Syst. Biol. 56(6):975–987, 2007

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150701670569

A Nonparametric Method for Accommodating and Testing Across-Site Rate Variation

JOHN P. HUELSENBECK,[1] AND MARC A. SUCHARD[2,3,4]

[1]*Department of Integrative Biology, University of California, Berkeley, CA 94720, USA; E-mail: johnh@berkeley.edu*
[2]*Department of Biomathematics and* [3]*Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095, USA and*
[4]*Department of Biostatistics, UCLA School of Public Health, Los Angeles, CA 90095, USA*

Abstract.—Substitution rates are one of the most fundamental parameters in a phylogenetic analysis and are represented in phylogenetic models as the branch lengths on a tree. Variation in substitution rates across an alignment of molecular sequences is well established and likely caused by variation in functional constraint across the genes encoded in the sequences. Rate variation across alignment sites is important to accommodate in a phylogenetic analysis; failure to account for across-site rate variation can cause biased estimates of phylogeny or other model parameters. Traditionally, rate variation across sites has been modeled by treating the rate for a site as a random variable drawn from some probability distribution (such as the gamma probability distribution) or by partitioning sites to different rate classes and estimating the rate for each class independently. We consider a different approach, related to site-specific models in which sites are partitioned to rate classes. However, instead of treating the partitioning scheme in which sites are assigned to rate classes as a fixed assumption of the analysis, we treat the rate partitioning as a random variable under a Dirichlet process prior. We find that the Dirichlet process prior model for across-site rate variation fits alignments of DNA sequence data better than commonly used models of across-site rate variation. The method appears to identify the underlying codon structure of protein-coding genes; rate partitions that were sampled by the Markov chain Monte Carlo procedure were closer to a partition in which sites are assigned to rate classes by codon position than to randomly permuted partitions but still allow for additional variability across sites. [Across-site rate variation; Bayesian estimation; Dirichlet process prior; Markov chain Monte Carlo.]

Sequence substitution rates depend upon the rate at which new alleles enter a population through mutation and the rate at which these alleles are fixed, replacing the other alleles that were previously present in the population. Importantly, the substitution rate can vary across an alignment of molecular sequences for a number of biological reasons. Functional constraints, for one, can differ on an alignment site-by-site basis, causing variation in the rate of substitution; sites that are more constrained by natural selection show fewer substitutions than sites that are less constrained or for which natural selection favors a nucleotide that is different than the one most frequently found in the population. Moreover, because the substitution rate depends upon two process—the mutation process and the fixation process—the substitution rate can vary across a sequence if the mutation rate varies. Finally, substitution rates can vary across an alignment of sequences for the simple reason that coalescence histories can vary across the sequences. Under a population genetics model such as the coalescence process with recombination, the history of coalescence can vary across sequences, with recombination events marking the boundaries between sites with different histories. If the amount of time represented by each coalescent history varies (e.g., if the time to the most recent common ancestor for each history varies), then substitution rates can vary across the sequences even if no selection is acting on the sequences and even if the mutation rate remains constant across the sequences.

For phylogenetic problems, the consensus of biologists favors variation in functional constraint across sequences as the main cause of among-site rate variation. The concern that the coalescence history can vary across the sequence is usually discounted for most phylogenetic studies. Indeed, a basic assumption of almost all phylogenetic analyses is that the same phylogeny underlies all of the sequences, so variation in rates caused by differences in coalescence histories is discounted a priori as a major cause of among-site rate variation. The assumption that the phylogenetic history, at least, is common to all of the sites in a phylogenetic analysis is probably a good one in most cases, because most phylogenetic analyses operate on different species that are separated by enough time that processes such as lineage sorting can be discounted. Variation in mutation rates, although documented in population studies (e.g., Arndt et al., 2005), is also thought to play a back-seat role to varying functional constraint as a cause of among-site variation in substitution rate. In this study, we concentrate on among-site variation in substitution rate that is caused by variation in functional constraints or, to the extent that it might occur, systematic variation in mutation rate. The types of among-site rate variation models we are concerned with affect all of the species at a particular site in the same way. We do not consider models such as the covariotide model, in which functional constraints can change over the evolutionary history for a particular site (Fitch and Markowitz, 1970; Tuffley and Steel, 1998; Galtier, 2001; Huelsenbeck, 2002). In this study, a high-rate site is always a high-rate site, remaining so over the entire history represented by the phylogeny.

Regardless of the cause of among-site variation in rates of substitution, it is critical to account for such variation in a phylogenetic analysis. Failure to do so can result in incorrect inference of phylogeny (Huelsenbeck and Hillis, 1993) as well as biased estimates of other model parameters (Wakeley, 1994). Along with the topology of the phylogeny, substitution rates are a critical component of all model-based phylogenetic analyses and appear as the branch lengths on a phylogenetic tree. Among-site rate variation is accommodated by assuming that the relative rate for the ith site, r_i, varies across the sequences according to some underlying parametric model. The ultimate effect of among-site rate

variation models is to multiply all of the branches of a tree for site i by the rate parameter r_i. In this manner, the branch length proportions are maintained on a site-by-site basis; a high-rate site, for example, has a tree with branch lengths that are all proportionally larger than a low-rate site.

Two general approaches have been pursued to model among-site rate variation. The first approach treats the relative rate for site i as a random variable drawn from a common across-sites gamma probability distribution (Uzzell and Corbin, 1971; Nei et al., 1976; Jin and Nei, 1990; Yang, 1993), a log normal probability distribution (Olsen, 1987), an inverse Gaussian probability distribution (Waddell and Steel, 1997), or a probability distribution in which some proportion of the sites remain invariable (Hasegawa et al., 1987). Mixtures of both invariant sites and randomly distributed rates are also successful. Gu et al. (1995) provide the first example of phylogenetic analysis using a mixture of a proportion of invariable sites and gamma-distributed rates model. Waddell and Steel (1997) consider extensions of the Gu et al. (1995) approach, including inverse Gaussian distributed rates. Yang (1995) considered a hidden Markov model of among-site rate variation in which the marginal rate at a site is gamma distributed but in which adjacent sites have potentially correlated rates, accounting for the observation that some genes harbor regions of high or low substitution rate. Finally, Kosakovsky-Pond and Frost (2005) considered a general discrete distribution for among-site rate variation and considered methods for allowing the parameters of such a model to be reliably estimated. This model is perhaps the most flexible model of among-site rate variation considered to date. The second approach for modeling among-site rate variation groups sites into classes and estimates the relative rate of substitution independently for each class. For protein-coding DNA sequences, sites are commonly grouped by codon position. The relative substitution rates are then estimated for the first-, second-, and third-codon position categories. Typically, one finds that third position sites have the highest substitution rate and second position sites have the lowest substitution rate; because of the redundancy of the genetic code, third position sites are more able to vary without changing the function of the protein. Although site-specific models like the one just described are usually applied to protein-coding DNA with partitioning being first-, second-, and third-codon positions, other ways of grouping sites into classes can also be applied. The most extreme partitioning scheme is to assign each site its own rate class (see Swofford et al., 1996; Nielsen, 1997); this means that the relative rate of substitution is independently estimated for each site in the sequences. Although on first inspection, such a model seems ideal—after all, each site probably does differ from other sites, at least slightly, in its overall rate of substitution—a site-specific model in which each site has an independently estimated rate parameter has poor statistical properties (Felsenstein, 2004). One must remain vigilant of the bias-variance trade-off that adding and removing parameters from statistical models engenders; completely independent site rates is one example of model overfitting. Occam's Razor suggests that the optimal model contains the fewest parameters requisite to appropriately account for the variation in the data.

We describe an alternative model of among-site rate variation that is akin to the site-specific model with different rates at each site but employs Occam's Razor as a guiding principle to randomly construct a parsimonious description of the rate variation process. Specifically, we consider the partitioning scheme, in which sites are assigned to rate categories, to be a random variable with a prior probability distribution that is described by the Dirichlet process prior model. The Dirichlet process prior has been used with increasing frequency in a Bayesian framework for modeling cases in which the data elements are drawn from a mixture of an unknown number of probability distributions. The Dirichlet process prior model allows the number of mixture components as well as the assignment of individual data elements to mixture components to vary. In the context of rate variation across sites, the Dirichlet process prior model places non-zero probability on an equal rates model (which occurs when all of the sites are assigned to the same rate class), some probability on the extreme case in which each site is placed in a rate class by itself, as well as probability on all of the other possible partitioning schemes that assign sites to rate classes. A priori, we can analytically calculate these weights and then compare them to a posteriori estimates to select among different rate variation models in a formal statistical framework.

