A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process

Nicolas Lartillot and Hervé Philippe

Canadian Institute for Advanced Research, Département de Biochimie, Université de Montréal, Montréal, Québec Canada

Most current models of sequence evolution assume that all sites of a protein evolve under the same substitution process, characterized by a 20×20 substitution matrix. Here, we propose to relax this assumption by developing a Bayesian mixture model that allows the amino-acid replacement pattern at different sites of a protein alignment to be described by distinct substitution processes. Our model, named CAT, assumes the existence of distinct processes (or classes) differing by their equilibrium frequencies over the 20 residues. Through the use of a Dirichlet process prior, the total number of classes and their respective amino-acid profiles, as well as the affiliations of each site to a given class, are all free variables of the model. In this way, the CAT model is able to adapt to the complexity actually present in the data, and it yields an estimate of the substitutional heterogeneity through the posterior mean number of classes. We show that a significant level of heterogeneity is present in the substitution patterns of proteins, and that the standard one-matrix model fails to account for this heterogeneity. By evaluating the Bayes factor, we demonstrate that the standard model is outperformed by CAT on all of the data sets which we analyzed. Altogether, these results suggest that the complexity of the pattern of substitution of real sequences is better captured by the CAT model, offering the possibility of studying its impact on phylogenetic reconstruction and its connections with structure-function determinants.

Introduction

Probabilistic methods are widely used in phylogenetic reconstruction. Their main advantage, compared to methods such as maximum parsimony, is to make all assumptions underlying the reconstruction explicit, while providing powerful and general techniques for validating those assumptions (Swofford et al. 1996; Sullivan and Swofford 2001). Given a stochastic model of evolution, these methods allow computation of the probability of observing the available data, conditional on a phylogenetic hypothesis (specified by a topology, branch lengths plus some other parameters). This probability is used as a measure of the likelihood of the corresponding hypothesis, and one then invokes the maximum likelihood (ML) principle, choosing the hypothesis for which the probability of observing the data is maximal.

Maximum likelihood is intuitively appealing, and mathematical theorems guarantee the asymptotic consistency of the method (Wald 1949). However, when applied to models that are too rich in parameters, ML can lead to over-fitting artifacts. This is true, for instance, when models with site-specific parameters are considered, a problem sometimes referred to as the "infinitely many parameter trap" (Felsenstein 2004). In practice, this limitation on the number of parameters restricts the range of models that the phylogeneticist, seeking more realism, would like to explore. An alternative probabilistic paradigm, the Bayesian method, has been introduced recently in the field of molecular phylogeny (Li 1996; Yang and Rannala 1997; Larget and Simon 1999; Huelsenbeck and Ronquist 2001). It was initially proposed as a practical alternative to ML, mainly on the grounds that it offers a natural measure of uncertainty, thereby avoiding the costly method of bootstrap resampling (Larget and Simon 1999). However, another advantage of Bayes, which is in our opinion much more fundamental, lies in its greater flexibility with respect to the model's dimensionality (Huelsenbeck et al. 2002; Rannala 2002). In contrast to ML, and through a different treatment of the nuisance parameters which are not directly estimated but integrated away, Bayes is able to deal with much higher dimensional models while offering several methods to test these models in the light of available data (Jeffreys 1935; Jaynes 2003). This flexibility opens new avenues of investigation, as it makes it possible to build more realistic models of sequence evolution by adjusting the dimensionality so that it reflects the complexity actually displayed by the data, rather than the limitations inherent in the method.

One of the ways by which restrictions are traditionally imposed on model dimensionality is by assuming that sites of an alignment are independent and identically distributed: that is, they are considered as independent realizations of the same substitution process, running along the branches of the underlying evolutionary tree (Felsenstein 1981). In certain circumstances, as in the case of pseudogenes, this assumption might be valid, but more often different nucleotides or different codons of a gene will evolve under very different selection pressures. The assumption of identical distribution is partially relaxed in the "rates across sites" models (Yang 1993, 1994), where the rate of evolution at each site is a random variable drawn from a gamma distribution. Such models yield a major improvement in statistical adequacy over the uniform rate model when applied to both nucleotide and protein data sets (Yang 1996). However, they still assume that all the remaining parameters of the model of evolution—i.e., the equilibrium frequencies and the relative rates of substitution among nucleotides or amino acids—are the same at all sites.

In practice, all these parameters are summarized into a 4×4 or 20×20 rate matrix. In the case of amino-acid alignments, empirical matrices are generally used, which have first been obtained by counting pairs of amino acids

Key words: phylogeny, Bayes, Dirichlet process mixtures, amino-acid replacement, Bayes factor, posterior predictive resampling.

E-mail: nicolas.lartillot@lirmm.fr

Mol. Biol. Evol. 21(6):1095–1109. 2004
DOI:10.1093/molbev/msh112
Advance Access publication March 10, 2004

at homologous positions in large sets of aligned proteins (Dayhoff, Eck, and Park 1972; Dayhoff, Schwartz, and Orcutt 1978; Jones, Taylor, and Thornton 1992). Matrices optimized by ML have also been proposed for mitochondrial (Adachi and Hasegawa 1996), chloroplast (Adachi et al. 2000), and nuclear (Whelan and Goldman 2001) proteins. More recently, an alternative, faster method of optimization has been introduced (Muller, Spang, and Vingron 2002), and a new method generalizing the use of empirical matrices has been proposed (Goldman and Whelan 2002).

A number of models for possible heterogeneity in the substitution pattern at distinct sites have already been proposed, both for nucleotide data and amino-acid data. In the case of nucleotides, the transition/transversion ratio was modeled as a site-specific random variable (Huelsenbeck and Nielsen 1999). As for proteins, a first approach has been proposed, in which substitutional heterogeneity is introduced through a set of eight to ten predefined categories, based on secondary structure and solvent accessibility considerations (Goldman, Thorne, and Jones 1996; Thorne, Goldman, and Jones 1996; Goldman, Thorne, and Jones 1998; Liò and Goldman 1999). Each category has its own rate matrix, optimized by ML on real data sets. This model was shown to be significantly supported by real sequences, yet it does not address the question of the extent of heterogeneity actually present in the data. Furthermore, it makes specific hypotheses about its determining factors. An alternative method has been proposed in which no prior constraints are specified between the substitution processes and other features of the protein, like the secondary structure (Koshi and Goldstein 1998; Koshi, Mindell, and Goldstein 1999; Koshi and Goldstein 2001). However, the substitution processes themselves are constrained to conform to a prior biochemical model. Although this approach was generalized (Dimmic, Mindell, and Goldstein 2000; Soyer et al. 2002), the total number of categories is predetermined and is kept small, still for dimensionality reasons. A more radical approach was taken by Bruno (Bruno 1996; Halpern and Bruno 1998), through a model in which the equilibrium frequencies of the 20 amino acids are distinct at each site of the data set. The resulting model seems to capture important features of the substitution process along protein sequences, but it requires a large number of taxa in order for the statistics at each column to be significant.

As mentioned above, the flexibility of the Bayesian paradigm with respect to model dimensionality makes it possible to build models assuming high levels of heterogeneity. Yet, this does not tell how the number of parameters can be adjusted properly to match the signal present in the data. A possibility is to use mixture models in which the dimensionality is not fixed a priori and is itself a free parameter of the inference. Such mixture models are being introduced in many fields of applied statistics (Escobar and West 1995; Green and Richardson 1998), including bioinformatics (Eskin, Grundy, and Singer 2001; Broet, Richardson, and Radvanyi 2002), but they have not yet been applied to molecular phylogenetics.

Here we propose a mixture model, CAT, which generalizes most of the previous approaches (Bruno 1996; Koshi and Goldstein 1998; Dimmic, Mindell, and Goldstein 2000). The model allows for a number K of classes, each of which is characterized by its own set of equilibrium frequencies, and lets each site "choose" the class under which its substitutional history is to be described. The model can be constrained, with the number of classes fixed to one, as in the standard one-matrix model, or such that each site is described by its own class. Alternatively, we can use a Dirichlet process prior (Ferguson 1973; Antoniak 1974) on the space of equilibrium frequencies, to let the total number of classes be a free variable of the inference. The posterior mean value is then a measure of the substitutional heterogeneity actually present in the alignment. We have implemented this model in a Markov chain Monte Carlo (MCMC) framework, allowing joint inference to be performed simultaneously on all the parameters of the model, including the mixture parameters, the rates at each site, the branch lengths, and the topology of the underlying phylogenetic tree. Using this model, we have conducted inferences on large real data sets and found that, in all cases, the level of heterogeneity is much higher than has been accounted for by previous mixture models. In addition, we show that the standard model based on one single empirical matrix conditioning the substitution process at all sites is outperformed by CAT.

Materials and Methods

Data and Trees

We used three alignments of amino-acid sequences as follows:

EF30-627

These data were obtained as a subset of a larger alignment of eukaryotic elongation factor 2 sequences. Thirty taxa were chosen to represent the diversity of the eukaryotic lineages, and their aligned sequences were retrieved. We removed all columns for which gaps were present or data were missing, leaving a total of 627 sites. A phylogenetic reconstruction under the JTT+F model, performed with the implementation described below, yielded a tree which we used as the fixed topology under which the most time-consuming inferences were conducted. This tree was identical to the posterior consensus tree obtained with the MrBayes 3.0 program (Huelsenbeck and Ronquist 2001) using the JTT matrix, a Dirichlet prior with a concentration parameter of 1 on the equilibrium frequencies, and an Invariant + Γ rate prior modeled by 16 discrete rate categories.

Ek55-1525

These data are a concatenation of the sequences of four cytoplasmic proteins (actin, EF-1α, α and β tubulins), sampled across 55 eukaryotic species (Baldauf et al. 2000). The alignment was kindly provided by Sandra Baldauf. We removed the diplomonads and trichomonads from our

analysis because of their long branches. When constrained, the topology was fixed as in Baldauf et al. (2000).

