A Site- and Time-Heterogeneous Model of Amino Acid Replacement

Samuel Blanquart, Nicolas Lartillot

A Site- and Time-Heterogeneous Model of Amino Acid Replacement

Samuel Blanquart and Nicolas Lartillot

Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, UMR 5506, CNRS-Université de Montpellier 2, Montpellier, France

We combined the category (CAT) mixture model (Lartillot N, Philippe H. 2004) and the nonstationary break point (BP) model (Blanquart S, Lartillot N. 2006) into a new model, CAT–BP, accounting for variations of the evolutionary process both along the sequence and across lineages. As in CAT, the model implements a mixture of distinct Markovian processes of substitution distributed among sites, thus accommodating site-specific selective constraints induced by protein structure and function. Furthermore, as in BP, these processes are nonstationary, and their equilibrium frequencies are allowed to change along lineages in a correlated way, through discrete shifts in global amino acid composition distributed along the phylogenetic tree. We implemented the CAT–BP model in a Bayesian Markov Chain Monte Carlo framework and compared its predictions with those of 3 simpler models, BP, CAT, and the site- and time-homogeneous general time–reversible (GTR) model, on a concatenation of 4 mitochondrial proteins of 20 arthropod species. In contrast to GTR, BP, and CAT, which all display a phylogenetic reconstruction artifact positioning the bees *Apis mellifera* and *Melipona bicolor* among chelicerates, the CAT–BP model is able to recover the monophyly of insects. Using posterior predictive tests, we further show that the CAT–BP combination yields better anticipations of site- and taxon-specific amino acid frequencies and that it better accounts for the homoplasies that are responsible for the artifact. Altogether, our results show that the joint modeling of heterogeneities across sites and along time results in a synergistic improvement of the phylogenetic inference, indicating that it is essential to disentangle the combined effects of both sources of heterogeneity, in order to overcome systematic errors in protein phylogenetic analyses.

Introduction

The "pruning" algorithm (Felsenstein 1981) was originally devised for data likelihood computation under the so-called F81 Markovian substitution process. It opened the way for probabilistic approaches in phylogenetics, first using maximum likelihood (ML), and subsequently using Bayesian analysis based on Markov Chain Monte Carlo (MCMC) sampling. Those full likelihood methods yield estimations of parameters, such as the topology, the branch lengths, or the rates of substitution, based on the observed data and on a set of assumptions formalized into a probabilistic model.

The original probabilistic evolutionary model was much simplified, essentially for practical reasons. In particular, strong assumptions were made concerning 1) the constancy of the overall rate of substitution across sites as well as along lineages, 2) the independence between positions along the sequence, and 3) the use of a single Markovian substitution process applied along all lineages as well as over all sites. Following this simplified but seminal version, many models relaxing those assumptions have been proposed.

The rate constancy assumption was relaxed by associating distinct gamma-distributed rates of substitution to each site (Yang 1994) and, furthermore, by allowing them to vary along lineages (Tuffley and Steel 1998; Huelsenbeck 1999; Galtier 2001). Site independence was partially dealt with by accounting for structural properties of biochemical sequences, for example, concerning the Watson–Crick pairs in RNA stems (Jow et al. 2002; Hudelot et al. 2003; Gibson et al. 2005), the codon positions in DNA genes (Goldman and Yang 1994), or the physical constraints implied by amino acid neighborhood in protein folding (Robinson et al. 2003; Rodrigue et al. 2005). We are in this work more concerned with the third assumption, especially in the case of protein sequences. This assumption, suggesting that a single Markovian process of substitution may be applied for all sites and at all times, implies that the equilibrium frequencies (i.e., the stationary probabilities) of the 20 amino acids are the same at all points along the sequence and along the tree. It thus implies that all taxa and all sites should display homogeneous state compositions, up to stochastic fluctuations. Yet, biological sequences do not display that property.

First, it has been shown that taxa may display heterogeneous state compositions or compositional biases. This concerns nucleotide (Jukes and Bhushan 1986; Montero et al. 1990; Bernardi 1993) as well as amino acid (Foster et al. 1997) sequence contents and goes against the stationary assumption. It has furthermore been shown that under homogeneous models, unrelated sequences sharing similar compositions may attract each other, yielding erroneous clustering in the phylogenies subsequently obtained (Lockhart et al. 1992; Lake 1994; Lockhart et al. 1994; Galtier and Gouy 1995; Yang and Roberts 1995; Foster and Hickey 1999; Mooers and Holmes 2000; Foster 2004). To address this problem, likelihood-based models assuming that changes in the substitution process equilibrium frequencies may occur along lineages have been proposed (Yang and Roberts 1995; Galtier and Gouy 1998; Foster 2004; Blanquart and Lartillot 2006; Boussau and Gouy 2006; Gowri-Shankar and Rattray 2007). Relaxing the stationary assumption, those models are denoted as nonstationary (although they are also nonhomogeneous along time). Importantly, they alleviate attraction artifacts due to similar compositional biases.

