurations is contrasted in turn with a modeling of nonsynonymous rate heterogeneity, using the *Dirichlet process* device (Ferguson 1973) as described by Huelsenbeck *et al.* (2006). From our analysis of three real data sets, our findings indicate that a mechanistic MG-style modeling strategy that explicitly recognizes uneven codon preferences, while accounting for a global background of nucleotide propensities and heterogeneous nonsynonymous substitution rates, tends to outperform other models (or match top performing models). We explore aspects of the posterior distributions under the top models and conduct posterior predictive assessments (Rubin 1984; Gelman *et al.* 1996, 2004) highlighting implicit properties of the models.

DATA

We used the following data sets, which we refer to here using a shorthand indicating the number of sequences and their length in number of codons: Globin *17-144*, 17 vertebrate sequences of the β-globin gene, described in Yang *et al.* (2000a); Lysin *25-134*, 25 abalone sperm lysin sequences, described in Yang *et al.* (2000b); and Hiv *22-99*, 22 sequences of the human immunodeficiency virus type 1 protease, described in Doron-Faigenboim and Pupko (2007). For computational reasons, we worked under a fixed tree throughout and, in all three cases, used the same topologies as those used in the works cited for each data set.

MODELS

The models considered here are defined according to continuous-time Markov processes, with the infinitesimal generators written as $Q = [Q_{ij}]$, specifying the

instantaneous rate of substitution from codon i to codon j—we write i_c (j_c) to refer to the nucleotide state at position c of codon i (j). The matrix Q (under the universal genetic code with stop codons excluded) is 61×61 (*i.e.*, $1 \le i, j \le 61$, and $1 \le c \le 3$). It can be significantly simplified by viewing all substitutions as resulting from point mutations, such that each entry either is set to 0 (when states differ by two or three nucleotides) or is constructed from low-dimensional components. In the next sections, we describe these components in detail, constructing the entries of Q following modeling approaches inspired from Muse and Gaut (1994) and Goldman and Yang (1994), with extensions proposed by Huelsenbeck *et al.* (2006) and Yang and Nielsen (2008). The models are not identical to those presented in these original works, but correspond to flexible generalizations, while allowing us to focus on the distinguishing features of interest.

MG-style models: We begin with the mechanistic modeling standpoint proposed by Muse and Gaut (1994), with a Markov generator given by

$$Q_{ij} \propto \begin{cases} \varrho_{i_c j_c} \varphi_{j_c}, & \text{if } \mathcal{A}, \\ \omega \varrho_{i_c j_c} \varphi_{j_c}, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where

$\mathcal{A}$: i and j are synonymous and differ only at codon position c;

$\mathcal{B}$: i and j are nonsynonymous and differ only at codon position c;

$\varrho = (\varrho_{ab})_{1 \le a,b \le 4}$ is a set of (symmetrical) nucleotide relative exchangeability parameters, with the (arbitrary) constraint $\sum_{1 \le a < b \le 4} \varrho_{ab} = 1$; $\varphi = (\varphi_a)_{1 \le a \le 4}$, with $\sum_{a=1}^{4} \varphi_a = 1$, represents a set of global nucleotide equilibrium propensities; and ω is the coefficient bearing on nonsynonymous rates, for now treated as a global parameter.

When $\omega = 1$, this model corresponds to the well-known GTR model invoked for nucleotide-level interpretations, but with the purifying constraint against all stop codons. Here, however, ω is always treated as a free parameter, and the model is referred to as MG-F1 $\times$ 4.

Following in the MG style, Yang and Nielsen (2008) introduced a model with parameters bearing on codon fitness. A convenient parameterization of equivalent dimensionality is given as

$$Q_{ij} \propto \begin{cases} \varrho_{i_c j_c} \varphi_{j_c} \left(\frac{\psi_j}{\psi_i}\right)^{1/2}, & \text{if } \mathcal{A}, \\ \omega \varrho_{i_c j_c} \varphi_{j_c} \left(\frac{\psi_j}{\psi_i}\right)^{1/2}, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $\psi = (\psi_j)_{1 \le j \le 61}$, with $\sum_{j=1}^{61} \psi_j = 1$, represents a set of 61 codon preference parameters (60 d.f.). The exponent $\frac{1}{2}$ ensures reversibility (see the appendix), although the model could be expanded at this level as well (see Goldman and Whelan 2002). As it is, entries corresponding to substitutions from an unpreferred codon to a preferred codon ($\psi_j/\psi_i > 1$) will be higher than entries corresponding to substitutions from a preferred to an unpreferred codon ($\psi_j/\psi_i < 1$), and in this way, an explicit account of global codon preference (CP) is included while maintaining an account of background nucleotide propensities. We refer to this model as MG-F1 $\times$ 4-CP.

Note that the codon preferences captured by ψ can be the result of several factors, including, for instance, global amino acid preferences. One way of assessing whether the CP model is capturing effects beyond those of global amino acid preferences is to compare it with a simplified version of the CP formulation, which accounts only for such features as given by

$$Q_{ij} \propto \begin{cases} \varrho_{i_c j_c} \varphi_{j_c}, & \text{if } \mathcal{A}, \\ \omega \varrho_{i_c j_c} \varphi_{j_c} \left(\frac{\phi_{f(j)}}{\phi_{f(i)}}\right)^{1/2}, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where $\phi = (\phi_k)_{1 \le k \le 20}$ is a 20-dimensional (19-d.f.) vector associated with amino acid preferences (AAP), and where $f(i)$ returns an index corresponding the amino acid encoded by codon i. As in the case of the CP model, entries corresponding to substitutions from an unpreferred amino acid to a preferred amino acid ($\phi_{f(j)}/\phi_{f(i)} > 1$) will thus be higher than entries corresponding to substitutions from a preferred to an unpreferred amino acid ($\phi_{f(j)}/\phi_{f(i)} < 1$). We refer to this model as MG-F1 $\times$ 4-AAP.

Finally, despite departing from the mechanistic modeling perspective, we also investigate the F3 $\times$ 4 configurations for the models defined in (1)–(3), by substituting φ appropriately with codon position-specific nucleotide propensity parameters, written as $\varphi^{(c)} = (\varphi_a^{(c)})_{1 \le a \le 4}$, where $\forall c, 1 \le c \le 3$, $\sum_{a=1}^{4} \varphi_a^{(c)} = 1$. The MG-F3 $\times$ 4 model is thus written as

$$Q_{ij} \propto \begin{cases} \varrho_{i_c j_c} \varphi_{j_c}^{(c)}, & \text{if } \mathcal{A}, \\ \omega \varrho_{i_c j_c} \varphi_{j_c}^{(c)}, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

the MG-F3 $\times$ 4-CP model as

$$Q_{ij} \propto \begin{cases} \varrho_{i_c j_c} \varphi_{j_c}^{(c)} \left(\frac{\psi_j}{\psi_i}\right)^{1/2}, & \text{if } \mathcal{A}, \\ \omega \varrho_{i_c j_c} \varphi_{j_c}^{(c)} \left(\frac{\psi_j}{\psi_i}\right)^{1/2}, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

and the MG-F3 $\times$ 4-AAP model as

$$Q_{jj} \propto \begin{cases} \varrho_{i_c j_c} \varphi^{(c)}_{j_c}, & \text{if } \mathcal{A}, \\ \omega \varrho_{i_c j_c} \varphi^{(c)}_{j_c} \left(\frac{\phi_{f(j)}}{\phi_{f(i)}}\right)^{1/2}, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

GY-style models: The models in the style proposed by Goldman and Yang (1994) have Markov generators specified as

$$Q_{jj} \propto \begin{cases} \varrho_{i_c j_c} \pi_j, & \text{if } \mathcal{A}, \\ \omega \varrho_{i_c j_c} \pi_j, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise,} \end{cases} \quad (7)$$

where $\pi = (\pi_i)_{1 \le i \le 61}$, with $\sum_{i=1}^{61} \pi_i = 1$, represents a 61-dimensional (60 d.f.) vector of codon stationary probabilities (distinct from ψ).