Mt45-3596

The complete coding sequences of the mitochondrial genomes of 45 mammals were aligned with each other, and the ambiguous regions of the alignment were removed, yielding a data matrix of 45 × 3,596 amino acids. The MrBayes program was run on this alignment, using the MTREV empirical matrix, with four discrete Γ-categories. The resulting majority-rule consensus (using the *allcompat* option) was used as the fixed topology.

Notation and Parameters

The available data are in the form of an alignment of P amino-acid sequences, of length N. Let i index the columns C_i, or sites, and j the branches. All of the models used in this work assume that the sequences are related by an unknown, unrooted phylogenetic tree t, with branch lengths $\beta_j \geq 0$, $j = 1, 2, \ldots, 2P - 3$. Sites have their own rates of substitution $r_i > 0$, such that $\frac{1}{N}\sum_{i=1}^{N} r_i = 1$. Gaps are treated as missing data.

Markovian Model of Amino-Acid Replacement

For all models, substitutions occur according to a Markov process running along the branches of the tree. Such a Markov process is characterized by a rate matrix $Q = [Q_{lm}]$ that can be expressed in terms of 20 stationary probabilities, or equilibrium frequencies, (π_l), $1 \leq l \leq 20$, $\sum_{l=1}^{20} \pi_l = 1$ and a set of relative rates, or exchangeability parameters, (ρ_{lm}), $1 \leq l, m \leq 20$, according to the relation

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

$$Q_{ll} = -\sum_{m \neq l} Q_{lm}. \tag{2}$$

The process is assumed to be reversible, $\rho_{lm} = \rho_{ml}$, and the matrix has been scaled so that the substitution rate is one, using the normalization constant

$$Z = 2 \times \sum_{1 \leq l < m \leq 20} \rho_{lm}\pi_l\pi_m. \tag{3}$$

With this scaling, branch lengths are measured in expected number of substitutions per site. From Q, one obtains the transition probability matrix $P(\nu) = [P_{lm}(\nu)]$, specifying the probability that amino-acid l changes into m over an evolutionary distance of ν, as $P(\nu) = e^{\nu Q}$.

Mixture Modeling

We proposed a model that is site-heterogeneous with respect to the substitution process, and which we called CAT (because in effect, it classifies sites into categories). Under the CAT model, sites are distributed according to a mixture of K distinct classes. Each class is characterized by its own substitution matrix Q^k, and the class to which each site belongs is specified by an allocation variable $z_i \in [1..K]$. The vector $z = (z_i)_{i \in [1..N]}$ is called the allocation vector.

For simplicity, we consider only mixtures of matrices $\{Q^k\}_{k=1..K}$ having all the same set of relative rates ρ, but different stationary probabilities π^k, so that the mixtures are actually defined on the space of stationary probabilities (the π-space):

$$Q^k_{lm} = \frac{1}{Z^k}\rho_{lm}\pi^k_m, \quad l \neq m \tag{4}$$

$$Q^k_{ll} = -\sum_{m \neq l} Q^k_{lm}, \tag{5}$$

where, in order to constrain the substitution rate, we invoke a normalization constant

$$Z^k = 2 \times \sum_{1 \leq l < m \leq 20} \rho_{lm}\pi^k_l\pi^k_m. \tag{6}$$

With this normalization, the branch lengths have the same meaning under all classes; i.e., they are measured in expected number of substitutions. Because a class is entirely characterized by its π-vector, we call the latter its *profile*.

The relative rates ρ_{lm} can be fixed to pre-specified values, allowing different models to be specified. The mathematically simplest model is obtained by setting all the relative rates equal to unity: sites are described by a mixture of Poisson processes. Alternative models were also considered, based on the relative rates of known empirical matrices: JTT (Jones, Taylor, and Thornton 1992) and MTREV (Adachi and Hasegawa 1996), depending on the data set under study. The corresponding models were called CAT-POISSON, CAT-JTT, and CAT-MTREV, respectively.

The relative rates can as well be considered as free parameters of the inference (CAT-GTR model). Because they are defined up to a multiplicative constant, which cancels out upon normalization of the rate matrix, we impose the formal constraint that

$$\sum_{1 \leq l < m \leq K} \rho_{lm} = 1. \tag{7}$$

Once site-specific rates, substitution matrices, and allocation variables are known, the likelihood at each site is computed using Felsenstein's pruning algorithm (Felsenstein 1981). In the case where the processes are Poisson, it is possible to perform mathematically equivalent computations by recoding the process at each site in such a way that all nonobserved amino acids at a given column are collectively considered as one single state. The stationary probability of this new state is the sum of the stationary probabilities of all non-observed amino acids. Felsenstein's pruning algorithm is in S^2P, where S is the number of states, and P the number of taxa. For amino-acid data without the recoding, $S = 20$. Under the recoding, S will be in general less than 20, and for highly conserved alignments it can often be as low as 2 or 3, which yields a significant increase in computational speed (up to a 50-fold increase in speed was observed in the case of the eukaryotic data set Ek55-1525).

Priors

We used uninformative priors on t, β, r and ρ, as follows:

- $t \sim$ *Uniform* over topologies
- $\beta_j \sim$ *Uniform*$[0, \beta_{max}]$, with $\beta_{max} = 100$
- $r \sim$ *Dirichlet*$(1, 1, \ldots, 1)$
  The resulting marginal distribution at each site is a γ distribution, with an α parameter of 1.
- $\rho \sim$ *Dirichlet*$(1, 1, \ldots, 1)$

In addition, we defined a prior on the mixtures, using a Dirichlet process (DP) (Ferguson 1973; Antoniak 1974). A DP prior is parameterized by a concentration parameter α, and a base distribution $G_0(\pi)$ defined on the π-space, the space of stationary probabilities. It can be described by the following procedure for randomly drawing a configuration from the prior (Neal 2000):

- Draw the number of classes K, together with a N-vector of allocation variables z, according to $p(z, K \mid \alpha)$.
- Draw K values i.i.d. in the π-space: $\forall k$, $\pi^k \sim G_0$

$p(z, K \mid \alpha)$ can be expressed as

$$p(z, K \mid \alpha) = \alpha^K \frac{\prod_{k=1}^{K} (\eta_k - 1)!}{\prod_{i=1}^{N} (\alpha + i - 1)}, \qquad (8)$$

where η_k stands for the ocupancy number of class k (i.e., the number of sites allocated to class k). Integrating this expression over z yields the marginal prior distribution on the number of classes, $p(K \mid \alpha)$, which has a complicated form, not reported here (Antoniak 1974; Escobar and West 1995). An intuitive formulation of its dependence on α is that, given α, the marginal prior probability that two different columns taken at random belong to the same class is $1/(1 + \alpha)$ (Neal 2000). Thus, α defines the level of heterogeneity a priori assumed, with higher values favoring a larger number of classes. We considered α as a free parameter, with a flat prior between two extreme values, and zero outside.

$$\alpha \sim Uniform[\alpha_{min}, \alpha_{max}]$$

We set α_{min} to 0.001, and α_{max} to 1,000. This prior on α induces a smooth marginal prior distribution on K, allowing for the posterior mean number of classes to be determined principally by the data. Finally, we used a flat Dirichlet prior as the base distribution on the sets of stationary probabilities:

$$G_0(\pi) = Dirichlet(1, 1, \ldots, 1)$$

In particular cases, one can also dispense with the DP prior and constrain the value of K. Two cases were considered:

- The one-matrix model: $K = 1$. This corresponds to the usual models: Poisson, JTT, mtREV and, when the relative rates ρ_{lm} are considered as free parameters, GTR. The stationary probabilities of the single substitution matrix can be either free or fixed to their empirical estimates. The models using empirical estimates are called Poisson+F, JTT+F, WAG+F, or mtREV+F, depending on the set of relative rates that are used (in the case of the GTR model, stationary probabilities are always considered as free parameters).
- The maximally heterogeneous model: $K = N$, where N is the number of sites. Each site evolves under its own substitution process. The N sets of 20 stationary probabilities are considered unknown parameters ($19N$ free parameters). The resulting model is called MAX, and under equal relative rates is similar to the model proposed by Bruno (1996).

Implementation and Monte Carlo Sampling

Gibbs Sampling of the Dirichlet Process

To visualize how a DP prior mixture model works, one can make explicit the probability of all possible allocations of a given site, conditional on the rest of the data. For any given site i, we denote collectively by z_{-i} and π_{-i} the allocation variables and the stationary probabilities at all other sites. In addition, for any class k, $\eta_{k,-i}$ stands for the number of sites other than i allocated to class k (i.e., the number of $j \in [1..N]$, $j \neq i$ such that $z_j = k$). Then, one has (Neal 2000):

$$p(z_i = z_j \text{ for some } j \neq i \mid C_i, z_{-i}, \pi_{-i}) = Z \eta_{z_j,-i} p(C_i \mid \pi^{z_j}). \qquad (9)$$

$$p(z_i \neq z_j \text{ for all } j \neq i \mid C_i, z_{-i}, \pi_{-i}) = Z\alpha \int p(C_i \mid \pi) G_0(\pi) d\pi. \qquad (10)$$

Here, Z is the appropriate normalizing constant that makes these probabilities sum to one. The dependence of $p(C_i \mid \pi)$, the likelihood at the ith. column, on all parameters other than the stationary probabilities π (like the topology and the branch lengths) has been omitted for notational simplicity.

