A Dirichlet Process Covarion Mixture Model and Its Assessments Using Posterior Predictive Discrepancy Tests

Yan Zhou, Henner Brinkmann, Nicolas Rodrigue, Nicolas Lartillot, Herve Philippe

A Dirichlet Process Covarion Mixture Model and Its Assessments Using Posterior Predictive Discrepancy Tests

Yan Zhou,[1] Henner Brinkmann,[1] Nicolas Rodrigue,[2] Nicolas Lartillot,[1] and Hervé Philippe*,[1]

[1]Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, CP 6128, Succursale Centre-Ville, Montréal, Québec, Canada

[2]Department of Biology, University of Ottawa, Ottawa, ON, Canada

***Corresponding author:** E-mail: herve.philippe@umontreal.ca.

Associate editor: Peter Lockhart

Abstract

Heterotachy, the variation of substitution rate at a site across time, is a prevalent phenomenon in nucleotide and amino acid alignments, which may mislead probabilistic-based phylogenetic inferences. The covarion model is a special case of heterotachy, in which sites change between the "ON" state (allowing substitutions according to any particular model of sequence evolution) and the "OFF" state (prohibiting substitutions). In current implementations, the switch rates between ON and OFF states are homogeneous across sites, a hypothesis that has never been tested. In this study, we developed an infinite mixture model, called the covarion mixture (CM) model, which allows the covarion parameters to vary across sites, controlled by a Dirichlet process prior. Moreover, we combine the CM model with other approaches. We use a second independent Dirichlet process that models the heterogeneities of amino acid equilibrium frequencies across sites, known as the CAT model, and general rate-across-site heterogeneity is modeled by a gamma distribution. The application of the CM model to several large alignments demonstrates that the covarion parameters are significantly heterogeneous across sites. We describe posterior predictive discrepancy tests and use these to demonstrate the importance of these different elements of the models.

Key words: heterotachy, covarion, phylogenetics, model violations, posterior predictive discrepancy.

Introduction

The ability to infer accurate phylogenies is becoming more and more important as the flow of genomic data produced increases. Bayesian Markov chain Monte Carlo (MCMC) methods to address this problem are now popular, as they more readily allow the development of sophisticated models of sequence evolution. This is particularly important because the accuracy of phylogenetic inference heavily depends on the quality of the underlying models (Lanave et al. 1984; Yang 1996; Whelan and Goldman 2001; Phillips et al. 2004; Lartillot et al. 2007). For instance, the long-branch attraction artifact (Felsenstein 1978) is reduced through the use of the CAT model (Lartillot et al. 2007; Philippe et al. 2007; Delsuc et al. 2008). This model allows for a heterogeneous substitution process across sites (in addition to the heterogeneity of rate across sites) using a Dirichlet process prior (Ferguson 1973; Antoniak 1974; Neal 2000; Lartillot and Philippe 2004). Dirichlet process priors are convenient nonparametric devices for modeling site-specific effects, while relaxing the strict assumptions of the underlying statistical law implied by more classical parametric priors (Richardson and Green 1997).

Heterotachy (Philippe and Lopez 2001; Lopez et al. 2002), which describes the fact that substitution rates vary not only across sites but also across time, has drawn the attention of many researchers (Lockhart et al. 1996; Tuffley and Steel 1998; Galtier 2001; Huelsenbeck 2002; Kolaczkowski and Thornton 2004; Spencer et al. 2005; Wang et al. 2007; Zhou et al. 2007). Heterotachy was first characterized by Fitch and coworkers (Fitch and Markowitz 1970; Fitch 1971; Miyamoto and Fitch 1995) and was then shown to be frequent (e.g., 95% of the variable cytochrome b positions are heterotachous in vertebrates [Lopez et al. 2002]). It has been shown that heterotachy potentially impedes phylogenetic inference (Lockhart et al. 1996; Lopez et al. 1999, 2002; Philippe et al. 2000; Inagaki et al. 2004; Kolaczkowski and Thornton 2004). For instance, an uneven distribution of invariant sites can positively mislead phylogenetic reconstruction (Lockhart et al. 1996). Based on their observations, Fitch and Markowitz proposed the covarion model of sequence evolution (Fitch and Markowitz 1970; Fitch 1971). The covarion hypothesis states that, at a given time, due to functional constraints, some sites are free to vary and other sites are not; and at a later time, due to changes in functional constraints, some sites that were free to vary earlier no longer accept substitutions (and vice versa). The covarion hypothesis naturally creates heterotachous patterns of evolution.

Several models have been proposed to handle heterotachy. Based on the covarion hypothesis, Tuffley and Steel (1998) proposed a Markov-modulated Markov model, in which a stochastic process describes the ON/OFF state changes along the tree, whereas another stochastic process describes the substitution process when sites are in the ON state. In a context with m observed states ($m = 4$ for nucleotide data, $m = 20$ for amino acid data), the overall

process is defined over 2*m states, because a given position can be either in the ON or OFF state.

Huelsenbeck implemented an improved variant of this covarion model that allows for substitution rate variation across sites (Huelsenbeck 2002). Galtier (2001) relaxed the constraint of ON and OFF states and proposed another form of Markov-modulated Markov covarion model: sites freely transit along the tree among different rate categories following a discrete gamma distribution. In each discrete gamma rate category, sites then follow the classical Markovian substitution process. However, this model does not allow for the OFF state. Wang et al. (2007) recently combined Tuffley and Steel's and Galtier's models and proposed a triple Markovian process: Sites are not only able to transit between ON and OFF states, in the ON state, they are also allowed to transit between different rate categories; in each rate category, they follow a classical Markov transition process for substitutions. Likelihood ratio tests demonstrated that this model has a better fit than all other covarion models (Wang et al. 2007). Nevertheless, the large size of the transition matrix (2*g* m ×2*g* m, g is the number of rate categories) for this triple Markovian process implies a heavy computational burden.

On the other hand, because the branch length is the expected number of substitutions, heterogeneity of substitution rates across branches and across sites can be modeled with different sites having different sets of branch lengths. Accordingly, Kolaczkowski and Thornton (2004, 2008) proposed a mixture branch length (MBL) model to handle heterotachy: The MBL model consists of a mixture of components with different sets of branch lengths. However, given a large number of species, the number of parameters increases rapidly with each new component. Indeed, the covarion model has been shown to have a better fit than the MBL and the homotachous models on several large real data sets (Zhou et al. 2007). One explanation for the poor performance of the MBL model is that most branches of the different MBL components are correlated, rendering them redundant except for a few branches. To address this issue, Pagel and Meade (2008) proposed to use a reversible-jump MCMC technique in order to detect which branches require a set of different lengths; as expected, only the most heterotachous regions of the tree require extra branch lengths to adequately describe the data. An alternative to the MBL model would be a breakpoint model in which all sites share the same branch lengths except for some branches in which a fair amount of sites have drastic changes in substitution rate (Gu 2001; Dorman 2007). Nevertheless, determination of breakpoints along the branches demands heavy computations and has its own technical difficulties (Gu 2001; Dorman 2007; Blanquart and Lartillot 2008).

The elegance of the covarion model is that it has only two parameters that try to recover heterotachous signals by integrating the history of transitions (or switches) between ON and OFF states over branches and sites. For instance, sites having less substitutions in one part of the tree can be assumed to stay longer in the OFF state; sites having more substitutions in another part of the tree would be interpreted as spending more time in the ON state. The current covarion model assumes that the switch rates between the ON and OFF are homogenous across sites and stationary along the tree (Tuffley and Steel 1998; Huelsenbeck 2002). However, due to variations in functional requirements along the sequences, some sites might stay in the ON state much longer than other sites, or switch between ON and OFF with frequencies different from other sites, such that the switch rates between ON and OFF and the mean time spent in the ON state could be significantly heterogeneous across sites. Moreover, using large data sets resulting from the concatenation of genes with divergent function increases the chance of heterogeneities across sites in phylogenetic inference (Rodriguez-Ezpeleta et al. 2007). One might therefore question whether applying a single set of covarion parameters on a heterogeneous data set might constitute a serious model violation. Therefore, testing whether the transition rates between ON and OFF vary among sites is of great interest.