Several options for π are available. First, it can be based on a set of global nucleotide propensity parameters according to

$$\pi_i \propto \varphi_{i_1} \varphi_{i_2} \varphi_{i_3}. \quad (8)$$

We refer to this model as GY-F1 × 4. Careful examination of this model reveals a number of peculiar properties, which seem undesirable. For instance, in a mutational context prone to events leading to A or T (*i.e.*, where the parameters φ_A and φ_T tend to be higher than φ_C and φ_G) a substitution from codon CGC to CTC would have a lower instantaneous rate than a substitution from codon ATA to AGA. In such a scenario, the rate of an event involving the second codon position depends on the nucleotide states at the first and third positions, which, in this case, leads to the higher rate for the substitution going against the mutational bias. From the mechanistic model construction described by the MG strategy, however, there are no obvious reasons for linking a change at the second position to the states at the first and third positions, unless this is mediated by selective effects at the codon level (*e.g.*, stop codons). Accordingly, for this same instance, the MG model displays the reverse situation, with the CGC to CTC substitution having a higher instantaneous rate than the ATA to AGA substitution in a manner consistent with the mutational bias.

Another similar choice for the GY models is to base π on codon-position-specific nucleotide equilibrium frequencies,

$$\pi_i \propto \varphi^{(1)}_{i_1} \varphi^{(2)}_{i_2} \varphi^{(3)}_{i_3}, \quad (9)$$

in which case we refer to the model as GY-F3 × 4. Note that the GY-F1 × 4 and MG-F1 × 4 models, as well as the GY-F3 × 4 and MG-F3 × 4 models, are respectively constructed from the exact same parameters; they also have the same limiting distributions and hence differ only in terms of their transient specifications (further details on this point are given in Yang and Nielsen 2008, as well as in the appendix). Finally, we consider the case where π is directly free—a full 61-dimensional (60-d.f.) vector—which we refer to as GY-F61.

The limiting distributions of all models are given in full in the appendix, along with further details specific to our implementation.

Priors: Our prior on branch lengths is *exponential*, with a mean determined by a hyperparameter, itself endowed with an exponential prior of mean 1. Adopting the approach presented by Huelsenbeck *et al.* (2006), our most general prior on nonsynonymous rate factors is the Dirichlet process (DP)—as an infinite mixture across sites—with hyperparameter α, modulating the assumed "graininess" of nonsynonymous heterogeneity; α is endowed with an exponential prior of mean 1. The Dirichlet process prior also utilizes a base measure, defining the probability distribution of each component; as in Huelsenbeck *et al.* (2006), we use $p(\omega) = 1/(1 + \omega)^2$, the probability density of the ratio of two identically distributed draws from an exponential. This same base prior is used when dispensing with the DP framework, with the model based on a single global ω-factor. All other parameters have flat Dirichlet priors on their respective state space.

MODEL COMPARISONS

Bayes factors: Given a data set D and a model M, specified by some high-dimensional parameter vector $\theta \in \Theta$, we wish to evaluate the *predictive probability*, or *marginal likelihood*, written as $p(D \mid M)$ and obtained by averaging the likelihood $p(D \mid \theta, M)$ over the prior $p(\theta \mid M)$:

$$p(D \mid M) = \int_\Theta p(D \mid \theta, M) p(\theta \mid M) d\theta. \quad (10)$$

When comparing two models of interest, M_0 and M_1, the Bayes factor (B_{01}) provides a measure of the *evidence* in favor of one model over the other:

$$B_{01} = \frac{p(D \mid M_1)}{p(D \mid M_0)} \quad (11)$$

(Jeffreys 1935). A Bayes factor $>$ ($<$) 1 is considered as evidence in favor of M_1 (M_0).

We used the model-switch thermodynamic integration framework described in Lartillot and Philippe (2006) to evaluate Bayes factors across all codon substitution models described above. The model-switch thermodynamic integration method is a computationally intensive Markov chain Monte Carlo procedure, and our strategy here has been to evaluate all log Bayes factors with respect to the GY-F1 × 4 model. Several log Bayes factors are computed from multiple independent model-switch schemes. For instance, the log Bayes factor between MG-F3 × 4-CP-DP and GY-F1 × 4 is assembled from

$$\ln \frac{p(D \mid \text{MG-F3} \times 4\text{-CP-DP})}{p(D \mid \text{GY-F1} \times 4)} = \ln \frac{p(D \mid \text{MG-F3} \times 4\text{-CP-DP})}{p(D \mid \text{MG-F3} \times 4\text{-CP})} + \ln \frac{p(D \mid \text{MG-F3} \times 4\text{-CP})}{p(D \mid \text{MG-F3} \times 4)} + \ln \frac{p(D \mid \text{MG-F3} \times 4)}{p(D \mid \text{GY-F3} \times 4)} + \ln \frac{p(D \mid \text{GY-F3} \times 4)}{p(D \mid \text{GY-F1} \times 4)}, \quad (12)$$

where each term is evaluated using a distinct model-switch scheme (see supplemental materials). Following Lartillot and Philippe (2006), we found the sampling error to be less problematic than the "thermic lag" of the method, which we thus made the focus of our MCMC settings. We ran each calculation in duplicate, using the quasi-static bidirectional method detailed in Lartillot and Philippe (2006, also see the supplemental material). Each pair of model-switch integrations produces two values, used here to construct an interval, and gives a crude sense of the precision of the settings. Note that the bidirectional scheme is employed as a framework for empirical MCMC tuning (see supplemental material) and not as a method for formal calculation of the Monte Carlo error.

As expected, the log Bayes factors can be made more precise when comparing models that are parametrically closest to the reference model (see solid bands in Figure 1). When computing log Bayes factors for the more complex models, involving several distinct model-switch schemes, the interval of the overall log Bayes factor against GY-F1 × 4 is constructed conservatively (to produce the broadest possible interval). This means that the interval (the error) grows with each step in model space. However, a careful model-space traversal appears necessary in practice, despite this growing error of multiple steps, as performing a direct integration between a high-dimensional model and the reference would require overcostly computations (see supplemental material). Still, in certain cases, an entirely unambiguous model ranking remains computationally prohibitive. For instance, with the Globin *17-144* data, the log Bayes factor of MG-F3 × 4-CP-DP against GY-F1 × 4, and the log Bayes factor of MG-F1 × 4-CP-DP against GY-F1 × 4, overlap with each other, which thus prevents us from clearly distinguishing the two models. In the present context, obtaining the required level of precision for distinguishing between log-marginal likelihoods that differ by a few units is relatively uninteresting and not worth the computational investment that would be needed when utilizing the present methods. Our objective here is rather to map out the main effects of different formulations in terms of overall model fit.