These two equations immediately suggest a Gibbs sampling algorithm, in which each site is taken in turn, and its allocation variable is reassessed according to these conditional probabilities. Note that the total number of classes K can change through this update. Thus, if the site had a class on its own, but is re-allocated to another already existing class, K will decrease by one. In the reverse situation, if site i were initially allocated to a class k such that $\eta_{k,-i} > 0$, and ends up allocated to a new class, then K will increase. In all other cases, K will remain constant. Upon iteration, and when combined with other updates (see below), this algorithm allows K to fluctuate across the whole range $[1..N]$.

What determines the equilibrium level reached by K? First, one can see from equation 9 that, not surprisingly, the choice among alternative allocations of site i to already existing classes is driven by the relative likelihoods of these allocations. In contrast, equation 10 shows that the probability of letting the site have a class on its own is mainly determined by how the *prior* expectation of the site's likelihood over all possible sets of stationary probabilities compares with likelihoods under already existing classes. This amounts to comparing the performance of

a *non-fitted* new profile to the currently available ones, *which have already been fitted* to the rest of the data. This asymmetry makes it difficult for a site to induce the creation of a new class, unless the pattern at the corresponding column of the alignment is sufficiently distinct from all other columns. Finally, the weights ($\eta_{z,-i}$ in equation 9, or α in equation 10) represent the net variations of the prior factor $P(z, K \mid \alpha)$. That the weight is α in equation 10 makes it apparent that α has a direct influence on the propensity of proposing new classes, and thus on the stationary value of K.

In our MCMC sampler, we rely on a slightly modified version of this update mechanism (Neal 2000), as follows:

- SwitchMode: In addition to the K currently existing classes, κ new classes are created, with profiles drawn from G_0. One site (i) is then chosen at random, and for every $k \in [1, K + \kappa]$, site i is allocated to class k, the corresponding likelihood for site i (L_i^k) is computed, and a Gibbs sampling over k is performed based on the probabilities

$$p^k = Z * w^k * L_i^k, \tag{11}$$

  where $w^k = \eta_{k,-i}$, if $\eta_{k,-i} > 0$, and $w^k = \alpha / \kappa$ otherwise. Different values of κ do not change the stationary distribution, but they yield faster convergence and better mixing. We empirically set κ to 5.

SwitchMode is used in combination with two other operators, one updating the stationary probabilities of one class at a time (StationaryMove) while keeping z and α constant, and another one proposing a new value for α.

- StationaryMove: one site is chosen at random, an update of the profile of the class to which it is allocated is proposed according to a Dirichlet distribution centered on the current value, with a tuning parameter of λ_S.
- Alpha: A random real value $\delta = \lambda_\alpha * (U - 0.5)$, with a tuning parameter $\lambda_\alpha > 0$, is added to α, with back-reflection in $[\alpha_{min}, \alpha_{max}]$, if required. The Hastings ratio equals 1, and the likelihood does not depend on α, so that the only nontrivial factor involved in this Metropolis update is the resulting change in $P(z, K \mid \alpha)$, yielding the acceptance probability:

$$r = Min\left(1, \frac{p(z, K \mid \alpha^*)}{p(z, K \mid \alpha)}\right) = Min\left(1, \frac{\alpha^{*K}}{\alpha^K} \prod_{i=1}^{N} \frac{\alpha + i - 1}{\alpha^* + i - 1}\right), \tag{12}$$

  where α and α^* stand for the current and the proposed values of α, respectively.

Topology and Branch Lengths

We used the Global and Local algorithms proposed by Larget and Simon (1999). In addition, we also devised a node-sliding operator, as well as three operators proposing new values for the branch lengths, while leaving the topology invariant:

- NodeSliding: we randomly pick an internal branch of the unrooted tree and then designate its two nodes u and v. The other two neighbors of u are called a and b, and those of v are called c and d, with equal probability. We then slide the branch $a - u$ along the segment $b - u - v - d$, by a distance $l = \lambda_N * (U - 0.5)$, where $U \sim Uniform[0, 1]$, and $\lambda_N > 0$, is a tuning parameter. If $l > 0$, the move is made toward d, and toward b otherwise. In addition, if the move pushes the branch out of the range defined by b and d, the excess is reflected back into the required interval. The Hastings ratio for this proposal is 1.
- OneBranch: one branch of the unrooted tree is chosen at random, and its length is multiplied by a random factor $r = e^{\lambda_0(U-0.5)}$, where $U \sim Uniform[0, 1]$, and $\lambda_0 > 0$, is a tuning parameter. The Hastings ratio equals r.
- AllBranch: all branch lengths are updated simultaneously, each β_j being multiplied by a distinct random factor $r_j = e^{\lambda_A(U_j-0.5)}$, where λ_A is a tuning parameter. The Hastings ratio is $\prod_j r_j$.
- Homothetic: all branch lengths are updated simultaneously, as in OneBranch, all being multiplied by the same random factor $r = e^{\lambda_H(U-0.5)}$. The Hastings ratio is r^{2P-3}.

Rates Across Sites

As in Larget and Simon (1999), we propose a new set of values according to a Dirichlet distribution centered on the current parameter value. We found that this Monte Carlo operator was more efficient if applied only on a subset of the rate vector. Specifically, a small number q of sites, $i_1, i_2, ..i_q$, are chosen at random, and a new set of values, according to a Dirichlet distribution restricted on $i_1, i_2 \ldots i_q$, with weights $\lambda_D r_{i_1}, \lambda_D r_{i_2}, \ldots, \lambda_D r_{i_q}$, was proposed as un update. λ_D is a tuning parameter. Empirically, we found that $q = 10$ gives a good mixing.

General MCMC Settings

During sampling, the different components of the hypothesis vector were updated separately, in random alternation, according to a grid of weights $(w_m)_{1 \le m \le M}$, where M is the total number of operators. Specifically, defining $W = \sum_{m=1}^{M} w_m$, we call a *cycle* a series of W iterations. For each iteration, m is drawn at random, according to $m \sim w_m/W$, and the corresponding operator is called to act on the current state. The weights (w_m) were determined empirically, like the tuning parameters, to optimize the mixing of the Markov chain. Both the weights and the tuning parameters are dependent on the model, as well as on the data set under which the inference is conducted (see table S1 in the Supplementary Material online).

For each run, the convergence was assessed by checking for the absence of long-term trends in a series of key monitor functions. Specifically, we monitored the log likelihood, the tree length, the entropy of the rate distribution, the number of classes, and the average profile entropy (class-occupancy weighted average). In addition, in most cases, at least two independent runs were performed, starting from different points of the parameter space taken at random, and their marginal properties (majority-rule consensus tree, class number, and composition) were compared.

Under CAT-Poisson, CAT-JTT, CAT-mtREV, and CAT-GTR, we did a total of 600,000 update cycles. The first 100,000 points were discarded, and we subsampled every 50 cycles after burn in. Under JTT+F and MTREV+F, the only free parameters are the site-specific rates of substitution and the branch lengths, which makes convergence much faster, so that we only did a total of 100,000 update cycles, removed the first 20,000, and subsampled every 10 cycles. Under GTR, we did a total of 200,000 update cycles, removed the first 40,000, and subsampled every 10 cycles.

Clustering

To identify the classes that are stable across a MCMC run, we pooled all the classes of all the points of the run (burn-in discarded), and defined clusters based on the degree of similarity between the corresponding profiles. Specifically, for each pair of classes (h, k), we computed the quadratic distance between their profiles,

$$d = \left[\sum_{l=1}^{20} (\pi_l^h - \pi_l^k)^2\right]^{\frac{1}{2}}, \quad (13)$$

and considered that two profiles belong to the same cluster if the distance separating them is below a predefined threshold $d_{max} = 0.01$.

Note that any given point of the sample may have several of its classes contained in some clusters, while it may not be represented in other clusters. A given cluster is referred to as a *stable cluster* if one and only one class is affiliated to this cluster for more than 80% of the points of the sample. For each stable cluster, one can compute the following:

- its mean occupancy number: specifically, the sum of all its classes' occupancy number divided by the sample size.
- mean profile: occupancy number weighted average of its classes' profiles.

A stable cluster can be interpreted as a class that can be identified unequivocally across the MCMC sampling, and thus, we will refer to stable clusters directly as *stable classes*.

We developed software for principal component analysis. Plots were visualized using *Gnuplot* (http://www.gnuplot.info). We wrote a program for visualizing profiles using a representation akin to sequence logos (Schneider and Stephens 1990), translating sets of stationary probabilities into PostScript files.

Posterior Predictive Resampling

One of the ways to evaluate the performances of alternative models is to compare their predictions on real and simulated data (Gelman, Meng, and Stern 1996). For instance, in the present context, we wish to compare the number of classes inferred by CAT on real data with that obtained on data simulated under CAT itself, or under a standard model, such as JTT. For these comparisons to be meaningful, it is important to simulate the replica under sensible parameter values. In the maximum likelihood framework, one would choose the ML estimate. In a Bayesian analysis, it is more customary to sample the replica from the posterior predictive distribution (Rubin 1984; Gelman, Meng, and Stern 1996; Gelman et al. 2004). Given a data set D, and a model M, parameterized by $h \in \Omega$, this distribution is defined as:

$$P(D^{rep} \mid M, D) = \int_\Omega P(D^{rep} \mid h, M) P(h \mid D, M) dh, \quad (14)$$

where

$$P(h \mid D, M) = \frac{P(D \mid h, M) P(h \mid M)}{P(D \mid M)} \quad (15)$$

is the posterior probability distribution over the parameters, induced by the data.

In practice, a collection of R replicas are obtained using the following procedure: first, run a MCMC under model M, with data D, discard the burn-in, and take R points regularly spaced in the remaining part of the chain $(h_r)_{1 \le r \le R}$. Next, for each r, simulate a replica D_r under h_r.