Our aim was to develop a model having different sets of covarion parameters (i.e., the switch rates between ON and OFF) for different sites. One possible solution is a mixture model with a number of components each possessing their own covarion parameters. Mixture models can be finite or infinite. For finite mixture models, the number of components is given a priori. Several finite mixture models have recently been proposed in phylogenetic analyses, for example, mixtures of substitution matrices (Pagel and Meade 2004) or the MBL model (Kolaczkowski and Thornton 2004; Spencer et al. 2005; Zhou et al. 2007). With finite mixture models, the number of components can be estimated by model comparison in the maximum likelihood framework (McLachlan and Peel 2000; Steel 2005; Zhou et al. 2007; Kolaczkowski and Thornton 2008) or by a posterior sampler using reversible-jump MCMC to sample through different dimensions of model space in the context of Bayesian methods (Green 1995). However, this estimation is difficult even under a fixed topology (Zhou et al. 2007), considering the changing dimensionality of the parameter space (Kolaczkowski and Thornton 2008). As an alternative to determining the number of components, an infinite mixture model can be applied. The most common approach to an infinite mixture model is using the Dirichlet process (Ferguson 1973; Neal 2000). The Dirichlet process is a nonparametric method to group observations that have similar behaviors and has been shown to successfully handle various heterogeneity problems in phylogenetic analysis (Lartillot and Philippe 2004; Huelsenbeck et al. 2006; Huelsenbeck and Andolfatto 2007; Huelsenbeck and Suchard 2007; Rodrigue, Lartillot, and Philippe 2008).

In this study, we develop the covarion mixture (CM) model, which is an infinite mixture model utilizing a Dirichlet process to handle the heterogeneities of the covarion parameters across sites in a Bayesian MCMC framework. We first study the heterogeneities of covarion parameters in real data sets. We then investigate the impact of the coexistence of different heterogeneities (rate of ON/OFF switch vs. rate of

substitution) on the inference of parameters. Finally, we assess the fit of models using posterior predictive discrepancy tests (Rubin 1984; Gelman et al. 1996).

Materials and Methods

Data Sets

Five amino acid alignments covering a wide range of site and taxon number were analyzed: 1) an opisthokont nuclear data set consisting of 17,912 sites and 63 species; 2) an animal nuclear data set consisting of 13,529 sites and 36 species; 3) an animal mitochondrial data set consisting of 2,373 sites and 116 species; 4) a vertebrate mitochondrial data set consisting of 3,478 sites and 136 species; and 5) a mammalian mitochondrial data set consisting of 3,559 sites and 53 species. The first two data sets are subsamples of the alignment of Lartillot and Philippe (2008) made to reduce the percentage missing data. The three other data sets are extracted from a large in-house alignment of complete holozoan proteomes and the unambiguously aligned regions were detected using GBlocks (Castresana 2000). Data sets are available at www.phylobayes.org in the phylobayes CM package. For all the data sets, constant sites are not included allowing significant reduction of the computation time for the CAT part of the model.

Furthermore, to perform posterior predictive discrepancy tests D^{H} (see below), several subgroups have been defined in four data sets: Arthropoda (36 species), Deuterostomia (45), and non-Bilateria (35) for animal mitochondrial data; Eutheria (24) and Metatheria (29) for mammal mitochondrial data; Teleostei (86), Gymnophiona (7), Caudata (26), Archeobratrachia (5), and Neobratrachia (12) for vertebrate mitochondrial data; Holozoa (33) and Fungi (30) for opisthokonts nuclear data.

Standard Huelsenbeck Covarion Model

For a given site i, the transition matrix for the Markov-modulated Markov process is (Tuffley and Steel 1998; Huelsenbeck 2002):

$$R = \begin{bmatrix} -S_{01}I & S_{01}I \\ S_{10}I & Q - S_{10}I \end{bmatrix}, \quad (1)$$

where I is the $m \times m$ identity matrix (m being the number of states; $m = 20$ for amino acids), Q is the $m \times m$ instantaneous rate matrix for substitution, S_{01} is the switch rate from OFF (0) to ON (1), and S_{10} is the switch rate from ON (1) to OFF (0). The stationary probabilities for ON and OFF, respectively, are $\pi_{ON} = S_{01}/(S_{01} + S_{01})$, $\pi_{OFF} = S_{10}/(S_{01} + S_{01})$. The stationary probability vector for the 2*m states is $(\pi_{OFF}\lambda, \pi_{ON}\lambda)$, where λ denotes the stationary frequency vector for m states.

When the rates are not uniform across sites and are assumed to follow a Γ distribution, the Q matrix, instead of the R matrix, is adjusted multiplicatively with a site-specific rate (i.e., rate across sites [RAS]) (Huelsenbeck 2002). In this way, the number of switches between ON and OFF is not proportional to the substitution rate.

The two parameters (S_{10} and S_{01}) specific to the covarion process can be transformed into another set of two parameters: the expected proportion of sites being the ON state π_{ON} ($\pi_{ON} = S_{01}/[S_{10} + S_{01}]$) along the tree and the average switch rate X ($X = 2S_{10}S_{01}/[S_{10} + S_{01}]$), which is the total number of switches between ON and OFF per branch length unit. This alternative set of parameters is useful to monitor the behavior of the covarion model and to make biological interpretations.

Infinite Mixture Model Using a Dirichlet Process

The Dirichlet process is a stochastic process, with which a number of distributions are dispensed under a Dirichlet distribution (Antoniak 1974; Escobar and West 1995). Supposing that observation i ($i = 1,\ldots, N$) is drawn from a mixture distribution over θ, the Dirichlet process can be realized with the following formula (Blackwell and MacQueen 1973):

$$\theta_i|\theta_1,\ldots,\theta_{i-1} \sim \frac{1}{i-1+\alpha}\sum_{j=1}^{i-1}\delta(\theta_j) + \frac{\alpha}{i-1+\alpha}G_0, \quad (2)$$

where $\delta(\theta)$ is the distribution centered at θ, α is a hyperparameter that controls the dispersion of the Dirichlet process, and G_0 is the base distribution. One application of the Dirichlet process is the prior for the infinite mixture model. The mixture model consists of K components that share the same base distribution G_0. By integration, the prior for c_i, with which site i is assigned to one component c, is

$$P(c_i = c|c_1,\ldots c_{i-1}) = \frac{n_{i,c} + \frac{\alpha}{K}}{n - 1 + \alpha}, \quad (3)$$

where $n_{i,c}$ is the number of sites in the component c to which site i is assigned (Neal 2000). The hyperparameter α influences the number of components. When the hyperparameter α is large, site i has a high probability of having a new component of its own; when α is small, site i is likely to be grouped with others.

CM Model

The CM model is basically a Huelsenbeck covarion model where the parameters S_{10} and S_{01} can vary across sites. More precisely, one Dirichlet process is defined on the parameter $\theta = (S_{10}, S_{01})$, and the base distribution G_0 is a joint of two independent exponentials of mean 1. To extensively explore the nature of the CM model, we define the prior for the hyperparameter α of the Dirichlet process as uniform in [0, 1,000]; therefore, the number of components in the CM model largely depends on the heterogeneities of the data.

Overall Models

The CAT model (Lartillot and Philippe 2004) is a mixture model allowing site-specific stationary probabilities using a Dirichlet process. In this paper, all the models are combined with the CAT model, because this model has generally a better fit than site-homogeneous models and is computationally relatively rapid (Lartillot and Philippe 2004, 2008).

We use the abbreviation COV for the standard one-component Huelsenbeck covarion model; CM for the

CM model; $+\Gamma$ for models with gamma-distributed rates discretized with four categories. Covarion model generally refers to both COV and CM models in the framework of the Huelsenbeck covarion model. Therefore, the CAT + CM + Γ model actually consists of two Dirichlet processes and handles three different site-specific heterogeneities (amino acid stationary probabilities, switch rates between ON and OFF, as well as the substitution rates in the ON state).

We recode states such that all nonobserved amino acids at a given column are treated as a single state (Lartillot and Philippe 2004). This recoding does not influence the likelihood calculation, that is, the likelihood is numerically identical to that obtained without the recoding. A fast algorithm (Galtier and Jean-Marie 2004) is used for the diagonalization of the matrix of a double Markovian process.

Posterior Estimation by MCMC

The parameters' posterior probability for data y is

$$P(z,\theta,\nu|y)=\frac{P(y|z,\theta,\nu)P(z)P(\theta)P(\nu)}{\int_{z,\nu,\theta}P(y|z,\nu,\theta)P(z)P(\nu)P(\theta)}, \quad (4)$$

where z is the allocation vector (c_1,c_2,. . .,c_n) that assigns site (1,. . .,n) to covarion components; θ is the switch rates S_{10} and S_{01}; ν is the rest of the parameters, such as branch length, $P(z)$ and $P(\theta)$ have been introduced in the CM model setting; other prior setting can be found in Lartillot and Philippe (2004).