Posterior predictive assessments: The calculation of Bayes factors allows for a quantitative ranking of a competing set of models. However, we want to explore how the models perform in absolute terms, by comparing the value of certain statistics computed on the real data with those computed on data replicates, generated by simulation under the model of interest. In the Bayesian framework, the simulation is typically conducted for each draw of a sample from the posterior, and the distribution of the statistic is referred to as the *posterior predictive distribution*. This designation is appropriate, in that the distribution consists of what we would expect to see under the model, in light of what has been learned from the data at hand. Discrepancies between posterior predictive distributions and observed values reveal weaknesses of a model, as previously explored in phylogenetic modeling contexts (*e.g.*, Bollback 2002; Nielsen 2002; Lartillot *et al.* 2007). A conventional way of summarizing the discrepancy between a posterior predictive distribution and the observed value of some statistic is to report the area of the distribution to the right of the observed value, otherwise referred to as the posterior predictive *P*-value (see, *e.g.*, Gelman *et al.* 2004); this can be computed as the proportion of replicates for which the statistic equals or exceeds the observed value. We illustrate this approach here using codon-position-specific nucleotide frequencies and codon entropy.

Note that formulating tests based on the posterior predictive *P*-values to "reject" models would require further calibration (since the data replicates were generated from unknown, previously estimated parameter values). In the case of the nucleotide frequencies, such a formal testing across all dimensions would require further correction (since dimensions are not independent). Nonetheless, these types of difficulties do not prevent us from performing simple graphical contrasts of features of true and replicated data (Gelman *et al.* 2004).

RESULTS AND DISCUSSION

Bayes factors: The series of log Bayes factors reported in Figure 1 reveals considerable differences in model performance, indicating the importance of conducting a careful examination of alternative parametric choices. We note that the MG-F1 × 4-CP-DP model is among the top ranking models for all three data sets. This result is somewhat expected. First, nonsynonymous rate heterogeneity has now been observed across numerous data sets (Yang 2006), and it thus seems reasonable to anticipate a good performance of the Dirichlet process framework proposed by Huelsenbeck *et al.* (2006). For the Globin *17-144* data set, we also tried a version of the DP model with a uniform hyperprior over the interval [0:1000] on α, in the MG-F1 × 4-CP-DP setting. With respect to the reference model, the resulting log Bayes factor interval of [237.8; 242.3] is overlapping with the top-ranking models, suggesting that model choice in

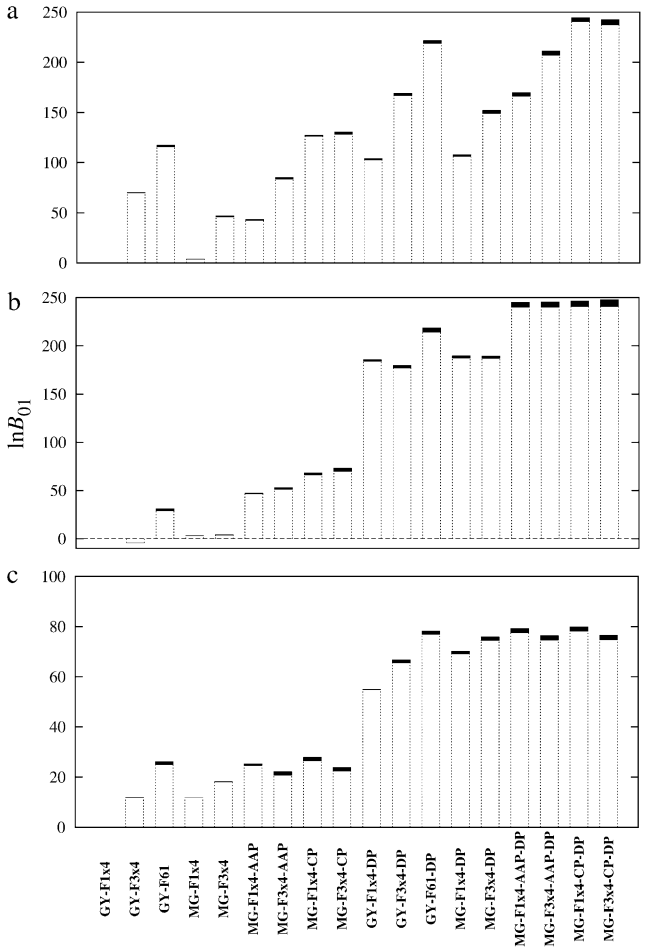

FIGURE 1.—Natural logarithm of the Bayes factor for all models taken with respect to to the GY-F1 × 4 model. The solid bands are representative of the overall precision of the calculations (broader bands indicate poorer precision). (a) The Globin *17-144* data set; (b) Lysin *25-134*; (c) Hiv *22-99*.

favor of the DP model is robust to the hyperprior; more work is still needed, however, to explore alternative prior structures on the Dirichlet process in this context, in particular regarding alternative base priors distributions. The other specifications of this top model are also reassuring, in the sense that adhering closely to the mechanistic perspective of teasing apart mutational features and selective constraints produces, at worst, a model of roughly equivalent performance to models lacking such a natural interpretation. In addition, all three data sets suggest uneven codon preferences, although such preferences appear to go well beyond amino acid preferences only in the case of the Globin *17-144*.

We next note that under the simpler settings of MG-style models, suppressing AAP or CP parameters, the F3 × 4 configuration is generally preferred over the F1 × 4 configuration for all three data sets. The periodic pattern of codon-position-specific nucleotide propensities is a feature expected from the structure of the genetic code, through selective constraints. Such an interpretation, however, is not accurately represented by expanding the nucleotide-level parameterization. Indeed, with the richer models, including the CP parameters in particular, the F3 × 4 configurations are only mildly preferred over the F1 × 4 configuration, and when invoking the Dirichlet process, modeling heterogeneous nonsynonymous rates, the numerical error no longer allows for a clear distinction between these two configurations (except for the Hiv *22-99* data set, which gives preference to the F1 × 4 configuration). We also observed the parametric redundancy of including both the F3 × 4 setting and the CP parameters to be susceptible to identifiability problems (see supplemental material).

The GY style of models based on the F1 × 4 and F3 × 4 configurations is generally disfavored over the MG-style counterparts (except for the Globin *17-144* data set, which gives favor to GY-F3 × 4 over MG-F3 × 4). Surprisingly, for the Lysin *25-134* data set, the simpler GY-F1 × 4 model is slightly preferred over the GY-F3 × 4 model. However, for all three data sets, the GY model based on F61 configurations outperforms the other GY-style models, as well as the simpler MG-style models. In the case of the Globin *17-144* data set, the contrast of the F61 configuration is even greater than that observed between homogeneous and heterogeneous models of nonsynonymous rates; for instance, the log Bayes factor of GY-F61 against GY-F1 × 4 is [115.8; 117.4] whereas that of GY-F1 × 4-DP against GY-F1 × 4 is [102.3; 104.2]. These results for the GY-F61 model are also indicative of uneven codon preferences. However, as previously mentioned, the codon preferences accounted for in this GY formulation are confounded with other features, including the background of nucleotide propensities, making the model less attractive on interpretive grounds. Accordingly, when contrasted with the richer MG formulations accounting for codon (or amino acid) preferences, the GY-F61 model is less attractive on quantitative grounds (except for Hiv *22-99*, in which case it matches the top MG-style models).