Numerical Evaluation of the Bayes Factor by Thermodynamic Integration

The most common Bayesian method of model comparison consists in computing the Bayes factor (Jeffreys 1935, 1961; Jaynes 2003). The Bayes factor in favor of model M_1 against model M_0, given the data D, is defined as

$$B_0^1 \equiv \frac{P(D \mid M_1)}{P(D \mid M_0)}. \quad (16)$$

$P(D \mid M_i)$, $i = 0, 1$, is the likelihood averaged over the prior (or *marginal* likelihood):

$$P(D \mid M_i) = \int_\Omega P(D \mid h, M_i) P(h \mid M_i) dh, \quad (17)$$

where h stands for the parameter vector, and Ω for the whole parameter space. Model M_1 will be supported over model M_0 if B_0^1 is greater than 1, which amounts to choosing the model that has the higher marginal likelihood. To compute Bayes factors, we used a numerical method based on an analogy with thermodynamics; it is called *thermodynamic integration* (Ogata 1989). The details of this method, also known as *path sampling* (Gelman 1998), are explained in the Appendix.

When more than two models are being compared, one can exploit the additivity property of the logarithm of the Bayes factor, i.e., $\forall M_0, M_1, M_2$, $\ln B_1^2 = \ln B_0^2 - \ln B_1^2$. Therefore, one only needs to compute the Bayes factors of each of the models with respect to the same reference model M_0, and compare their values directly. In the present study, Poisson+F was used as the reference.

All source code files and alignments are available upon request.

Results

We first conducted inferences under the CAT-Poisson Model (i.e., a free number of Poisson processes), on the elongation factor 2 data set, EF30-627. The topology was

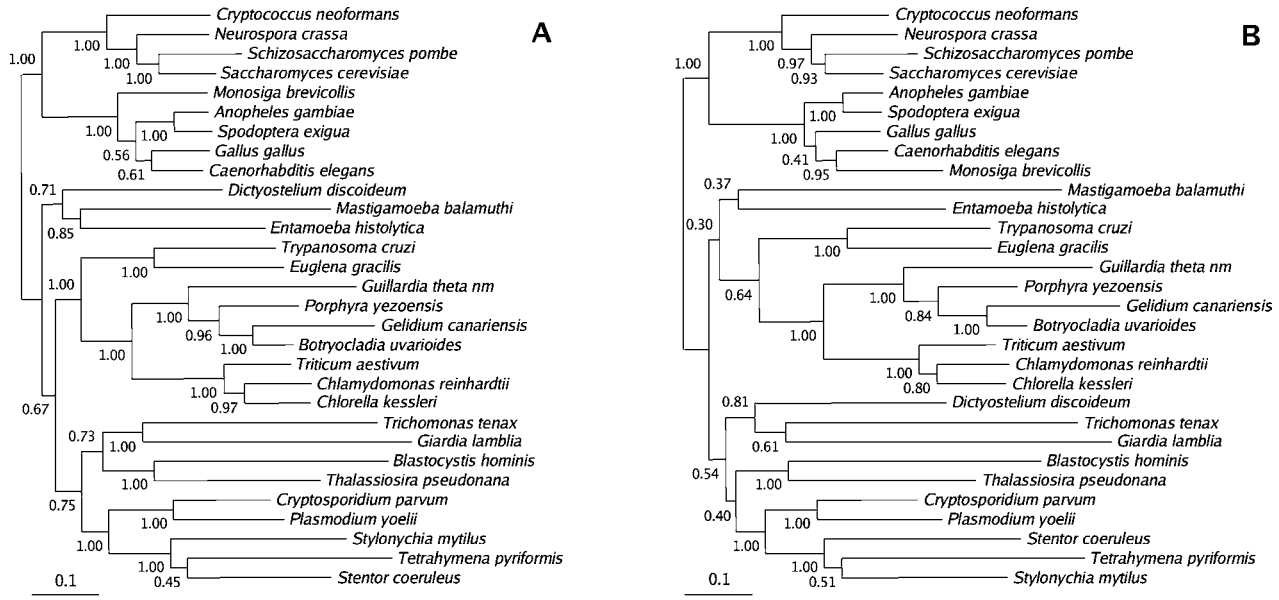

FIG. 1.—Majority rule consensus phylogenies obtained with the models JTT+F (*A*) and CAT-POISSON (*B*). Branch lengths are proportional to the expected number of substitutions per site. Node support values are equal to the posterior probabilities.

constrained according to the posterior consensus tree found by running our program on this data set using the JTT+F model (fig. 1*A*), while all other parameters (branch lengths, site-specific rates, class number and profiles, as well as the allocations of sites to modes) were left unconstrained.

Figure 2*A* shows the evolution of the value of K, the number of classes, during the elongation of two independent MCMC chains. The chains were initialized at $K = 1$ and $K = N$, respectively. Irrespective of the starting point, K converged to a well-defined interval centered on 28 (28.4 ± 3.5). The histograms of the frequencies estimated from the two independent chains are very similar (fig. 2*B*), the corresponding 95% credibility interval being [22,35] and [21,36]. Substitutional homogeneity across sites is thus rejected by our model ($p(K = 1 \mid D) \simeq 0$). On the other hand, the maximally heterogeneous situation is also excluded by the present analysis ($p(K = N \mid D) \simeq 0$). In fact, the mean posterior number of classes, as inferred from CAT-POISSON, is low, when compared to the total number of columns of the alignment (an average of one class per 22 sites). It should be stressed that, even with as few as 28 classes, the model is dealing with a relatively high number of additional parameters, compared to JTT+F (specifically, $28 \times 19 = 532$ free continuous parameters for the profiles, and $N = 627$ discrete parameters for the allocation vector, where N is the length of the alignment).

We performed the same analysis without constraining the topology. We found a topology (fig. 1*B*) similar to that obtained using the JTT+F model (fig. 1*A*). A few differences are present, however: under CAT-POISSON, the choanoflagellate *Monosiga* is found within the metazoan clade, and not as its sister group. In addition, *Dictyostelium* is clustered with diplomonads and trichomonads. Finally, stramenopiles became a sister group of the alveolates, instead of the diplomonads and trichomonads. In spite of these differences, however, the mean number of classes ($\langle K \rangle = 27.9 \pm 3.6$) is close to that obtained when the topology is constrained, suggesting that the analysis performed under CAT-POISSON is robust against phylogenetic uncertainty.

Interestingly, the inferred total length of the tree is strongly model dependent: under the constrained topology, we observed a posterior mean number of 8.13 ± 0.21 substitutions per site under CAT-POISSON, versus 7.46 ± 0.14 under JTT+F (table 1).

Stable Class Identification

Any two points of a given Markov chain obtained under the CAT model may differ in all respects (class number, site to class allocations, class specific profiles, as well as branch lengths and site-specific rates), and there is no a priori obvious way of identifying a given class throughout the sample. On the other hand, there may be some stable patterns. To identify them, we took 1,000 regularly spaced points from a given run and pooled all the classes observed at each of these points. Each class is characterized by its profile, and can be assimilated to a vector in a 20-dimensional space. A principal component analysis (PCA) on the set of pooled classes is shown in figure 3*A*. Two kinds of patterns can be seen: first, a broad set of dots scattered more or less homogeneously across a large region in the center of the projection, and second, a series of 10 to 12 dense clouds. To further characterize this distribution, we used a clustering method (see *Materials and Methods*), and identified a series of 11 stable classes (i.e., classes that can be unequivocally identified across the sample). The mean profile and the mean occupation number of each identified stable class is shown in figure 3*D*, and a projection of their profiles back onto the PCA (fig. 3*A*, large crosses) clearly shows that these stable classes correspond to the dense clouds.

The profiles of the stable classes are biochemically reasonable (fig. 3*D*). For instance, there are classes with

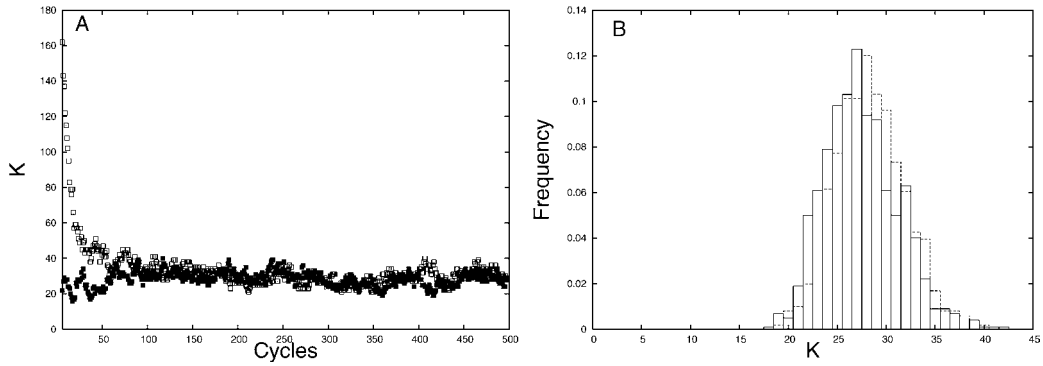

FIG. 2.—*A*. Traceplot showing the evolution of K, the number of modes, during the elongation of two independent MCMCs (only the first 500 cycles are shown). The initial conditions were $K = 1$ (circles) and $K = N$ (squares), with N standing for the number of sites. *B*. Histograms displaying the frequencies at which each value of K was observed along the two chains referred to in *A*, starting from $K = 1$ (solid lines), and $K = N$ (dashed lines).

negatively charged residues (D and E), with positively charged residues (K and R), or with aromatic (F and Y) residues. Note the diversity of alternative classes belonging to the same biochemical category: there are two hydrophobic classes, one with the amino acids I and V, and one favoring L and M. Finally, some amino acids are found in several stable classes, like S, which belongs to both (A,S) and (S,T) classes; the most extreme example is that of alanine (A), which is present in five different classes, (A,S), (A,G), (A,P), (A,C,S,T,V), and (A,D,E,G,K,Q,N,S,T,V). Thus, according to these results, the same amino acid can undergo different types of substitutions depending on the context. The same cluster analysis was performed without constraining the topology and gave essentially identical results (see fig. S1, in the Supplementary Material online).

