Bayesian Analysis of Partitioned Data

Brian Moore[1,2], Jim McGuire[1,3], Fredrik Ronquist[4], and John P. Huelsenbeck[1]

[1]*Department of Integrative Biology, University of California, Berkeley*
3060 VLSB #3140, Berkeley, CA 94720-3140, U.S.A.

[2]*Department of Evolution and Ecology, University of California, Davis*
Storer Hall, One Shields Avenue, Davis, CA 95616, U.S.A.

[3]*Museum of Vertebrate Zoology, University of California, Berkeley*
3101 VLSB #3160, Berkeley, CA 94720-3160, U.S.A.

[4]*Swedish Museum of Natural History,*
Box 50007, SE-104 05 Stockholm, Sweden

To whom correspondence should be addressed:

Brian R. Moore
University of California, Davis
Department of Evolution and Ecology
Storer Hall, One Shields Avenue
Davis, CA 95616
U.S.A.

Phone: 530-752-7104
E-mail: `brianmoore@ucdavis.edu`

Abstract.—Variation in the evolutionary process across the sites of nucleotide sequence alignments is well established, and is an increasingly pervasive feature of data sets composed of gene regions sampled from multiple loci and/or different genomes. Inference of phylogeny from these data demands that we adequately model the underlying process heterogeneity; failure to do so can lead to biased estimates of phylogeny and other parameters. Traditionally, process heterogeneity has been accommodated by first assigning sites to data partitions based on relevant prior information (reflecting codon positions in protein-coding DNA, stem and loop regions of ribosomal DNA, etc.), and then estimating the phylogeny and other model parameters under the resulting mixed model. Here, we consider an alternative approach for accommodating process heterogeneity that is similar in spirit to this conventional mixed-model approach. However, rather than treating the partitioning scheme as a fixed assumption of the analysis, we treat the process partition as a random variable with a Dirichlet process prior model. We apply this method to several empirical data sets, and compare our results to those estimated previously using conventional mixed-model selection criteria based on Bayes factors. We find that estimation under the Dirichlet process prior model discovers novel partition schemes that may more effectively balance error variance and estimation bias, while rendering phylogenetic inference more robust to process heterogeneity by virtue of integrating estimates over all possible partition schemes.

[Bayesian phylogenetic inference; Dirichlet process prior; Markov chain Monte Carlo; maximum likelihood; partitioned analyses; process heterogeneity.]

It is widely acknowledged that the pattern of nucleotide substitution across an alignment of gene sequences can exhibit heterogeneity, and that this variation can potentially cause problems for phylogenetic analysis unless the variability is accommodated. Deviations from a homogeneous substitution process include both simple *rate heterogeneity* (i.e., among-site rate variation) stemming from site-to-site differences in selection-mediated functional constraints, systematic differences in mutation rate, etc., or may involve more fundamental *process heterogeneity*, where the sites in an alignment are evolving under qualitatively different evolutionary processes. In the worst case, the phylogenetic tree relating the species in an alignment may vary across sites as a result of lineage sorting, hybridization, or horizontal gene transfer. Even when all of the sites in an alignment share a common phylogenetic history, however, other aspects of their evolutionary process may differ. Process heterogeneity might occur within a single gene region (e.g., between stem and loop regions of ribosomal sequences), or among gene regions in a concatenated alignment (e.g., comprising multiple nuclear loci and/or gene regions sampled from different genomes). Here we focus on inference scenarios where the sites of an alignment share a common phylogenetic history but where two or more *process partitions* in the data (sensu Bull et al., 1993) may otherwise differ with respect to the underlying process of molecular evolution.

Failure to accommodate process heterogeneity is known to adversely impact phylogeny estimation, causing biased estimates of the tree topology and nodal support (Brandley et al., 2005; Brown and Lemmon, 2007), estimates of branch lengths and divergence times (Marshall et al., 2006; Poux et al., 2008; Vendetti et al., 2008), and estimates of other model parameters (Nylander et al., 2004; Pagel and Meade, 2004).

To avoid these problems, investigators typically adopt a Bayesian 'mixed-model' approach (Ronquist and Huelsenbeck, 2003) in which the sequence alignment is first parsed into a number of partitions that are intended to capture plausible process heterogeneity within the data, then a substitution model is specified for each predefined process partition (using a given model-selection criterion, such as the hierarchical likelihood ratio test or the Akaike Information Criterion), and finally the phylogeny and other parameters are estimated under the resulting composite model. In this approach, therefore, the partition scheme is an assumption of the inference (i.e., the parameter estimates are conditioned on the specified mixed model), and the parameters of each process partition are independently estimated. For most sequence alignments, several (possibly many) partitioning schemes of varying complexity are plausible *a priori*, which therefore requires a way to objectively identify the partitioning scheme that balances estimation bias and error variance associated with under- and over-parameterized mixed models, respectively. Increasingly, mixed-model selection is based on Bayes factors (e.g., Suchard et al., 2001), which involves first calculating the

marginal likelihood under each candidate partition scheme and then comparing the ratio of the marginal likelihoods for the set of candidate partition schemes (Brandley et al., 2005; Nylander et al., 2004; McGuire et al., 2007).

There are several potential concerns associated with the use of this mixed-model selection approach. As a practical matter, it may not be feasible to evaluate all (or even most) plausible partition schemes for a given alignment. Evaluating all plausible mixed models quickly becomes computationally prohibitive, as each candidate partition scheme requires a full MCMC analysis to estimate the marginal likelihood. Consequently, this approach may result in the specification of a suboptimal composite model. Moreover, there is some concern that the most common technique for approximating the marginal likelihood of a (mixed) model — the harmonic mean estimator of Newton and Raftery (1994) — may be biased toward the inclusion of superfluous parameters, leading to the selection of over-partitioned composite models (Sullivan and Joyce, 2005; Brown and Lemmon, 2007). [Note that other marginal likelihood estimators — based on the Savage-Dickey ratio (Suchard et al., 2001) or thermodynamic integration (Lartillot and Philippe, 2006) — appear to avoid this bias, but are restricted to the evaluation of nested partition schemes or entail a substantially increased computational burden, respectively.] More generally, it has been argued that the application of Bayes factors to phylogenetic inference problems may not sufficiently penalize the inclusion of superfluous parameters (via the prior terms), which again may lead to selection of over-parameterized partition schemes (e.g., Pagel and Meade, 2004; Sullivan and Joyce, 2005). In light of these concerns, it is interesting to note that all of the empirical applications of Bayes factors have found very strong support for the most complex candidate partition schemes evaluated in those studies.

Here, we propose an alternative approach for accommodating process heterogeneity that treats the partition scheme, in which nucleotide sites are assigned to process partitions, as a random variable with a prior probability distribution that is specified by the Dirichlet process prior model. The Dirichlet process prior is a nonparametric model often used in Bayesian analysis of clustering problems where the data elements (e.g., nucleotide sites) are drawn from a mixture of an unknown number of probability distributions (e.g., the various evolutionary processes). The Dirichlet process prior model allows both the number of mixture components and the assignment of individual data elements to the set of mixture components to vary. This approach has recently been applied to several phylogenetic problems, such as identifying heterogeneity in the process of amino-acid replacement in protein sequences (Lartillot and Philippe, 2004), detecting sites under positive selection in protein-coding sequences (Huelsenbeck et al., 2006), and accommodating among-site substitution rate variation in nucleotide sequences (Huelsenbeck and Suchard, 2007). The Dirichlet

process prior model provides a natural means for accommodating process heterogeneity in nucleotide sequences because it allows us to specify a non-zero prior probability on all possible partition schemes, ranging from a uniform model (in which all sites are assigned to the same data partition) to a saturated model (in which each site is assigned to a unique data partition). These prior weights on partitions are first calculated analytically and then compared to their corresponding posterior probability estimates to provide a formal statistical framework for assessing the ability of partition schemes to capture process heterogeneity within the sequence data.

We first provide a more detailed description of the Dirichlet process prior model, and then describe how this approach can be used to identify the number and composition of partitions that best capture process heterogeneity. We apply this method to several empirical data sets, and then compare results under this approach to those obtained using conventional mixed-model selection based on Bayes factors.

Methods

Overview

In this study, we assume that substitutions occur according to the general time reversible (GTR) substitution model first described by Tavaré (1986) with gamma-distributed rate variation across sites (Yang, 1993, 1994). The GTR model assumes that substitutions occur along the branches of the phylogenetic tree according to the following matrix of rates

$$\mathbf{Q} = \begin{pmatrix} \cdot & r_{AC}\pi_C & r_{AG}\pi_G & r_{AT}\pi_T \\ r_{AC}\pi_A & \cdot & r_{CG}\pi_G & r_{CT}\pi_T \\ r_{AC}\pi_A & r_{CG}\pi_C & \cdot & r_{GT}\pi_T \\ r_{AC}\pi_A & r_{CT}\pi_C & r_{GT}\pi_G & \cdot \end{pmatrix} \mu$$

where the nucleotides are in the order A, C, G, T. The GTR model has six exchangeability parameters, $\mathbf{r} = (r_{AC}, r_{AG}, r_{AT}, r_{CG}, r_{CT}, r_{GT})$, that allow for rate biases between nucleotides and four nucleotide frequency parameters, $\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$, that are the stationary probabilities of the process. (The parameter μ in the above rate matrix is a scaling factor chosen such that the mean rate of substitution is one.) Rate variation across sites is modeled by assuming that the rate at a particular site in the sequence is a random variable drawn from a mean-one gamma distribution with parameter α. The parameters of the entire model for a data set with S species can then be summarized as follows:

Tree topology	τ
Branch-length proportions	$\mathbf{p} = (p_1, p_2, \ldots, p_{2S-3})$
Tree length	T
Exchangeability parameters	$\mathbf{r} = (r_{AC}, r_{AG}, r_{AT}, r_{CG}, r_{CT}, r_{GT})$
Nucleotide frequencies	$\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$
Gamma shape parameter	α