Posterior distributions: To highlight a few aspects of the MG-F1 × 4-CP-DP model, which was among the top performing for all three data sets, we display posterior distributions of parameters (obtained using plain MCMC sampling). Our focus is on the combination of background nucleotide propensities with global codon preference parameters, but we contrast the distributions with those obtained under the simpler model suppressing CP parameters, as well as under the GY-F61-DP model. We also explore the impacts of different models on the detection of positive selection.

Figure 2 displays the 95% credibility intervals of the posterior distributions of the global nucleotide propensity parameters for each data set. The solid lines correspond to the interval obtained under MG-F1 × 4-DP, whereas the dashed lines are obtained under MG-F1 × 4-CP-DP. The distributions are far more diffuse under the CP version, although their general locations appear

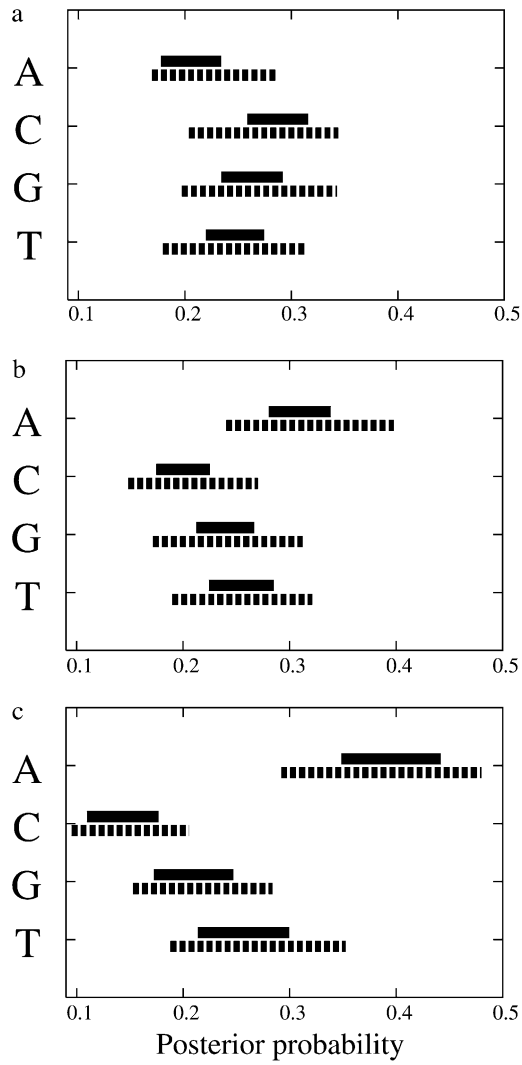

FIGURE 2.—Ninety-five percent credibility intervals of global nucleotide propensity parameters obtained under MG-F1 × 4-DP (solid lines) and under MG-F1 × 4-CP-DP (dashed lines). (a) The Globin*17-144* data set; (b) Lysin*25-134*; (c) Hiv*22-99*.

similar. This is the well-known feature of increasingly diffuse distributions obtained when expanding the form of models and that we observed for the other parameters of the models as well (not shown).

Figure 3 displays the 95% credibility intervals of the posterior distributions of codon preference parameters under the MG-F1 × 4-CP-DP model, in comparison with the codon stationary parameters under GY-F61-DP model. Note that the parameters do not have the same interpretation under the two models. For the GY-F61-DP model, the parameters correspond to the limiting distribution of the Markov process; but for the MG-F1 × 4-CP-DP model, the limiting distribution of the Markov process is based on both codon preference parameters and nucleotide propensity parameters, and hence the graphical comparison displayed here does not correspond to a formal statistical testing. The GY-F61-DP model leads to tighter 95% credibility intervals than the MG-F1 × 4-CP-DP, which concords with its lower dimensionality. However, the overall distributions of the two sets of parameters are reasonably similar, with some degree of overlap in all cases. Thus both models appear to be capturing similar overall trends. The distributions suggest pronounced overall codon preferences for Globin*17-144*, but milder preferences for Lysin*25-134* and Hiv*22-99*. This corroborates well with our computed Bayes factors, which, for instance, indicate that for Lysin*25-134* and Hiv*22-99*, the improvement brought about by the CP parameters is less important than for the Globin*17-144* data. Observing the distributions for the Globin*17-144* data set in detail, we find that the parameter values appear to capture long observed tendencies of codon preferences on similar data, such as the elevated use of CTG for encoding leucine, GTG for valine, or GGC for glycine; indeed, these were some of the first observations stimulating research into the causes of codon preferences (*e.g.*, FITCH 1980; MODIANO *et al.* 1981; KIMURA 1983).

We contrasted the conclusions of the GY-F61-DP, MG-F1 × 4-DP, and MG-F1 × 4-CP-DP models with regard to the central application of such models: the inference of amino acid positions having undergone positive selection. Under the DP settings, the posterior probability of a site being under positive selection can be computed from the proportion of draws from a sample (obtained via plain MCMC sampling) found to be in a class $\omega > 1$, as described in HUELSENBECK *et al.* (2006). We first note that for the Globin*17-144*, focusing on posterior probabilities at 0.9, 0.95, and 0.99 cutoff levels, the MG-F1 × 4-DP and MG-F1 × 4-CP-DP models infer sites under positive selection at each level, whereas the GY-F61-DP model infers no sites at either level (Table 1). The list of sites under positive selection under the three models considered also differs for the other two data sets (Table 1). The computed Bayes factor can be used to weigh the conclusions of different models and indeed could form the basis of a model averaging inference of positive selection. It will be important to conduct a broader empirical study of the impacts of these and other parametric choices on the detection of positive selection.

Posterior predictive assessments: To investigate whether different nucleotide frequencies observed at the three different codon positions can be a result of codon preferences, we computed the codon-position-specific frequencies of each nucleotide on data replicates, generated by simulation under parameters sampled from the posterior distribution. Focusing on the Globin*17-144* data set, Figure 4 displays the posterior predictive distribution of the frequency of nucleotide A at the three positions under the MG-F1 × 4-DP model (solid lines), the MG-F1 × 4-CP-DP model (thick dashed lines), and the GY-F61 model (thin dashed lines). The MG-F1 × 4-DP model leads to closely matching distributions at the three positions, from the definition of the