The Dirichlet process prior is often described in context to the "Chinese restaurant problem," a description that provides an intuitive explanation of the process using a hypothetical example of seating patrons in a restaurant. One imagines a queue of patrons waiting outside the entrance of a Chinese restaurant. The restaurant, besides serving wonderful Chinese cuisine, has a countably infinite number of tables (and each table can seat an infinite number of people). The patrons enter the restaurant one at a time. The first patron sits at some arbitrary table. Obviously this occurs with probability one. The next patron can either sit at the same table as the first, with probability $\frac{1}{1+\chi}$, or at an unoccupied table with remaining probability $\frac{\chi}{1+\chi}$. In general, patron i sits at occupied table k with probability $\frac{\eta_k}{i+\chi}$ or at an unoccupied table with probability $\frac{\chi}{i+\chi}$. (η_k is the number of people currently sitting at table k.) After all of the patrons have entered the restaurant, the patrons will occupy some number of tables K, and the occupancy of the tables can be noted. Importantly, the probability of the number of tables and the specific pattern of people sitting at tables can be calculated, and this probability does not depend upon the order in which patrons entered the restaurant. For the purposes of this paper, we replace "patrons" with the word alignment "site." Sites that "sit at the same table" all share a common relative substitution rate. There are other ways to describe the Dirichlet process prior (e.g., a stick-breaking description for the proportion of data elements

that are assigned to each cluster). Besides being used on a limited basis in phylogenetics (Lartillot and Philippe, 2004; Huelsenbeck et al., 2006) and population genetics, where the sampling procedure underlies the Ewens' sampling formula (Ewens et al., 1998), the Dirichlet process prior has been of general use in Bayesian analysis of mixture models.

METHODS

Our description of the Dirichlet process prior model for among-site rate variation involves many parameters. We provide a list of most of the parameters used in this paper in Table 1. Throughout, we use $f(\cdot)$ or $g(\cdot)$ to describe a probability or probability density/mass function. The identity of the probability distribution should be clear from its arguments.

TABLE 1. A description of most of the parameters used in this study.

Parameter	Description
S	Number of species/sequences in the alignment
N	Length of sequences in the alignment
$\mathbf{Y}$	$S \times N$ matrix containing the alignment
$\mathbf{y}_i$	Column vector containing the sequence characters at site i in the alignment
τ	Unrooted tree topology
$B(S)$	Number of unrooted trees possible for S species
t_b	Length of branch b expressed in terms of expected number of substitutions per site
$\mathbf{v}$	Vector containing the branch lengths for a tree
T	Tree length (sum of the branch lengths, $\sum_{b=1}^{2S-3} t_b$)
ϕ_b	Proportion of the tree length accounted for by the branch b, $\phi_b = t_b/T$
$\boldsymbol{\phi}$	Vector containing the branch proportions for a tree
q_{uv}	Instanteous rate of change from nucleotide u to nucleotide v, $u, v \in \{A, C, G, T\}$
$\mathbf{Q}$	Matrix containing the instantaneous rates of change between nucleotides
θ_{uv}	Relative rate of substitution between nucleotides u and v, $u < v \in \{A, C, G, T\}$
$\boldsymbol{\theta}$	Vector containing the relative substitution rates
π_u	Stationary frequency of nucleotide u, where $u \in \{A, C, G, T\}$
$\boldsymbol{\pi}$	Vector containing the four nucleotide frequencies
μ	Scaling parameter for the rate matrix, ensuring that the average rate of substitution is one
r	Rate multiplier (usually rate multipliers are chosen such that the mean rate value is one)
p	Proportion of invariable sites
α	Shape parameter of the gamma probability distribution
β	Scale parameter of the gamma probability distribution
λ	Rate parameter of the exponential probability distribution
K	Number of rate classes
$\sigma(i)$	Rate class assignment for site i
$\boldsymbol{\sigma}$	Allocation vector, containing the rate class assignments for all sites
η_k	Number of sites assigned to rate class k
χ	Concentration parameter of the Dirichlet process prior
$\mathcal{B}_x$	Bell number for x elements
$\mathcal{S}_1(\cdot, \cdot)$	Stirling number of the first kind
$\mathcal{S}_2(\cdot, \cdot)$	Stirling number of the second kind
ψ_1	Tuning parameter used in modified LOCAL proposal mechanism
ψ_2	Tuning parameter used in Dirichlet proposal mechanism for base frequencies
ψ_3	Tuning parameter used in Dirichlet proposal mechanism for substitution rates
ψ_4	Tuning parameter used when modifying the rate of a rate class
κ	Number of auxiliary rate classes used when updating the allocation vector

Data

We assume that the biologist has properly aligned DNA sequences from S species. The alignment is represented as an $S \times N$ matrix of nucleotides, $\mathbf{Y}$, where N is the length of the sequences in the alignment.

The following provides an example of an alignment of $S = 4$ sequences, each $N = 10$ nucleotides in length:

$$\begin{array}{ll} \text{Species 1} & \text{TTTTCTGATG} \\ \text{Species 2} & \text{TTTTCTGATG} \\ \text{Species 3} & \text{TCTTCAGACG} \\ \text{Species 4} & \text{TGTTCAGAAA} \end{array} \quad (1)$$

Each column of the alignment is called a site; this example alignment has $N = 10$ sites, which can be represented as the column vectors: $\mathbf{y}_1 = (\text{TTTT})^{\text{t}}$, $\mathbf{y}_2 = (\text{TTCG})^{\text{t}}$, $\mathbf{y}_3 = (\text{TTTT})^{\text{t}}$, $\mathbf{y}_4 = (\text{TTTT})^{\text{t}}$, $\mathbf{y}_5 = (\text{CCCC})^{\text{t}}$, *etc.*

Phylogenetic Model

Our phylogenetic model consists of a tree representing the genealogical relationships of the species and a model of character change that describes how the characters (nucleotides in this case) change over evolutionary time on the phylogenetic tree to produce the observed data $\mathbf{Y}$ at the tree's S tips. We imagine that all of the possible phylogenetic trees have been labeled, $\tau_1, \tau_2, \tau_3, \ldots, \tau_{B(S)}$, where $B(S)$ is the number of possible trees for S species. In this study, we use a time-reversible model of character change (which will be described in more detail below); as a consequence, we consider only unrooted phylogenetic trees. The unrooted trees that are inferred in this paper can be rooted using other criteria, such as the outgroup criterion. There are a total of $B(S) = (2S-5)!!$ possible unrooted phylogenetic trees for S species (Schröder, 1870).

Every unrooted phylogenetic tree has $2S-3$ branches. Ideally, the lengths of the branches are represented in terms of real time; for instance, the time between speciation events on the tree. However, because it is difficult to infer the branch lengths on a phylogenetic tree in terms of time, these lengths are usually measured in terms of the number of substitutions per site that are expected to occur along each branch. We denote the collections of branch lengths as $\mathbf{t} = (t_1, \ldots, t_{2S-3})$. The tree length is the sum of the branch lengths: $T = \sum_{b=1}^{2S-3} t_b$.

We assume that nucleotides change along the tree according to a continuous-time Markov chain. This is a typical assumption of phylogenetic analysis. A continuous-time Markov chain can be completely described by knowledge of the instantaneous rates of change between nucleotides. Here, we assume that substitutions occur according to the general time-reversible (GTR) model of DNA substitution, which has

instantaneous rates:

$$\mathbf{Q} = \{q_{uv}\} = \begin{pmatrix} — & \theta_{AC}\,\pi_C & \theta_{AG}\,\pi_G & \theta_{AT}\,\pi_T \\ \theta_{AC}\,\pi_A & — & \theta_{CG}\,\pi_G & \theta_{CT}\,\pi_T \\ \theta_{AG}\,\pi_A & \theta_{CG}\,\pi_C & — & \theta_{GT}\,\pi_T \\ \theta_{AT}\,\pi_A & \theta_{CT}\,\pi_C & \theta_{GT}\,\pi_G & — \end{pmatrix} \times \mu \quad (2)$$

and was first described by Tavaré (1986). The entry in the uth row and vth column of the matrix specifies the infinitesimal rate of change from nucleotide u to nucleotide v. The diagonal entries of $\mathbf{Q}$, here shown with a dash, are specified such that each row sums to zero. The frequency of nucleotide u under stationarity of the substitution process is denoted π_u. The commonly used DNA substitution models all have the unusual feature that the stationary frequencies of the nucleotides are built directly into the rate matrix. The sum of the four base frequencies must, of course, equal one. We also impose the constraint that the sum of the six rate parameters equals one: $\theta_{AC} + \theta_{AG} + \theta_{AT} + \theta_{CG} + \theta_{CT} + \theta_{GT} = 1$. To ensure that branch lengths on the tree are expressed in terms of expected number of substitutions per site, the substitution rate matrix must be rescaled such that the expected substitution rate is one for a unit branch length. This is achieved by setting $\mu = -1/\sum_{u\in(A,C,G,T)} \pi_u q_{uu}$. It also means that we cannot estimate the absolute rates of substitution, only their relative rates. The substitution rate parameters are contained in the vector $\boldsymbol{\theta} = (\theta_{AC}, \theta_{AG}, \theta_{AT}, \theta_{CG}, \theta_{CT}, \theta_{GT})$. Similarly, the nucleotide frequency parameters are contained in the vector $\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$.

Likelihood

With the addition of two more assumptions—that substitutions at different sites and along different branches are independent of one another—we can calculate the likelihood. The likelihood is proportional to the probability of the observed data, $\mathbf{Y}$, conditional on the unknown parameters of the model $(\tau, \mathbf{t}, \boldsymbol{\theta}, \boldsymbol{\pi})$. We use the pruning algorithm described by Felsenstein (1981) to analytically calculate the likelihood $f(\mathbf{y}_i|\tau, \mathbf{t}, \boldsymbol{\theta}, \boldsymbol{\pi})$ of each site i in the alignment.