The CAT-POISSON model was applied to data sets of larger size, consisting in the concatenation of four nuclear proteins from 55 eukaryotes (Ek55-1525), and of the mitochondrial proteins of 45 mammals (Mt45-3596). We found significant support in favor of heterogeneity in all cases (table 1), as shown both by the posterior mean number of classes (26 and 35) and by the number of stable classes identified by clustering (14 and 22).

Biochemically reasonable classes are found in all cases (see figures S2 and S3 in the Supplementary Material online), and they bear many similarities with each other across the three data sets, in particular between EF30-627 and Ek55-1525, suggesting that the number and the composition of the classes inferred by CAT-POISSON correspond to generic properties of the substitution patterns in proteins. More significant differences are observed between mitochondrial proteins and the other two data sets. This might reflect that mitochondrial proteins are mostly transmembrane proteins, whereas the proteins included in the two other data matrices are exclusively cytosolic factors. This is in agreement with the fact that the MTREV is markedly different from other empirical substitution matrices (Adachi and Hasegawa 1996).

Posterior Predictive Checks

We compared the inferences conducted on EF30-627 with the results obtained from data sets of the same size, but simulated under various conditions. First, for data simulated under one single Poisson process and analyzed under CAT-POISSON, one class was recovered with significant probability ($P = 0.65$). Next, we simulated data under CAT-POISSON, drawing the parameters of the simulation from the posterior predictive distribution (10 replica, see *Materials and Methods*). The posterior mean number of classes found on the simulated data sets ($\langle K \rangle = 23.1 \pm 2.7$) is slightly lower than the value found on the real data set ($\langle K \rangle = 28.4 \pm 3.5$), suggesting that, as an estimate of the number of classes, $\langle K \rangle$ is biased downward. However, the number of identified stable classes is very similar ($K_{SM} = 11.20 \pm 1.08$ versus $K_{SM} = 11$), as well as the underlying profiles (not shown). A PCA projection of the classes obtained in one of these simulations (fig. 3*B*) does not display significant differences with the analysis

Table 1
Estimates of Class Number and Tree Length Under CAT-POISSON, MAX-POISSON and JTT+F

	CAT-POISSON				MAX-POISSON	JTT+F
Data Set	$\langle \alpha \rangle$[a]	$\langle K \rangle$[b]	K_{SM}[c]	$\langle TL \rangle_{CAT}$[d]	$\langle TL \rangle_{MAX}$[e]	$\langle TL \rangle_{JTT+F}$[f]
EF30-627	6.3 ± 1.6	28.4 ± 3.5	11	8.13 ± 0.21	7.06 ± 0.14	7.46 ± 0.14
Ek55-1525	4.7 ± 1.3	26.4 ± 3.7	14	5.36 ± 0.10	4.82 ± 0.07	4.78 ± 0.08
Mt45-3596	5.4 ± 1.1	35.3 ± 3.2	22	7.54 ± 0.10	5.90 ± 0.05	5.91 ± 0.05

[a] Mean posterior value of the Dirichlet Prior concentration parameter.
[b] Mean posterior number of classes.
[c] Number of stable classes detected by clustering.
[d] Mean posterior tree length under CAT.
[e] Mean posterior tree length under MAX-POISSON.
[f] Mean posterior tree length under JTT+F.

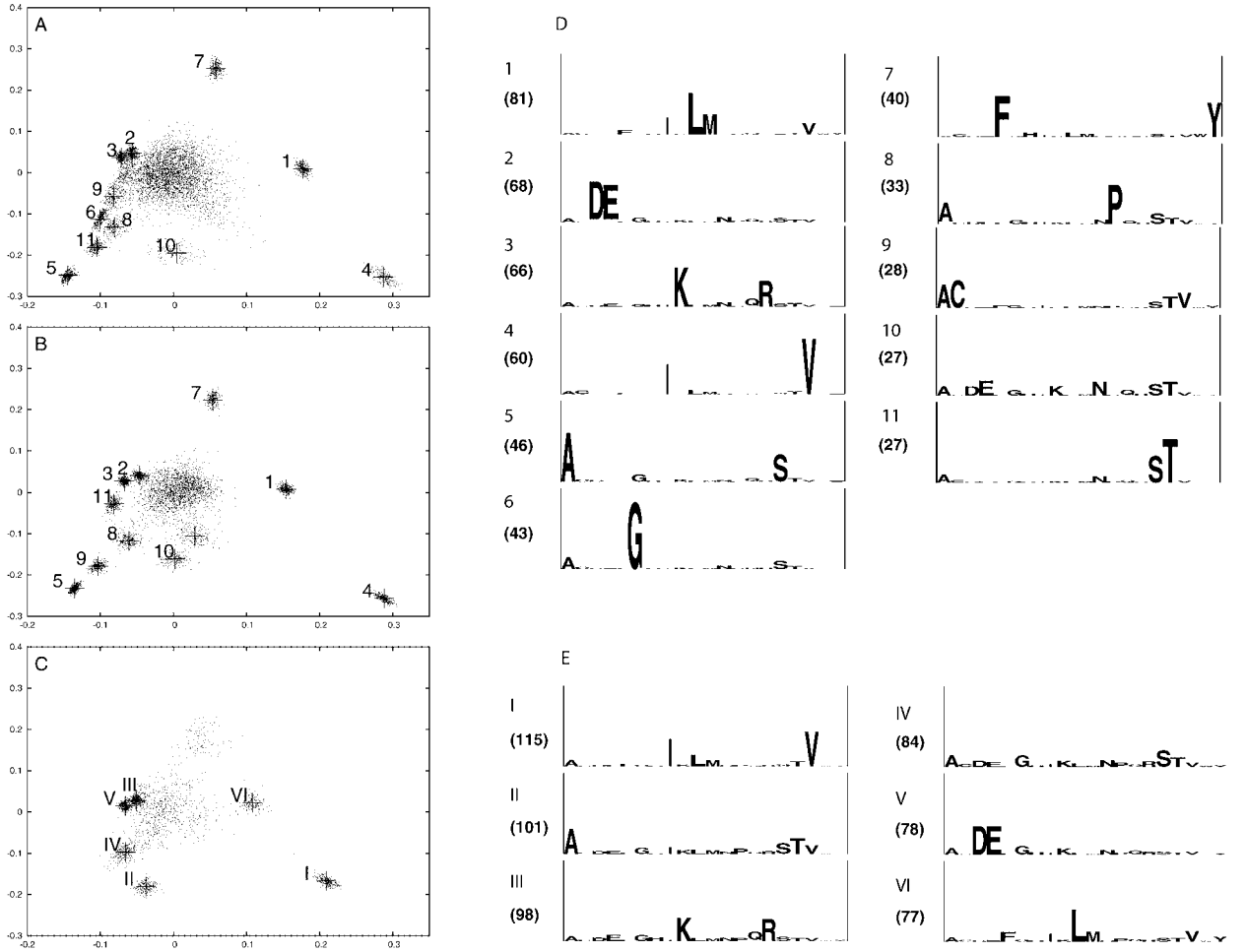

FIG. 3.—*A*, *B*, *C*. Principal Component analysis of the profiles obtained under the CAT model, using *A*, real sequences (EF30-627); *B*, data simulated under CAT (SimuCAT30-627); *C*, data simulated under JTT (SimuJTT30-627). To make comparison easier, the profiles in *B* and *C* are projected onto the 2-dimensional subspace of $\mathbf{R}^{20}$ defined by the principal components computed in *A*. First axis accounts for 17.3%, and second axis for 12.7%, of the total variance in *A* and *B*, and 15.3% and 11.3% of the variance in *C*. *D*. Profiles of the 11 stable modes obtained with EF30-627. The one-letter amino-acid code is used. Font height is proportional to the frequency of the corresponding amino-acid. Modes are reported on panels *A* and *B*. *E*. Profiles of the six stable modes obtained with JTTSimu30-627. Modes are reported on panel *C*.

performed on the real data set (fig. 3*A*). Thus, in spite of a slight bias, the mixtures inferred by CAT-POISSON appear to be quite robust to posterior predictive resampling.

In contrast, when the same experiment is performed on alignments that have been obtained by simulation under the JTT+F model, a much lower level of heterogeneity is detected ($\langle K \rangle = 13.7 \pm 1.6$, n = 10 replica). Upon clustering, an average of $K_{SM} = 6.8 \pm 0.8$ stable classes is obtained. The profiles of these stable classes display significant differences with those observed for real data (fig. 3*C* and *E*): for instance, there is only one class containing A, S, and T, in place of the two (A,S) and (S,T) classes found on real sequences, and the (F,Y) class has disappeared altogether.

CAT Inferences Using Non-Poissonian Rate Matrices

The previous experiment suggests that the information contained in the JTT matrix is able to account for only part of the substitutional heterogeneity across sites present in real data. To investigate this problem further, we set the relative rates underlying our mixture model equal to those of the JTT matrix. (in the case of mitochondrial proteins, we used the MTREV coefficients instead). Our argument is that, if one single empirical matrix (JTT) is sufficient to account for the substitutional heterogeneity of a particular data set, then an inference conducted under the CAT-JTT model on this data set should yield a nonvanishing posterior probability of having one single class.