Concerning the homogeneity assumption along the sequence, it is readily observed that a given site of a protein alignment does not display all possible amino acids, but only a particular subset, generally characterized by similar biochemical properties (e.g., small hydrophobic and aliphatic amino acids I, V, and L, or polar and positively charged ones, K and R). This was in a first step accounted for using more general substitution processes than F81.

Key words: phylogeny, MCMC, nonstationary, mixture, posterior predictive, model violation, LBA.

E-mail: samuel.blanquart@lirmm.fr.

Mol. Biol. Evol. 25(5):842–858. 2008
doi:10.1093/molbev/msn018
Advance Access publication January 29, 2008

Those more general processes allow for higher exchange rates between biochemically equivalent amino acids, encoded using general time–reversible (GTR) matrices, either set to empirical values, such as JTT (Jones et al. 1992) or WAG (Whelan and Goldman 2001), or considered as free parameters in the so-called GTR model (Lanave et al. 1984; Tavaré 1986; Barry and Hartigan 1987; Rodriguez et al. 1990). They have been extensively used and were shown efficient for improving model fitness and phylogenetic accuracy. However, they appear not to be sufficient to accurately catch the observed site-specific biochemical preferences (Lartillot and Philippe 2004). In particular, they were shown to inaccurately handle saturation effects due to multiple substitutions among highly exchangeable amino acids (Lartillot et al. 2007), which in certain cases may be responsible for long branch attraction (LBA) artifacts (Felsenstein 1978). In contrast, site-specific saturation effects were shown to be better handled by models explicitly encoding the variations of the biochemical properties along the sequences into the stationary probabilities of the amino acid replacement process, using site-specific (Bruno 1996; Halpern and Bruno 1998) or mixture (Dimmic et al. 2000; Lartillot and Philippe 2004) models. Those kinds of models encode the site-specific biochemical constraints more precisely and importantly were shown to alleviate LBA artifacts (Lartillot et al. 2007).

Altogether, those models outperform the standard time- and sequence-homogeneous model and improve the phylogenetic reconstruction accuracy. However, now that it has been shown that proteins indeed display variations of their evolutionary behavior both across lineages and along their sequence, it is also obvious that, in general, they will display these 2 properties simultaneously. Yet, until now, these 2 features have never been considered jointly in the frame of 1 single time- and sequence-heterogeneous model. As we had previously proposed 2 models, handling site specificities (Lartillot and Philippe 2004) and nonstationary sequence evolution (Blanquart and Lartillot 2006), respectively, named CAT and BP, we now propose to combine them into 1 single nonstationary mixture model which we called CAT–BP.

In the CAT model, sites are clustered into K categories characterized by their own processes of amino acid replacement. The K processes are simple F81 processes, thus entirely defined by a vector of equilibrium frequencies over the 20 amino acids. These processes are applied along all branches of the tree and are therefore assumed to be stationary and homogeneous. The resulting model may consequently be affected by artifacts induced by similar compositional biases shared between unrelated taxa. As an illustration, let us consider the 2 positively charged amino acids Lysine and Arginine, K and R, which are respectively encoded in the standard genetic code by codons K = {AAA, AAG} and R = {CGT, CGC, CGA, CGG, AGA, AGG}. In case of a global drift of a genome towards AT richness, a protein site functionally constrained to accept only K or R will more likely display a state K (Foster et al. 1997; Singer and Hickey 2000). Whatever the GC-content, on the other hand, such a site is likely to remain under the same selective constraint throughout the tree, that is, to always accept exclusively K or R. This suggests that the site-specific substitution processes underlying the CAT mixture model should be modulated along lineages, so as to accommodate compositional shifts, while keeping a certain biochemical identity over the whole tree. In addition, on a given lineage, a global modulation of the amino acid content has to be applied simultaneously and in a coherent fashion to all K substitution processes defined by the mixture. More specifically, in the latter example of a globally increasing content of K (and a concomitant decrease of R) induced by AT richness, all categories of the mixture have to adapt in a correlated way their stationary probabilities for states K and R. Finally, compositional biases at the amino acid level are not only driven by underlying nucleotidic biases but also by more general environmental conditions, such as temperature (Lobry and Chessel 2003; Singer and Hickey 2003; Lobry and Necsulea 2006), halophily adaptation (Kennedy et al. 2001; Fukuchi et al. 2003), or even more specific ecological life styles (Bogatyreva et al. 2006; Das et al. 2006; Tekaia and Yeramian 2006). This suggests that the most general kind of compositional shift along time should be a priori considered and not only shifts induced by nucleotide compositional variations.