Assuming that substitutions are independent across sites, the likelihood of the complete alignment is the product of the site likelihoods:

$$\mathcal{L}(\tau, \mathbf{t}, \boldsymbol{\theta}, \boldsymbol{\pi}) = f(\mathbf{Y}|\tau, \mathbf{t}, \boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{N} f(\mathbf{y}_i|\tau, \mathbf{t}, \boldsymbol{\theta}, \boldsymbol{\pi}). \quad (3)$$

With the likelihood function in hand, parameter estimation can be performed using the method of maximum likelihood or using Bayesian inference.

Across-Site Rate Variation

To this point, our description of the phylogenetic model assumed that substitution rates were equal across sites. Rate variation across sites has been introduced into phylogenetic models using two general strategies. The first method a priori partitions the sites into different categories, and then estimates the rates independently for each partition. Such a priori partitioned models are often called site-specific models. We identify these models by appending "+SS" to the substitution model name (e.g., GTR+SS). For a quintessential example of a site-specific model, imagine that the sequence alignment derives from protein-coding DNA. A common partitioning scheme for protein-coding DNA is to categorize sites by codon position. All of the first position sites are assumed to share a rate multiplier r_1. Similarly, all second position sites share a rate multiplier r_2 and third position sites have the rate multiplier r_3. To ensure that branch lengths on the tree remain expressed in terms of expected number of substitutions per site, the rate multipliers are constrained such that their mean rate is one; that is, the rates follow the constraint that $(r_1 n_1 + r_2 n_2 + r_3 n_3)/n = 1$, where n_k for $k = 1, 2, 3$ is the number of sites assigned to partition k. The likelihood under a site-specific model is straightforward to calculate. In essence, the branch lengths are all multiplied by the rate parameter $r_{\sigma(i)}$, where $\sigma(i)$ is the fixed mapping from site i to partition $\sigma(i) = 1$, 2, or 3 depending on whether i is at a first-, second-, or third-codon position, respectively. The likelihood for the site i becomes $f(\mathbf{y}_i|\tau, \mathbf{t} \times r_{\sigma(i)}, \boldsymbol{\theta}, \boldsymbol{\pi})$. The rate multiplier $r_{\sigma(i)}$ functions as a scaling factor for all of the branch lengths: $\mathbf{t} \times r_{\sigma(i)} = (t_1 r_{\sigma(i)}, \ldots, t_{2S-3} r_{\sigma(i)})$. Under the site-specific model, all branches are affected in the same way; all branches for a given site expand or contract proportionally. This proportional effect on all branches is also true for other models for among-site rate variation.

The other general strategy for accommodating rate variation across sites is to treat the rate at a given site as a random variable drawn from some common across-sites parameteric probability distribution. Two models are commonly used—the proportion of invariable sites model and the gamma model. For the proportion of invariable sites model, the rate multiplier is $r = 0$ with probability p and is $r = 1/(1-p)$ with probability $1 - p$ (Hasegawa et al., 1987). The likelihood for site i is then integrated over the two possibilities: $f(\mathbf{y}_i|\tau, \mathbf{t}, \boldsymbol{\theta}, \boldsymbol{\pi}, p) = p \times f(\mathbf{y}_i|\tau, \mathbf{t} \times 0, \boldsymbol{\theta}, \boldsymbol{\pi}) + (1-p) \times f[\mathbf{y}_i|\tau, \mathbf{t}/(1-p), \boldsymbol{\theta}, \boldsymbol{\pi}]$.

For the gamma model, the rate multipliers r_i are assumed to be drawn independently and identically distributed (iid) from a gamma probability distribution (Yang, 1993). Each rate multiplier, then, has prior probability density

$$f(r_i|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} {r_i}^{\alpha-1} e^{-\beta r_i} \quad (4)$$

for $r_i > 0$. The gamma function $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} \mathrm{d}t$. For natural numbers, $\Gamma(z) = (z-1)!$ The probability distribution $f(\cdot|\alpha, \beta)$ has mean α/β and variance α/β^2. Again, to ensure that the branch lengths remain expressed in terms of expected number of substitutions per site, the probability distribution is chosen such that the mean rate

of substitution is one. This is achieved by fixing $\alpha = \beta$. The likelihood for site i is calculated by marginalizing over all possible rate multiplier values

$$f(\mathbf{y}_i|\tau, \mathbf{t}, \boldsymbol{\theta}, \boldsymbol{\pi}, \alpha) = \int_0^\infty f(\mathbf{y}_i|\tau, \mathbf{t} \times r, \boldsymbol{\theta}, \boldsymbol{\pi})\, f(r|\alpha, \alpha)\, \mathrm{d}r. \tag{5}$$

In practice, it is difficult to calculate the integral necessary to average over all rates for the gamma model. Instead, the gamma distribution is arbitrarily discretized into an a priori fixed number of categories, and the mean or median realized value for each category is used to represent all of the possible rates in that category (Yang, 1994). Hence, the integration over all possible rates simplifies to a practical summation over the rate categories. Similar to the nomenclature for the site-specific model, we specify the proportion of invariable sites and gamma among-site rate variation models by appending "+I" and "+Γ," respectively, to the substitution model name. Many phylogenetic studies use a mixture of the site-specific, proportion of invariable sites, and gamma rate variation models. Using mixtures often provides a significant improvement of the fit of the phylogenetic model to the observed data.

Bayesian Inference of Phylogeny

Bayesian inference of phylogeny is based upon the posterior probability distribution of the model parameters. The posterior distribution is the probability mass function of the model parameter conditional on the data. Via Bayes theorem,

$$f(\tau, \mathbf{t}, \boldsymbol{\theta}, \boldsymbol{\pi}|\mathbf{Y}) = \frac{f(\mathbf{Y}|\tau, \mathbf{t}, \boldsymbol{\theta}, \boldsymbol{\pi})\, f(\tau, \mathbf{t}, \boldsymbol{\theta}, \boldsymbol{\pi})}{f(\mathbf{Y})}, \tag{6}$$

where $f(\cdot|\mathbf{Y})$ is the posterior probability distribution of the parameters, $f(\mathbf{Y}|\cdot)$ is the likelihood, $f(\cdot)$ is the prior probability distribution, and $f(\mathbf{Y})$ is the marginal likelihood of the data (Mau, 1996; Li, 1996; Rannala and Yang, 1996; Mau and Newton, 1997; Yang and Rannala, 1997; Larget and Simon, 1999; Mau et al., 1999; Newton et al., 1999). We have already described how the likelihood is calculated. Bayesian analysis introduces a prior probability distribution on the parameters of the phylogenetic model. The prior probability distribution represents the biologist's beliefs about the parameter(s) before collecting the observations. The following represent commonly assummed priors on the phylogenetic parameters and is the default condition in the program MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003):

$$\begin{aligned}
\tau &\sim \text{Uniform on all } B(S) \text{ trees},\\
t_b &\sim \text{Exponential}(\lambda) \text{ for } b = 1, \ldots, 2S-3,\\
\boldsymbol{\theta} &\sim \text{Dirichlet}(1, 1, 1, 1, 1, 1), \text{ and}\\
\boldsymbol{\pi} &\sim \text{Dirichlet}(1, 1, 1, 1).
\end{aligned} \tag{7}$$

That is to say, all phylogenetic trees are assumed to be equally probable a priori. Moreover, the branch lengths on each tree are assumed to be iid exponentially distributed random variables with mean $1/\lambda$ (here, we set $\lambda = 10$), the nucleotide frequencies are assumed to be drawn from a flat Dirichlet probability distribution, and the rate parameters of the substitution model are assumed to be random variables also drawn from a flat Dirichlet probability distribution. In the above specification, we have not included among-site rate variation model parameters, such as the proportion of invariable sites parameter p or the gamma shape parameter α. We consider these priors further after the development of an alternative parameteriation that naturally lends itself to our novel among-site rate variation model formulation.

A Dirichlet Process Model for Across-Site Rate Variation

There is an alternative parameterization of the phylogenetic model that is exactly equivalent to the scheme given above. This alternative parameterization relies on the following facts about exponential, gamma, and Dirichlet probability distributions. First, the sum of M exponential(λ) random variables is a gamma probability distribution with parameters $\alpha = M$ and $\beta = \lambda$. To see this, the exponential distribution has probability density $f(x|\lambda) = \lambda e^{-\lambda x}$. Summing M exponential random variables generates an analytically tractable convolution integral, leading to

$$f(y|M, \lambda) = \frac{\lambda^M}{\Gamma(m)} y^{M-1} e^{-\lambda y} \tag{8}$$

where $y = x_1 + x_2 + \ldots + x_M$. Second, imagine dividing each of the M exponential random variables by their gamma-distributed sum. Doing this for the mth exponential random variable with realized value x_m generates $\phi_m = x_m/y$, where y is the realized sum of the M exponential random variables. Random variable ϕ_m is the proportion of the sum represented by x_m. The joint probability distribution for $(\phi_1, \ldots, \phi_M)$ follows a flat Dirichlet probability distribution. Hence, an alternative parameterization of our prior model in Equation (7) is

$$\begin{aligned}
\tau &\sim \text{Uniform over trees},\\
\phi_1, \ldots, \phi_{2S-3} &\sim \text{Dirichlet}(1, \ldots, 1),\\
T &\sim \text{Gamma}(2S-3, \lambda),\\
\boldsymbol{\theta} &\sim \text{Dirichlet}(1, 1, 1, 1, 1, 1), \text{ and}\\
\boldsymbol{\pi} &\sim \text{Dirichlet}(1, 1, 1, 1),
\end{aligned} \tag{9}$$

where ϕ_b is the proportion of total tree length T allocated to branch b: specifically, $t_b = \phi_b \times T$. To complete the specification, let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{2S-3})$.