We estimate parameters of the phylogenetic model in a Bayesian framework. In a Bayesian analysis, inferences are based upon the posterior probability distribution of the parameters. The posterior probability is calculated using Bayes's theorem as

$$f(\tau, \mathbf{p}, T, \mathbf{r}, \boldsymbol{\pi}, \alpha \,|\, \mathbf{X}) = \frac{f(\mathbf{X} \,|\, \tau, \mathbf{p}, T, \mathbf{r}, \boldsymbol{\pi}, \alpha) f(\tau, \mathbf{p}, T, \mathbf{r}, \boldsymbol{\pi}, \alpha)}{f(\mathbf{X})}$$

where $\mathbf{X}$ represents the alignment(s) of DNA sequences (see below). Bayes's theorem states that the posterior probability distribution of the parameters $[f(\cdot \,|\, \mathbf{X})]$ is equal to the likelihood $[f(\mathbf{X} \,|\, \cdot)]$ times the prior probability distribution of the parameters $[f(\cdot)]$ divided by the marginal likelihood $[f(\mathbf{X})]$. The likelihood is calculated by under the GTR model and using numerical methods first described by Felsenstein (1981). Note that Bayesian inference treats the parameters as random variable, which therefore requires the specification of a prior probability distribution on those parameters. The prior is interpreted as the biologist's beliefs about the parameters before the sequence data were collected. In this study, we assume the following prior probability distributions for the parameters of the phylogenetic model:

$$\begin{aligned}
\tau &\sim \text{Discrete Uniform}(1, \ldots, B(S)) \\
\mathbf{p} &\sim \text{Dirichlet}(1, 1, \ldots, 1) \\
T &\sim \text{Gamma}(2S - 3, \lambda_1) \\
\mathbf{r} &\sim \text{Dirichlet}(1, 1, 1, 1, 1, 1) \\
\boldsymbol{\pi} &\sim \text{Dirichlet}(1, 1, 1, 1) \\
\alpha &\sim \text{Exponential}(\lambda_2)
\end{aligned}$$

where $B(S)$ is the number of unrooted trees possible for S species.

Modern phylogenetic studies often involve the simultaneous analysis of multiple data elements that may have evolved under qualitatively different evolutionary processes. For example, phylogenetic studies are increasingly based on multiple gene regions: How should one account for potential heterogeneity in the parameter values among the various genes in such analyses? In the program MrBayes (Ronquist and Huelsenbeck, 2003), one can allow the parameters of the model to be independently estimated or arbitrarily constrained in the different data subsets (in this case genes).

For example, one could allow the tree-length parameter, T, to be shared (i.e., 'linked') across all genes which makes the biological assumption that the rate of substitution in the genes is essentially equal. Alternatively, one could allow the tree-length parameter to be 'unlinked', such that this parameter is estimated independently for each gene, or constrain it to be equal for some genes but potentially different for others. This same reasoning applies to the other parameters in the model. The difficulty with this approach, however, is to find the best scheme among a bewildering number of alternative ways of linking and unlinking parameters across partitions.

In this study, we consider the use of a Dirichlet process prior (DPP) model to identify an optimal means of accommodating heterogeneity in parameters. Specifically, we apply independent DPP models to all of the parameters of the model, with the exception of the tree topology and branch-length proportions which we assume are shared across all of the data subsets. In the following, we describe the details of this model and the analyses performed in this study.

Data Partitions

Bull et al. (1993) introduced the notion of *process partitions* as divisions of characters into two or more subsets, with each subset evolving under potentially different rules. We assume that the biologist has identified subdivisions within the sequence data, and partitioned these data accordingly. The partitions might reflect knowledge of the structure of the gene, dividing the data, for example, according to coding versus non-coding DNA, stem versus loop regions of ribosomal genes, or by codon position for protein-coding genes. For genomic data, the partitions might be by gene. In the most extreme case, the biologist might assign every character (column in an alignment) to its own subset.

We assume that the biologist has correctly aligned sequences from S organisms. We further assume that the biologist has identified partitions in the complete alignment, dividing the alignment into K subsets. Consider, for example, the following alignment of $S = 5$ protein-coding sequences:

Human	`CTGACTCCTGAGGAGAAGTCTGCCGTTACT...`
Cow	`CTGACTGCTGAGGAGAAGGCTGCCGTCACC...`
Mouse	`CTGACTGATGCTGAGAAGTCTGCTGTCTCT...`
Marsupial	`TTGACTTCTGAGGAGAAGAACTGCATCACT...`
Chicken	`TGGACTGCTGAGGAGAAGCAGCTCATCACC...`

which are the first 30 sites of five of the sequences of a partial alignment of β-globin DNA sequences (Yang et al., 2000). This alignment can be partitioned by codon-position, which would result in

$K = 3$ subsets, which we label $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$:

$$\mathbf{X}_1 = \begin{pmatrix} \texttt{CACGGATGGA}\ldots \\ \texttt{CAGGGAGGGA}\ldots \\ \texttt{CAGGGATGGT}\ldots \\ \texttt{TATGGAATAA}\ldots \\ \texttt{TAGGGACCAA}\ldots \end{pmatrix} \quad \mathbf{X}_2 = \begin{pmatrix} \texttt{TCCAAACCTC}\ldots \\ \texttt{TCCAAACCTC}\ldots \\ \texttt{TCACAACCTC}\ldots \\ \texttt{TCCAAAAGTC}\ldots \\ \texttt{GCCAAAATTC}\ldots \end{pmatrix} \quad \mathbf{X}_3 = \begin{pmatrix} \texttt{GTTGGGTCTT}\ldots \\ \texttt{GTTGGGTCCC}\ldots \\ \texttt{GTTTGGTTCT}\ldots \\ \texttt{GTTGGGCCCT}\ldots \\ \texttt{GTTGGGGCCC}\ldots \end{pmatrix}$$

for the first, second, and third codon partitions, respectively. The complete alignment is denoted $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3)$.

Combinatorics of Partitions

The parameters of the phylogenetic model can be assigned to data subsets in a large number of possible ways. Consider, as an example, a simple case in which the data have been divided into three subsets and a single parameter, θ, is to be assigned to the various subsets. The parameter can be constrained to be the same in all three subsets (which is equivalent to subdividing the data, and then ignoring the subsets when estimating parameters), allowed to be potentially different in each subset, or constrained to be equal among some subsets but different in others. The table, below, enumerates the possible partitions of three subsets:

$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$		RGF
θ_1	θ_1	θ_1	$\rightarrow$	(1,1,1)
θ_1	θ_1	θ_2	$\rightarrow$	(1,1,2)
θ_1	θ_2	θ_1	$\rightarrow$	(1,2,1)
θ_1	θ_2	θ_2	$\rightarrow$	(1,2,2)
θ_1	θ_2	θ_3	$\rightarrow$	(1,2,3)

where the parameter subscript indicates the differentially estimated parameters and the partition is also described according to the restricted growth function (RGF) notation (Stanton and White, 1986). The total number of partitions for K elements is described by the Bell numbers (Bell, 1934). The Bell number for K elements is the sum of the Stirling numbers of the second kind:

$$\mathcal{B}(K) = \sum_{i=1}^{K} \mathcal{S}(K, i),$$

where the Stirling numbers of the second kind are given by the following equation:

$$\mathcal{S}(n, k) = \frac{1}{k!} \sum_{i=0}^{k-1} (-1)^i \binom{k}{i} (k - i)^n.$$

Dirichlet Process Prior

The Dirichlet Process Prior (DPP) model is a probability model on partitions (Ferguson, 1973; Antoniak, 1974). This model contains a singe parameter, called the concentration parameter (χ), that controls how probability mass is assigned to different partitions. For the simple example in which there are $K = 3$ subsets, the DPP model distributes probability to each partition:

RGF	Prob.
(1,1,1)	1/3
(1,1,2)	1/6
(1,2,1)	1/6
(1,2,2)	1/6
(1,2,3)	1/6

(For this example, we assume $\chi = 1$.) More generally, the probability of a particular partition of K elements containing k subsets is

$$p(\mathbf{z}, k \,|\, \chi, K) = \chi^k \frac{\prod_{i=1}^{k}(\eta_i - 1)!}{\prod_{i=1}^{K}(\chi + i - 1)},$$

where $\mathbf{z}$ is the allocation vector specifying how elements are assigned to subsets of the partition, k is the number of subsets in the partition (or the 'degree' of the partition), and η_i is the number of elements assigned to subset i. For the case in which $K = 3$ the values for these parameters are

RGF	$\mathbf{z}$	K	k	$\eta_\cdot$
(1,1,1)	(1,1,1)	3	1	$\eta_1 = 3$
(1,1,2)	(1,1,2)	3	2	$\eta_1 = 2$, $\eta_2 = 1$
(1,2,1)	(1,2,1)	3	2	$\eta_1 = 2$, $\eta_2 = 1$
(1,2,2)	(1,2,2)	3	2	$\eta_1 = 1$, $\eta_2 = 2$
(1,2,3)	(1,2,3)	3	3	$\eta_1 = 1$, $\eta_2 = 1$, $\eta_3 = 1$

The DPP model contains one additional component: a probability distribution (sometimes referred to as the 'base distribution', denoted G_0) for the parameter assigned to each subset. A complete description of the process, then, specifies the concentration parameter and the base distribution for the parameter. The DPP model is often used in Bayesian analysis for clustering. The elements of the partition are the objects to be clustered. Typically, the objects to be clustered correspond to data elements. The subsets of the partition denote the clustering; the subsets specify which elements are grouped together in the same cluster. Importantly, under the DPP model the number of clusters is a random variable. The DPP model does not force the biologist to pre-specify the number of clusters *a priori*.

The Chinese Restaurant Process is a useful metaphor that has been used to describe the DPP model (Aldous, 1985; Pitman, 2002). The metaphor works as follows: Consider a Chinese restaurant with a countably infinite number of tables. Patrons enter the restaurant one at a time and randomly choose a table at which to sit. The ith patron will choose to sit at table m at which η_m patrons are already seated with probability $\frac{\eta_m}{i-1+\chi}$; alternatively, this patron will choose some unoccupied table with probability $\frac{\chi}{i-1+\chi}$. The 'restaurant patrons' in the metaphor are replaced by data elements. Data elements that are 'seated at the same table' share a common parameter, θ, which is drawn from the probability distribution G_0 (i.e., $\theta \sim G_0$). Note that tables with large numbers of patrons will tend to attract new patrons.