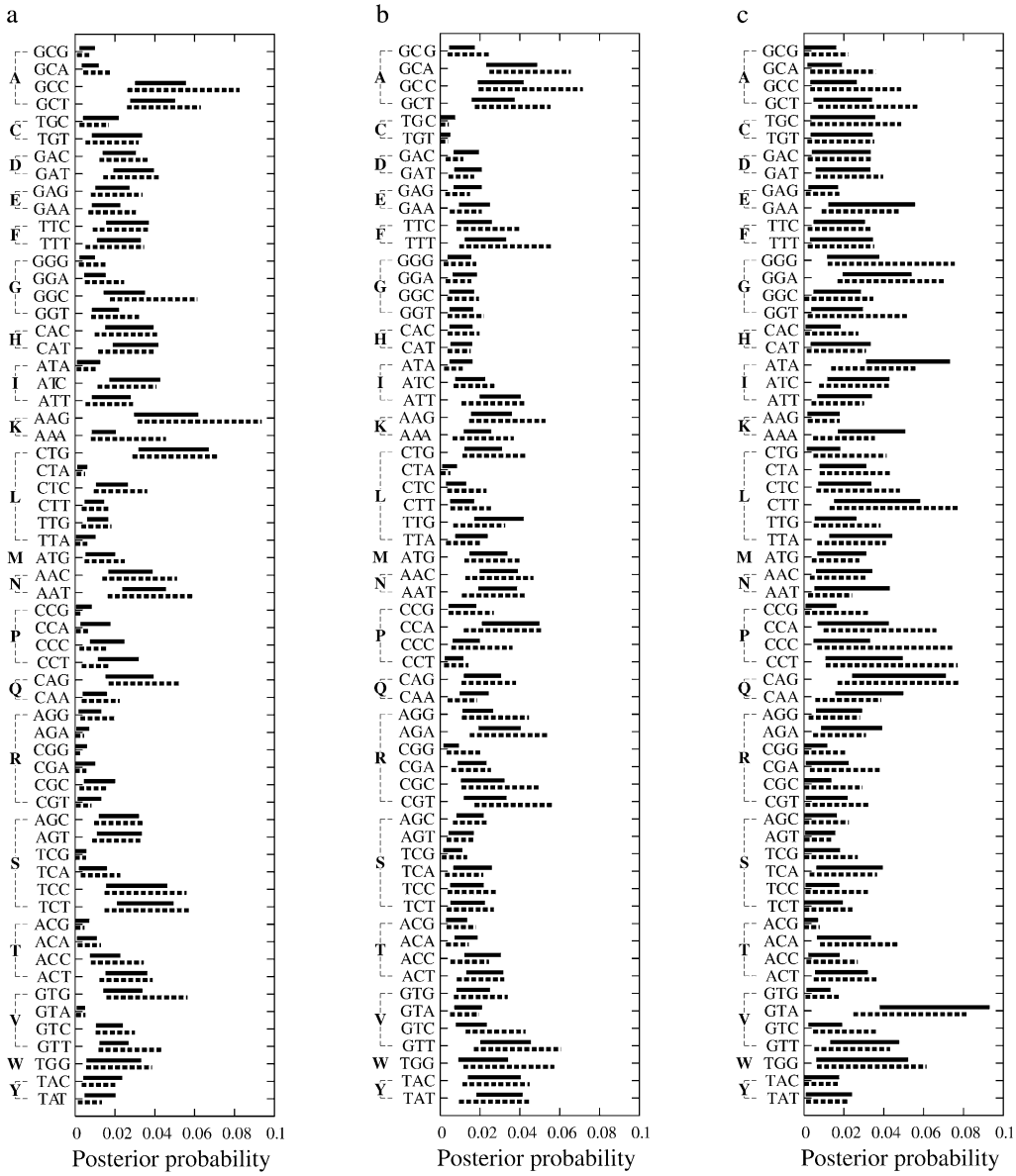

FIGURE 3.—Ninety-five percent credibility intervals of codon stationary probabilities under the GY-F61-DP model (solid lines) and codon preference parameters under the MG-F1 × 4-CP-DP model (dashed lines). (a) The Globin *17-144* data set; (b) Lysin *25-134*; (c) Hiv *22-99*.

model (although the exclusion of stop codons induces mild shifts). In Figure 4, b and c, the posterior predictive *P*-values of 0 and 1 are indicators of problems with the MG-F1 × 4-DP model. The MG-F1 × 4-CP-DP model, in contrast, leads to markedly different distributions at the three positions, in a manner approaching the observed frequencies. The GY-F61-DP model is also capable of producing this effect, although graphical comparisons for the three other nucleotides suggest that it does not perform as well as the MG-F1 × 4-CP-DP model (supplemental material).

In another posterior predictive check, we computed the relative frequency of codons in alignments (real and replicated), from which we then evaluated the overall unevenness of the frequency profile from the codon entropy. In Figure 5, we find, as expected, that the MG-F1 × 4-DP model induces a relatively even predictive codon frequency profile (from the definition of the model), leading to a high codon entropy (solid-line histogram). The *P*-value of 1 indicates that the model is inadequate. In contrast, under the MG-F1 × 4-CP-DP model we obtain a more uneven profile, and thus a lower codon entropy (thick-dashed-line histogram); the *P*-value is 0.528, and the model is not deemed problematic under this test. The GY-F61-DP model leads to an intermediate codon entropy (thin-dashed-line histogram), which poorly anticipates the empirical value, but is nonetheless much closer than the MG-F1 × 4-DP model.

These two posterior predictive checks are indicative of tensions in the GY-F61-DP model configuration. The overall model construction strategy of the GY approach may be unable to adequately capture codon preferences without inducing distorted nucleotide propensities. Although the GY-F61-DP model finds a compromise to this tension that performs better than simpler models,

TABLE 1

Amino acid sites under positive selection

Data	Model	Sites
Globin *17-144*	GY-F61-DP	—
	MG-F1 × 4-DP	7, *48*, 50, *54*, 67, 85, 123,
	MG-F1 × 4-CP-DP	7, *11*, 50, *67*, 85, 123
Lysin *25-134*	GY-F61-DP	*2*, *3*, 4, 6, 7, 9, 10, 11, 12, 14, 32, 33, 36, *37*, 41, 44, 64, 67, 68, 70, 74, 83, 86, 87, 100, *106*, *107*, 113, 115, *116*, 120, *123*, 126, 132
	MG-F1 × 4-DP	4, 6, 7, 9, 10, 11, 12, 14, 32, 33, 36, 40, 41, 44, *45*, 64, 67, 68, 70, 74, 75, *82*, 83, 86, 87, 100, 106, *107*, 113, 115, 119, 120, 126, 127, 132
	MG-F1 × 4-CP-DP	4, 6, 7, 9, 10, 12, *14*, 32, 33, 36, 37, 41, 44, 64, 67, 68, 70, 74, 75, 83, *86*, 87, 100, 106, 113, 115, 119, 120, 123, 126, *127*, 132
Hiv *22-99*	GY-F61-DP	54, 37, 63
	MG-F1 × 4-DP	*10*, *12*, *32*, 33, 37, 41, 46, *47*, 50, 54, 63, 82
	MG-F1 × 4-CP-DP	*10*, *32*, *33*, 37, *50*, 54, 63

Numbers in italics are at the 0.9 level, those in regular type are at the 0.95 level, and those underlined are at the 0.99 level.

in its ability to phenomenologically capture overall codon preference trends, it is outperformed by the MG-F1 × 4-CP-DP model, which explicitly decouples codon preferences from nucleotide propensities.

CONCLUSIONS AND FUTURE DIRECTIONS

In recent years, numerous codon substitution models have been proposed, subscribing to either the GY or the MG perspective. GY-style models have been extended in several ways, for instance, to account for heterogeneous nonsynonymous rates (*e.g.*, Nielsen and Yang 1998; Yang *et al.* 2000a; Huelsenbeck and Dyer 2004; Huelsenbeck *et al.* 2006) or to recognize differences in the types of nonsynonymous substitutions, either based on various amino acid distance metrics (*e.g.*, Yang *et al.* 1998), by partitioning amino acids into predefined classes (Sainudiin *et al.* 2005; Wong *et al.* 2006), by incorporating information from an empirically derived amino acid replacement matrix (Doron-Faigenboim and Pupko 2007), or even based on full empirical modeling strategies (Kosiol *et al.* 2007). In parallel, MG-style models have been extended in similar ways as well, such as the models recognizing heterogeneous synonymous and nonsynonymous substitution rates across codon sites (Kosakovsky Pond and Muse 2005), models incorporating partition-based amino acid exchange propensities (Schadt and Lange 2002), or models accounting for dependence between codon positions due to protein tertiary structure (Robinson *et al.* 2003).