We assume all sites are independent, so that the likelihood of the parameters for data y is the product of the likelihood at each site. A site-specific likelihood is conditional on a covarion component of which a site is assigned to

$$P(y|z,\theta,\nu)=\prod_{i=1}^{N}P(y_i|c_i,\theta,\nu). \quad (5)$$

MCMC is applied to obtain the posterior distribution over the parameters. In order to obtain a quick convergence, Gibbs sampling is applied with the help of auxiliary components for the Dirichlet process mixture model according to algorithm 8 described by Neal (2000).

Two independent chains are run to check the convergence of the chains. The MCMC chains are considered to reach convergence when the plots for all variables (e.g., likelihood value and number of covarion components) from different independent chains show the same posterior distributions. The posterior estimations of the parameters are the expectations of these parameters under the posterior distribution. For instance, the posterior estimation of site-specific S_{01} and S_{10} in the CM model is the mean of S_{01} and S_{10} for each site in the posterior distribution.

Events Mapping along the Tree

The substitutions and switches between ON and OFF can be studied using stochastic mapping. We use the data augmentation method for the stochastic mapping described by Rodrigue, Philippe, and Lartillot (2008). Briefly, applying uniformization, the Markov process is transformed into a Poisson process that allows for virtual substitutions (from one state to itself), and the waiting time for a substitution event no longer depends on the current state of the process. In the case of our study, the "events" for mapping refer to amino/nucleotide substitutions and switches between ON and OFF. Therefore, the size of the Markov matrix on which we apply the uniformization procedure is $2^*m \times 2^*m$ (for amino acid, $m = 20$), and we map events among 2^*m states. After removing the virtual events, we have the information about the number of substitutions in ON states, the number of switches between ON and OFF, and the time spent in ON and OFF states, for each site and each branch. These mappings are then used for constructing posterior predictive discrepancy tests.

The Posterior Predictive Distribution

Supposing φ is the parameter vector of the model, a series of posterior predictive data sets y^{pp} are simulated with values of φ drawn from the posterior distribution (i.e., conditional on the observed data set y^{obs}), such that the marginal probability of the posterior predictive data y^{pp} is

$$P(y^{pp}|y^{obs},\text{Model})=\int P(y^{pp}|\varphi)P(\varphi|y^{obs},\text{Model})\,d\varphi. \quad (6)$$

For the double Dirichlet processes model, that is, the CAT + CM model, a site would be simulated simultaneously with both the CAT component and the CM component to which this site belongs in the posterior distribution. Therefore, the simulation would reflect any interactions between the two different mixture models, if such interactions exist.

Multiple replications are generated for each φ. Here, 200 data points in the posterior samples are collected for each MCMC chain, and for each data point 5 replications were applied to generate the posterior predictive data sets. In the following, the posterior predictive distribution will be taken as our null distribution (Rubin 1984; Gelman et al. 1996).

Posterior Predictive Discrepancies Assessments

The classical P value of a test statistic T for data y is defined as

$$p(y,\text{Model})=P(T(Y)\geq T(y)|\text{Model}), \quad (7)$$

where T is a pivotal statistic, which is not dependent on any unknown parameters, and the data Y are sampled under the null distribution.

In the presence of nuisance parameters or in the context of Bayesian estimation, the parameter φ is not known or "fixed." Therefore the P value is defined as

$$p(y,\text{Model},\varphi)=P(T(Y)\geq T(y)|\varphi,\text{Model}), \quad (8)$$

where the test statistic T is dependent on the unknown parameter φ. In this case, the null distribution $T(Y)|\varphi$ is hard to know.

Because y^{pp} are simulated under the posterior distribution, the distribution of $T(y^{pp})$ can be taken as a null distribution (Rubin 1984). More specifically, Gelman et al. (1996) introduced posterior predictive discrepancy variable

$D(y, \varphi)$, which is a parameter-dependent statistic to measure the distance between the data y and the posited model. The posterior predictive discrepancy variable $D(y, \varphi)$ is actually a function of both the data and the parameters of the model. We are interested in the location of $D(y^{obs}, \varphi)$ in the distribution of $D(y^{PP}, \varphi)$ (null distribution). Therefore, the P value is defined as the probability that $D(y^{PP}) \geq D(y^{obs})$ in the posterior distribution:

$$p(y^{obs}, \text{Model}) = \int P(D(y^{PP}) \geq D(y^{obs})|\varphi) \times P(\varphi|y^{obs}, \text{Model})\text{d}\varphi. \quad (9)$$

The P value of the posterior predictive discrepancy based on MCMC can be obtained in a straightforward way by counting how many $D(y^{PP})$ are larger than $D(y^{obs})$. A low P value indicates a poor fit of the model to the data.

In order to check model fit with different aspects, different discrepancy variables can be constructed. In this study, we construct three discrepancy variables D^R, D^H, and D^O, based on three different aspects with variables R, H, and O that are devised to study substitution rate across sites, within-site substitution rate variation and the proportion of time for sites spent in the ON state, respectively. All the posterior predictive discrepancy variables in this study are constructed according to formula (10). Supposing a discrepancy variable regarding the variable v, D^v is

$$D^v(y, \varphi) = \frac{1}{N}\sum_{i=1}^{N}\frac{(v_i^m - v_i^e)^2}{v_i^e}, \quad (10)$$

where the "observed" value (v_i^m) of variable v for site i is computed on a mapping, whereas the expected value v_i^e is analytically derived based on the model.

The Discrepancy Variable D^R for Rate Heterogeneity. We construct a discrepancy variable D^R based on the difference between the number of observed substitutions along the tree and the number of substitutions expected by the model. Hence,

$$D^R(y, \varphi) = \frac{1}{N}\sum_{i=1}^{N}\frac{(R_i^m - R_i^e)^2}{R_i^e}, \quad (11)$$

where R_i^m is the total number of substitutions at site i, which is directly available from a mapping; R_i^e is the number of substitutions expected by the model for site i, and its value is equal to $\pi_{ON,i}{}^*B^*r_i$, the product of the site-specific proportion of being ON (π_{ON}) (for noncovarion model, $\pi_{ON,i} = 1$), the tree length (B) and site-specific substitution rate r_i (for non-RAS model, $r_i = 1$).

The Discrepancy Variable D^H for Heterotachy. Heterotachy can be revealed as heterogeneity of within-site substitution rates in different monophyletic groups (Miyamoto and Fitch 1995; Lopez et al. 1999). We therefore assess models using the discrepancy statistic D^H:

$$D^H(y, \varphi) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{p}\frac{(H_{ij}^m - H_{ij}^e)^2}{H_{ij}^e}, \quad (12)$$

where p is the number of groups; H_{ij}^m is the number of substitutions mapped in monophyletic group j for site i; H_{ij}^e is the number of substitutions expected by the model in monophyletic group j for site i, and its value is $\pi_{ON,i}{}^*B_j{}^*r_i$, of which B_j is the tree length of group j.

The Discrepancy Variable D^O for the "ON" State Behavior. To refine the assessment of various covarion models, we focus on a third statistic, D^O, which considers the relative time a site spent in the ON state:

$$D^O(y, \varphi) = \frac{1}{N}\sum_{i=1}^{N}\frac{(O_i^m - O_i^e)^2}{O_i^e}, \quad (13)$$

where O_i^m is (time in ON state)/(time in ON state + time in OFF state) obtained by the mapping, O_i^e is $\pi_{ON,i}$, which is estimated by the model.

The covarion models and the posterior predictive discrepancy tests introduced in this paper are implemented in PhyloBayes (Lartillot et al. 2009) and are available at www.phylobayes.org.

Results

CM Model

The CM model was applied on the five real data sets. Virtually identical posterior estimations from two independent chains show a good convergence of the MCMC on the Dirichlet process (supplementary fig. S1, table S1, Supplementary Material online). For instance, the posterior estimates of covarion parameters (i.e., S_{10} and S_{01}) for a given site are comparable.

Figure 1 shows the histogram of the number of components (K_{cov}) in the posterior distribution for the opisthokont nuclear data set. Although K_{cov} is variable (from 5 to 21), it is never equal to 1 in the posterior distribution. So the standard covarion model, which is a special case of the CM model with $K_{cov} = 1$, has a negligible posterior probability. This is confirmed by all data sets we have analyzed so far, which have an average number of components from 8 to 28 (table 1).