We apply the DPP model to all of the phylogenetic model parameters, except for the topology and branch lengths of the tree (however, see Ané et al., 2007). Specifically, we first identify process partitions in the data and then allow the specific phylogenetic model parameters, such as the nucleotide frequencies, gamma shape parameter, or substitution rates, to cluster across partitions according to a DPP model prior.

Markov chain Monte Carlo

We adopt a Bayesian approach for estimating model parameters. In a Bayesian analysis, inferences are based on the posterior probability distribution of the model parameters, which is obtained using Bayes' theorem as

$$\Pr(\boldsymbol{\theta} \,|\, \mathbf{X}) = \frac{\Pr(\mathbf{X} \,|\, \boldsymbol{\theta}) \times \Pr(\boldsymbol{\theta})}{\Pr(\mathbf{X})}$$

where $\boldsymbol{\theta}$ is a vector containing the parameters of the phylogenetic model and $\mathbf{X}$ is the DNA sequence alignment. $\Pr(\boldsymbol{\theta} \,|\, \mathbf{X})$ is the posterior probability distribution of the parameters, $\Pr(\mathbf{X} \,|\, \boldsymbol{\theta})$ is the likelihood, and $\Pr(\boldsymbol{\theta})$ is the prior probability distribution on the parameters. The normalizing constant in Bayes' theorem, $\Pr(\mathbf{X})$, is calculated by marginalizing over the parameters. For a phylogenetic model, the normalizing constant (also called the 'marginal likelihood') involves a summation over all possible trees and integration over all possible combinations of model parameters. In practice, the marginal likelihood cannot be calculated analytically.

We use Markov chain Monte Carlo (MCMC; Metropolis et al., 1953; Hastings, 1970) to approximate the posterior probability distribution of the phylogenetic parameters. The general idea is to construct a Markov chain that has as its state space the possible values for the model parameters and a stationary probability distribution that is the target distribution of interest (i.e., the posterior probability distribution of the model parameters). Samples taken from the chain represent valid samples from thgoverning the shape ofe posterior distribution. MCMC works by iterating the

following steps many thousands or millions of times: (1) Call the current state of the Markov chain θ. If this is the first cycle of the Markov chain, then initialize θ to some value (perhaps by picking values from the prior probability distribution). (2) Propose a new value for the model parameter θ, called θ'. The proposal mechanism must be stochastic, with a probability of $f(\theta \to \theta')$ of choosing the proposed state θ' from the current state θ. The probability of the reverse proposal, from θ' to θ, which is not actually made in computer memory, is $f(\theta' \to \theta)$. (3) Calculate the probability of accepting θ' as the next state of the Markov chain

$$R = \min\left(1, \frac{\Pr(\mathbf{X} \mid \theta')}{\Pr(\mathbf{X} \mid \theta)} \times \frac{\Pr(\theta')}{\Pr(\theta)} \times \frac{f(\theta' \to \theta)}{f(\theta \to \theta')}\right)$$

(4) Generate a uniform(0,1) random variable called u. If $u < R$, then accept the proposed state and set $\theta = \theta'$. (5) Go to step 1.

We update the parameters of the phylogenetic model one-at-a-time. The proposal mechanisms for most of the phylogenetic model parameters, such as the tree topology, base frequencies, gamma shape parameter, and substitution rate parameters, have been discussed elsewhere and are now rather standard in the field. However, the proposal mechanism for the partitions with a Dirichlet process prior model merit further discussion. Neal (2000) discussed several possible MCMC updates for DPP models. We used 'Algorithm 8' from that paper to update the partition schemes. To return to the Chinese Restaurant metaphor, we imagine that we have instantiated a restaurant in computer memory, with tables representing subsets of the partition and patrons who are seated at tables representing the sites (or, as currently implemented, sets of sites that are determined by the biologist's initial division of the data into process partitions). The MCMC procedure works as follows. First, pick a patron and remove it from the table. If the patron was the only individual at that table, then remove the table from computer memory. Otherwise, decrease the count of the number of patrons at the table by one (e.g., decrease η_i). Calculate the likelihood when the patron is seated at the mth remaining table in computer memory (L_m). Also, calculate the likelihood when the patron is seated at each of the κ 'auxiliary' tables. These auxiliary tables are instantiated for the purpose of allowing the patron to be seated at a new (unoccupied) table. The parameter values for the auxiliary tables are drawn from the prior probability distribution for the parameter. The probability of seating the patron at the mth table in computer memory at which η_m patrons are already seated is $C \times \eta_m \times L_m$ and the probability that the patron is seated at the kth auxiliary table is $C \times (\chi/\kappa) \times L_k$, where C is a normalizing constant. After the patron is reseated at a previously existing or a new (auxiliary) table, the unoccupied auxiliary tables are deleted from computer memory. One MCMC cycle involves a scan of all the patrons, with the above update mechanism applied to each.

The methods described in this paper—which are intended to automate the phylogenetic analysis of partitioned data sets—have been implemented in the freely available application, *AutoParts*. This command-line program enables a fully hierarchical Bayesian phylogenetic analysis of partitioned data, where the input is an alignment of nucleotide sequences and two or more data subsets that are specified to capture biologically plausible sources of process heterogeneity in the sequence data (such as different gene/genomic regions, codon positions, etc.). The program uses MCMC to approximate the marginal posterior probability distribution of all parameters (tree topology, branch lengths, and substitution model parameters) while simultaneously integrating over all possible partition schemes under the DPP model.

Summarizing Partitions

In a Bayesian MCMC analysis, posterior probabilities of model parameters are based upon the MCMC samples. For example, the posterior probability of a particular phylogenetic tree is approximated as the fraction of the time that the Markov chain visited that particular tree. One of the challenges in an MCMC analysis of phylogeny is how to best summarize the information on the joint posterior probability distribution of the model parameters contained in the MCMC samples. For some simple model parameters, such as α—the parameter that governs the shape of the gamma distribution used to describe among-site substitution-rate variation—one can summarize the results of an MCMC analysis by constructing a frequency histogram of the sampled α values; this frequency histogram represents an approximation of the marginal posterior probability distribution for that parameter. The marginal distribution can be further summarized in a number of ways, by presenting, for example, the mean of the distribution, or an interval that contains 95% of the posterior probability. The tree parameter presents some difficulties in a Bayesian MCMC analysis, as it is not a standard model parameter (see Yang et al., 1995). Consequently, there is no natural way to represent the results of a Bayesian MCMC analysis, except perhaps to simply report the posterior probabilities of individual trees. That said, there are several useful methods that are commonly used to summarize the trees sampled using MCMC. For example, the majority-rule consensus tree is often used to summarize the results of an MCMC analysis (in the same way that bootstrap samples are summarized; Felsenstein, 1985).

Our MCMC procedure also samples the partitions of the model parameters. Partitions, like trees, are not a standard model parameter and it is not clear how best to summarize the information contained in an MCMC sample of partitions. One possibility is to simply report the posterior probabilities of individual partitions. This method, however, has problems when the posterior probability is spread across many partitions and the Markov chain may then explore many thousands or

millions of partitions. In this study, we summarize the MCMC samples of partitions using the idea of a 'mean partition'. The mean partition, denoted $\bar{\sigma}$, is the partition that minimizes the sum of the squared distances to all of the sampled partitions. We use a distance on partitions first described by Gusfield (2002). Gusfield's partition distance between two partitions can be interpreted as the minimum number of elements that must be removed from both partitions to make the induced partitions the same. For example, the partitions $\sigma_1 = (1, 1, 1, 1, 2, 2)$ and $\sigma_2 = (1, 1, 1, 2, 2, 2)$ can be made the same by removing the fourth element resulting in the induced partition $(1, 1, 1, 2, 2)$; in this example, the partition distance is $d_p = 1$.

Empirical Examples

Data sets.—We examined five data sets in this study that span a range of size (in terms of the number of species, S, and sequence length, L) and number of data subsets, K (Table 1). First, we examined an alignment of $S = 32$ gall wasp sequences from Nylander et al. (2004). The gall wasp alignment ($L = 3080$) comprised $K = 11$ data subsets: each of the three protein-coding genes — including two nuclear genes (elongation factor 1α [EF1α] and long-wavelength opsin [LWRh]) and one mitochondrial gene (cytochrome oxidase c subunit I [COI]) — were partitioned by codon position, and the the 28S ribosomal sequence was partitioned by stem and loop regions. Second, we examined an alignment of $S = 60$ skink sequences from Brandley et al. (2005). The skink alignment ($L = 2678$) comprised $K = 11$ data subsets: mitochondrial sequences included the NADH dehydrogenase subunit 1 (ND1) gene, which was partitioned by codon position, and three individual tRNA partitions for tRNA^{GLU}, tRNA^{ILE}, and tRNA^{GLN}; ribosomal sequences included separate partitions for 12S rRNA, 16S rRNA, and tRNA^{MET}. Third, we examined an alignment of $S = 164$ hummingbird and relevant outgroup sequences from McGuire et al. (2007). The hummingbird alignment ($L = 4122$) comprised $K = 9$ data subsets: mitochondrial sequences included the entire NADH dehydrogenase subunit 2 (ND2) and about half of NADH dehydrogenase subunit 4 (ND4), which were each partitioned by codon position, with a separate partition for the tRNA on the flanking regions; nuclear sequences included separate partitions for intron 5 of the adenylate kinase gene (AK1) and intron 7 of the beta brinogen (BFib) gene. Fourth, we examined an alignment of $S = 86$ cichlid sequences from **?**. The cichlid alignment ($L = 1710$) comprised $K = 7$ data subsets: the two protein-coding genes — including the mitochondrial gene cytochrome oxidase subunit I (COI) and the nuclear gene *Tmo-4C4* (*Tmo*) — were individually partitioned by codon position, and one ribosomal partition was included for the small subunit RNA gene, 16S. Finally, we examined an alignment of $S = 54$ flowering dogwood sequences from Xiang et al. (2006). The dogwood alignment ($L = 1991$) comprised $K = 4$ data subsets: chloroplast sequences from the

protein-coding region of the maturase K (matK) gene were partitioned by codon position, and one nuclear partition was included for sites from the internal transcribed spacer (ITS) gene region.