We performed a CAT-JTT inference both on EF30-627 and on a data set drawn from the JTT+F induced posterior predictive distribution. Whereas the probability of recovering a single class is high in the case of the simulated data ($P = 0.6$), it is virtually 0 for real data (fig. 4, and table 2). Surprisingly, the posterior mean number of classes is even greater with CAT-JTT ($\langle K \rangle = 51.6 \pm 8.3$) than with CAT-POISSON ($\langle K \rangle = 28.4 \pm 3.5$), although when a cluster-analysis is performed, a lower number of stable classes than in CAT is obtained ($K_{SM} = 5$). The same analysis was conducted on Ek55-1525, and on the mitochondrial data set Mt45-3596, (in the latter case,

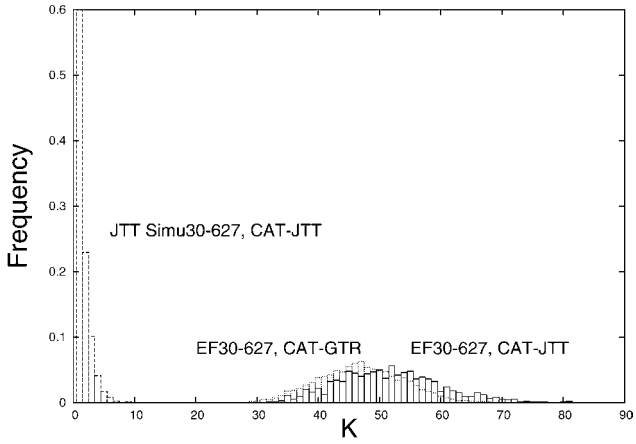

FIG. 4.—Histograms showing the estimated posterior probability distribution of K, the number of modes under the CAT-JTT for data simulated under JTT+F (dashed lines), under CAT-JTT for the real data (EF30-627, solid lines), and under CAT-GTR for the real data (dotted lines).

using MTREV as the empirical matrix). In both cases, the posterior probability of the data being described by one single class is virtually zero, and the posterior mean number of classes is also higher under CAT-JTT or CAT-MTREV (table 2), than under CAT-POISSON (table 1).

In a last experiment, the relative rates were considered as unknown parameters (although they were still constrained to be the same for all classes). The results were qualitatively the same as with CAT-JTT or CAT-MTREV, with a posterior mean number of classes of $\langle K \rangle = 47 \pm 7$ inferred on EF30-627 (fig. 4). Thus, even under non-uniform relative rates, a significant part of the heterogeneity among sites has to be further accounted for by forcing the mixture to contain more than one class. In other words, models based on one single matrix are far from capturing the substitutional heterogeneity of real data to its full extent.

Bayesian Model Comparison

Bayes factor evaluation is the most common method of model comparison in Bayesian inference (Gelman 1998; Jaynes 2003), and has already been applied in phylogenetics (Suchard, Weiss, and Sinsheimer 2001). Using the numerical method called *thermodynamic integration* (Ogata 1989, see Appendix), we computed the Bayes factor between POISSON+F, taken as the reference model, and CAT-POISSON, MAX-POISSON, and a series of empirical matrix models (JTT+F, WAG+F, and MTREV+F; table 3). In the case of EF30-627, we also included the general model GTR in the analysis.

Our results first allow comparison of the one-matrix models among themselves: as expected (Whelan and Goldman 2001), WAG+F performs better (and MTREV+F worse) than JTT+F on cytosolic proteins (EF30-627 amd Ek55-1525). Interestingly, JTT+F and WAG+F are better than GTR on EF30-627, suggesting that the relative rates of known empirical matrices are close to optimal for this data set. Irrespective of their relative performances, however, the one-matrix models are all outperformed by CAT-POISSON. For example, the support in favor of CAT-POISSON with respect to WAG+F on EF30-627 is of $1{,}932 - 1{,}760 = 172$ natural units of logarithm, which corresponds to 0.27 natural log units per site. This means that, on average, each column is $e^{0.27} = 1.32$ times better explained by CAT-POISSON than by WAG+F. In contrast to CAT-POISSON, MAX-POISSON does not seem to be favored over one-matrix models: for all three data sets, MAX-POISSON was by far the worst-performing model after POISSON+F (table 3).

Table 2
Estimates of the Number of Classes Under CAT-JTT and CAT-MTREV

Data Set	$P(K = 1)$	$\langle K \rangle$	K_{SM}
SimuJTT30-627	0.60	1.7 ± 1.1	1
EF30-627	0	51.6 ± 8.3	5
SimuJTT55-1525	0.64	1.8 ± 1.4	1
Ek55-1525	0	45.7 ± 6.0	14
Mt45-3596	0	70.9 ± 6.6	16

Discussion

Characterizing Substitutional Heterogeneity Using a Dirichlet Process Prior Model

We have developed a Bayesian model of sequence evolution, CAT, that can account for substitutional heterogeneity across the sites of protein sequences. In contrast to already existing mixture models designed for the same purpose (Goldman, Thorne, and Jones 1996; Thorne, Goldman, and Jones 1996; Goldman, Thorne, and Jones 1998; Koshi and Goldstein 1998; Koshi, Mindell, and Goldstein 1999; Liò and Goldman 1999; Dimmic, Mindell, and Goldstein 2000; Koshi and Goldstein 2001), all of which have a fixed number of classes, the number of substitutional classes in CAT is itself a free parameter. In this way, it will adapt to the substitutional complexity of the underlying data set, and it provides an estimate of this complexity through the posterior mean number of classes.

Inferences using the CAT model on real protein data sets uncover a high level of heterogeneity, much higher than what is assumed by all other mixture models that have been proposed thus far. To give a numerical estimate, under CAT-POISSON, we always found a posterior mean number of classes greater than 20, whereas in all other models, the number of classes is always set to a value lower than 10. Moreover, the difference between the posterior mean number of classes found upon posterior predictive resampling and the estimate obtained on the real data set (23.1 ± 2.7 versus 28.4 ± 3.5) reveals the existence of a bias, suggesting that the actual level of heterogeneity is probably even higher than what we have observed.

When looking at the classes' profiles and occupancy numbers, one sees that there actually exist markedly different kinds of classes. On the one hand, using a simple clustering algorithm, we were able to identify a restricted set of classes that are stable along a given Markov chain, and are represented by many sites across the sequences. The profiles of these classes (fig. 3*D*) correspond to reasonable biochemical features, although they are more diverse and more refined than the basic biochemical

Table 3
Model Comparison. Natural Logarithm of the Bayes Factor Between Poisson+F and All Other Models

Data Set	JTT+F	WAG+F	mtREV+F	GTR	CAT-Poisson	MAX-Poisson
EF30-627	1,631	1,760	1,374	1,550	1,932	807
Ek55-1525	3,324	3,628	2,848	—	3,808	1,112
Mt45-3596	10,089	8,970	10,943	—	12,389	855

categories generally used. Strikingly, most of them are markedly peaked, giving a significant weight to only two or three amino acids, suggesting the presence of a strong, site-specific, selection pressure at the amino-acid level, which would be one of the dominant forces shaping the substitution process. Another feature is that a given amino acid can belong to several distinct classes, indicating that context-dependent amino-acid exchangeability is an important aspect of protein evolution. Note that this property cannot be accounted for by the standard one-matrix models. Similar observations have been made previously, using a parsimony-based approach (Lopez 1997).

On the other hand, the difference between the posterior mean number of classes and the number of stable classes (28.4 $\pm$ 3.5 versus 11 in the case of EF30-627), reveals the existence of a large series of classes which are not easily identifiable. Closer examination indicates that most of these classes have a low occupancy number; some of them appear in a transitory fashion along a given run, and others, although persistent, do not always have the same sites affiliated to them. These classes should be interpreted with caution: they might not correspond to actual substitutional categories, and they might instead be considered as formal mathematical objects, relevant only through their marginal average influence on the substitution processes operating at the corresponding sites.

The fact that some classes are stable, while some other are not, offers an interesting insight into the flexibility of the Bayesian mixture models such as CAT-Poisson: obviously, the set of unstable classes, meaningful only through their average influence on the site-specific substitution processes, is a typically Bayesian feature, and it does not have an equivalent in existing heterogeneous models implemented in the ML framework. Incidentally, this suggests that the level of heterogeneity that these ML models assume should be compared not to the posterior mean number of classes found under CAT-Poisson, but rather, to the number of stable classes. This number provides a much more conservative estimate of the level of heterogeneity; yet, we found it to be consistently higher than 10, ranging from 11 to 22, confirming that the heterogeneity has been underestimated thus far.

Comparison with Other Heterogeneous Models

Thorne, Goldman, and co-workers propose to associate amino-acid replacement patterns with protein secondary structure and solvent accessibility (Goldman, Thorne, and Jones 1996; Thorne, Goldman, and Jones 1996; Goldman, Thorne, and Jones 1998; Liò and Goldman 1999). In practice, their model implements a mixture of 10 classes, corresponding to five types of secondary structure and two levels of solvent accessibility. Using parametric bootstrap, they show that their mixture model improves significantly on the standard homogeneous model. Their approach is very different from ours, in that we did not pre-specify what could determine the variations of the substitution pattern among sites. The results seem to be accordingly quite distinct: this is particularly obvious in the composition of the 10 substitutional classes estimated by Thorne, Goldman, and co-workers, which are all characterized by very broad distributions on all 20 amino acids, not as peaked as the classes proposed by CAT. To further compare the two approaches, we made a preliminary investigation of the correlation between the secondary structure and the substitution profiles inferred by CAT-Poisson at each site of the elongation factor 2 data set: we identified the sites that were allocated to one of the stable classes at a frequency higher than 80%, and we looked at their distribution across the sequence (see table S2 in the Supplementary Material online). We found that although α-helices are enriched in classes such as (K,R) and (D,E), and β-sheets are enriched in classes like (I,V) and (F,Y), all of the classes but one are represented in the two kinds of secondary structure, as well as in the remaining parts of the sequence. As for solvent accessibility, correlations were uncovered, but also here, most of the classes are represented in both buried and exposed categories. Thus, the heterogeneity uncovered in the present work cannot be explained only in terms of secondary structure and solvent accessibility. More generally, it seems that our approach and that of Thorne, Goldman, and co-workers do not work at the same level: CAT would focus on local, specific, biochemical requirements, which can differ widely even between two adjacent residues, belonging to the same secondary structure, whereas the model of Thorne, Goldman, and co-workers averages over this low-level heterogeneity, to capture more global modulations of the amino-acid replacement pattern, correlated to local secondary structure determinants.