The posterior distributions of site specific S_{10} and S_{01} for the opisthokont data are shown in figure 2. As expected, there is a great heterogeneity across sites. S_{01} varies from ~0.4 to ~1 and S_{10} varies from ~0.5 to ~2.5. The other four data sets confirm that the covarion parameters vary significantly across sites (table 2, supplementary fig. S2, Supplementary Material online).

Comparisons of Real Data Sets and Their COV and CM Simulated Counterparts

To further validate the CM model, data sets were simulated under COV and CM models using the parameters estimated from real data sets (see, posterior predictive distribution in Materials and Methods). The CM model was then applied on these two types of simulated data sets to compare the results with the original real data sets.

Data sets simulated under CM yield similar posterior distributions for the number of components of the mixture with those obtained under the original real data sets (fig. 1

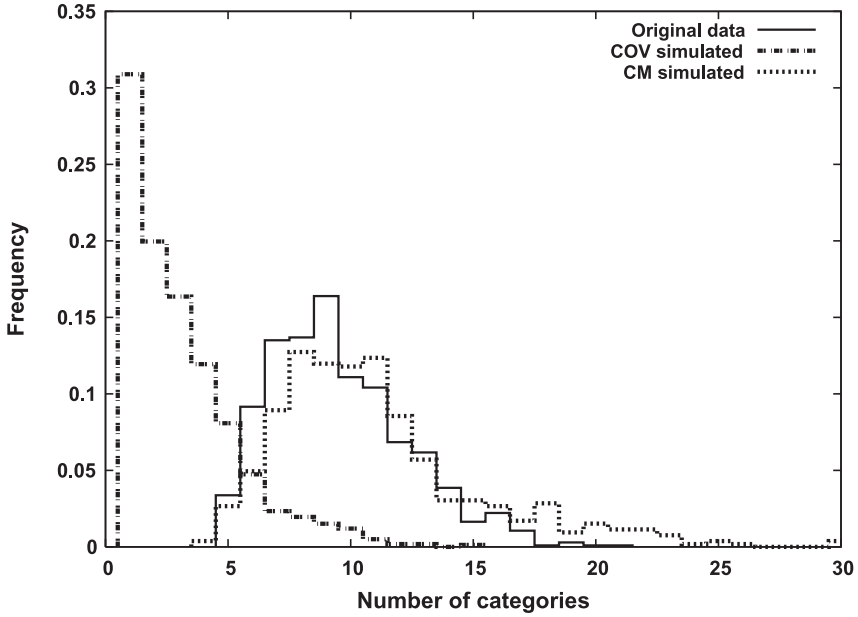

Fig. 1. Histograms of the number of CM components (K_{cov}) inferred by the Dirichlet process from the posterior distributions of the Opisthokont alignment and the corresponding data sets simulated with COV and CM models.

and supplementary fig. S2, Supplementary Material online). The average number of components for simulated CM data and for real data are always much higher than ones for simulated COV data, of which the values of K_{cov} are close to 1 and generally less than three (table 1).

The distributions of the CM and the COV simulated data for site-specific S_{10} and S_{01} were also studied (supplementary fig. S3, Supplementary Material online, table 2). For the simulated CM data set, S_{01} and S_{10} varied widely and their mean and variance are quite similar to the ones obtained from the original real data sets. In contrast, for the simulated COV data sets, most sites were concentrated in a narrow strip around the COV original simulated values, and variances of covarion parameters are more than 10 times smaller than ones for real data sets. These simulations demonstrate that the CM model is efficient in detecting the heterogeneity of the covarion parameters when data are heterogeneous and does not artificially inflate it when data are homogeneous.

Biological Meaning of the CM Model

We have observed significant heterogeneities of the covarion parameters across sites in all alignments we have analyzed so far. Because these data sets consist of different genes, one would be interested to know whether the covarion parameters are heterogeneous among genes. Figure 3 displays the mean and variance of the covarion parameters S_{10} and S_{01} for the 12 genes of the mammal mitochondrial data sets. Although covarion parameters vary widely within each gene, their mean seems also to vary among genes. To rigorously determine whether the values of site-specific covarion parameters S_{10} and S_{01} are affected by the genes, multivariate analysis of variance was performed on the three mitochondrial data sets. All the four statistics (Wilks' Lambda, Pillai's Trace, Hotelling–Lawley Trace, and Roy's Greatest Root) strongly reject the null hypothesis that the covarion parameters are not affected by the genes ($P \leq 0.0001$). Interestingly, the S10 (S01) parameters are higher (lower) for proteins of complexes III and IV (cytb, cox1, cox2 and cox3) than for other proteins, which is not observed for animal and vertebrate mitochondrial data sets (supplementary figs. S4 and S5, Supplementary Material online). This could be related to the well-documented acceleration of these genes in primates (Schmidt et al. 2001), which has also been detected by mixture of branch length model (Zhou et al. 2007).

The effect of the CM model and other models on tree lengths was also studied (table 3). As expected, the most complex models are better at detecting multiple substitutions; hence they display a greater tree length. In particular, trees inferred by the CM model are the longest. We speculate that the CM model, showing longer branches, has a greater capacity to capture multiple substitutions than other models via an effective categorization of sites mainly in the ON state through the use of site-specific switch rates between ON and OFF. Moreover, as expected, all models

Table 1. Number of Components for Covarion Parameters Inferred by the CM Model (Mean ± Standard Deviation [SD]).

	Opisthokont Nuclear	Animal Nuclear	Animal Mitochondrial	Vertebrate Mitochondrial	Mammal Mitochondrial
Original data	9.6 ± 2.8	8.6 ± 4.1	14.3 ± 5.0	28.7 ± 8.3	9.9 ± 5.2
CM simulated	11.0 ± 4.2	6.7 ± 2.4	13.2 ± 5.8	24.5 ± 11.1	7.5 ± 4.1
COV simulated	3.1 ± 2.3	3.6 ± 2.2	2.50 ± 1.71	2.6 ± 2.2	2.5 ± 1.7