The alignments of the gall wasps, skinks, and hummingbirds are of particular interest in the current context, as these data sets were previously used by Brandley et al. (2005), Nylander et al. (2004), and McGuire et al. (2007), respectively, to evaluate and choose among a limited set of partitioning schemes using Bayes factors.

Exploring the impact of the concentration parameter on phylogenetic inferences.—We performed two primary series of analyses to assess the sensitivity of phylogenetic inference to specification of the concentration parameter of the Dirichlet process prior. In the first series of analyses, the concentration parameter, χ, was fixed to a set of values that centered the prior mean for the number of process partitions, $E(K)$, on a range of values that were appropriate to each data set. For example, in our analyses of an alignment with $K = 11$ data subsets, we fixed the concentration parameter to a range of values such that the prior mean for the number of partitions spanned the entire range of possible values — specifically such that $E(K) = 1.2$, 2, 4, 6, 8, 10, 10.8 — whereas our analyses of an alignment with $K = 9$ data subsets specified a set of values for χ such that $E(K) = 1.2$, 2, 4, 6, 8, 8.8, and so on. The second series of analyses treated the concentration parameter as a random variable that was governed by a gamma hyperprior. In these analyses, the value of the gamma hyperprior was specified such that the prior mean for the number of partitions spanned the same range of values as those considered in the analyses that fixed the prior mean on the number of partitions.

MCMC analyses.—We estimated the posterior probability distribution of trees for each of the five study groups under the Dirichlet process prior model using the Markov chain Monte Carlo (MCMC) algorithms implemented in AutoParts. Each MCMC analysis was initiated from random starting trees and the chains were run for $2\text{x}10^7$ cycles and thinned by sampling every 1000 cycles. In total, 224 separate MCMC analyses were performed: Four independent MCMC chains were run for each of the five alignments, and for each alignment we performed analyses under the appropriate range of values for the concentration parameter, and for each value of the concentration parameter we performed replicate analyses where the concentration parameter was treated as fixed or as a random variable (as described above).

Assessing MCMC performance.—Convergence of each chain to the target distribution was inferred by plotting time series for (and graphing the marginal probabilities of) sampled parameter values using Tracer v.1.4.1 (Rambaut and Drummond, 2007). We also assessed convergence by comparing inferences from the four replicate MCMC analyses for each alignment-by-concentration parameter combination to ensure that the marginal posterior probabiltiy estimates from the inde-

pendent chains were effectively identical. Finally, we used the '*comparetree*' function implemented in MrBayes v.3.1.2 (Ronquist and Huelsenbeck, 2003) to assess the correspondence of clade probabilities estimated by the set of independent chains. The adequacy of sampling intensity was assessed using the estimated sample size (ESS) diagnostic implemented in Tracer v.1.4.1 (Rambaut and Drummond, 2007).

Validation of MCMC algorithms.—To provide a benchmark for for inference under the DPP model, and to validate the MCMC algorithms, we estimated the posterior probability distribution of trees for each of the five study groups under a uniform GTR+Γ substitution model using both the MCMC algorithms implemented in AutoParts and the proven Metropolis-coupled Markov chain Monte Carlo (MCMCMC or MC^3) algorithms implemented in MrBayes v.3.1.2 (Ronquist and Huelsenbeck, 2003). In the analyses performed using AutoParts, we enforced a uniform substitution model by setting the proposal probabilities for new partitions to zero, such that the MCMC did not consider partition schemes of $K > 1$. For the unpartitioned analyses using MrBayes, we enforced a uniform substitution model such that each model parameter was estimated from all of the sites within a given alignment.

In total, we performed 40 of these baseline analyses: Four independent MCMC chains were run for each of the five alignments, and for each alignment we performed analyses under a uniform GTR+Γ substitution model using both AutoParts and MrBayes. Details of the MC^3 analyses are as described above; however, each of the replicate MC^3 runs performed using MrBayes comprised four incrementally heated chains (where the parameter governing the 'temperature' of the heated chains ranged from $0.2 - 0.1$) that were initiated from random starting trees. Performance of the MC^3 analyses was assessed as described previously; however, convergence of the Metropolis-coupled chains additionally entailed running 'paired'-MC^3 analyses (i.e., two independent, parallel runs of four incrementally heated chains each) and monitoring the average standard deviation of split frequencies sampled by the paired cold chains.

Results and Discussion

Performance of MCMC algorithms.—In general, the MCMC algorithms used to estimate the joint posterior probability density of trees and other parameters under the Dirichlet process prior model appear to have performed well for each of the five alignments, whether we enforced uniform substitution models (for the purposes of algorithm validation), or when jointly estimating the process partitions with the concentration parameter either fixed to a specified value or governed by a gamma hyperprior.

Under baseline inference scenarios — where a uniform GTR+Γ substitution model was enforced — independent estimates obtained using the MCMC algorithms implemented in AutoParts and the MC^3 algorithms implemented in MrBayes were virtually identical (see *Supplementary Material*). The convergence of estimated marginal posterior probability densities of all model parameters — including the overall marginal log likelihoods and clade probabilities — suggests that the MCMC algorithms used to implement the Dirichlet process prior model performed appropriately under these simplified inference conditions.

Under more complex inference scenarios — where inferences were integrated over all possible partition schemes under the Dirichlet process prior model — the MCMC algorithms again performed well. Time series plots of all model parameters sampled by each of the four independent MCMC analyses under each of the 56 unique concentration-parameter treatments that we explored for the five study alignments appeared to stabilize within five million cycles. Accordingly, parameter estimates were based on samples drawn during the final 15 million cycles of each chain.

Moreover, independent chains inferred virtually identical marginal posterior probability densities for all parameters — including those of the GTR substitution model, the α-shape parameter used to model among-site substitution rate variation, as well as the clade posterior probabilities, and the mean process partition scheme for each of the five empirical alignments — providing additional evidence that the MCMC algorithms used to implement the Dirichlet process prior model successfully converged to the stationary distributions for these data sets.

The MCMC algorithm appears to have mixed well over the joint posterior densities for the five study data sets: acceptance rates of proposal mechanisms for all parameters were within the target window $(20 - 60\%)$, and the marginal posterior probability densities for all parameters were typically tight and focused. Samples drawn during the stationary phase of independent chains were pooled, which provided adequate sampling intensity for all estimated parameters for all five data sets (i.e., all ESS values $>> 10^3$).

Sensitivity of inferences to the concentration parameter.—The two primary series of analyses explored alternate approaches for specifying the concentration parameter of the Dirichlet process prior model. Whether we used a fixed value of χ or a placed a gamma hyperprior on χ to center the prior mean on the number of process partitions at a particular value [e.g., $E(K) = 4$], inferences were virtually indistinguishable for all five of the study data sets (see *Supplementary Material*). This result is particularly encouraging: for most analyses, there will typically be little empirical basis for committing to a particular fixed value of χ, which might therefore require several independent analyses under a range of values for $E(K)$. However, our findings suggest that a vague hyperprior can be placed on χ so that it can be estimated from the data (or equivalently, inferences can

be integrated over a suitably broad prior distribution for χ) in the course of a single MCMC analysis. Because inferences under the two approaches were virtually identical, the remainder of the discussion will focus exclusively on results obtained under fixed values of χ.

Some specific results.—Before describing more general aspects of our findings, we will first focus on results for a particular (albeit fairly typical) case. Specifically, estimates of the α parameter governing the shape of gamma distribution used to model variation in substitution rates across sites in the hummingbird data set (Figure 1). The marginal posterior probability density of the α-shape parameter is independently estimated for each of the $K = 9$ data partitions (Figure 1A), where the density for each data partition is an amalgam of 24 overlain estimates—that is, the four replicate MCMC analyses performed under each of the six values for the prior mean on the number of process partitions [$E(K) = 1.2$, 2, 4, 6, 8, 8.8] that we explored for the hummingbird data set.

Four aspects of these results warrant comment. First, for each data partition, the marginal posterior probability densities estimated by the four independent MCMC analyses *under each particular value* of $E(K)$ are virtually identical, indicating that the chains successfully converged to the stationary distributions.

Second, the marginal posterior probability densities of the α-shape parameter estimated from each data partition *over the entire range of values* for $E(K)$ are equally indistinguishable, suggesting that these parameter estimates are robust to (mis)specification of the concentration parameter of the Dirichlet process prior model.

Third, the marginal posterior probability densities are typically focused and unimodal in form, suggesting both that the MCMC algorithms mixed well over these parameters, and that the pre-specified data partitions successfully captured process heterogeneity in the sequence alignment. A notable exception is apparent for the second-position sites of the ND4 gene (Figure 1B): the marginal posterior probability density for this data partition is clearly bimodal, indicating residual heterogeneity in the degree of among-site substitution rate variation within this data partition. Because the data partitions are pre-specified by the investigator, the latent structure within the second-position sites of the ND4 gene is not accessible to the Dirichlet process prior model.

Finally, the set of marginal posterior probability densities for the nine data partitions appear to cluster in three distinct groups: the first cluster comprises data partitions for tRNA and the first- and second-position sites of the ND2 and ND4 genes (with relatively high levels of among-site substitution rate variation), the second cluster comprises partitions for the AK1 and BFib introns (with intermediate levels of among-site substitution rate variation), and the final cluster includes the two data partitions for the third-position sites of the ND2 and ND4 genes (with relatively low levels of among-site substitution rate variation). Indeed, this allocation of data partitions among process

partitions is identical to the mean partition scheme identified by the Dirichlet process prior model. For convenience, we summarize these results using the graphical convention depicted in Figure 1C. This plot reveals the posterior estimate for the number of process partitions, $E(K \mid \mathbf{X}) = 3$, and the inferred mean partition scheme for this parameter, $E(\bar{\sigma} \mid \mathbf{X})$, which summarizes the assignment of nine pre-specified data partitions among the three process partitions for the α-shape parameter. The number in the upper left of the panel indicates the number of partition schemes in the 95% credible set (i.e., the 95% HPD) for this parameter.