In light of all of these developments, the present study effectively takes a step back, to reassess the core motivation underlying codon-based models, namely, the formulation of a biologically meaningful parameterization that disentangles the different factors bearing on the overall substitution process. Our results indicate that a mechanistically founded attempt of teasing apart nucleotide propensities and amino acid or codon preferences in the MG style tends to surpass, or at least match, the optimal GY-style model. These results suggest that future modeling investigations should consider incorporating any extensions in the MG context.

Quantitative model comparisons based on Bayes factors would be of particular interest to evaluate how some of the recently proposed models mentioned above compare with the models studied here. For instance, the models proposed in Robinson *et al.* (2003), which can account for dependencies between codon sites due to structural constraints at the protein level, could be included into the scope of models evaluated. Working at the amino acid level, we previously proposed a variation of the thermodynamic method for evaluating Bayes factors with models accommodating a general dependence across sites (Rodrigue *et al.* 2006), which could be transposed to the codon level in a straightforward way. A broad range of model extensions are also evident from the AAP and CP approaches proposed in this work: given that these richer MG-style models lead to an improved overall fit, models based on mixtures of AAP or CP parameters could also be studied, so as to capture site-specific preferences. To this end, the Dirichlet process prior, applied here to model nonsynonymous rate heterogeneity across sites (Huelsenbeck *et al.* 2006), could also be applied to the AAP parameters or to the CP parameters. Indeed, the necessary MCMC operators for manipulating such models have been described previously (Lartillot and Philippe 2004). It might also be of interest to extend these models to nonstationary nucleotide propensities, as well as to nonstationary amino acid or codon preferences, on the

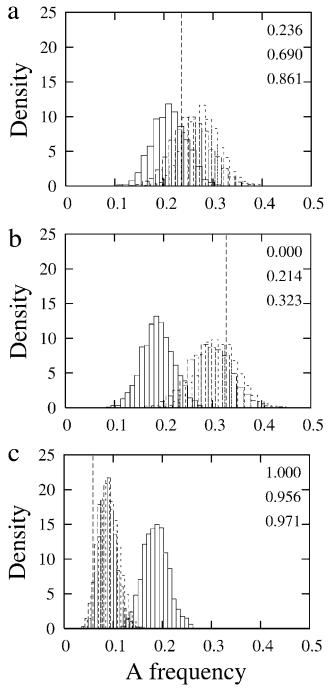

FIGURE 4.—Posterior predictive distributions of the frequency of nucleotide A at the first (a), second (b), and third (c) codon positions under the MG-F1 × 4-DP (solid-line histogram), MG-F1 × 4-CP-DP (thick-dashed-line histogram), and GY-F61-DP (thin-dashed-line histogram) models. The observed values (computed on the real alignment) are displayed as dashed vertical lines. In the top-right corner, the first value is the posterior predictive *P*-value under the MG-F1 × 4-DP model, the second value is that under the MG-F1 × 4-CP-DP model, and the third is that under the GY-F61-DP model.

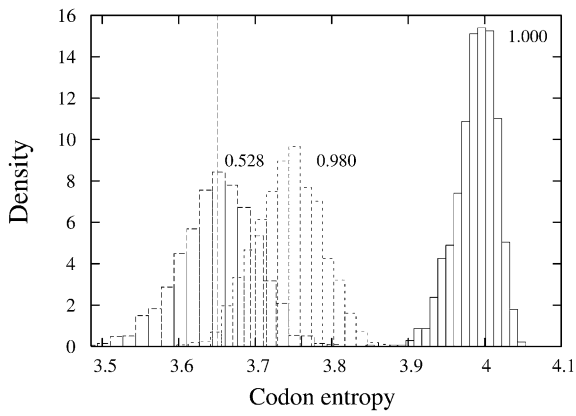

FIGURE 5.—Posterior predictive distribution of the codon entropy for the Globin*17-144* data set. The solid-line histogram corresponds to the MG-F1 × 4-DP model, the thick-dashed-line histogram corresponds to the MG-F1 × 4-CP-DP model, and the thin-dashed-line histogram corresponds to the GY-F61-DP model. The value observed on the true alignment is displayed as a dashed vertical line, and the posterior predictive *P*-value is written next to each histogram.

basis of ideas presented in BLANQUART and LARTILLOT (2006) and NIELSEN *et al.* (2007).

We note that analogous quantitative model comparisons would be difficult in a frequentist framework. The nonnested form of the models of interest complicates traditional frequentist tests, in which case most practitioners rely on criteria such as the Akaike information criterion (AIC) (AKAIKE 1974). However, the AIC (and other similar criteria) relies on maximum-likelihood parameter estimates, as well as the log-likelihood value at this optimal point. Performing maximum-likelihood estimation and log-likelihood calculations will likely become increasingly difficult as richer nonanalytical models are proposed (RODRIGUE *et al.* 2007), such as the Dirichlet process models or models with dependence across sites (ROBINSON *et al.* 2003).

Extending comparisons over many data sets will be of prime importance to determine if the results presented here can be generalized. However, such a project would entail significant computational costs using the model-switch thermodynamic methods employed here (several log Bayes factor calculations required over 40 days on an Intel P4 3.2 GHz desktop computer). We are currently investigating a marginal-likelihood estimator combining sigmoidal model-switch schemes (LEPAGE *et al.* 2007) with the Laplace method for integrals, along the lines described in RODRIGUE *et al.* (2007). Combinations with other Bayes factor calculation schemes (*e.g.*, SUCHARD *et al.* 2001; CHOI *et al.* 2007) are also possible, and the availability of such a suite of computational devices should enable a broad empirical study and help uncover and quantify the main factors bearing on protein-coding sequence evolution.

We thank Stéphane Aris-Brosou for his feedback on an early version of the manuscript, as well as Rasmus Nielsen and two anonymous reviewers for thoughtful comments. We also thank the Réseau Québécois de calcul de haute performance for computational resources. This work was supported by Génome Québec, 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, the Natural Sciences and Engineering Research Council of Canada, the Canadian Institute for Advanced Research, the Canadian Research Chair Program, the Centre National de la Recherche Scientifique (through the Action Concertée Incitative–Informatique, Mathématique, Physique en Biologie Moléculaire Model-Phylo funding program), and the 60ème Commission Franco-Québécoise.

LITERATURE CITED

AKAIKE, H., 1974 A new look at the statistical model identification. IEEE Trans. Automat. Contr. **AC-19:** 716–723.

BLANQUART, S., and N. LARTILLOT, 2006 A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Mol. Biol. Evol. **23:** 2058–2071.

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

CHOI, S. C., A. HOBOLTH, D. M. ROBINSON, H. KISHINO and J. L. THORNE, 2007 Quantifying the impact of protein tertiary structure of molecular evolution. Mol. Biol. Evol. **24:** 1769–1782.