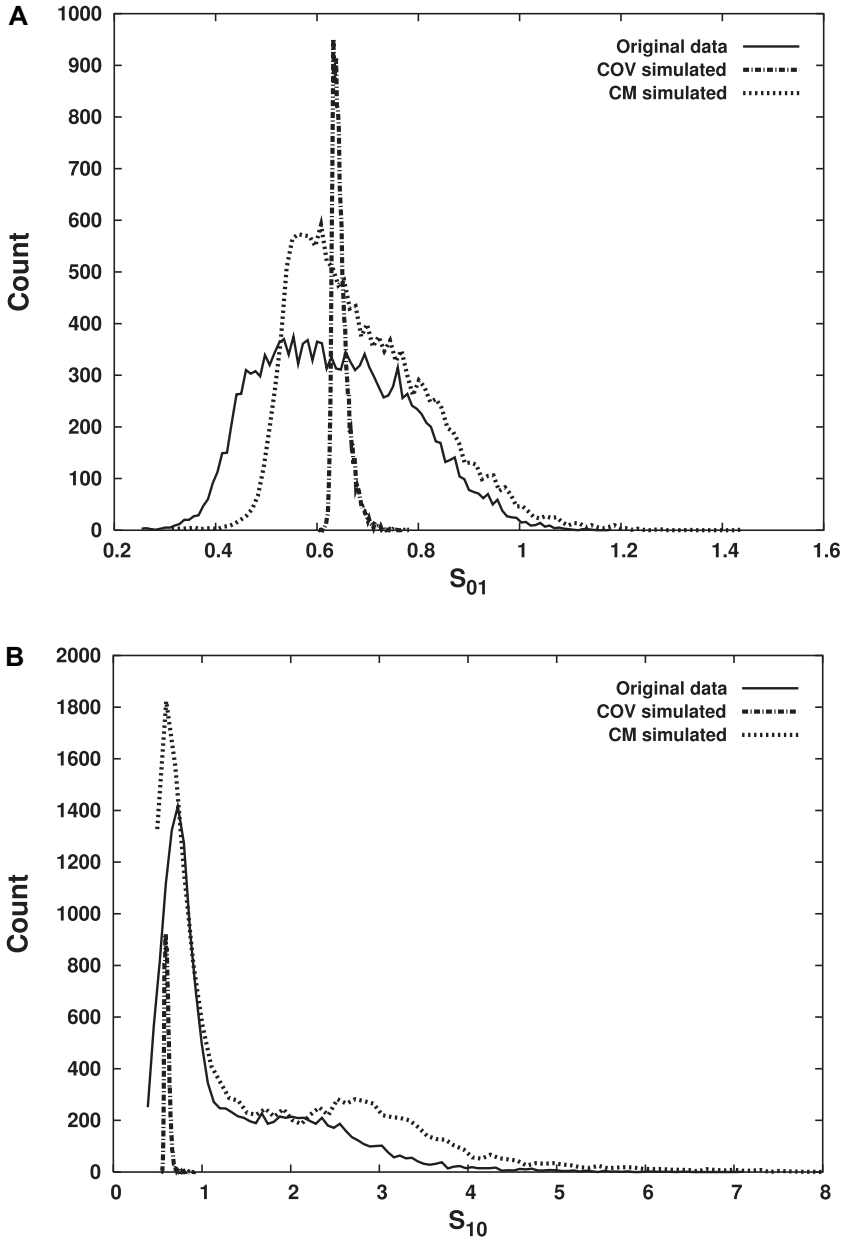

FIG. 2. Distributions of the posterior estimates of site-specific S_{01} (*A*) and S_{10} (*B*) for the 15,435 sites of the opisthokont nuclear alignment and its CM and Covarion simulated counterparts.

including the Γ distribution infer longer branches than models without it; similarly, all the covarion models have longer branches than homotachous models by allowing sites to switch between ON and OFF.

Interactions between the Discrete Gamma Rate Model and the Huelsenbeck Covarion Models

For the animal nuclear data set, we compared the estimated α value of the discrete gamma distribution for rates

Table 2. Covarion Parameter Values (Mean ± SD) for Real and Simulated Data Sets Inferred by the CM Model.

		Opisthokont Nuclear	Animal Nuclear	Animal Mitochondrial	Vertebrate Mitochondrial	Mammal Mitochondrial
Original data	S_{10}	**1.34 ± 0.89**	**1.06 ± 0.50**	**0.83 ± 0.24**	**1.46 ± 0.42**	**0.76 ± 0.14**
	S_{01}	*0.64 ± 0.14*	*0.67 ± 0.14*	*0.91 ± 0.31*	*1.07 ± 0.42*	*0.78 ± 0.12*
CM simulated	S_{10}	**1.72 ± 1.31**	**0.96 ± 0.35**	**0.77 ± 0.23**	**1.26 ± 0.32**	**0.85 ± 0.27**
	S_{01}	*0.70 ± 0.13*	*0.65 ± 0.13*	*0.89 ± 0.27*	*1.16 ± 0.45*	*0.76 ± 0.08*
COV simulated	S_{10}	**0.60 ± 0.01**	**0.48 ± 0.02**	**0.50 ± 0.01**	**0.65 ± 0.01**	**0.59 ± 0.02**
	S_{01}	*0.64 ± 0.004*	*0.57 ± 0.01*	*0.69 ± 0.02*	*0.79 ± 0.01*	*0.77 ± 0.02*

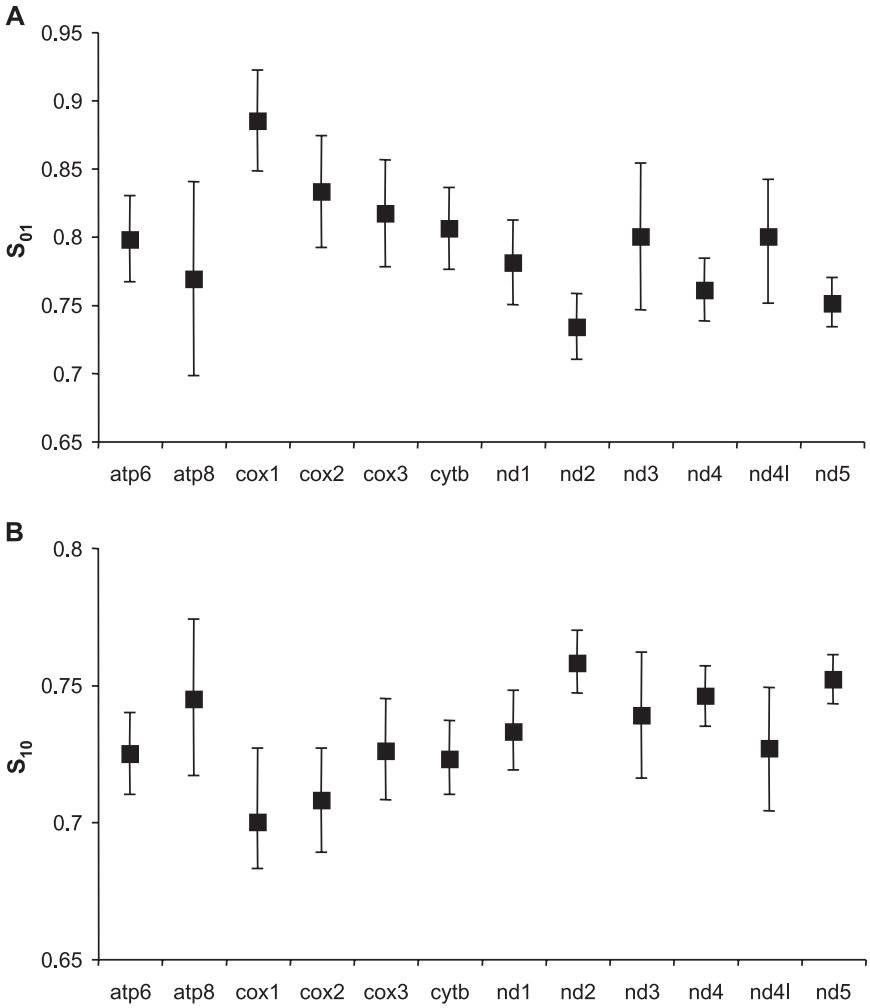

Fig. 3. Distributions of the means and 95% confidence interval for the site-specific S_{01} (*A*) and S_{10} (*B*) covarion parameters across genes. The names of the 12 genes for the mammalian mitochondrial data set have been indicated.

across sites under different models (table 4). Interestingly, when the covarion process is introduced, the discrete gamma rates become less heterogeneous across sites than under a noncovarion model: The shape parameter α for the discrete gamma rates increases from 1.58 to 2.65. When the heterogeneity of the covarion process across sites is considered, the estimated heterogeneity of rates becomes even less pronounced: The value of α increases further to 3.35. Such interactions are expected because a covarion process can mimic rate variation across sites by letting each site spend a longer or shorter time in the ON and OFF state (e.g., a site with a long time spent in the OFF state can be assumed to be a very slow evolving site). On the other hand, taking the heterogeneities of substitution rate across sites into account influences the inference of the covarion parameters (table 5).

Because the covarion and RAS modeling approaches interact with each other, one would be interested in 1) whether covarion signals and/or the heterogeneities of substitution rates across sites can be recovered under different models; 2) how the estimations of covarion and/or substitution rates across sites signals are affected under different models. For simplicity, the results are shown only with the one-component covarion model for the animal nuclear

Table 3. Tree Lengths (±SD) Inferred by Different Models.

Model/Data	Opisthokont Nuclear	Animal Nuclear	Animal Mitochondrial	Vertebrate Mitochondrial	Mammal Mitochondrial
CAT	11.11 ± 0.03	9.62 ± 0.04	19.51 ± 0.12	18.78 ± 0.08	9.71 ± 0.12
CAT + Γ	11.97 ± 0.05	10.42 ± 0.07	21.46 ± 0.17	21.08 ± 0.16	10.60 ± 0.15
CAT + COV	13.68 ± 0.07	11.91 ± 0.10	24.49 ± 0.21	23.19 ± 0.21	11.88 ± 0.16
CAT + COV + Γ	14.52 ± 0.10	12.44 ± 0.11	25.49 ± 0.28	24.98 ± 0.28	12.18 ± 0.18
CAT + CM	13.65 ± 0.13	11.73 ± 0.19	24.43 ± 0.25	23.84 ± 0.29	12.38 ± 0.24
CAT + CM + Γ	15.26 ± 0.19	13.24 ± 0.19	26.65 ± 0.65	26.15 ± 0.57	12.64 ± 0.32

Table 4. Posterior Estimation of α Value for the Discrete Gamma Rates by Various Models for the Animal Nuclear Data Set.