General aspects of the results.—Inspection of the number of process partitions comprising the mean partition schemes provides insight into the level and nature of process heterogeneity within the sequence alignments. For example, the α-shape parameter was inferred to have the largest number of process partitions across all five data sets (Figures 2–6), which is consistent with previous empirical (e.g., Nylander et al., 2004) and theoretical (e.g., Yang, 1996; Huelsenbeck and Rannala, 2004) results that have identified among-site substitution rate variation as a critically important modeling component. By contrast, the rank order in the number of process partitions — and the corresponding level of process heterogeneity — inferred for the other parameters (base frequencies, substitution rate, and tree length) varied among the five sequence alignments.

Several generalities regarding the impact of the concentration parameter merit comment. First, even when a substantial proportion ($> 95\%$) of the prior probability was placed on a homogeneous substitution model [i.e., where $E(K) \approx 1$; the left-most columns in Figures 2–6], the mean partition schemes for all data sets were nevertheless inferred to have $E(K \mid \mathbf{X}) > 1$, suggesting that these alignments are very difficult to explain using a uniform model.

Second, inferences of the mean partition schemes for the five data sets differed somewhat in their sensitivity to the prior on the number of partitions, $E(K)$. Specifically, the mean partition scheme inferred for skinks ($K = 11$), hummingbirds ($K = 9$), and cichlids ($K = 7$) were largely stable over the entire range of priors placed on the number of process partitions (Figures 3, 4, and 5, respectively), whereas those for gall wasps ($K = 11$) and dogwoods ($K = 4$) exhibited greater sensitivity to the prior on $E(K)$, such that the number of process partitions increased as more prior probability was placed on more parameter-rich partition schemes (Figures 2 and 6, respectively). Even in these cases, however, sensitivity of the mean partition schemes to the priors on $E(K)$ for the dogwood and gall wasp data sets was restricted to one or two parameters, respectively, while the inferred process partitions for the other parameters remained stable over the entire range of $E(K)$.

Third, the pattern and level of uncertainty in estimates of the mean partition schemes — reflected in the size of the 95% HPD of $E(\bar{\sigma} \mid \mathbf{X})$ — varied somewhat with the prior mean on the

number of process partitions across individual parameters and among the five data sets. In the predominant pattern (exhibited by 40% of the parameters), the level of uncertainty in the mean partition scheme was uniform over the range of priors on $E(K)$; e.g., the size of the 95% HPD for the mean partition scheme of the tree-length parameter in the gall wasp alignment included 1–2 partition schemes over $E(K) = (1, 2, 4, 6, 8, 10, 11)$ (Figure 2). In the second pattern (exhibited by 25% of the parameters), uncertainty in the mean partition scheme was greatest at intermediate values of $E(K)$; e.g., the 95% HPD for the α-shape parameter in the gall wasp alignment peaks at $E(K) = 8$ (Figure 2). This pattern may partly reflect the increased number of possible partition schemes at intermediate values of $E(K)$. For example, the 11 data partitions in the gall wasp alignment can be combined in 1023 unique partition schemes with 2 process partitions, 179,487 unique partition schemes with 6 process partitions, and 11,880 unique partition schemes with 8 process partitions. In the final pattern (exhibited by the remaining 35% of the parameters), uncertainty in the mean partition scheme scaled linearly with the prior mean on the number of partition schemes; e.g., the 95% HPD of the substitution-rate parameters in the gall wasp alignment peaks at $E(K) \approx 11$ (Figure 2). As $E(K)$ increases, more prior mass is placed on partition schemes that include a greater number of process partitions, increasing the probability that data partitions will be distributed over a greater number of process partitions, such that fewer data are available to estimate each parameter. Accordingly, the marginal posterior probability densities for each parameter are apt to become more diffuse (especially for smaller data partitions and/or those with limited variation), which impedes the ability of the Dirichlet process prior to unambiguously assign these marginal posterior probability densities to process partitions, with a corresponding increase in the uncertainty associated with the inferred mean partition schemes.

Finally, posterior estimates of the mean partition schemes were associated with considerable uncertainty for these data: on average, the 95% credible set for the partition schemes included 340 process partitions (i.e., calculated as the mean size of the summed credible sets for the four parameters — the columns in Figures 2–6 — averaged over all values of $E(K)$ over all five alignments). The generally large number of plausible partition schemes emphasizes the appeal of integrating inferences of phylogeny (and other model parameters) over this important source of uncertainty (rather than conditioning inferences on any particular partition scheme).

Comparison to Alternative Approaches for Accommodating Process Heterogeneity

Partition-scheme selection based on Bayes factors.—The Dirichlet process prior approach for accommodating process heterogeneity is similar in spirit to the conventional partition-scheme selection that relies on Bayes factors. Three of the five alignments evaluated in the present study (for

gall wasps, skinks, and hummingbirds) have previously been subjected to extensive mixed-model selection using Bayes factors, which affords the opportunity to compare these two approaches. Below we consider three aspects of this comparison.

First, in every case, estimation under the Dirichlet process prior model discovered novel mean partition schemes that were not evaluated in the previous studies that relied upon Bayes factors. We emphasize that the selected studies are notable for the relatively large number of partition schemes that they considered. Nevertheless, the need to perform a full MCMC analysis under each candidate partition scheme (in order to estimate the corresponding marginal likelihood) entails substantial computational overhead, such that even these exceptionally thorough studies necessarily considered only a small fraction of the possible mixed-model space. Specifically, the gall wasp study evaluated 0.0013% (9 of 678,570) of the partition schemes possible for an alignment with $K = 11$ data partitions; the skink study evaluated 0.043% (9 of 21,147) of the partition schemes possible for an alignment with $K = 9$ data partitions; and the hummingbird study evaluated 0.043% (9 of 21,147) possible partition schemes for an alignment with $K = 9$ data partitions. The ability of the Dirichlet process prior model to integrate inference of phylogeny over the entire range of partition schemes in the course of a single MCMC analysis greatly increases the probability of estimating under the partition scheme that best captures process heterogeneity within the data, thereby minimizing biases associated with model misspecification.

Second, the dimensionality of the partition schemes inferred by the two approaches differed substantially. As in all other applications, mixed-model selection using Bayes factors identified the most parameter-rich candidate model for the gall wasp, skink, and hummingbird alignments. By contrast, the mean partition schemes inferred under the Dirichlet process prior approach were considerably more parsimonious. Specifically, the mean/selected partition schemes identified under the Dirichlet process prior/Bayes factor approaches included 32/99 free parameters for the gall wasp alignment, 27/98 free parameters for the skink alignment, and 29/108 free parameters for the hummingbird alignment, respectively. There are three plausible explanations for the discrepancy in mixed-model selection under these two approaches: (1) Bayes factor selection based on the comparison of marginal likelihoods estimated using the harmonic mean (Newton and Raftery, 1994) may be biased toward the inclusion of superfluous parameters and selection of over-partitioned mixed models (Pagel and Meade, 2004; Sullivan and Joyce, 2005; Brown and Lemmon, 2007; McGuire et al., 2007); (2) the Dirichlet process prior model has a tendency for clusters to self propagate, such that data partitions might be attracted to a smaller number of larger process partitions, which may lead to the selection of under-partitioned composite models; and (3) the candidate partition schemes evaluated by means of Bayes factors are separated by substantial gaps in dimensionality

(owing to the limited sampling from the pool of all possible partition schemes), which may lead to the selection of over-partitioned composite models.

We explored the latter (and, we believe, most plausible) explanation via a series of analyses of the gall wasp data set. We approximated the joint posterior probability density of phylogeny and other parameters by means of MC^3 algorithms implemented in MrBayes under the mean partition scheme identified by the Dirichlet process prior analyses, and also under several partition schemes that were adjacent in parameter space to this partition scheme (i.e., some with slightly more and some with slightly fewer parameters than the mean partition scheme). The details by which these analyses were performed and post-processed are identical to those described previously in the sections *MCMC analyses* and *Assessing MCMC performance*, respectively. For each of these analyses we approximated the marginal log likelihood using the harmonic-mean estimator, and then used Bayes factor to select among the set of candidate partition schemes. The set of marginal likelihoods estimated from these experiments with the gall wasp data set were compared using Bayes factors, which selected the same (mean) partition scheme identified by the Dirichlet process prior analyses (results not shown).

The final point of comparison between the Dirichlet process prior and Bayes factor approaches pertains to difference in error variance. A corollary of the substantial reduction in model complexity under the Dirichlet process prior model — achieved by reduction of superfluous parameters — is a corresponding reduction in the error variance and associated uncertainty of parameter estimates. For example, the level of uncertainty associated with the topology parameter — reflected in the size of the 95% credible set of trees — estimated under the Dirichlet process prior/Bayes factor approaches included 927/1513 trees for the gall wasp alignment, $2,488/3106$ trees for the skink alignment, and $3,961/4672$ trees for the hummingbird alignment, respectively. The reduction in estimation error (particularly the topology and branch-length parameters) under the Dirichlet process prior model should decrease uncertainty not only in estimates of phylogeny, but also in phylogeny-based inferences focused on the study of character evolution, rates of diversification, biogeographic history, etc.

Extensions.—The Dirichlet process prior model we describe provides a versatile approach for accommodating process heterogeneity across sites of nucleotide sequence alignments, and so should improve estimates of phylogeny from those data. Importantly, this framework can readily be extended in various ways by relaxing some of the current assumptions. For example, we have assumed that the process partitions evolve under the generalized time-reversible (GTR) substitution model. In principle, it is simple to integrate over partition schemes while simultaneously averaging over a range of less general substitution models (e.g., HKY, F81) for each process partition (cf.

Huelsenbeck et al., 2004). Although such an extension is feasible, previous simulations studies (e.g. Huelsenbeck and Rannala, 2004) lead us to believe that it is preferable to estimate under the most complex model available (GTR), provided that acceptable MCMC performance can be achieved. Accordingly, the general dictum of Bayesian inference — vigilant diagnosis of MCMC performance — is particularly crucial for estimation of phylogeny (and other model parameters) under high-dimensional mixture models, such as the Dirichlet process prior approach we have described.