The CAT model is much more closely related to the model proposed by Koshi and Goldstein (Koshi and Goldstein 1998; Koshi, Mindell, and Goldstein 1999; Koshi and Goldstein 2001). In a first version of their model, these authors imposed restriction on the set of possible amino-acid profiles, by defining their model exclusively in terms of hydrophobicity and size. These two physical-chemical properties probably play an important role in the shaping of the classes that we have obtained, but they are not sufficient to account for their diversity. In particular, we have found classes that are more clearly defined by the electric charge (D,E), or by the aromaticity (F,Y) (fig. 3*D*). This restriction was relieved in subsequent

work (Dimmic, Mindell, and Goldstein 2000; Soyer et al. 2002), and the resulting model is more similar to CAT, except for the limitation in the total number of classes, which was set to 5. Another potentially important difference is that Koshi and Goldstein's model assumes that all sites belonging to the same class have also the same rate of substitution, while according to the CAT model, substitutional classes and site-specific overall rates of substitution are a priori independent variables. The fact that all stable classes that we observed on real data contain both fast and slow evolving sites (not shown) tends to give support to our prior choice.

Evaluation of the Model's Performance

The models investigated in the present work were compared through the evaluation of the Bayes factor. Note that the Bayes factor, as the ratio of the marginal likelihoods under two alternative models, could seem at first to be tantamount to a likelihood ratio. However, it is not equivalent to it. For instance, JTT+F, being nested within GTR, would necessarily yield a lower *maximal* likelihood than GTR. Yet, its *marginal* likelihood is larger than that of GTR ($\ln BF = 81$ natural log units in favor of JTT+F; table 3). This illustrates that the Bayes factor implicitely penalizes higher dimensional models. In fact, this penalty can be estimated asymptotically, yielding the Bayesian Information Criterion (Schwartz 1978):

$$\ln BF \simeq BIC = \Delta \ln \hat{L} - \frac{1}{2} k \ln N, \qquad (18)$$

where k is the difference of the number of parameters of the two models, and N is the number of columns of the alignment. This formula is valid only for i.i.d models, and it cannot be used to compare, for instance, GTR with CAT. However, in the general case, the Bayes factor can always be considered as a tradeoff between the relative likelihood scores of the two models and the informational cost of their respective parameterization.

The Bayes factor favors CAT-Poisson over standard one-matrix models, for all of the data sets which we have analyzed. It should be emphasized that an empirical matrix such as JTT has a significant prior advantage over CAT-Poisson, because it incorporates external information bearing on biochemical realism. In contrast, the CAT-Poisson model is completely naive, in the sense that the prior is totally uninformative. Nevertheless, this lack of prior biochemical knowledge does not seem to impair the performance of CAT-Poisson, presumably because this latter model is able to extract a substantial amount of equivalent knowledge from the data set, as confirmed by the biochemical relevance of the classes' profiles. Importantly, JTT+F is rejected even on quite small data sets, like EF2, which implies that the amount of information sufficient for CAT to outperform JTT+F is low. To verify this, we computed the Bayes factor under a smaller data set, made of the first third of EF30-627 (209 sites), and with a subset of five species. In this case, we observed that the Bayes factor was now in favor of JTT+F ($\ln B = -6.9$ natural units), indicating that the amount of information necessary for CAT-Poisson to be fit is not reached on this highly reduced data matrix.

In contrast to CAT-Poisson, MAX-Poisson is rejected when compared to JTT+F. A plausible explanation is that MAX is penalized by its very large number of parameters (19 for each site), which cannot all be correctly determined by the information contained in the data.

Sensitivity to the Prior Distribution and to the Model's Assumptions

It is known that sensitivity to the prior is an important issue when dealing with parameter-rich models (Huelsenbeck et al. 2002; Rannala 2002). In the present work, there are two main directions along which the prior should be tested. First, the prior on K, the number of classes, is currently defined through the prior on α, the mixture concentration parameter. For α, we have only tried a flat prior distribution, although we could test other possibilities. More generally, we could dispense with the Dirichlet process prior, and work on more general mixture models that make it possible to work directly with the prior on K (Green and Richardson 1998). Second, we could also test the base distribution on the π-space. Thus far, we have chosen a flat Dirichlet distribution, but this could be generalized to any Dirichlet distribution by varying the center of the distribution, or by modulating its concentration parameter.

The assumption that the substitution process at each site is Poisson is another potential limitation; in particular, it does not take the codon structure into account. In principle, it is easy to dispense with this simplifying assumption. Our current implementation already allows the use of any predefined set of relative rates. Similarly, one could imagine a version of the CAT model formulated at the codon level, and in which a single set of relative rates of codon substitution would be combined with class-specific amino-acid acceptance profiles. However, the computations under unequal relative rates of exchange between amino acids can be as much as 50 times slower than under a Poisson process, a factor which is even greater in the case of the codon models. Further algorithmic development is therefore needed in order to proceed in this direction.

Impact on Phylogenetic Inference

Inferences about class number and composition conducted on EF30-627 under a fixed tree are very similar to those where the phylogeny is also a free parameter (see fig. S1 in the Supplementary Material online). Likewise, we observed that the number and the profiles of the classes inferred for the mitochondrial data set were not very sensitive to the exact tree used as a constraint (not shown). This suggests that our analysis is robust with respect to the phylogeny, a fact which is not surprising, as similar conclusions have been reached by authors studying the sensitivity of rate across site parameters (Yang 1995), or of tests of model comparison (Posada and Crandall 2001), to the underlying phylogeny.

Conversely, however, the models investigated in this work differ in some of their predictions, in particular, in their estimates of the evolutionary distance between sequences and of the phylogenetic relationships. First, the tree length under CAT-Poisson is 10% to 20% higher than under JTT+F, suggesting that mutational saturation of the sequences could be better accounted for by CAT-Poisson than by JTT+F. This would not be completely surprising (Miyamoto and Fitch 1996). Mutational saturation—i.e., a high number of convergences and reversions throughout the sequences—will be all the more frequent as the substitution process at any given site is, on average, confined to a very restricted set of amino acids. This seems to be exactly the kind of situation inferred by CAT-Poisson. Importantly, these homoplasies will be more easily missed by models that underestimate site-specific restrictions of the set of admissible residues. It is quite plausible that JTT+F leads to such an underestimation, and this would explain the discrepancies observed between JTT+F and CAT-Poisson, concerning the number of substitutions along the tree.

Second, there are a few differences in the phylogenetic reconstruction obtained under CAT-Poisson and under JTT+F (fig. 1). One of the groupings found under CAT-Poisson, the monophyly of chromalveolates (stramenopiles + alveolates), is reasonable, and has been advocated by other studies (Baldauf et al. 2000; Fast et al. 2001; Bapteste et al. 2002), whereas others, like the polyphyly of amoebas and the position of the choanoflagellate *Monosiga* within the animal clade, are more questionable (Bapteste et al. 2002; Lang et al. 2002). More work is still needed to evaluate the performances of CAT in phylogenetic reconstruction, an issue we are currently investigating. The present results already illustrate that a better statistical fit does not necessarily imply a more reliable tree (Sullivan and Swofford 2001). In any case, CAT might allow better inference of evolutionary distances, and for that reason, it could be used for molecular datings. In addition, it can be a promising tool for analyzing the relationships between site-specific substitution patterns and structure-function determinants.

Appendix: Numerical Evaluation of the Bayes Factor by Thermodynamic Integration

In what follows, we consider two models, M_0 and M_1, which are defined on the same set of parameters, $h \in \Omega$ (note that this condition is purely formal, because parameters specific to one of the models, say M_0, can be included in the parameter vector of the other model, even if they are not involved in the computation of the likelihood under M_1). The Bayes factor in favor of model M_0, with respect to model M_1, is defined as the ratio of the marginal likelihoods:

$$B_{01} = \frac{P(D \mid M_1)}{P(D \mid M_0)}. \tag{19}$$

A parallel can be drawn between Bayesian inference and statistical physics as follows. Given one of the models, M_i, define an energy function under M_i as

$$E_i(h) = -\ln P(D \mid h, M_i) - \ln P(h \mid M_i). \tag{20}$$

According to Bayes's theorem

$$P(h \mid D, M_i) = \frac{P(D \mid h, M_i)P(h \mid M_i)}{P(D \mid M_i)} = \frac{1}{Z_i} e^{-\frac{E_i(h)}{kT}}, \tag{21}$$

where $kT = 1$, and $Z_i = P(D \mid M_i)$ plays here the role of a normalization factor. Thus, the energy is defined so that the posterior distribution is a Boltzmann distribution at a "temperature" $kT = 1$: this distribution can be sampled from using the classical MCMC methods, and the resulting Markov chain will thus be equivalent to a physical system at thermal equilibrium, fluctuating between all its accessible microscopic states. The free energy of such a system is related to the normalization factor Z_i, which is called the partition function in physics, by the relation

$$F_i = -kT \ln Z_i. \tag{22}$$

Because Z_i is identical to the marginal likelihood, one has

$$log\ B_{01} = log\ \frac{Z_1}{Z_0} = F_0 - F_1. \tag{23}$$

Hence, choosing the model of largest marginal likelihood amounts to choosing that of lowest free energy.