Model	α Value for the Discrete Gamma Rate (±SD)
CAT + Γ	1.58(±0.04)
CAT + COV + Γ	2.65(±0.11)
CAT + CM + Γ	3.36(±0.25)

alignment. Similar results were obtained with the CM model. Briefly, data sets were simulated with parameter values drawn from posterior distributions of the animal nuclear data for the three models: CAT + Γ, CAT + COV (the Tuffley and Steel model) and CAT + COV + Γ (the Huelsenbeck model), respectively. Subsequently, each simulated data set was analyzed with all these three models (table 6).

Simulated CAT + Γ Data Set. The CAT + Γ model recovered the original value of α for the discrete Γ distribution. With the CAT + COV + Γ model, the original α value was also recovered; however, S_{10} became extremely small, and π_{ON} was close to one. In other words, sites spent most of the time in the ON state, and no covarion signal was detected. In the absence of the RAS model, the CAT + COV model captured part of the RAS signal (S_{01}: 0.71 and S_{10}: 0.30).

Simulated CAT + COV Data Set. The CAT + COV model recovered the original value for the covarion parameters. With a CAT + COV + Γ model, the covarion model parameters were also recovered, and as expected, the RAS signal became very weak with α reaching 25. However, if a discrete gamma rate model is applied on the data that only contain covarion signal, the covarion signal would be considered as a RAS signal by the CAT + Γ model: $\alpha = 2.0$.

Simulated CAT + COV + Γ Data Set. The CAT + COV + Γ model recovered the value of α for the discrete gamma-rate distribution as well as the covarion parameters. This suggests that the two types of signals can in principle be distinguished. When the CAT + Γ model was applied, α was estimated at 1.48, below the true value (2.49), suggesting that the discrete Γ model takes both RAS and covarion signals as RAS signal. Similarly, when the CAT + COV model was applied on the data set, the estimation of the covarion parameters was influenced by the RAS signal contained in the data: S_{10} was increased from 0.43 to 0.61.

Altogether, these experiments suggest that although in practice the RAS and heterotachy signals are strongly influenced by each other, in principle they are identifiable.

Table 5. Posterior Estimation of S_{10} and S_{01} by Various Covarion Models for the Animal Nuclear Data Set.

Model	S_{10}(±SD)	S_{01}(±SD)
CAT + COV	0.56(±0.02)	0.54(±0.01)
CAT + COV + Γ ($\alpha = 2.65 \pm 0.11$)	0.45(±0.02)	0.57(±0.01)

Posterior Predictive Discrepancy Assessments of the Rate Heterogeneity across Sites

Posterior predictive discrepancy was used with the D^R statistic, which measures the ability of a model to handle the heterogeneity of rate across sites (table 7). As expected, the CAT model, which assumes uniform substitution rate across sites is rejected ($P < 0.01$). The CAT + Γ model is not rejected for the animal/opisthokont nuclear and mammal mitochondrial data sets ($P \geq 0.05$) but is slightly rejected for the other two data sets (animal–vertebrate mitochondrial data sets, $0.01 < P < 0.05$). Yet CAT + Γ has a better fit than CAT with the respect of substitution rate across sites. The CAT + COV and CAT + COV + Γ models are rejected for all the data sets ($P < 0.01$). Interestingly, the CAT + CM and CAT + CM + Γ models show a good fit with all the data sets ($P \geq 0.05$). Remarkably, the CAT + CM model fully handles an evolutionary property (RAS signal) for which it has not been designed (i.e., being designed to handle heterotachy signal). Results of D^R tests suggest that the discrete gamma model is outperformed by the CM model for handling the heterogeneities of rate across sites.

Posterior Predictive Discrepancy Assessments of Heterotachy at the Level of Monophyletic Groups

The D^H test indicates how well a model reflects heterotachy at the level of the monophyletic groups (table 8). As expected, the noncovarion models (i.e., CAT/CAT + Γ) are rejected for all the data sets ($P < 0.01$). Surprisingly, the CAT + COV model is also unable to deal with heterotachy ($P < 0.01$). Except for the mammal mitochondrial data ($P = 0.14$), the D^H test shows that the CAT + COV + Γ cannot reflect heterotachous properties observed in the alignments ($P < 0.01$). However, it shows that the CM/CM + Γ models cannot be rejected for all the real data sets we analyzed ($P \geq 0.05$). This demonstrates that all the analyzed models in our study, except for the CM and CM + Γ models, are unable to reflect heterotachous signals at the level of monophyletic groups.

Posterior Predictive Discrepancy Assessments of the ON State Behavior

The CAT + CM and CAT + CM + Γ models appear indistinguishable for the D^R and D^H tests. However, the Γ model seems necessary, otherwise, the estimated α value of the Γ distribution for CAT + CM + Γ model, which is currently only 3.36 (table 4), would be as high as for the simulated CAT + COV data, about 25 (table 6). To further investigate this point, the discrepancy tests D^O were designed based on the average time a given site spent in the ON state along the tree (table 9).

Both CAT + COV and CAT + CM models with uniform substitution rate are rejected ($P < 0.05$). However, the CAT + COV + Γ model is not rejected ($P \geq 0.05$) for all real data sets except for the vertebrate mitochondrial alignment ($P < 0.01$). Furthermore, CAT + CM +Γ model has a good fit for all the five alignments ($P \geq 0.05$), and

Table 6. Posterior Estimation of α Value for the Discrete Gamma Rate, S_{10}, and S_{01} for the Three Simulated Data Sets.

Simulated Data	Model	α of Discrete Γ(±SD)	S_{01}(±SD)	S_{10}(±SD)
CAT + Γ	CAT + Γ	1.53(±0.03)	NA	NA
	CAT + COV + Γ	1.53(±0.03)	1.14 (±0.98)	0.01(±0.02)
	CAT + COV	NA	0.71(±0.02)	0.30(±0.01)
CAT + COV	CAT + COV	NA	0.56(±0.01)	0.59(±0.02)
	CAT + COV + Γ	25.18(±0.08)	0.55(±0.01)	0.59(±0.02)
	CAT + Γ	2.0(±0.04)	NA	NA
CAT + COV + Γ	CAT + COV + Γ	2.7(±0.10)	0.58 (±0.01)	0.44±(0.01)
	CAT + COV	NA	0.57(±0.01)	0.61(±0.02)
	CAT + Γ	1.48(±0.03)	NA	NA

The original value of the parameters for the simulated data sets:
CAT + Γ simulated data set: $\alpha = 1.57$.
CAT + COV simulated data set: S_{01}: 0.52 and S_{10}: 0.55.
CAT + COV + Γ simulated data set: $\alpha = 2.49$, S_{01}: 0.55, and S_{10}: 0.43.

the *P* values are always higher than those for CAT + COV + Γ model. This implies that in contrast to the discrete gamma rate models, the uniform substitution rate models show poor fit when assessed with the discrepancy statistic D^O. One possible explanation for the poor fit is that the covarion with uniform substitution rate models try to deal with RAS signals in the data with covarion parameters, and consequently, the covarion parameters are likely to be misestimated.

Discussion

Posterior Predictive Tests

In classical statistical tests, the test statistics are completely free of unknown variables. Thus, the null distribution (e.g., X^2, *F* distribution) is a well-defined distribution, say without any uncertainty. However, sometimes, due to the presence of nuisance parameters, the statistics are dependent on parameters of unknown value; or due to a small sample size, the assumed distribution is not valid anymore; or in the Bayesian framework, estimations are not a single set of optimal values but a posterior distribution. Therefore, the corresponding statistical tests for assessing models are conditional on parameters with unknown values. In all of these cases, their distributions are hard to describe analytically. In these situations, simulations are often used to obtain the null distribution. For instance, instead of taking X^2 distribution as the null distribution, a null distribution is simulated for small data sets (Roff and Bentzen 1989).

In the case of our study, the number of substitutions within different subgroups depends on the branch lengths of the groups, site-specific substitution rates, stochastic mapping with the ON and OFF states along the tree, etc. Posterior predictive data naturally give a solution to the simulation of the null distribution on the unknown parameters because the statistic for posterior predictive data and the observed data share the same distribution of unknown parameters. The advantage of posterior predictive discrepancy tests is that they relax the restriction on the distribution under the null hypothesis for the statistical tests and allow any parameter-dependent statistics. For instance, Gelman et al. (1996) extended the classic goodness of fit model to the Bayesian framework and introduced the posterior predictive discrepancy, which is a parameter-dependent version of the classical statistic, to assess models. Protassov et al. (2002) suggested posterior predictive likelihood ratio tests to compare nested models.