As currently implemented, the Dirichlet process prior model requires that the biologist define the set of data partitions prior to the analysis; once specified, these data partitions become an inexorable assumption of the analysis. The success of the method will therefore partly depend on our prior knowledge regarding likely patterns of process heterogeneity within sequence alignments. In a few cases that we examined here (e.g., among-site substitution rate variation in second-position sites of the ND4 gene in the hummingbird alignment; Figure 1B), the pre-specified data partitions appeared to harbor residual process heterogeneity that could not be accommodated by the Dirichlet process prior model. The Dirichlet process prior approach can be extended to accommodate this scenario simply by relaxing the very strong prior on the set of pre-specified data partitions ($Pr = 1$) by means of proposal mechanisms that enable subsets of sites within each data partition to be split (and merged) so as to capture any latent process heterogeneity within the predetermined data partitions.

Our approach accommodates variation in the evolutionary process across the sites of an alignment, but nevertheless assumes that these sequences share a common phylogenetic history. For some data sets—particularly multiple nuclear loci sampled from closely related and/or recently diverged species—the assumption of a single phylogenetic history is likely to be violated by processes such lineage sorting and deep coalescence. Fortunately, it is straightforward to extend the Dirichlet process prior model developed here to accommodate this aspect of process heterogeneity by simply treating the tree topology like other parameters in the model (see, e.g., Ané et al., 2007). That is, rather than assuming that all data partitions share a common tree-topology parameter, we could treat the number of topologies and the assignment of data partitions to tree topologies as random variables. We are optimistic that the Dirichlet process prior approach as currently described will provide biologists with an efficient and robust method for accommodating process heterogeneity in estimation of phylogeny, and that future extensions of this framework will prove fruitful for addressing related problems in phylogenetic biology.

Acknowledgments

This research was supported by grants from the NSF (DEB-0445453) and NIH (GM-069801)

awarded to J.P.H., FR was supported by VR grant 2007-.

References

Aldous, D. 1985. Exchangeability ad related topics. Pages 1–198 *in* École d'Été de Probabilités de Saint-Flour XIII-1983 Springer, Berlin.

Ané, C., B. Larget, D. A. Baum, S. D. Smith, and A. Rokas. 2007. Bayesian estimation of concordance among gene trees. Molecular Biology and Evolution 24:412–426.

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

Bell, E. T. 1934. Exponential numbers. American Mathematics Monthly 41:411–419.

Brandley, M. C., A. Schmitz, and T. W. Reeder. 2005. Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. Systematic Biology 54:373–390.

Brown, J. M. and A. R. Lemmon. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. Systematic Biology 56:643–655.

Bull, J. J., J. P. Huelsenbeck, C. W. Cunningham, D. L. Swofford, and P. J. Waddell. 1993. Partitioning and combining data in phylogenetic analysis. Systematic Biology 42:384–397.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution 17:368–376.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791.

Ferguson, T. S. 1973. A bayesian analysis of some nonparametric problems. Annals of Statistics 1:209–230.

Gusfield, D. 2002. Partition-distance: a problem and class of perfect graphs arising in clustering. Information Processing Letters 82:159–164.

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

Huelsenbeck, J. P., S. Jain, S. W. D. Frost, and S. L. K. Pond. 2006. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. Proceedings of the National Academy of Science, U.S.A. 103:6263–6268.

Huelsenbeck, J. P., B. Larget, and M. E. Alfaro. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Molecular Biology and Evolution 21:1123–1133.

Huelsenbeck, J. P. and B. Rannala. 2004. Frequentist properties of bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Systematic Biology 53:904–913.

Huelsenbeck, J. P. and M. Suchard. 2007. A nonparametric method for accommodating and testing across-site rate variation. Systematic Biology 56:975–987.

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

Lartillot, N. and H. Philippe. 2006. Computing Bayes factors using theromodynamic integration. Systematic Biology 55:195–207.

Marshall, D. C., C. Simon, and T. R. Buckley. 2006. Accurate branch length estimation in partitioned Bayesian analyses requires accommodation of among-partition rate variation and attention to branch length priors. Systematic Biology 55:993–1003.

McGuire, J. A., C. C. Witt, D. L. Altshuler, and J. V. Remsen. 2007. Phylogenetic systematics and biogeography of hummingbirds: Bayesian and maximum likelihood analyses of partitioned data and selection of an appropriate partitioning strategy. Systematic Biology 56:837–856.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. Journal of Chemical Physics 21:1087–1092.

Neal, R. M. 2000. Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics 9:249–265.

Newton, M. A. and A. E. Raftery. 1994. Approximate Bayesian inference by the weighted likelihood bootstrap. Journal of the Royal Statistical Society, B 56:3–48.

Nylander, J. A. A., F. Ronquist, J. P. Huelsenbeck, and J. L. N. Aldrey. 2004. Bayesian phylogenetic analysis of combined data. Systematic Biology 53:47–67.

Pagel, M. and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Systematic Biology 53:571–581.

Pitman, J. 2002. Combinatorial stochastic processes. Technical Report 621, University of California at Berkeley, lecture notes for St. Flour Summer School .

Poux, C., O. Madsen, J. Glos, W. W. de Jong, and M. Vences. 2008. Molecular phylogeny and divergence times of Malagasy tenrecs: Influence of data partitioning and taxon sampling on dating analyses. BMC Evolutionary Biology 8:102.

Rambaut, A. and A. J. Drummond. 2007. Tracer v1.4. http://beast.bio.ed.ac.uk/Tracer.

Ronquist, F. and J. P. Huelsenbeck. 2003. Mrbayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Stanton, D. and D. White. 1986. Constructive Combinatorics. Springer-Verlag, New York.

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

Sullivan, J. and P. Joyce. 2005. Model selection in phylogenetics. Annual Review of Ecology, Evolution, and Systematics 36:445–466.

Tavaré, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lectures in Mathematics in the Life Sciences 17:57–86.

Vendetti, C., A. Meade, and M. Pagel. 2008. Phylogenetic mixture models can reduce node-density artifacts. Systematic Biology 58:286–293.

Xiang, Q. Y., D. T. Thomas, W. Zhang, S. R. Manchester, and Z. Murrell. 2006. Species level phylogeny of the genus *cornus* (Cornaceae) based on molecular and morphological evidence — implications for taxonomy and Tertiary intercontinental migration. Taxon 55:9–30.

Yang, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Molecular Biology and Evolution 10:1396–1401.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. Journal of Molecular Evolution 39:306–314.

Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. Journal of Molecular Evolution 42:587–596.

Yang, Z., N. Goldman, and A. E. Friday. 1995. Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. Systematic Biology 44:384–399.

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

Figure Legends

Figure 1. Inferred patterns of among-site substitution rate variation within the hummingbird sequence alignment. (A) Marginal posterior probability densities for the α-shape parameter of the discrete gamma model have been estimated for each of the 9 predefined data partitions. (B) In contrast to the unimodal marginal densities for other data partitions, the posterior density for the second-position sites of the ND4 gene is distinctly bimodal, suggesting process heterogeneity within this data partition. (C) Analyses under the Dirichlet process prior model provide estimates both for the number of process partitions, $E(K \mid \mathbf{X}) = 3$ (represented as rows in the panel), as well as the mean partition scheme, $E(\bar{\sigma} \mid \mathbf{X})$ (depicted by the assignment of the data-partition symbols to their respective rows).

Figure 2. Inferred patterns of process heterogeneity within the gall-wasp sequence alignment inferred under the Dirichlet process prior model. Each row corresponds to a parameter of the model, and each column corresponds to a specified value for the prior mean on number of process partitions in the sequence alignment, $E(K)$. Accordingly, each panel summarizes estimates for a particular parameter under a specified value of $E(K)$. The number of occupied rows in a panel corresponds to the estimated number of process partitions, $E(K \mid \mathbf{X})$, and the distribution of data partitions (depicted as symbols) among the rows of a panel specifies the inferred mean partition scheme, $E(\bar{\sigma} \mid \mathbf{X})$. The number in the upper left of each panel specifies the size of the 95% credible set of partition schemes.

Figure 3. Inferred patterns of process heterogeneity within the skink sequence alignment inferred under the Dirichlet process prior model. Graphical conventions as detailed in Figure 2.

Figure 4. Inferred patterns of process heterogeneity within the hummingbird sequence alignment inferred under the Dirichlet process prior model. Graphical conventions as detailed in Figure 2.

Figure 5. Inferred patterns of process heterogeneity within the cichlid sequence alignment inferred under the Dirichlet process prior model. Graphical conventions as detailed in Figure 2.

Figure 6. Inferred patterns of process heterogeneity within the dogwood sequence alignment inferred under the Dirichlet process prior model. Graphical conventions as detailed in Figure 2.

Figure 7. Partition schemes estimated under the Dirichlet process prior model (left column) and those based on pairwise Bayes factor comparisons (right column) for analyses of the gall wasp (top row), skink (middle row), and hummingbird (bottom row) sequence alignments. The arithmetic

mean of the marginal log likelihood is reported in the upper left of each panel; the size of the 95% credible set of trees is reported in the upper right.