Building upon the parameterization in Equation (9), let T_i equal the site-specific tree length for site i. A model without rate variation restricts $T_1 = \cdots = T_N = T$. Here, we nonparametrically relax this assumption and allow

the tree lengths to differ across sites according to a Dirichlet process prior model. Each site i exists in one of K rate categories, where K is a priori unknown and each rate category differs in its tree length $T_{[k]}$ for $k \in \{1, \ldots, K\}$. To inform the assignment of sites to rate categories, we return to the vector of mappings $\boldsymbol{\sigma} = (\sigma(1), \ldots, \sigma(N))$; however, under the Dirichlet process prior, the mappings are random with $\sigma(i) \in \{1, \ldots, K\}$. For example, for $N = 10$ sites, one possible assignment vector is

$$\boldsymbol{\sigma} = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1). \tag{10}$$

This mapping vector assigns all sites in the same rate category and represents a model with equal rates across sites. For illustration, some other possible mapping vectors include

$$\begin{aligned}
\boldsymbol{\sigma} &= (1, 2, 1, 2, 1, 3, 1, 1, 3, 1),\\
\boldsymbol{\sigma} &= (1, 2, 3, 1, 2, 3, 1, 2, 3, 1),\\
\boldsymbol{\sigma} &= (1, 2, 1, 2, 2, 1, 1, 2, 1, 1), \text{ and}\\
\boldsymbol{\sigma} &= (1, 2, 3, 4, 5, 6, 7, 8, 9, 10).
\end{aligned} \tag{11}$$

The final partitioning, above, shows the least parsimonious partitioning scheme, in which each site falls into a different rate class. These vectors represent a very small minority of all possible partitionings. For example, $N = 10$ sites yields a total of 115,975 ways to partition sites to rate classes. In this paper, we label partitions according to the restricted growth function notation of Stanton and White (1986); elements are sequentially numbered with the constraint that the index numbers for two sites are the same if they are found in the same rate category. If K is fixed, then the total number of ways to partition N sites among K categories is given by the Stirling numbers of the second kind:

$$\mathcal{S}_2(N, K) = \frac{1}{K!}\sum_{j=0}^{K-1}(-1)^j\binom{K}{j}(K - j)^N. \tag{12}$$

Under the Dirichlet process prior model, K is random and can range from one to N. Therefore, the total number of ways to partition N sites among rate categories is a sum of the Stirling numbers of the second kind. This sum is called the Bell number (Bell, 1934):

$$\mathcal{B}_N = \sum_{K=1}^{N}\mathcal{S}_2(N, K). \tag{13}$$

The number of possible ways to assign sites to rate categories can become very large. For example, for sequences of length $N = 1000$, there are a total of $\mathcal{B}_{1000} \approx 3 \times 10^{1927}$ possible ways to assign sites to rate categories!

The Dirichlet process prior treats both the number of rate clases (K) and the assignment of sites to rate classes ($\boldsymbol{\sigma}$) as random variables (Antoniak, 1974; Ferguson, 1973). With this in mind, the Dirichlet process prior model for rate variation across sites can be formally described as follows. First, K and $\boldsymbol{\sigma}$ are both drawn a priori from the probability distribution

$$f(\boldsymbol{\sigma}, K|\chi, N) = \chi^K\frac{\prod_{k=1}^{K}(\eta_k - 1)!}{\prod_{i=1}^{N}(\chi + i - 1)}, \tag{14}$$

where η_k is the number of sites assigned to rate category k and χ is the "concentration" parameter for the Dirichlet process prior model. After the sites have been assigned to K rate classes, a tree length $T_{[k]}$ is assigned to each category k by drawing $T_{[k]}$ from a gamma probability distribution with shape $2S - 3$ and scale λ.

The paramount parameter of the Dirichlet process prior model is its concentration parameter χ. Generally speaking, χ determines how clumpy the process becomes. If χ is small, then the Dirichlet process prior model tends to favor fewer classes. In fact, the prior probability of K categories is

$$f(K|\chi, N) = \frac{\mathcal{S}_1(N, K)\chi^K}{\prod_{i=1}^{N}(\chi + i - 1)}, \tag{15}$$

where $\mathcal{S}_1(\cdot, \cdot)$ is the absolute value of the Stirling numbers of the first kind. One obtains Equation (15) by integrating Equation (14) over all possible partitions $\boldsymbol{\sigma}$ for K categories. The prior probability of two sites i_1 and i_2 finding themselves grouped together into the same rate class is

$$f(\sigma(i_1) = \sigma(i_2)|\chi) = \frac{1}{1 + \chi}. \tag{16}$$

This formulation clearly shows that when χ is small, two sites are more likely to find themselves in the same rate class than when χ is large.

Markov Chain Monte Carlo

We wish to infer the parameters of the phylogenetic model in a Bayesian framework. We cannot feasibly calculate the posterior probability distribution of the parameters analytically, because doing so involves large summations and integrals that cannot be solved analytically. We resort to Markov chain Monte Carlo (MCMC) to approximate the posterior probability distribution of the phylogenetic model parameters (Metropolis et al., 1953; Hastings, 1970). The general goal is to construct a Markov chain whose state space reflects the parameter space of the phylogenetic model and has a stationary distribution that is the posterior probability distribution of interest. Samples then taken from this Markov chain at stationarity are valid, albeit dependent, samples from the posterior probability distribution of the phylogenetic parameters (Tierney, 1994). We base inference on the samples taken from the Markov chain. For example, the fraction of the time a specific phylogenetic tree is sampled estimates the marginal posterior probability of that tree.

We implement a variant of MCMC employing the Metropolis-Hastings algorithm. The general idea is to (1) propose a new state (combination of model parameters) for the Markov chain; (2) calculate the probability of accepting the proposed state as the next state of the Markov chain; and (3) accept or reject the proposed state according to the acceptance probability calculated in step 2. This procedure is repeated many times. Green (2003) describes a flexible method for constructing complex proposal mechanisms that keeps the calculation of the acceptance probability simple. Specifically, a new state $\mathbf{x}'$ is proposed as the next state of the Markov chain. The generation of the new state involves the generation of possibly many random numbers $\mathbf{u}$ from an arbitrary probability distribution $g(\mathbf{u})$. The new state is a deterministic function of the random numbers and the original state $\mathbf{x}$ of the Markov chain, $\mathbf{x}' = \mathbf{h}(\mathbf{x}, \mathbf{u})$. The reverse move from $\mathbf{x}'$ to $\mathbf{x}$—a move that is not actually made in computer memory—is imagined through another set of random numbers $\mathbf{u}'$ drawn from a possibly different probability distribution $\mathbf{g}'(\mathbf{u}')$, where $\mathbf{x} = \mathbf{h}'(\mathbf{x}', \mathbf{u}')$. The probability of accepting the proposed state $\mathbf{x}'$ as the next state of the Markov chain is

$$R = \min\left(1, \text{Likelihood Ratio} \times \text{Prior Ratio} \times \frac{\mathbf{g}'(\mathbf{u}')}{\mathbf{g}(\mathbf{u})} \times \left|\frac{\partial(\mathbf{x}', \mathbf{u}')}{\partial(\mathbf{x}, \mathbf{u})}\right|\right). \quad (17)$$

The last factor is called the Jacobian J of the transform to $\mathbf{x}'$ and $\mathbf{u}'$ with respect to $\mathbf{x}$ and $\mathbf{u}$; it is introduced to balance the change of variables that may occur when proposing new states for the Markov chain. In Appendix 1, we provide details on our parameter proposals. Of particular interest are the proposals to update the number of rate categories K and site rate class assignments $\boldsymbol{\sigma}$. We make available source code for the DPP rate-variation model by request to interested readers.

Data Analysis

We analyzed four alignments of protein-coding DNA sequences: (1) an alignment of β-globin sequences sampled from $S = 17$ vertebrates ($N = 432$; Yang et al., 2000); (2) an alignment of cytochrome oxidase I (COI) sequences sampled from $S = 15$ gopher species ($N = 379$; Hafner et al., 1994); (3) an alignment of COI sequences sampled from $S = 17$ louse species ($N = 379$; Hafner et al., 1994); and (4) an alignment of cytochrome *b* sequences sampled from $S = 31$ mammalian species ($N = 1140$; Larget and Simon, 1999).