Like the classical *P* value, the posterior predictive *P* value gives the risk information if we reject the null hypothesis. Thus, a high *P* value does not automatically imply the model is accepted; rather, it implies that there is no evidence to reject the model. Therefore, one should apply as many discrepancy tests with various aspects as possible to exclude unfit models. However, the statistic applied should be critical to reflect the difference between the data and the model. For instance, the D^R statistic, which accounts for the site-specific substitution rate, indicates the poor fit of the uniform substitution model, whereas it is unable to indicate the poor model fitness due to heterotachy.

Compared with other model selection methods in the Bayesian framework (e.g., cross validation [Aki and Jouko 2002; Lartillot et al. 2007; Blanquart and Lartillot 2008], Bayes factor using thermodynamic integration [Lartillot and Philippe 2006]), the posterior predictive test is affordable for the current computational system. Yet one cannot

Table 7. The *P*-Value of the Posterior Predictive Discrepancy Test D^R Considering the Number of Substitutions along the Entire Tree.

Model/Data	Opisthokont Nuclear	Animal Nuclear	Animal Mitochondrial	Vertebrate Mitochondrial	Mammal Mitochondrial
CAT	<0.01	<0.01	<0.01	<0.01	<0.01
CAT + Γ	0.11	0.05	0.04	0.04	0.1993
CAT + COV	<0.01	<0.01	<0.01	<0.01	<0.01
CAT + COV + Γ	<0.01	<0.01	<0.01	<0.01	<0.01
CAT + CM	0.29	0.41	0.40	0.77	0.65
CAT + CM + Γ	0.73	0.56	0.32	0.83	0.46

Table 8. The *P* Value of the Posterior Predictive Discrepancy Test D^H Considering the Number of Substitutions in Different Monophyletic Groups.

Model/Data	Opisthokonts Nuclear	Animal Mitochondrial	Vertebrate Mitochondrial	Mammal Mitochondrial
CAT	<0.01	<0.01	<0.01	<0.01
CAT + Γ	<0.01	<0.01	<0.01	<0.01
CAT + COV	<0.01	<0.01	<0.01	<0.01
CAT + COV + Γ	<0.01	<0.01	<0.01	0.14
CAT + CM	0.51	0.71	0.88	0.76
CAT + CM + Γ	0.66	0.39	0.86	0.52

rank models globally based on posterior predictive discrepancy tests, which actually take a role of analytical tools on the fitness of the model. Nevertheless, in the case of our study, because the COV model and the CM model are nested, the posterior distribution of K_{cov}, well above one, allows rejecting the COV model in favor of the CM model.

Coexistence of Rate Variation across Site and Heterogeneities of Covarion Parameters

Our studies show that the covarion parameters across sites are significantly heterogeneous. For instance, contrary to data sets simulated under the COV model, covarion parameters vary a lot in real data sets (fig. 2, table 2, supplementary fig. S3, Supplementary Material online). Considering this heterogeneity, relaxing the homogeneity of covarion parameters over sites improves the model fit. The posterior predictive discrepancy tests with respect to the heterotachy signal (i.e., D^H test and D^O test) show that the CM models, which allow for heterogeneities of S_{10} and S_{01} across sites, have better fits than COV models.

In real data sets, heterogeneities exist not only in covarion parameters but also in many other parameters, for example, substitution rates and stationary probabilities. Different models have been developed to specifically handle different types of heterogeneities. However, we see that heterogeneous models also attempt to handle other types of heterogeneities, which are not their original targets (table 6). For instance, the CM model can accommodate rate variation across sites, without RAS being specified, by allowing various values of π_{ON} among sites: Slow sites would have high π_{OFF} (or high S_{10}), and fast sites would have high π_{ON} (high S_{01}). However, the covarion parameters are not particularly devised for site-specific substitution rates, and thus they might not be able to recover such heterogeneities of the substitution rate efficiently. Figure 4 shows that the π_{OFF} is negatively correlated with the substitution rate only when substitution rates are small (<1), but is slightly positively correlated when rates are high (>1). Moreover, in attempting to address both RAS and heterotachy signals simultaneously, inferences under the pure CM model may be misleading. The posterior predictive discrepancy test D^O suggested a poor model fit for the CM with a uniform rate model. In the CM + Γ model, each site is assigned to a substitution rate mainly aiming at representing the average selective pressure over the whole tree; the CM part of the model then functions as an adjustor to distribute the variation of the substitutions along the tree via two parameters, π_{ON} (the proportion of being in ON) and *X* (the scattering level of switches along the tree).

A straightforward way to combine the RAS and covarion model is using Galtier's version of the covarion model (Galtier 2001). However, assuming four categories of rates, the dimension of the transition matrix in the Markov chain would be $4*m \times 4*m$ ($m = 20$ for amino acid data), which is very difficult to handle currently in terms of computation time, but might be helpful in the future with the advance of computer technology.

In phylogenetic analyses, different models have been developed to handle different types of heterogeneities. In this paper, we caution that different models handling different types of heterogeneities might interact with each other and that these interactions might impair inferences if not appropriately handled.

Application of the Dirichlet Process

The nonparametric mixture model using a Dirichlet process is an efficient method to handle heterogeneities in the data (Escobar and West 1995; Neal 2000; Lartillot and Philippe 2004; Huelsenbeck et al. 2006; Huelsenbeck and Suchard 2007; Rodrigue, Lartillot, and Philippe 2008). We verified that the Dirichlet process is able to handle both homogeneous and heterogeneous data. For

Table 9. The *P* Value of the Posterior Predictive Discrepancy Test D^O Considering the Proportion of Time Per Site in the ON State of the Covarion Process.

Model/Data	Opisthokonts Nuclear	Animal Nuclear	Animal Mitochondrial	Vertebrate Mitochondrial	Mammal Mitochondrial
CAT + COV	<0.01	<0.01	<0.01	<0.01	<0.01
CAT + COV + Γ	0.24	0.24	0.06	<0.01	0.07
CAT + CM	<0.01	<0.01	<0.01	<0.01	<0.01
CAT + CM + Γ	0.64	0.48	0.56	0.55	0.30

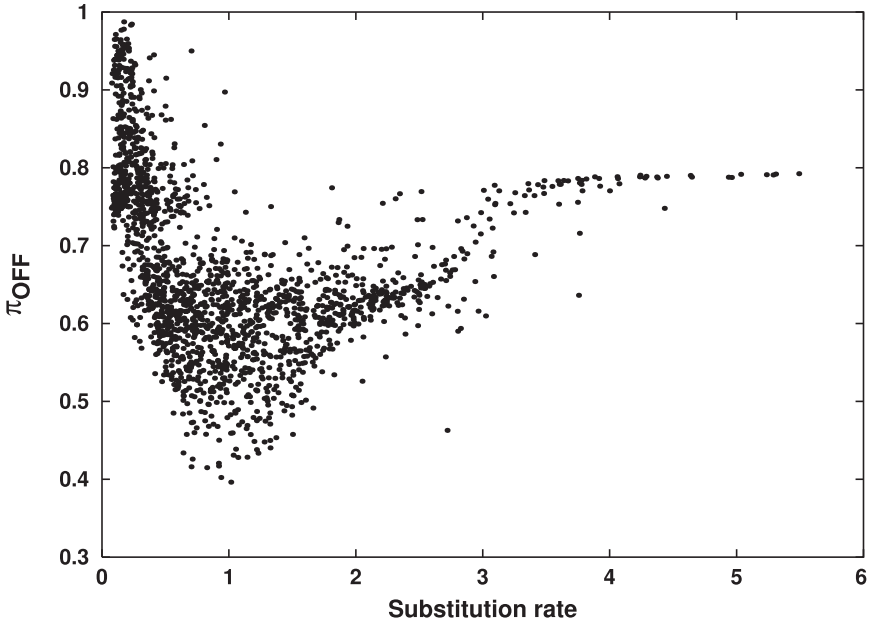

FIG. 4. Plot of site-specific continuous rate inferred by CAT + Γ model against the site-specific π_{off} inferred by the CAT + CM model for the opisthokont nuclear data set.

simulated homogeneous data, most sites share a similar value of the covarion parameters, and the number of components is very low. For simulated heterogeneous data, the Dirichlet process mixture model is able to recover the shape of the heterogeneous distribution, and the number of components is close to that of the real data. From this, we can conclude that the CM model is much better than the one-component covarion model for real data.