FIGURE 1

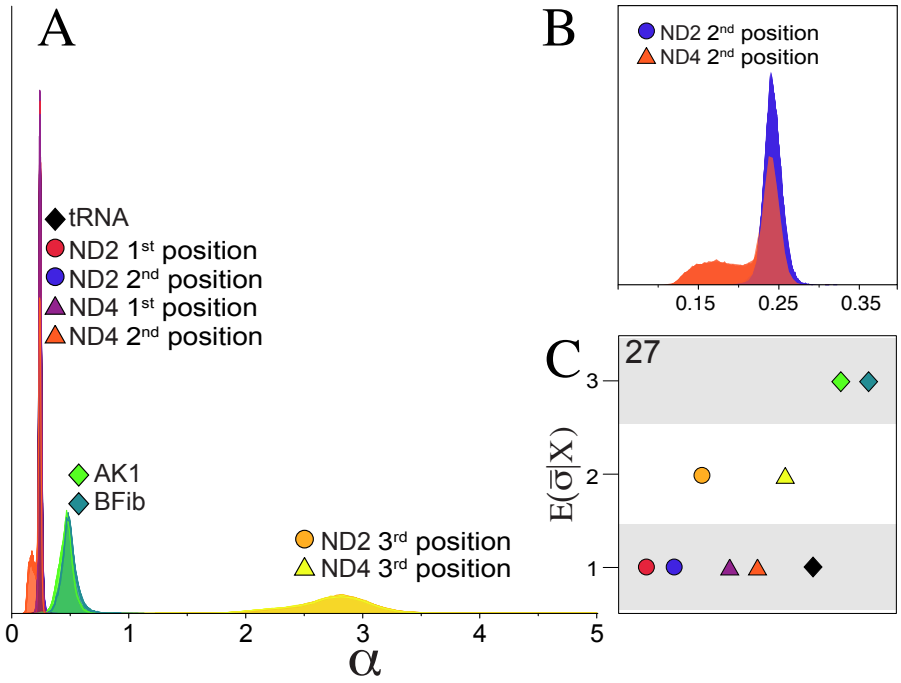

A
tRNA
ND2 1st position
ND2 2nd position
ND4 1st position
ND4 2nd position
AK1
BFib
ND2 3rd position
ND4 3rd position
0
1
2
3
4
5
α
B
ND2 2nd position
ND4 2nd position
0.15
0.25
0.35
C
27
E(σ̄|X)
3
2
1

FIGURE 2

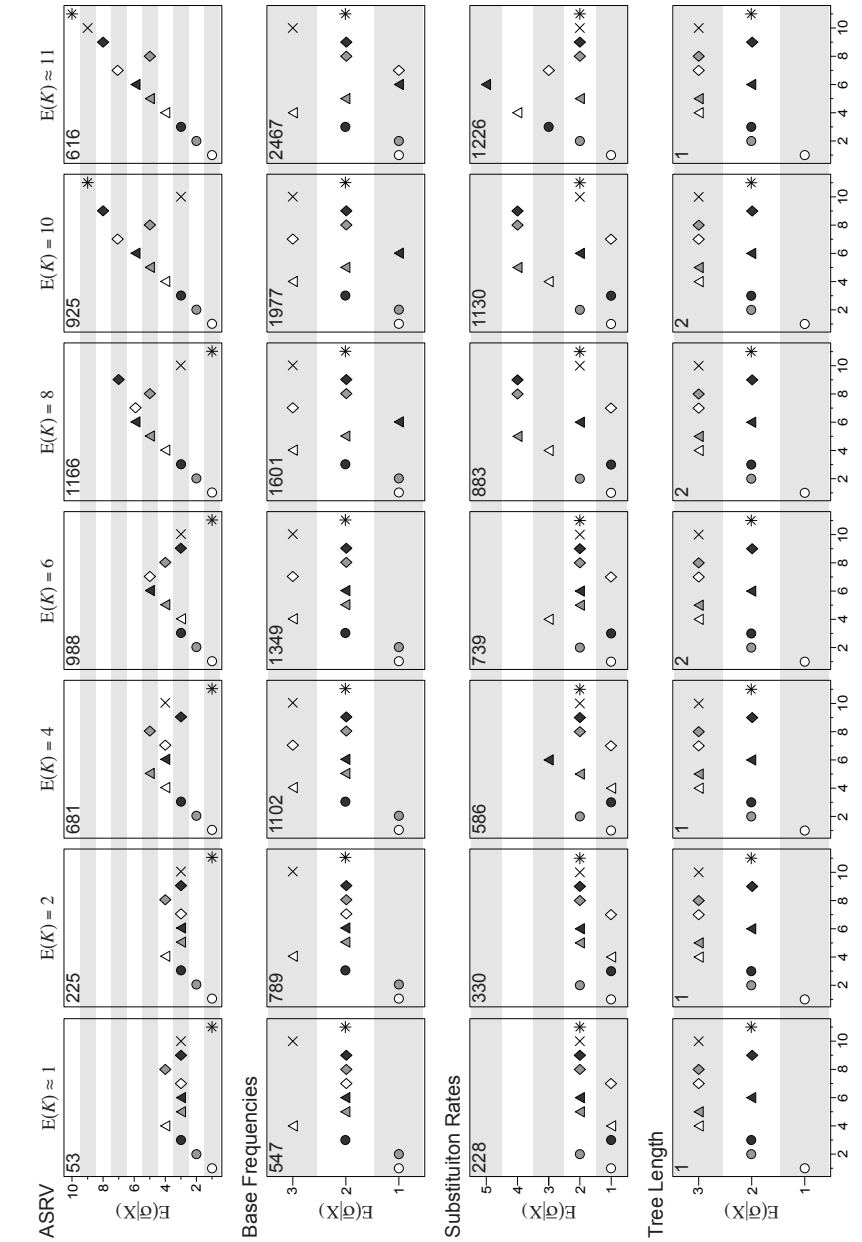

ASRV
Base Frequencies
Substituiton Rates
Tree Length
E(K) ≈ 1
E(K) = 2
E(K) = 4
E(K) = 6
E(K) = 8
E(K) = 10
E(K) ≈ 11
E(σ̄|X)
53
225
681
988
1166
925
616
547
789
1102
1349
1601
1977
2467
228
330
586
739
883
1130
1226
1
2
COI
EF1α
LWRh
28S
1st position
2nd position
3rd position
stem region
loop region

FIGURE 3

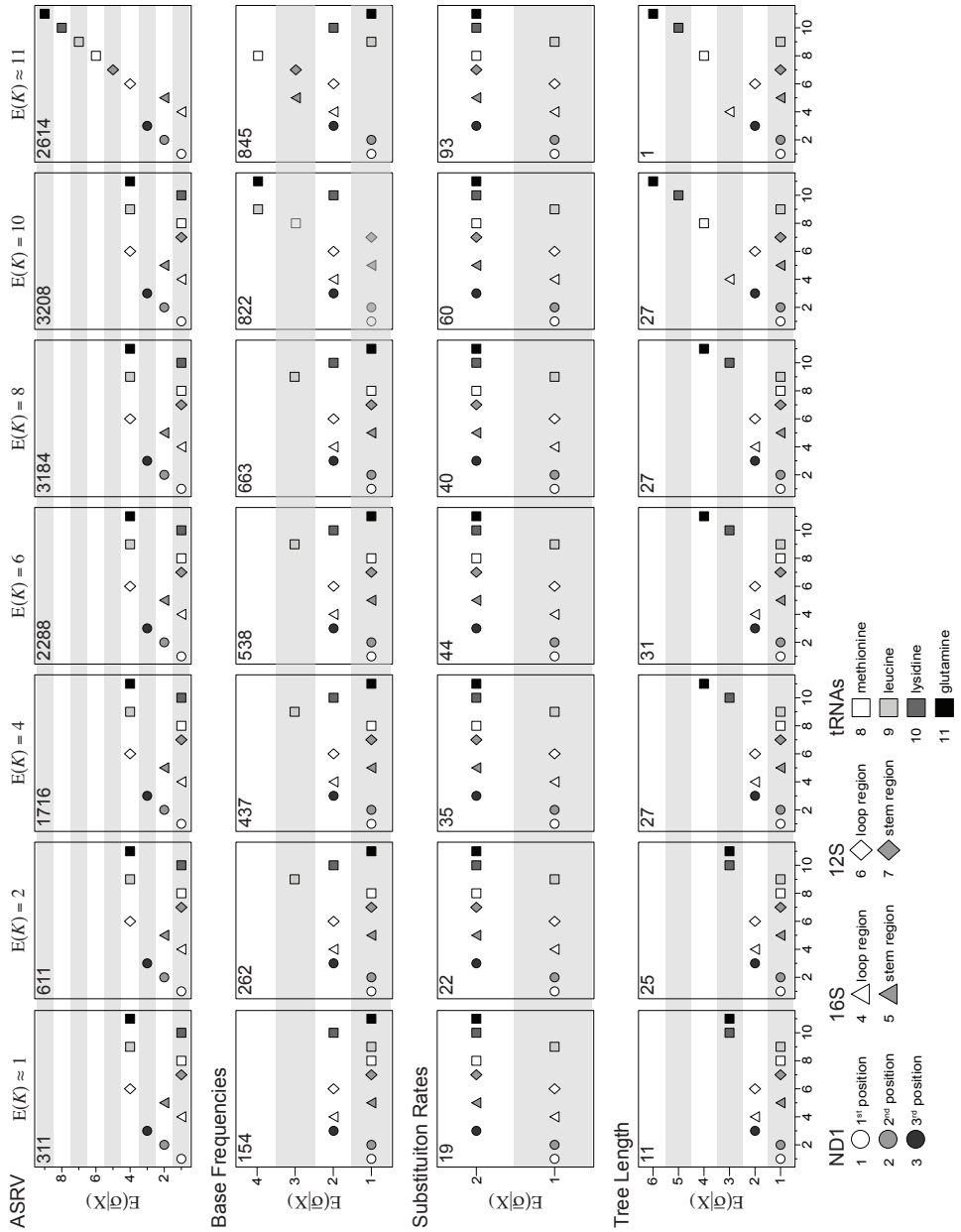

ASRV
Base Frequencies
Substituiton Rates
Tree Length
E(K) ≈ 1
E(K) = 2
E(K) = 4
E(K) = 6
E(K) = 8
E(K) = 10
E(K) ≈ 11
E(σ̄|X)
311
611
1716
2288
3184
3208
2614
154
262
437
538
663
822
845
19
22
35
44
40
60
93
11
25
27
31
27
27
1
ND1
16S
12S
tRNAs
1 1st position
2 2nd position
3 3rd position
4 loop region
5 stem region
6 loop region
7 stem region
8 methionine
9 leucine
10 lysidine
11 glutamine