All analyses were performed under the general time-reversible (GTR) model of DNA substitution. We analyzed the data under five different rate models; we implemented an equal rates GTR, GTR+SS, and GTR+Γ models for among-site rate variation, as well as a mixture model (GTR+SS+Γ) and the Dirichlet process prior model (GTR+DPP). The Dirichlet process prior requires that the concentration parameter χ realizes some value. One approach, not pursued here, is to place a hyperprior on the concentration parameter. Often, a gamma prior is placed on χ. The approach we pursue is to choose χ such that the prior mean of the number of rate categories is $E(K) = 3$, $E(K) = 5$, $E(K) = 10$, and $E(K) = 20$. Doing this allows us to investigate the robustness of results to choice of χ. The specific values of χ used to achieve these expectations are shown in Table 2.

TABLE 2. Concentration parameters χ of the Dirichlet process prior. For four prior expected number of rate categories $E(K)$, we report the requisite χ.

		χ			
Gene	N	$E(K) = 3$	$E(K) = 5$	$E(k) = 10$	$E(k) = 20$
Vertebrate β-globin	432	0.32	0.68	1.71	4.20
Gopher COI	379	0.33	0.70	1.76	4.35
Lice COI	379	0.33	0.70	1.76	4.35
Mammalian cytochrome *b*	1140	0.28	0.58	1.41	3.34

We analyzed each alignment two times, running a Markov chain for a total of one million cycles for each analysis. Updates of the allocation vector were attempted about 10% of the time. Each update of the allocation vector involved scanning all sites, attempting to reassign each site to a rate class. Hence, the total number of MCMC updates was much larger than one million (e.g., there were over 100 million updates performed for each analysis of the mammalian cytochrome *b* alignment). We discarded samples taken during the first half of a million cycles as the burn-in for the Markov chain before it reaches stationarity.

RESULTS AND DISCUSSION

Model Choice

We compared the fit of the Dirichlet process prior model for among-site rate variation to several alternative models for accommodating rate variation. For the four data sets examined in this study, the Dirichlet process prior model fitted the data substantially better than the alternative models. Table 3 shows the estimated log marginal likelihoods for five rate-variation models (Equal, +Γ, +SS, +SS+Γ, and +DPP) for all four

TABLE 3. The log marginal likelihoods for the different rate-variation models examined in this study. The subscript in DPP_i indicates the prior mean of the number of classes for the Dirichlet process prior model for among-site rate variation.

	Data set			
Model	Vertebrate β-globin	Gopher COI	Lice COI	Mammalian cytochrome *b*
Equal	−3950	−2430	−3456	−12714
$+\Gamma_4$	−3840	−2215	−3005	−11486
+SS	−3871	−2117	−2905	−11476
$+SS+\Gamma_4$	−3812	−2112	−2913	−11251
DPP_3	−3720	−2062	−2815	−10756
DPP_5	−3724	−2062	−2821	−10745
DPP_{10}	−3722	−2067	−2806	−10729
DPP_{20}	−3723	−2071	−2797	−10745

alignments. The marginal likelihood for model M_ℓ is obtained by integrating the model likelihood over all possible combinations of model parameters, weighing each combination by its prior probability under model ℓ. In a Bayesian framework, model choice is guided by comparing marginal likelihoods. The marginal likelihood of a model is the likelihood of the data integrated against all possible parameter values under the prior distribution. Specifically, their ratio is the Bayes factor (BF) in favor of the model in the numerator against the model in the denominator. We estimated the marginal likelihoods using the harmonic mean estimator of Newton and Raftery (1994). This estimator is consistent, although it can be adversely affected by rare posterior samples with small likelihood values. In situations where the differences between marginal likelihoods are more modest, more efficient BF estimation techniques are available; these include the Savage-Dickey ratio for nested models (Suchard et al., 2001), reversible-jump Markov chain Monte Carlo to sample the joint parameter space for competing models (Green, 1995), and bridge sampling via thermodynamic integration (Lartillot and Philippe, 2006). We obtain the log BFs by taking the differences between log marginal likelihoods in Table 3. In this light, BFs $\gg 1$ or $\ll 1$ still offer strong conclusions, whereas BFs ≈ 1 should be viewed with caution. After the DPP models, the next best model for all four data sets is either the SS or SS+Γ model.

Marginal Posterior Probability Distribution of Rates

To examine the nonparametric behavior of rate variation, we examined the marginal posterior probability density of the rates for each alignment site. Figure 1 shows the marginal posterior probability distribution of the tree length T for five arbitrarily chosen sites taken from the vertebrate β-globin alignment. The column vectors containing the information at sites 5, 30, 66, 156, and 341 are

$$\begin{aligned}
\mathbf{y}_5 &= (\mathrm{C,C,C,C,C,C,C,C,C,C,C,C,C,C,C,C})^{\mathrm{t}},\\
\mathbf{y}_{30} &= (\mathrm{T,T,T,T,T,C,C,C,C,T,T,T,T,C,C,T,T})^{\mathrm{t}},\\
\mathbf{y}_{66} &= (\mathrm{T,T,T,T,T,T,T,T,T,T,C,T,A,G,C,C})^{\mathrm{t}},\\
\mathbf{y}_{156} &= (\mathrm{T,T,T,T,T,T,T,C,T,C,C,C,C,C,C,T,C})^{\mathrm{t}}, \text{ and}\\
\mathbf{y}_{341} &= (\mathrm{A,A,A,A,A,G,G,G,G,A,A,A,A,C,C,C,C})^{\mathrm{t}}
\end{aligned} \qquad (18)$$

and are variant for all sites except site 5. The sequence characters in these data columns fall in the order: human, tarsier, bushbaby, hare, rabbit, cow, sheep, pig, elephant seal, rat, mouse, hamster, marsupial, duck, chicken, *Xenopus laevis*, and *Xenopus tropicalis*. The marginal posterior probability distribution of T is often multimodal. For example, for site 30, the probability distribution of T has two or three modes under the Dirichlet process prior model. These multiple modes are real; they are repeatably obtained from MCMC analyses that start from different random combinations of parameters. The marginal distribution is unimodal for the equal-rate and site-specific-rate models; in fact, under the equal-rate model, the probability distribution for all sites is identical (as it must be under an equal-rate model). Similarly, the marginal probability distribution of T is identical for all sites of the same codon position for the site-specific model. We implemented the gamma model for among-site rate variation with four discrete rate categories. This implementation detail is reflected in the marginal probability distribution of rates under the models that assume gamma-distributed rate variation (i.e., the +Γ and +SS+Γ models), where there are four modes, reflecting the mean rate for each rate category.

Similar to the gamma model, the Dirichlet process prior model for among-site rate variation also relies on discrete classes of rates. However, the assignment of sites to specific classes is now a random variable, as are the realized rate values for each class. Which rate class a site finds itself in depends not only upon the distribution of nucleotides among species for that site (e.g., the pattern of nucleotides at the site) but also upon the details of the tree and branch lengths.

Although the marginal posterior probability distribution for rates under the Dirichlet process prior model, and indeed for the other rate variation models, appears to be reliably approximated using the MCMC procedure, one should take care not to overinterpret the distributions. The marginal posterior probability distribution of rates is likely to be strongly influenced by details of the prior model and implementation. For example, the four modes in the marginal distribution of rates under the gamma model results because we implemented the gamma model with exactly four discrete categories (Yang, 1994), not for any biological reason. Similarly, the Dirichlet process prior model has a discrete number of rate classes to describe how rates vary across the sequences. The multiple modes in the marginal posterior probability distribution for rates under the Dirichlet process prior is likely caused by the fact that the model has a discrete number of rate classes; the rate class a site finds itself assigned to is a compromise between the optimal rate for that site and how well the rate for other sites can be explained by the rate class.

Codon Structure

All of the alignments examined in this study were of protein-coding DNA sequences. Common practice a priori partitions these data by codon position. Our modeling representation for codon partitioning is

$$\boldsymbol{\sigma}_{(\mathrm{codon})} = (1, 2, 3, 1, 2, 3, 1, 2, 3, \ldots). \qquad (19)$$

A comparison of the marginal likelihoods (Table 3) shows that a model with rates partitioned according to codon position (the +SS model in this paper) does not explain the data as well as a model in which the partition is considered a random variable (the +DPP model of this paper). Although the codon partition does not fit as well as other partitions that were sampled during the course of the MCMC analysis, one can ask whether the codon partition is in some sense closer to the sampled partitions than, say, a random partition of the sites.