As discussed above, the CM model can generally take care of RAS signals when the RAS model is not available. More interestingly, the CAT + CM model even performs better than the CAT + Γ for the D^R test for some data sets. This is because the Dirichlet process is more efficient at handling heterogeneities of data than is the four category discrete gamma distribution. We expect that the site-specific substitution rate model using the Dirichlet process will have a much better fit than the classical discrete gamma rate model (Huelsenbeck and Suchard 2007).

The posterior predictive discrepancies tests confirm that the CAT + CM + Γ model is able to model the RAS signals as well as heterotachous signals. However, we are unable to show a better phylogenetic inference due to convergence problems when treating the topology as a free parameter; when several MCMC are independently run, all the nuisance parameters converge to similar values, and the topologies are highly similar, except a few nodes, which are precisely the ones of interest (unpublished results). Convergence problems may have several causes. One possible reason is the inefficiency of the MCMC sampling. For instance, we observed that sometimes two components have similar values of the covarion parameters. One solution to improve the MCMC for the Dirichlet process mixture model is using a "split–merge" algorithm (Jain and Neal 2000), which allows merging similar components and splitting a heterogeneous component into several components. This might be insufficient because strong correlations may exist between tree topology and preferred CM configurations. In fact, the CM model, being more flexible than currently available heterotachy models, may lead to situations of lack of identifiability with respect to the tree topology, such as demonstrated on theoretical grounds under more general heterotachy settings (Matsen and Steel 2007).

Heterotachous Models

The switch rates between ON and OFF for a given site could also change over time. In the current CM model, the values of these switch rates are assumed to be constant over the entire tree. Therefore, if in a tree the variation across time is only present in a few branches, the CM model might not be able to infer these variations solidly. One solution to this problem is to improve the taxon sampling, such that the variation signal becomes large enough for the CM model. The other possibility is to have a model that allows switch rates between ON and OFF to vary across sites and time, using for instance a breakpoint approach (Huelsenbeck et al. 2000; Blanquart and Lartillot 2008). Nevertheless, such a complex model would result in a heavy computational burden. In this context, the CM model can be combined with an MBL model, where reversible-jump techniques are used to reduce the number of branch lengths to infer (Pagel and Meade 2008). In such a case, some sites can have different branch lengths due to drastic, but rare, changes of substitution rates and follow a uniform CM model for most of the time. Implementing all of these approaches in a single encompassing statistical framework, which allows evaluation of their relative performance, would constitute a worthy direction for future work.

Acknowledgments

We acknowledge Génome Québec, the Canadian Research Chair and the Natural Sciences and Engineering Research Council of Canada for financial support, and the Réseau Québecois de Calcul de Haute Performance for computational resources.

References

Aki V, Jouko L. 2002. Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Comput.* 14: 2339–2468.

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

Blackwell D, MacQueen JB. 1973. Ferguson distributions via Polya Urn schemes. *Ann Stat.* 1:353–355.

Blanquart S, Lartillot N. 2008. A site- and time-heterogeneous model of amino acid replacement. *Mol Biol Evol.* 25:842–858.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17:540–552.

Delsuc F, Tsagkogeorga G, Lartillot N, Philippe H. 2008. Additional molecular support for the new chordate phylogeny. *Genesis* 46:592–604.

Dorman KS. 2007. Identifying dramatic selection shifts in phylogenetic trees. *BMC Evol Biol.* 7Suppl 1:S10.

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

Felsenstein J. 1978. Cases in which parsimony of compatibility methods will be positively misleading. *Syst Zool.* 27:401–410.

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

Fitch WM. 1971. Rate of change of concomitantly variable codons. *J Mol Evol.* 1:84–96.

Fitch WM, Markowitz E. 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.

Galtier N, Jean-Marie A. 2004. Markov-modulated Markov chains and the covarion process of molecular evolution. *J Comput Biol.* 11:727–733.

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

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

Gu X. 2001. Maximum-likelihood approach for gene family evolution under functional divergence. *Mol Biol Evol.* 18:453–464.

Huelsenbeck JP. 2002. Testing a covariotide model of DNA substitution. *Mol Biol Evol.* 19:698–707.

Huelsenbeck JP, Andolfatto P. 2007. Inference of population structure under a Dirichlet process model. *Genetics* 175:1787–1802.

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

Huelsenbeck JP, Larget B, Swofford D. 2000. A compound poisson process for relaxing the molecular clock. *Genetics* 154: 1879–1892.

Huelsenbeck JP, Suchard MA. 2007. A nonparametric method for accommodating and testing across-site rate variation. *Syst Biol.* 56:975–987.

Inagaki Y, Susko E, Fast NM, Roger AJ. 2004. Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF-1alpha phylogenies. *Mol Biol Evol.* 21: 1340–1349.

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

Kolaczkowski B, Thornton JW. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 431:980–984.

Kolaczkowski B, Thornton JW. 2008. A mixed branch length model of heterotachy improves phylogenetic accuracy. *Mol Biol Evol.* 25:1054–1066.

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

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol.* 7Suppl. 1:S4.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3. A Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25:2286–2288.

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

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

Lartillot N, Philippe H. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. *Philos Trans R Soc Lond B Biol Sci.* 363:1463–1472.

Lockhart PJ, Larkum AW, Steel M, Waddell PJ, Penny D. 1996. Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. *Proc Natl Acad Sci U S A.* 93:1930–1934.

Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. *Mol Biol Evol.* 19:1–7.

Lopez P, Forterre P, Philippe H. 1999. The root of the tree of life in the light of the covarion model. *J Mol Evol.* 49:496–508.

Matsen FA, Steel M. 2007. Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Syst Biol.* 56:767–775.

McLachlan GJ, Peel D. 2000. Finite mixture models. Wiley, New York.

Miyamoto MM, Fitch WM. 1995. Testing the covarion hypothesis of molecular evolution. *Mol Biol Evol.* 12:503–513.

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

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

Pagel M, Meade A. 2008. Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. *Philos Trans R Soc Lond B Biol Sci.* 363:3955–3964.

Philippe H, Brinkmann H, Martinez P, Riutort M, Baguna J. 2007. Acoel flatworms are not platyhelminthes: evidence from phylogenomics. *PLoS ONE.* 2:e717.

Philippe H, Germot A, Moreira D. 2000. The new phylogeny of eukaryotes. *Curr Opin Genet Dev.* 10:596–601.

Philippe H, Lopez P. 2001. On the conservation of protein sequences in evolution. *Trends Biochem Sci.* 26:414–416.

Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. *Mol Biol Evol.* 21: 1455–1458.

Protassov R, van Dyk DA, Connors A, Kashyap VL, Siemiginowska A. 2002. Statistics, handle with care: detecting multiple model components with the likelihood ratio test. *Astrophys J.* 571: 545–559.

Richardson S, Green PJ. 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J R Stat Soc Ser B (Methodological).* 59:731–792.

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

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

Rodriguez-Ezpeleta N, Philippe H, Brinkmann H, Becker B, Melkonian M. 2007. Phylogenetic analyses of nuclear, mitochondrial, and plastid multigene data sets support the placement of Mesostigma in the Streptophyta. *Mol Biol Evol*. 24:723–731.

Roff DA, Bentzen P. 1989. The statistical analysis of mitochondrial DNA polymorphisms: chi 2 and the problem of small samples. *Mol Biol Evol*. 6:539–545.

Rubin DB. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann Statist*. 12:1151–1172.

Schmidt TR, Wu W, Goodman M, Grossman LI. 2001. Evolution of nuclear- and mitochondrial-encoded subunit interaction in cytochrome c oxidase. *Mol Biol Evol*. 18:563–569.

Spencer M, Susko E, Roger AJ. 2005. Likelihood, parsimony, and heterogeneous evolution. *Mol Biol Evol*. 22:1161–1164.

Steel M. 2005. Should phylogenetic models be trying to 'fit an elephant'. *Trends Genet*. 21:307.

Tuffley C, Steel M. 1998. Modeling the covarion hypothesis of nucleotide substitution. *Math Biosci*. 147:63–91.

Wang HC, Spencer M, Susko E, Roger AJ. 2007. Testing for covarion-like evolution in protein sequences. *Mol Biol Evol*. 24:294–305.

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

Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol*. 11:367–372.

Zhou Y, Rodrigue N, Lartillot N, Philippe H. 2007. Evaluation of the models handling heterotachy in phylogenetic inference. *BMC Evol Biol*. 7:206.