DORON-FAIGENBOIM, A., and T. PUPKO, 2007 A combined empirical and mechanistic codon model. Mol. Biol. Evol. **24:** 388–397.

FELSENSTEIN, J., 2007 Trees of genes in populations, pp. 3–29 in *Reconstruction Evolution, New Mathematical and Computational Advances*, edited by O. GASCUEL and M. STEEL. Oxford University Press, London/New York/Oxford.

Ferguson, T. S., 1973 A Bayesian analysis of some nonparametric problems. Ann. Stat. **1:** 209–230.

Fitch, W. M., 1980 Estimation the total number of nucleotide substitutions since the common ancestors of a pair of homologous genes: comparison of several mehtods and three beta hemoglobin messenger RNA's. J. Mol. Evol. **16:** 153–209.

Gelman, A., X. L. Meng and H. Stern, 1996 Posterior predictive assessment of model fitness via realized discrepancies. Stat. Sin. **6:** 733–807.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin, 2004 *Bayesian Data Analysis*. Chapman & Hall/CRC Press, London/New York/Cleveland, OH/Boca Raton, FL.

Goldman, N., and S. Whelan, 2002 A novel use of equilibrium frequencies in models of sequence evolution. Mol. Biol. Evol. **19:** 1821–1831.

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

Huelsenbeck, J. P., and K. A. Dyer, 2004 Bayesian estimation of positively selected sites. J. Mol. Evol. **58:** 661–672.

Huelsenbeck, J. P., S. Jain, S. W. D. Frost and S. L. K. Pond, 2006 A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. Proc. Natl. Acad. Sci. USA **103:** 6263–6268.

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

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge/London/New York.

Kosakovsky Pond, S. L., and S. V. Muse, 2005 Site-to-site variation of synonymous substitution rates. Mol. Biol. Evol. **22:** 2375–2385.

Kosiol, C., I. Holmes and N. Goldman, 2007 An empirical codon model for protein sequence evolution. Mol. Biol. Evol. **24:** 1464–1479.

Lanave, C., G. Preparata, C. Saccone and G. Serio, 1984 A new method for calculating evolutionary substitution rates. J. Mol. Evol. **20:** 86–93.

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

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

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

Lepage, T., D. Bryant, H. Philippe and N. Lartillot, 2007 A general comparison of relaxed molecular clock models. Mol. Biol. Evol. **24:** 2669–2680.

Modiano, G., G. Battistuzzi and A. G. Motulsky, 1981 Nonrandom patterns of codon usage and nucleotide substitutions in human alpha- and beta-globin genes: An evolutionary strategy reducing the rate of mutations with drastic effects? Proc. Natl. Acad. Sci. USA **78:** 1110–1114.

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

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

Nielsen, R., and Z. Yang, 1998 Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148:** 929–936.

Nielsen, R., V. L. B. DuMont, M. J. Hubisz and C. F. Aquadro, 2007 Maximum likelihood estimation of ancestral codon usage bias parameters in Drosophila. Mol. Biol. Evol. **24:** 228–235.

Robinson, D. M., D. T. Jones, H. Kishino, N. Goldman and J. L. Thorne, 2003 Protein evolution with dependence among codons due to tertiary structure. Mol. Biol. Evol. **18:** 1692–1704.

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

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

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

Sainudiin, R., W. S. W. Wong, K. Yogeeswaran, J. Nasrallah, Z. Yang *et al.*, 2005 Detecting site-specific physicochemical selective pressures: applications to the class-I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system. J. Mol. Evol. **60:** 315–326.

Schadt, E., and K. Lange, 2002 Codon and rate variation models in molecular phylogeny. Mol. Biol. Evol. **19:** 1534–1549.

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

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

Thorne, J. L., S. C. Choi, J. Yu, P. G. Higgs and H. Kishino, 2007 Population genetics without intraspecific data. Mol. Biol. Evol. **24:** 1667–1677.

Wong, S. W., R. Sainudiin and R. Nielsen, 2006 Identification of physicochemical selective pressure on protein encoding nucleotide sequences. BMC Bioinform. **7:** 148.

Yang, Z., 2006 *Computational Molecular Evolution* (Oxford Series in Ecology and Evolution). Oxford University Press, Oxford.

Yang, Z., and R. Nielsen, 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., R. Nielsen and M. Hasegawa, 1998 Models of amino acid substitutions and applications to mitochondrial protein evolution. Mol. Biol. Evol. **15:** 1600–1611.

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

Yang, Z., W. J. Swanson and V. D. Vacquier, 2000b Maximum likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. Mol. Biol. Evol. **17:** 1446–1455.

Communicating editor: R. Nielsen

APPENDIX

In our implementation the entries of Q are based on two sets of specifications: a 61-dimensional vector of stationary probabilities, π, and a set of *transient specifications*, written as $\rho = (\rho_{ij})_{1 \le i,j \le 61}$ and combined according to

$$Q_{ij} \propto \rho_{ij}\pi_j, \quad i \neq j \tag{A1}$$

$$Q_{ii} = -\sum_{j \neq i} Q_{ij}. \tag{A2}$$

In this appendix, we write out in full the stationary probabilities under the models, as well as the full transient specifications, and give an example of the detailed balance check.

Stationary probabilities: First, expanding (8) for the stationary distribution under GY-F1 × 4, we have

$$\pi_i = \frac{\varphi_{i_1}\varphi_{i_2}\varphi_{i_3}}{\sum_{j=1}^{61}\varphi_{j_1}\varphi_{j_2}\varphi_{j_3}}. \tag{A3}$$

Similarly, with GY-F3×4, we have

$$\pi_i = \frac{\varphi^{(1)}_{i_1}\varphi^{(2)}_{i_2}\varphi^{(3)}_{i_3}}{\sum_{j=1}^{61}\varphi^{(1)}_{j_1}\varphi^{(2)}_{j_2}\varphi^{(3)}_{j_3}}. \tag{A4}$$

The stationary probability under GY-F61 is already entirely specified and the models MG-F1 × 4 and MG-F3 × 4 have the same stationary distributions as (A3) and (A4), respectively.

Under the MG-F1 × 4-CP model, the stationary probability is given by

$$\pi_i = \frac{\varphi_{i_1}\varphi_{i_2}\varphi_{i_3}\psi_i}{\sum_{j=1}^{61}\varphi_{j_1}\varphi_{j_2}\varphi_{j_3}\psi_j}, \tag{A5}$$

and under the MG-F1 × 4-AAP model it is given by

$$\pi_i = \frac{\varphi_{i_1}\varphi_{i_2}\varphi_{i_3}\phi_{f(i)}}{\sum_{j=1}^{61}\varphi_{j_1}\varphi_{j_2}\varphi_{j_3}\phi_{f(j)}}. \tag{A6}$$

The stationary distributions under the MG-F3 × 4-CP and MG-F3 × 4-AAP models follow directly as

$$\pi_i = \frac{\varphi^{(1)}_{i_1}\varphi^{(2)}_{i_2}\varphi^{(3)}_{i_3}\psi_i}{\sum_{j=1}^{61}\varphi^{(1)}_{j_1}\varphi^{(2)}_{j_2}\varphi^{(3)}_{j_3}\psi_j} \tag{A7}$$

and

$$\pi_i = \frac{\varphi^{(1)}_{i_1}\varphi^{(2)}_{i_2}\varphi^{(3)}_{i_3}\phi_{f(i)}}{\sum_{j=1}^{61}\varphi^{(1)}_{j_1}\varphi^{(2)}_{j_2}\varphi^{(3)}_{j_3}\phi_{f(j)}}, \tag{A8}$$

respectively.