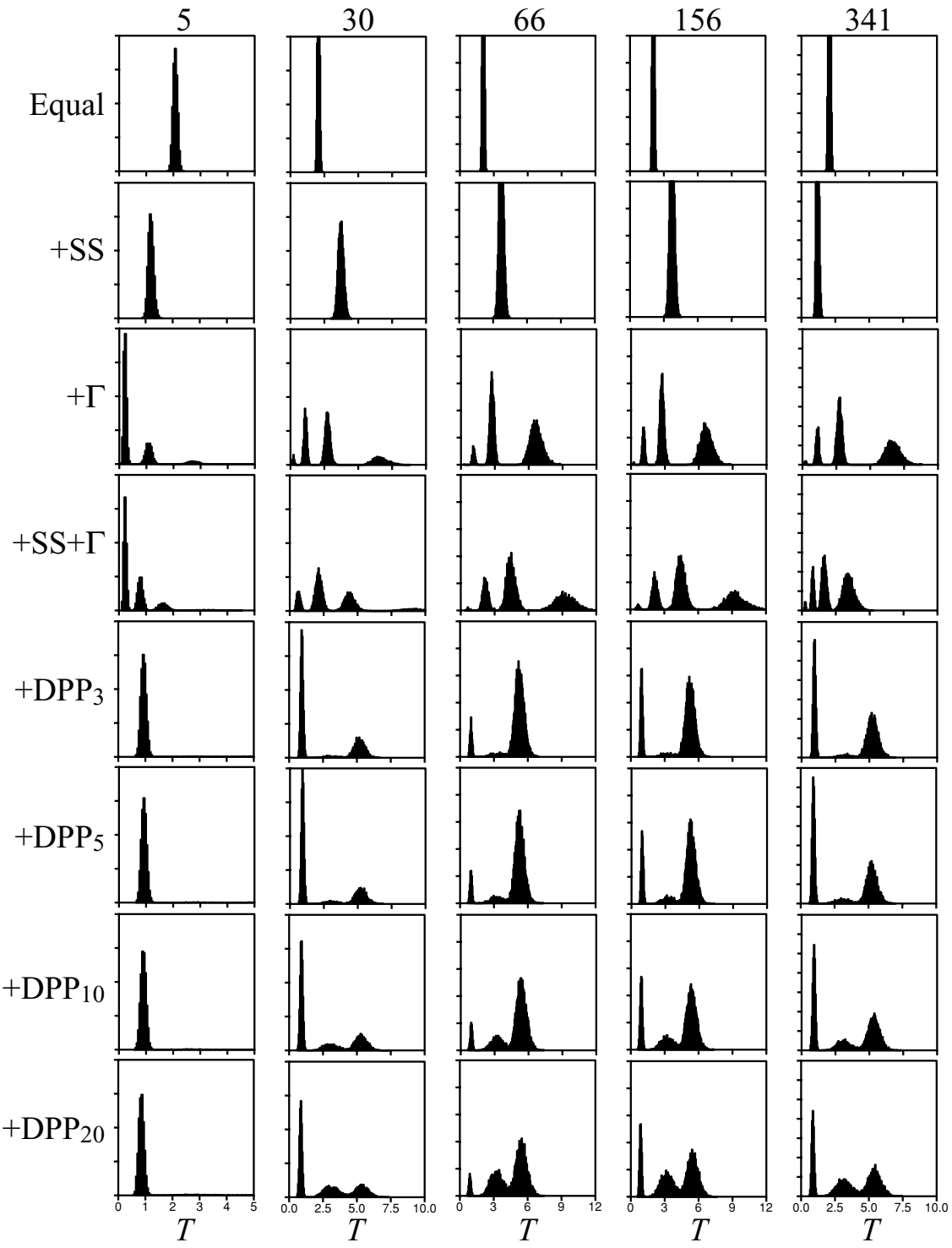

FIGURE 1. The marginal posterior probability density of the tree length T for five vertebrate β-globin sites under eight different rate variation models.

We explore the possibility that the partitions sampled under the Dirichlet process prior model are "closer" to the traditional codon partition than a random partitioning of the sites. We use a distance on partitions described by Gusfield (2002). Gusfield's distance between two partitions $d(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$ is the minimum number of elements that need to be deleted from the partitions to make the induced partitions equal. Equivalently, the distance between partitions is the minimum number of elements that must be moved from one cluster to another to make

the partitions the same. During the course of the MCMC analysis under the Dirichlet process prior model, a large number of partitions are sampled. These sampled partitions represent the current state of the Markov chain, describing how sites at that chain step are partitioned to rate classes. We will label the sampled partitions $\boldsymbol{\sigma}_{(p)}$ for $p = 1, \ldots, P$, the total number of sampled partitions. For each sampled partition, we also create a randomly permuted version, in which the elements are randomly assigned to clusters. This permutation procedure maintains the number of rate categories for the partition, but randomly assigns sites to rate classes. For example, consider the sampled partition

$$\boldsymbol{\sigma} = (1, 2, 1, 1, 2, 3, 1, 4, 1, 2). \quad (20)$$

Possible permutations of this partition include

$$\begin{aligned}
\boldsymbol{\sigma}' &= (1, 2, 1, 2, 2, 3, 4, 2, 1, 2),\\
\boldsymbol{\sigma}' &= (1, 2, 2, 2, 2, 3, 2, 4, 1, 1),\\
\boldsymbol{\sigma}' &= (1, 2, 2, 3, 4, 2, 4, 4, 2, 2),\\
\boldsymbol{\sigma}' &= (1, 1, 2, 1, 3, 3, 4, 1, 1, 3), \text{ and}\\
\boldsymbol{\sigma}' &= (1, 1, 2, 3, 4, 2, 2, 2, 1, 2). \quad (21)
\end{aligned}$$

We label $\boldsymbol{\sigma}'_{(p)}$ as a permutation of the original sample $\boldsymbol{\sigma}_{(p)}$. For each of the four alignments, we calculate the statistic

$$\frac{1}{P}\sum_{p=1}^{P} \mathrm{I}[d(\boldsymbol{\sigma}_{(\mathrm{codon})}, \boldsymbol{\sigma}_{(p)}) - d(\boldsymbol{\sigma}_{(\mathrm{codon})}, \boldsymbol{\sigma}'_{(p)}) > 0], \quad (22)$$

where $\mathrm{I}(\cdot)$ is the indicator function. This statistic is the fraction of the time the sampled partition is closer to the codon partition than a random permutation of the partition; for a long MCMC chain, this fraction converges to the probability that the data partitioning is closer to codon partitioning than random.

Table 4 shows the results of our analysis. The sampled partitions are closer to the codon partition than are random permutations of the sampled partitions. For example, about 97% to 99% of the sampled partitions for the alignments of COI for gophers and lice are closer to the codon partition than a random partition of the sites. This suggests that the Dirichlet process prior model we implemented appropriately captures the codon structure present in the data without a priori specification.

TABLE 4. The probability that the Dirichlet process prior partitionings are closer to a codon partitioning than random permutations.

	Data set			
Model	Vertebrate β-globin	Gopher COI	Lice COI	Mammalian cytochrome b
DPP_3	0.76	0.98	0.99	0.66
DPP_5	0.76	0.98	0.99	0.66
DPP_{10}	0.75	0.98	0.98	0.66
DPP_{20}	0.75	0.98	0.97	0.65

In addition to codon partitioning, the Dirichlet process prior model also appropriately identifies the differences in rates across codon sites. Table 5 summarizes the mean of the marginal posterior probability distribution of tree length T for first-, second-, and third-codon positions. As expected, the second position sites have the lowest rate of substitution and the third position sites have the highest rate.

Nonparametric Across-Site Rate Variation

The Dirichlet process prior model for among-site rate variation is similar in many ways to the site-specific model. The main difference between the Dirichlet process prior and the site-specific models concerns how the sites are partitioned to rate classes. For the site-specific model, the biologist must determine the partitioning scheme before the analysis; the partitioning scheme that is chosen, whether it partitions sites by codon position or some other biologically relevant aspect of the data, becomes an inescapable assumption of the analysis. Under the Dirichlet process prior model, on the other hand, the partitioning scheme is considered to be a random variable. Our method is also similar to the mixture models proposed by Pagel and Meade (2004, 2005), with the main difference being the question addressed (substitution rate variation versus variation in the rate matrix of the continuous-time Markov chain model of character change) and the prior probability distribution on partitions.

Our MCMC implementation of the Dirichlet process prior model for among-site rate variation depends upon algorithm 8 of Neal (2000). The MCMC appears to perform well for the four data sets we explore. However, the potential space of partitioning schemes which the MCMC chain explores is huge; for any particular data set, our implementation may fail. Other proposal mechanisms can be implemented for this problem, including some that involve split-and-merge proposals, or "sweetened" split-and-merge proposals, in which sites in a cluster are either split into two clusters or the elements in two clusters are merged into one (Jain and Neal, 2004, 2005).

The Dirichlet process prior model for among-site rate variation explains the pattern of rate variation in the four data sets we analyzed better than any other existing model. We do not know, of course, whether this will remain true for other data sets. However, we are hopeful that the Dirichlet process prior model will continue to do well for future data sets, as the Dirichlet process prior is very flexible. Importantly, the Dirichlet process prior allows sites with similar rates to be grouped together into the same rate class. This is not true for commonly used models for among-site rate variation. For example, consider a site-specific model, with sites partitioned by codon position. All sites at the same codon position are treated the same, regardless of any additional patterns of variation at the sites (Buckley et al., 2001); an invariant third-codon site is grouped together with third-codon position sites for which all four nucleotides are observed.

TABLE 5. The marginal posterior mean of the tree length T at first, second, and third codon positions. The mean tree length for the codon positions are formatted as "(first, second, third)."