FIGURE 4

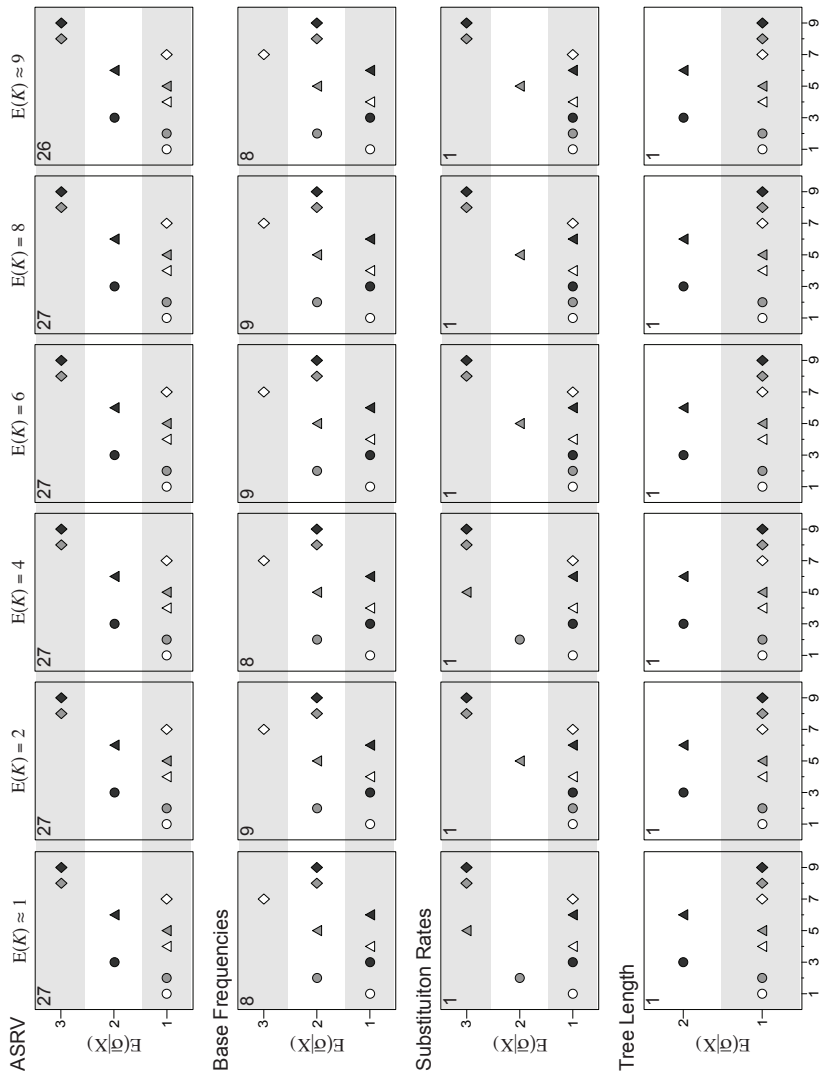

ASRV
Base Frequencies
Substituiton Rates
Tree Length
E(K) ≈ 1
E(K) = 2
E(K) = 4
E(K) = 6
E(K) = 8
E(K) ≈ 9
E(σ̄|X)
ND2
ND4
1 1st position
2 2nd position
3 3rd position
4 1st position
5 2nd position
6 3rd position
7 tRNA
8 AK1
9 BFib

FIGURE 5

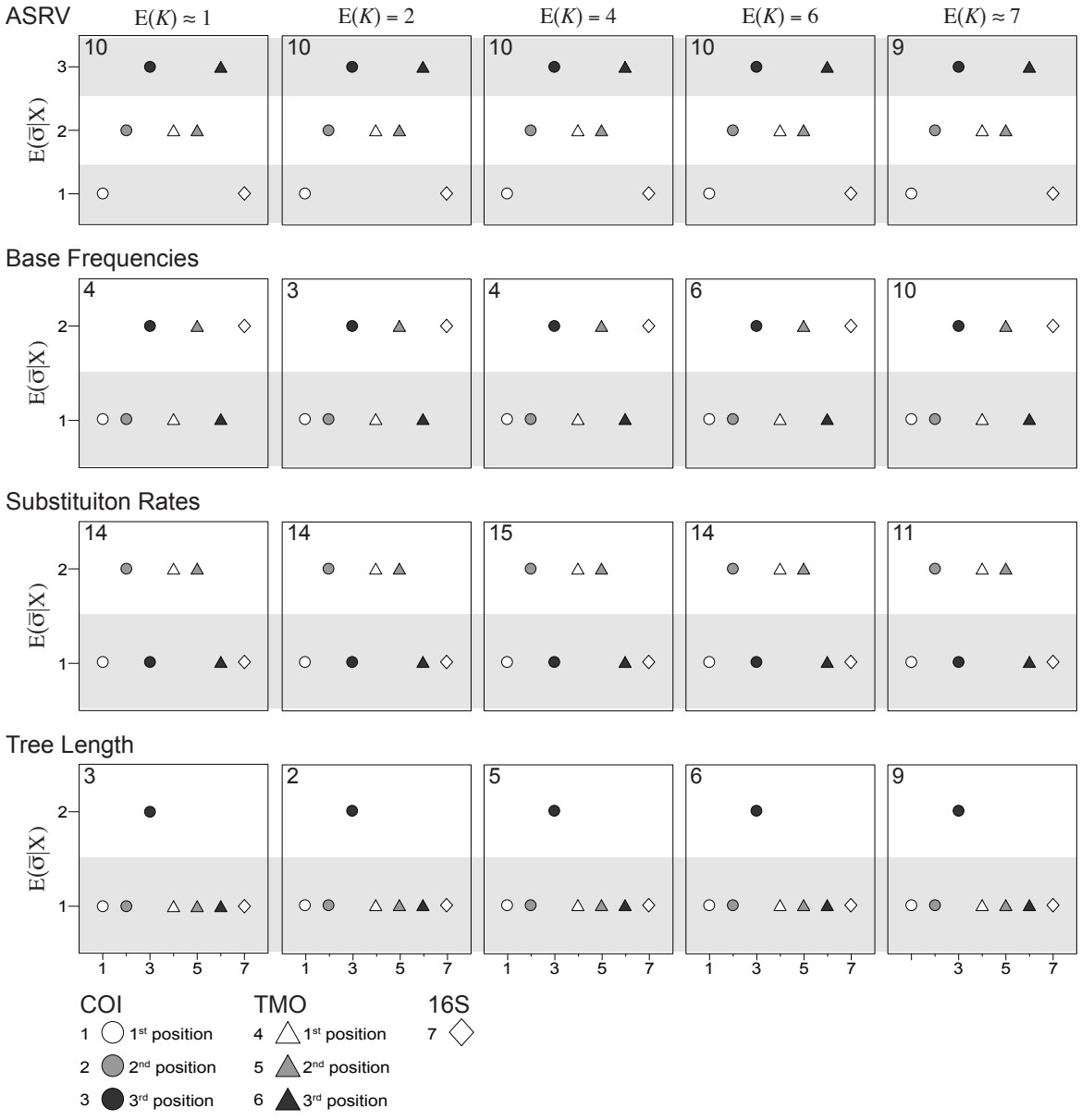
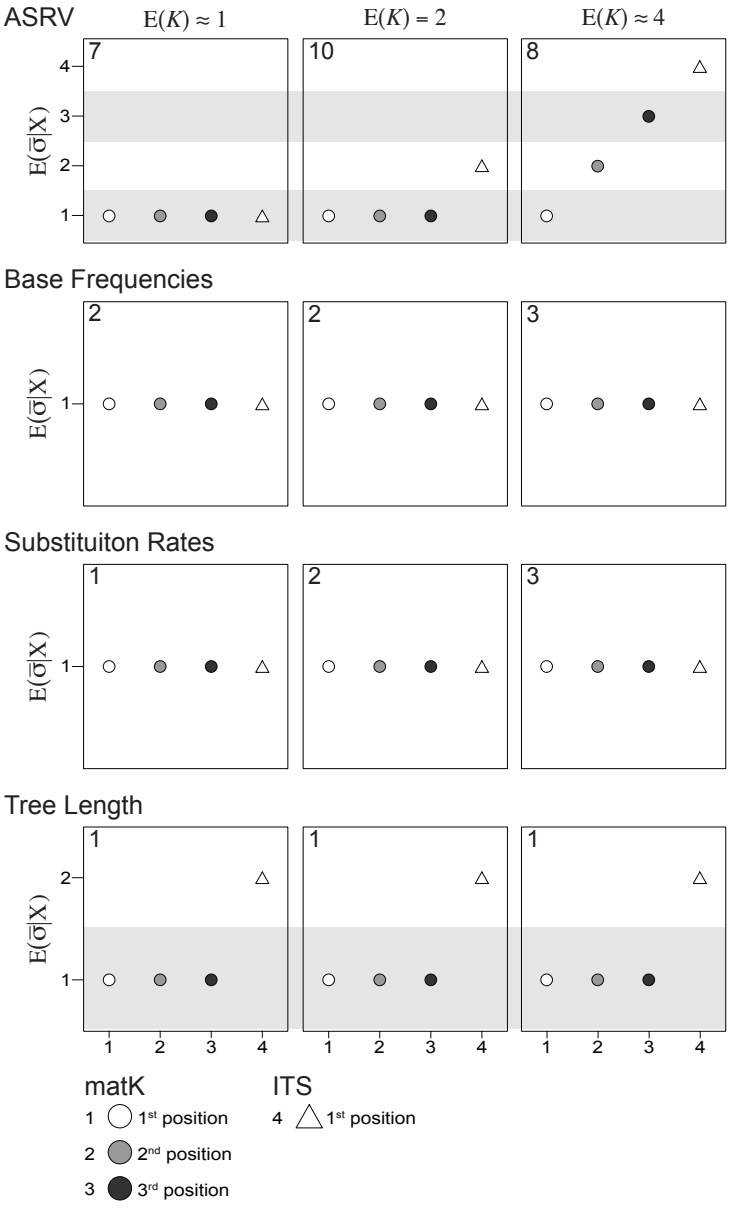

ASRV
E(K) ≈ 1
E(K) = 2
E(K) = 4
E(K) = 6
E(K) ≈ 7
E(σ̄|X)
Base Frequencies
Substituiton Rates
Tree Length
COI
1 1st position
2 2nd position
3 3rd position
TMO
4 1st position
5 2nd position
6 3rd position
16S
7

FIGURE 6


ASRV
E(K) ≈ 1
E(K) = 2
E(K) ≈ 4
7
10
8
E(σ̄|X)
Base Frequencies
2
2
3
Substituiton Rates
1
2
3
Tree Length
1
1
1
matK
1 1st position
2 2nd position
3 3rd position
ITS
4 1st position