Transient specifications: In the case of GY-type models, the transient specification is simply (7) without the π_j factor. In the case of the MG-F1 × 4 model, we have

$$\rho_{ij} = \begin{cases} \frac{\varrho_{i_{c'}j_c}}{\varphi_{j_{c'}}\varphi_{j_{c''}}}Z, & \text{if } \mathcal{A}, \\ \frac{\omega\varrho_{i_{c'}j_c}}{\varphi_{j_{c'}}\varphi_{j_{c''}}}Z, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise,} \end{cases} \tag{A9}$$

where c' and c'' are the two constant codon positions, and Z is the normalizing factor of the stationary distribution (in this case $Z = \sum_{j=1}^{61}\varphi_{j_1}\varphi_{j_2}\varphi_{j_3}$). Note that this latter Z factor is not needed when scaling Q. Once again, substituting φ_{j_c} with $\varphi^{(c)}_{j_c}$, and the appropriate Z, yields the transient specification for MG-F3 × 4.

For the MG-F1 × 4-CP model, the transient specification is given by

$$\rho_{ij} = \begin{cases} \frac{\varrho_{i_{c'}j_c}}{\varphi_{j_{c'}}\varphi_{j_{c''}}\sqrt{\psi_i\psi_j}}Z, & \text{if } \mathcal{A}, \\ \frac{\omega\varrho_{i_{c'}j_c}}{\varphi_{j_{c'}}\varphi_{j_{c''}}\sqrt{\psi_i\psi_j}}Z, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise,} \end{cases} \tag{A10}$$

and the specification of MG-F1 × 4-AAP is given by

$$\rho_{ij} = \begin{cases} \frac{\varrho_{i_{c'}j_c}}{\varphi_{j_{c'}}\varphi_{j_{c''}}\phi_{f(j)}}Z, & \text{if } \mathcal{A}, \\ \frac{\omega\varrho_{i_{c'}j_c}}{\varphi_{j_{c'}}\varphi_{j_{c''}}\sqrt{\phi_{f(i)}\phi_{f(j)}}}Z, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise.} \end{cases} \tag{A11}$$

As always, substituting φ_{j_c} with $\varphi_{j_c}^{(c)}$, and the appropriate Z, yields the transient specifications for the F3 $\times$ 4 versions of (A10) and (A11).

We have now fully specified π and ρ used in Equation A1. We can see that upon substituting stationary and transient specifications appropriately into (A1), the models defined in the main body of the text are obtained. For instance, for a nonsynonymous substitution under the MG-F1 $\times$ 4-CP model, we have

$$\rho_{ij}\pi_j = \frac{\omega\varrho_{i_c j_c}}{\varphi_{j_{c'}}\varphi_{j_{c''}}\sqrt{\psi_i\psi_j}} Z \times \frac{\varphi_{j_1}\varphi_{j_2}\varphi_{j_3}\psi_j}{Z} \tag{A12}$$

$$= \frac{\omega\varrho_{i_c j_c}\varphi_{j_c}\psi_j}{\sqrt{\psi_i\psi_j}} \tag{A13}$$

$$= \frac{\omega\varrho_{i_c j_c}\varphi_{j_c}\sqrt{\psi_j}\sqrt{\psi_j}}{\sqrt{\psi_i}\sqrt{\psi_j}} \tag{A14}$$

$$= \omega\varrho_{i_c j_c}\varphi_{j_c}\left(\frac{\psi_j}{\psi_i}\right)^{1/2}, \tag{A15}$$

corresponding to the entry obtained from (2).

Checking the detailed balance: The models studied here all satisfy the equality $\pi Q = 0$, and are time reversible, satisfying the equality $Q_{ij}\pi_i = Q_{ji}\pi_j$. These developments are lengthy, and so we display only one example, for the detailed balance check under MG-F1 $\times$ 4-CP in the case where i and j differ at one nucleotide position, implying a nonsynonymous substitution,

$$Q_{ij}\pi_i = Q_{ji}\pi_j \tag{A16}$$

$$\frac{\omega\varrho_{i_c j_c}\varphi_{j_1}\varphi_{j_2}\varphi_{j_3}\psi_j}{\varphi_{j_{c'}}\varphi_{j_{c''}}\sqrt{\psi_i\psi_j}}\varphi_{i_1}\varphi_{i_2}\varphi_{i_3}\psi_i = \frac{\omega\varrho_{j_c i_c}\varphi_{i_1}\varphi_{i_2}\varphi_{i_3}\psi_i}{\varphi_{i_{c'}}\varphi_{i_{c''}}\sqrt{\psi_j\psi_i}}\varphi_{j_1}\varphi_{j_2}\varphi_{j_3}\psi_j \tag{A17}$$

$$\frac{\omega\varrho_{i_c j_c}\cancel{\varphi_{j_1}\varphi_{j_2}\varphi_{j_3}\psi_j}}{\varphi_{j_{c'}}\varphi_{j_{c''}}\sqrt{\psi_i\psi_j}}\varphi_{i_1}\varphi_{i_2}\varphi_{i_3}\psi_i = \frac{\omega\varrho_{j_c i_c}\varphi_{i_1}\varphi_{i_2}\varphi_{i_3}\psi_i}{\varphi_{i_{c'}}\varphi_{i_{c''}}\sqrt{\psi_j\psi_i}}\cancel{\varphi_{j_1}\varphi_{j_2}\varphi_{j_3}\psi_j} \tag{A18}$$

$$\frac{\omega\varrho_{i_c j_c}}{\varphi_{j_{c'}}\varphi_{j_{c''}}\sqrt{\psi_i\psi_j}}\cancel{\varphi_{i_1}\varphi_{i_2}\varphi_{i_3}\psi_i} = \frac{\omega\varrho_{j_c i_c}\cancel{\varphi_{i_1}\varphi_{i_2}\varphi_{i_3}\psi_i}}{\varphi_{i_{c'}}\varphi_{i_{c''}}\sqrt{\psi_j\psi_i}} \tag{A19}$$

$$\frac{\omega\varrho_{i_c j_c}}{\cancel{\varphi_{j_{c'}}\varphi_{j_{c''}}\sqrt{\psi_i\psi_j}}} = \frac{\omega\varrho_{j_c i_c}}{\cancel{\varphi_{i_{c'}}\varphi_{i_{c''}}\sqrt{\psi_j\psi_i}}} \tag{A20}$$

$$\cancel{\omega}\varrho_{i_c j_c} = \cancel{\omega}\varrho_{j_c i_c} \tag{A21}$$

$$\varrho_{i_c j_c} = \varrho_{j_c i_c}, \tag{A22}$$

where the array ϱ is symmetrical, satisfying the equality.

Finally, we mention here that we follow the practice proposed by Huelsenbeck *et al.* (2006) and scale Q matrices such that branch lengths represent the expected number of synonymous substitutions per codon site, although we have also tried the model comparisons without any scaling of Q (such that branch lengths have no meaningful units) and obtained essentially identical results (not shown).