Let us now define the β-energy as

$$E_\beta(h) = (1 - \beta)E_0(h) + \beta E_1(h) \tag{24}$$

and the corresponding β-probability as

$$p_\beta(h) = \frac{1}{Z_\beta} e^{-E_\beta(h)}, \tag{25}$$

normalized by

$$Z_\beta = \int_\Omega e^{-E_\beta(h)}\, dh. \tag{26}$$

Note that for $\beta = 0$ ($\beta = 1$), p_β reduces to the posterior density under M_0 (M_1). Thus, the set $(p_\beta)_{0 \le \beta \le 1}$ defines a continuous path in the space of probability distributions, connecting the posterior distributions under M_0 and M_1.

As before, we can define

$$F_\beta = -\ln Z_\beta. \tag{27}$$

Taking the derivative of F_β with respect to β yields

$$\begin{aligned} \frac{\partial F_\beta}{\partial \beta} &= -\frac{1}{Z_\beta}\frac{\partial Z_\beta}{\partial \beta} = \frac{1}{Z_\beta}\int_\Omega \frac{\partial E_\beta}{\partial \beta} e^{-E_\beta(h)}\, dh \\ &= \int_\Omega (E_1(h) - E_0(h))\frac{1}{Z_\beta} e^{-E_\beta(h)}\, dh \\ &= <E_1 - E_0>_\beta, \end{aligned} \tag{28}$$

where $<\cdot>_\beta$ stands for the expectation with respect to p_β (or β-average).

This yields the integral formula

$$\ln B_{01} = F_0 - F_1 = -\int_0^1 <E_1 - E_0>_\beta\, d\beta, \tag{29}$$

which suggest a simple method for estimating $\ln B_{01}$ (Ogata 1989): for any $\beta \in [0, 1]$, a sample from $p_\beta(h)$ can

be obtained by running a MCMC. On such a sample, say $(h_\beta^l)_{1\leq l\leq L}$, β-averages can be computed using the usual asymptotic relation:

$$<f>_\beta \simeq \frac{1}{L}\sum_{l=1}^{L} f(h_\beta^l). \quad (30)$$

In particular, an estimate of $< E_1 - E_0 >_\beta$ can be computed in this way for any given value of β. Repeating this computation for a series of values of β regularly spaced over the unit interval, and approximating the integral by the Simpson procedure, yields an estimate of ln B_{01}. Evidently, the quality of the estimate will depend both on the length of each MCMC run and on the number of points of the interpolation. In the present work, we found that six points (β = 0, 0.2, 0.4, 0.6, 0.8, 1.0) were sufficient to give a precision of about 10%. As for the length of the MCMC run, 100,000 cycles turn out to be sufficient, yielding a precision of 1% on the estimate of $< E_1 - E_0 >_\beta$.

Supplementary Material

The following materials relevant to this article are available online: table 1. Settings of the MCMC runs; table 2. Distribution of the sites allocated to stable classes, according to secondary structure and solvent accessibility; fig. 1. CAT-Poisson analysis on the elongation factor data set EF30-627, under constrained (A,C) and unconstrained (B,D) topology; fig. 2. CAT-POISSON analysis on the eukaryotic data set Ek55-1525. fig. 3. CAT-Poisson analysis on the mammal mitochondrial data set Mt45-3596.

Acknowledgments

We are grateful to David Bryant, Olivier Gascuel, Franz Lang, Jeffrey Thorne, and two anonymous referees for their useful comments on the manuscript. We wish to thank Sandra Baldauf for making available the alignment of four eukaryotic proteins, and Béatrice Philippe for her help in determining the relations between classes and secondary structure on the elongation factor data set. This work was supported by the Canadian Research Chair, and by a starting fund from the University of Montréal.

Literature Cited

Adachi, J., and M. Hasegawa. 1996. Model of amino-acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Evol. **42**:459–468.

Adachi, J., P. J. Wadell, W. Martin, and M. Hasegawa. 2000. Plastid genome phylogeny and a model of amino-acid substitution for proteins encoded by chloroplast DNA. J. Mol. Evol. **50**:348–358.

Antoniak, C. E. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann. Statistics **2**:1152–1174.

Baldauf, S. L., A. J. Roger, I. Wenk-Siefert, and W. F. Doolittle. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science **290**:972–977.

Bapteste, E., H. Brinkmann, J. A. Lee, D. V. Moore, C. W. Sensen, P. Gordon, L. Durufle, T. Gaasterland, P. Lopez, M. Muller, and H. Philippe. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. Proc. Natl. Acad. Sci. USA **99**:1414–1419.

Broet, P., S. Richardson, and F. Radvanyi. 2002. Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. J. Comp. Biol. **9**:671–683.

Bruno, W. J. 1996. Modeling residue usage in aligned protein sequences via maximum likelihood. Mol. Biol. Evol. **13**:1368–1374.

Dayhoff, M., R. V. Eck, and C. M. Park. 1972. A model of evolutionary change in proteins. Pp. 88–89 *In* M. Dayhoff, ed., Atlas of Protein Sequence and Structure, National Biomedical Research Foundation, Washington, D.C.

Dayhoff, M., R. Schwartz, and B. Orcutt. 1978. A model of evolutionary change in proteins. Pp. 345–352 *In* M. Dayhoff, ed., Atlas of Protein Sequence and Structure, National Biomedical Research Foundation, Washington, D.C.

Dimmic, M. W., D. P. Mindell, and R. A. Goldstein. 2000. Modeling evolution at the protein level using an adjustable amino acid fitness model. Pac. Symp. Biocomput. **5**:18–29.

Escobar, M., and M. West. 1995. Bayesian density estimation and inference using mixtures. J. Am. Stat. Assoc. **90**:577–588.

Eskin, E., W. N. Grundy, and Y. Singer. 2001. Using mixtures of common ancestors for estimating the probabilities of discrete events in biological sequences. Bioinformatics **17**:S65–S73.

Fast, N. M., J. C. Kissinger, D. S. Roos, and P. J. Keeling. 2001. Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. Mol. Biol. Evol. **18**:418–426.

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

———. 2004. Inferring phylogenies. Sinauer Associates Inc., Sunderland, Mass.

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

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

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

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. Bayesian Data Analysis, 2nd edition. Chapman and Hall/CRC.

Goldman, N., J. Thorne, and D. Jones. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics **149**:445–458.

Goldman, N., J. L. Thorne, and D. T. Jones. 1996. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. J. Mol. Biol. **263**:196–208.

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

Green, P. J., and S. Richardson. 1998. Modelling heterogeneity with and without the Dirichlet process. Technical report, University of Bristol, Bristol, U.K.

Halpern, A. L., and W. J. Bruno. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol. Biol. Evol. **15**:910–917.

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

Huelsenbeck, J. P., and R. Nielsen. 1999. Variation in the pattern of nucleotide substitution across sites. J. Mol. Evol. **48**:86–93.

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

Jaynes, E. 2003. Probability Theory. The logic of science. Cambridge University Press, Cambridge, U.K.

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

———. 1961. Theory of Probability. Oxford University Press.

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

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

Koshi, J. M., and R. A. Goldstein. 1998. Models of natural mutations including site heterogeneity. Proteins **32**:289–295.

———. 2001. Analyzing site heterogeneity during protein evolution. Pac. Symp. Biocomput. pp. 191–202.

Koshi, J. M., D. P. Mindell, and R. A. Goldstein. 1999. Using physical-chemistry–based substitution models in phylogenetic analyses of HIV-1 subtypes. Mol. Biol. Evol. **16**: 173–179.

Lang, B. F., C. O'Kelly, T. Nerad, M. W. Gray, and G. Burger. 2002. The closest unicellular relatives of animals. Curr. Biol. **12**:1773–1778.

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

Li, S. 1996. Phylogenetic tree construction using Markov chain Monte Carlo. PhD dissertation, Ohio State University, Columbus, Ohio.

Liò, P., and N. Goldman. 1999. Using protein structural information in evolutionary inference: transmembrane proteins. Mol. Biol. Evol. **16**:1696–1710.

Lopez, P. 1997. Analyse phylogénétique de grands alignements de protéines: vers une classification des sites? Master degree dissertation, Université Paris XI, Paris, France.

Miyamoto, M. M., and W. M. Fitch. 1996. Constraints on protein evolution and the age of Eubacteria/Eukaryote split. Syst. Biol. **45**:568–575.

Muller, T., R. Spang, and M. Vingron. 2002. Estimating amino-acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. Mol. Biol. Evol. **19**:8–13.

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

Ogata, Y. 1989. A Monte Carlo method for high dimensional integration. Numerishe Mathematik **55**:137–157.

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

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

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

Schneider, T. D., and R. M. Stephens. 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. **18**:6097–6100.

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

Soyer, O., M. W. Dimmic, R. R. Neubig, and R. A. Goldstein. 2002. Using evolutionary methods to study G-protein coupled receptors. Pac. Symp. Biocomput. pp. 625–636.

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

Sullivan, J., and D. L. Swofford. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site variation and nucleotide substitution pattern are violated? Syst. Biol. **50**:723–729.

Swofford, D., G. P. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. In Molecular Systematics. Sinauer Associates, Sunderland, Mass.

Thorne, J. L., N. Goldman, and D. T. Jones. 1996. Combining protein evolution and secondary structure. Mol. Biol. Evol. **13**:666–673.

Wald, A. 1949. Note on the consistency of maximumm likelihood. Ann. Math. Stat. **20**:595–601.

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

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

———. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39**:306–314.

———. 1995. A space-time process model for the evolution of DNA sequences. Genetics **139**:993–1005.

———. 1996. Among site variation and its impact on phylogenetic analyses. Trends Ecol. Evol. **11**:367–370.

Yang, Z., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Mol. Biol. Evol. **14**:717–724.

Associate Editor

Accepted February 10, 2004