	Data set			
Model	Vertebrate β-globin	Gopher COI	Lice COI	Mammalian cytochrome b
DPP_3	(2.18, 1.73, 3.33)	(0.74, 0.34, 3.87)	(1.31, 0.48, 7.81)	(2.35, 0.64, 8.79)
DPP_5	(2.18, 1.74, 3.31)	(0.74, 0.34, 3.86)	(1.30, 0.48, 7.77)	(2.35, 0.64, 8.82)
DPP_{10}	(2.17, 1.73, 3.30)	(0.74, 0.34, 3.83)	(1.30, 0.49, 7.82)	(2.36, 0.64, 8.76)
DPP_{20}	(2.16, 1.72, 3.25)	(0.74, 0.35, 3.70)	(1.26, 0.48, 7.80)	(2.34, 0.64, 8.82)

This problem can be reduced if one used a mixture of the site-specific and gamma rate-variation models. However, even under the SS+Γ model, considerable structure remains in the position rates. The Dirichlet process prior model captures this additional variation easily.

The Dirichlet process prior model we describe can be improved in a number of ways. For example, under the current model, the tree length for a rate class is a random variable drawn from a gamma($2S - 3$, λ) probability distribution. However, the current implementation makes it difficult to accommodate sites that have very low rates (e.g., invariant sites). It should be possible to augment tree lengths with a strictly invariant class. Here, we consider the tree length as a random variable with length equal to zero with probability p and drawn from a gamma($2S - 3$, λ) probability distribution with probability $1 - p$. This would represent a mixture of the proportion of invariable sites and Dirichlet process prior models. Employing a Dirichlet process mixture model may allow a better fit to patterns of among-site rate variation (MacEachern and Müller, 1998).

ACKNOWLEDGMENTS

J.P.H. was supported by NSF grant DEB-0445453 and NIH grant GM-069801. M.A.S. was supported by NIH grant GM-068955.

REFERENCES

Antoniak, C. E. 1974. Mixtures of Dirichlet processes with applications to non-parametric problems. Ann. Stat. 2:1152–1174.

Arndt, P. F., T. Hwa, and D. A. Petrov. 2005. Substantial regional variation in substitution rates in the human genome: Importance of GC content, gene density, and telomere-specific effects. J. Mol. Evol. 6:748–763.

Bell, E. T. 1934. Exponential numbers. Am. Math. Monthly 41:411–419.

Buckley, C. Simon, and G. K. Chambers. 2001. Exploring among-site variation models in a maximum likelihood framework using empirical data: Effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. Syst. Biol. 50:67–86.

Ewens, W. J., and S. Tavaré. 1998. The Ewens sampling formula. Pages 230–234 in Encyclopedia of statistical science (S. Kotz, C. B. Read, and D. L. Banks, eds). Wiley, New York.

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

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts.

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

Fitch, W. M., and E. Markowitz. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. 4:579–593.

Galtier, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol. Biol. Evol. 18:866–873.

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

Green, P. J. 2003. Trans-dimensional Markov chain Monte Carlo. Pages 179–198 *in* Highly structured stochastic systems (P. J. Green, N. L. Hjort, and S. Richardson, eds.). Oxford University Press, Oxford, UK.

Gu, Y., Y. X. Fu, and W. H. Lu. 1995. Maximum likelihood estimation of the heterogeneity of substitution rates among nucleotide sites. Molecular Biology and Evolution 12:546–557.

Gusfield, D. 2002. Partition-distance: A problem and class of perfect graphs arising in clustering. Inform. Process. Lett. 82:159–164.

Hafner, M. S., P. D. Sudman, F. X. Villablanca, T. A. Spradling, J. W. Demastes, and S. A. Nadler. 1994. Disparate rates of molecular evolution in cospeciating hosts and parasites. Science 265:1087–1090.

Hasegawa, M., H. Kishino, and T. Yano. 1987. Man's place in Hominoidea as inferred from molecular clocks of DNA. J. Mol. Evol. 26:132–147.

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

Holder, M. T., P. O. Lewis, D. L. Swofford, and B. Larget. 2005. Hastings ratio of the LOCAL proposal used in Bayesian phylogenetics. Syst. Biol. 54:961–965.

Huelsenbeck, J. P. 2002. Testing a covariotide model of DNA substitution. Mol. Biol. Evol. 19:698–707.

Huelsenbeck, J. P., and D. M. Hillis. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42:247–270.

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

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

Jain, S., and R. Neal. 2000. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. J. Comput. Graph. Stat. 13:158–182.

Jain, S., and R. Neal. 2005. Splitting and merging components of a nonconjugate Dirichlet process mixture model. Technical Report 0507, Department of Statistics, University of Toronto.

Jin, L., and M. Nei. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. 7:82–102.

Kosakovsky Pond, S. L., and S. D. W. Frost. 2005. A simple hierarchical approach to modeling distributions of substitution rates. Mol. Biol. Evol. 22:223–234.

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

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

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

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

MacEachern, S.N., and P. Müller. 1998. Estimating mixtures of Dirichlet process models. J. Comput. Graph. Stat. 7:223–238.

Mau, B. 1996. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. PhD dissertation, University of Wisconsin, Madison.

Mau, B., and M. Newton. 1997. Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. J. Comput. Graph. Stat. 6:122–131.

Mau, B., M. Newton, and B. Larget. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics 55:1–12.
Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. W. Teller, and E. Teller. 1953. Equations of state calculations by fast computing machines. J. Chem. Phys. 21:1087–1091.
Neal, R. M. 2000. Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Stat. 9:249–265.
Nei, M., R. Chakraborty, and P. A. Fuerst. 1976. Infinite allele model with varying mutation rate. Proc. Nat. Acad. Sci. USA 73:4164–4168.
Newton, M. A., and A. E. Raftery. 1994. Approximate Bayesian inference by the weighted likelihood bootstrap. J. R. Stat. Soc. Ser. B 56:3–48.
Newton, M., B. Mau, and B. Larget. 1999. Markov chain Monte Carlo for the Bayesian analysis of evolutionary trees from aligned molecular sequences. Pages 143–162 *In* Statistics in molecular biology (F. Seillier-Moseiwitch, T. P. Speed, and M. Waterman, eds.). Monograph Series of the Institute of Mathematical Statistics. IMS, Hayward, California.
Nielsen, R. 1997. Site-by-site estimation of the rate of evolution and the correlation of rates in mitochondrial DNA. Syst. Biol. 46:346–353.
Olsen, G. J. 1987. Earliest phylogenetic branchings: Comparing rRNA-based evolutionary trees inferred with various techniques. Cold Spring Harbor Symp. Quant. Biol. 52:825–837.
Pagel, M., and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Sys. Biol. 53:571–581.
Pagel, M., and A. Meade. 2005. Mixture models in phylogenetic inference. Pages 121–142 *in* Mathematics of evolution and phylogeny (O. Gascuel, ed.). Oxford University Press, Oxford, UK.
Rannala, B., and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J. Mol. Evol. 43:304–311.
Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.
Schröder, E. 1870. Vier combinatorische Probleme. Z. Math. Phys. 15:361–376.
Stanton, D., and D. White. 1986. Constructive combinatorics. Springer-Verlag, New York.
Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics, 2nd edition (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer Associates, Sunderland, Massachusetts.
Tavaré, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Pages 57–86 *in* Lectures in mathematics in the life sciences, volume 17.
Tierney, L. 1994. Markov chains for exploring posterior distributions. Ann. Stat. 22:1701–1762.
Tuffley, C., and M. Steel. 1998. Modeling the covarion hypothesis of nucleotide substitution. Math. Biosci. 147:63–91.
Uzzell, T., and K. W. Corbin. 1971. Fitting discrete probability distributions ot evolutionary events. Science 172:1089–1096.
Waddell, P. J. and M. A. Steel. 1997. General time reversible distances with unequal rates across sites. Mol. Phylogenet. Evol. 8:398–414.
Wakeley, J. 1994. Substitution rate variation among sites and the estimation of transition bias. Mol. Biol. Evol. 11:436–442.
Yang, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396–1401.
Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. 39:306–314.
Yang, Z. 1995. A space-time process model for the evolution of DNA sequences. Genetics 139:993–1005.
Yang, Z., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte carlo method. Mol. Biol. Evol. 14:717–724.
Yang Z., R. Nielsen, N. Goldman, and A. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449.

First submitted 15 November 2006; reviews returned 28 January 2007; final acceptance 7 June 2007
Associate Editor: Thomas Buckley

APPENDIX 1

Here we describe in detail the MCMC proposal mechanisms for parameters in the Dirichlet process prior model for among-site rate variation developed in this paper.

Tree.—We use a variant of the "LOCAL" proposal mechanism of Larget and Simon (1999; also see Holder et al., 2005) to propose new trees. The proposal mechanism works as follows. First, an internal branch of the phylogenetic tree is chosen at random. Also chosen at random are two branches, incident to each end of the randomly chosen internal branch. Together, these random choices result in three contiguous branches. Figure 2 shows the area of rearrangement and uses a notation similar to that in Holder et al. (2005). The three branches that were randomly chosen are on the path between nodes a and c, and are referred to as the "backbone" of the move. These three branches represent a proportion of the total tree length $\hat{\phi}_1 = \phi_{ai} + \phi_{ij} + \phi_{ic}$. The branches outside of the backbone, those not randomly chosen, have a proportion of the tree length $\hat{\phi}_2 = 1 - \hat{\phi}_1$. For the example shown in Figure 1, which has only four tips, $\hat{\phi}_2 = \phi_{ib} + \phi_{jd}$. Second, we draw new values for $\hat{\phi}_1'$ and $\hat{\phi}_2'$ from a beta probability distribution with parameters $\hat{\phi}_1\psi_1$ and $\hat{\phi}_2\psi_1$, centered at the original proportions with tunable precision ψ_1. Finally, one of the two branches that are incident to the backbone, either the branch ib or the branch jd is chosen at random and detached from the backbone. That branch is then randomly reattached to the backbone, along the modified path that now has proportion ϕ'_{ac}.

The modified LOCAL proposal mechanism involves the generation of six random variables: (1) u_1, the random choice of internal branch; (2) u_2, the random choice of one of the two branches incident to one end of the randomly chosen internal branch; (3) u_3, the random choice of one of the two branches incident to the other end of the randomly chosen internal branch; (4) u_4, the new value for $\hat{\phi}_1$, which is drawn from a beta probability distribution; (5) u_5, the random choice of one of the two branches that are incident to the backbone (the three branches chosen with u_1, u_2, and u_3); and (6) u_6, the random proportion of the distance from node a to node c that the branch chosen using u_5 now falls. The probability/probability density of these six random variables is as follows:

$$\begin{aligned} g_1(u_1) &= \frac{1}{S-3}, \\ g_2(u_2) &= \frac{1}{2}, \\ g_3(u_3) &= \frac{1}{2}, \\ g_4(u_4) &= \frac{\Gamma(\psi_1)}{\Gamma(\hat{\phi}_1\psi_1)\Gamma(\hat{\phi}_2\psi_1)} u_4^{\hat{\phi}_1\psi_1-1}(1-u_4)^{\hat{\phi}_2\psi_1-1}, \\ g_5(u_5) &= \frac{1}{2} \text{ and} \\ g_6(u_6) &= 1. \end{aligned} \tag{23}$$

As in Holder et al. (2005), we assume without loss of generality that branch jd is chosen to randomly move along the backbone.

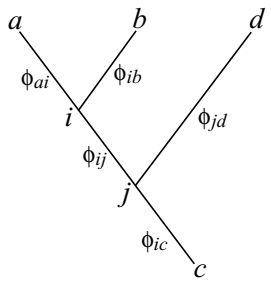

FIGURE 2. The area of rearrangement for the modified LOCAL proposal mechanism. The branch length proportions are denoted ϕ and are changed by the LOCAL proposal.

Moreover, we will follow the parameterization suggested by Holder et al. (2005) and measure path lengths from node a on the phylogenetic tree of Figure 2. Also for illustrative purposes, we drop dependence on the substitution process parameters temporarily. Consider the original state $\mathbf{x} = (\phi_{ai}, \phi_{aj}, \phi_{ac}, \phi_{ib}, \phi_{jd})$ and the proposed state $\mathbf{x}' = (\phi'_{ai}, \phi'_{aj}, \phi'_{ac}, \phi'_{ib}, \phi'_{jd})$. We obtain the imagined random variables $\mathbf{u}' = (u'_1, u'_2, u'_3, u'_4, u'_5, u'_6)$ to return to the original state from the six drawn random variables $\mathbf{u} = (u_1, u_2, u_3, u_4, u_5, u_6)$, via

$$\begin{aligned}
\phi'_{ai} &= \frac{\phi_{ai}}{\phi_{ac}} u_4, \\
\phi'_{aj} &= u_6 u_4 \\
\phi'_{ac} &= u_4 \\
\phi'_{ib} &= \frac{\phi_{ib}}{1 - \phi_{ac}} (1 - u_4) \\
\phi'_{jd} &= \frac{\phi_{jd}}{1 - \phi_{ac}} (1 - u_4) \\
u'_1 &= u_1
\end{aligned}$$

$$\begin{aligned}
u'_2 &= u_2 \\
u'_3 &= u_3 \\
u'_4 &= \phi_{ac} \\
u'_5 &= u_5 \\
u'_6 &= \frac{\phi_{aj}}{\phi_{ac}}
\end{aligned} \tag{24}$$

The Jacobian is the absolute value of the matrix of partial derivatives and is

$$J = \frac{u_4^2 (1 - u_4)^{2S-6}}{\phi_{ac}^2 (1 - \phi_{ac})^{2S-6}}, \tag{25}$$

and the acceptance probability for our modification of the LOCAL tree proposal mechanism becomes

$$R = \min\left(1, \text{Likelihood Ratio} \times 1 \times \frac{\Gamma(p_1\psi_1)\Gamma(p_2\psi_1) u_4'^{\,p'_1\psi_1 - 1} (1 - u'_4)^{p'_2\psi_1 - 1}}{\Gamma(p'_1\psi_1)\Gamma(p'_2\psi_1) u_4^{\,p_1\psi_1 - 1} (1 - u_4)^{p_2\psi_1 - 1}} \times \frac{u_4^2 (1 - u_4)^{2S-6}}{\phi_{ac}^2 (1 - \phi_{ac})^{2S-6}}\right). \tag{26}$$

The prior ratio is one, because all unrooted phylogenetic trees have equal probability and all combinations of branch length proportions have equal prior probability under a flat Dirichlet prior.

Base frequencies.—We propose new base frequencies by randomly drawing from a Dirichlet($\psi_2 \times \boldsymbol{\pi}$) where ψ_2 is a tunable precision parameter. The acceptance probability for a proposal of new nucleotide frequencies is

$$R = \min\left(1, \text{Likelihood Ratio} \times 1 \times \frac{\Gamma(\pi_A\psi_2)\Gamma(\pi_C\psi_2)\Gamma(\pi_G\psi_2)\Gamma(\pi_T\psi_2) \pi_A^{\pi'_A\psi_2 - 1} \pi_C^{\pi'_C\psi_2 - 1} \pi_G^{\pi'_G\psi_2 - 1} \pi_T^{\pi'_T\psi_2 - 1}}{\Gamma(\pi'_A\psi_2)\Gamma(\pi'_C\psi_2)\Gamma(\pi'_G\psi_2)\Gamma(\pi'_T\psi_2) \pi_A'^{\pi_A\psi_2 - 1} \pi_C'^{\pi_C\psi_2 - 1} \pi_G'^{\pi_G\psi_2 - 1} \pi_T'^{\pi_T\psi_2 - 1}}\right). \tag{27}$$

Again, the prior ratio is one because all combinations of nucleotide frequencies have equal probability under a flat Dirichlet prior.

Substitution rates.—We use the same proposal mechanism for substitution rate proportions as we do for nucleotide frequencies but with tunable precision parameter ψ_3. All other aspects of the move are similar, including the acceptance probability.

Rate classes.—We implemented a Gibbs sampling procedure with auxiliary components (Neal, 2000, algorithm 8) to update the allocation vector $\boldsymbol{\sigma}$. The method works as follows. First, we choose random site i and remove it from the current set of K rate classes. Assume that $\sigma(i) = k$. If site i is in a rate class by itself, then we remove that rate class from the set of all classes, decreasing K by one. Otherwise, η_k decreases by one. Let $K^{(-i)}$ denote the number of rate classes assigned to rate class k that remain after the removal of site i. We label the tree lengths for these classes $c_1, \ldots, c_{K^{(-i)}}, T_{[1]}, \ldots, T_{[K^{(-i)}]}$. We then construct κ auxiliary rate classes, with the rate for each class freshly drawn from the prior on the tree length T. We label these auxiliary tree lengths $T_{[K^{(-i)}+1]}, \ldots, T_{[K^{(-i)}+\kappa]}$. In the Gibbs portion of the step, we now reassign site i to new rate class k' with probability

$$f(\sigma(i) = k') = \begin{cases} b \times \eta_{k'}^{(-i)} f(\mathbf{y}_i | \tau, \phi \times T_{[k']}, \boldsymbol{\pi}, \boldsymbol{\theta}) & : \quad 1 \le k' \le K^{(-i)} \\ b \times (\chi/\kappa) f(\mathbf{y}_i | \tau, \phi \times T_{[k']}, \boldsymbol{\pi}, \boldsymbol{\theta}) & : \quad K^{(-i)} < k' \le K^{(-i)} + \kappa \end{cases} \tag{28}$$

where b is an easy-to-calculate normalizing constant. After this update, we discard any rate classes not associated with at least one site. We implement this Gibbs sampling component-wise across all N sites to arrive at a new K and $\boldsymbol{\sigma}$. For the analyses in this paper, we chose $\kappa = 5$.

We also implemented a proposal mechanism that changes the tree length $T_{[k]}$ for each rate class k using a random rate multiplier. The new tree length proposal $T'_{[k]} = T_{[k]} e^{\psi_4(u - 1/2)}$, where u is a uniform(0,1) random variable and ψ_4 is a tuning parameter. The acceptance probability for this proposal is

$$R = \min\left(1, \text{Likelihood Ratio} \times e^{\lambda(T_{[k]} - T'_{[k]})} \times \frac{T'_{[k]}}{T_{[k]}}\right). \tag{29}$$

We found that the MCMC worked well when the adjustable tuning parameters of the proposal mechanisms took the following values: $\psi_1 = 100$, $\psi_2 = 100$, $\psi_3 = 100$, and $\psi_4 = \log_e(1.1)$.