The CAT–BP model introduced here achieves this using the compound stochastic process formalism, first described by Huelsenbeck et al. (1999) and used in a nonstationary context in the BP model (Blanquart and Lartillot 2006). According to that formalism, N break points randomly appear along the lineages according to a Poisson process. Each break point introduces a discrete change in the global composition of the sequence. It has here been adapted so that each break point now causes a simultaneous and coherent modulation of the stationary probabilities of the K categories of the mixture. Thanks to this, the overall model still allows for site-specific biochemical constraints, although it is no more stationary.

In this paper, we introduce the model as well as the details related to its MCMC implementation. We compare the behavior of this model to that of GTR, CAT, and BP, on a concatenation of mitochondrial proteins of 20 arthropods. We show that all models infer a high level of saturation on this data set and that GTR, CAT, and BP all produce the same phylogenetic reconstruction artifact, positioning 2 hymenopterans (bees) among chelicerates, where they are attracted by a tick. These 2 bees, as well as the tick, are among the most AT rich and are furthermore the fastest evolving among the 20 taxa represented in the alignment, suggesting an artifact resulting from combined effects of fast evolution, site saturation, and compositional biases. Interestingly, only the CAT–BP combination is able to correctly place the 2 bees back within insects. Furthermore, using posterior predictive tests (Meng 1994; Gelman et al. 1996; Bollback 2002; Nielsen and Huelsenbeck 2002; Lartillot and Philippe 2004), we show that, unlike other investigated models, CAT–BP simultaneously accounts for the taxon- and site-specific distributions of amino acids observed in this alignment.

Methods

Notation and Parameters

A set of aligned sequences is available in form of a data matrix $\mathbf{D}$ of J sequences of I sites. Phylogenetic

relationships between the J sequences are represented by a rooted phylogenetic tree, denoted as τ, whose internal nodes represent speciation events. A length is associated to each branch. Let $\mathbf{t} = (t_j)$, where $j \in [1, \ldots, 2J - 2]$ are branch indices, denote the set of branch lengths. Additionally, sites have their own rate of substitution, $\mathbf{r} = (r_i)$, where $i \in [1, \ldots, I]$, distributed according to a Gamma distribution of variance $\frac{1}{\alpha}$, discretized into 4 categories.

Markovian Substitution Process

Substitutions (or amino acid replacements) at a given site are modeled by a Markovian process operating on a state space of size S ($S = 20$ for protein sequences). The most general Markovian substitution process is usually denoted as GTR (Lanave et al. 1984; Tavaré 1986; Barry and Hartigan 1987; Rodriguez et al. 1990). It involves a set of S stationary probabilities $\pi = (\pi_l)$, where $l \in [1, \ldots, S]$, and a set of relative exchange rates $\rho = (\rho_{lm})$, where $l, m \in [1, \ldots, S]$. The stochastic kernel (rate matrix) of the resulting substitution process, usually denoted as $\mathbf{Q}$, is obtained by combining π and ρ as follows:

$$\mathbf{Q}_{lm} = \pi_m \rho_{lm}, \quad l \neq m, \tag{1}$$

$$\mathbf{Q}_{ll} = -\sum_{l \neq m} \mathbf{Q}_{lm}, \tag{2}$$

where we have defined $\rho_{lm} = \rho_{ml}$ for all $l > m$. Given $\mathbf{Q}$, one can compute the probability $p_{l \to m}(t)$ of a substitution from a state l to a state m after an evolutionary time of $r_i t_j$ (along a branch of length t_j and for a site of rate r_i):

$$p_{l \to m}(r_i t_j) = \left[e^{r_i t_j \mathbf{Q}}\right]_{lm}. \tag{3}$$

We also use in this work the F81 process (Felsenstein 1981), which is the particular case, where $\rho_{lm} = 1 \ \forall l, m$. All relative exchange rates being equal, equation (3) simplifies into

$$p_{l \to m}(r_i t_j) = e^{-r_i t_j} \delta_{lm} + \left(1 - e^{-r_i t_j}\right) \pi_m, \tag{4}$$

where δ_{lm} is the Kronecker operator ($\delta_{lm} = 1$, if $l = m$, 0 otherwise). Note that equation (4) is the uniformized version of the F81 process (i.e., waiting times do not depend on the current state, and virtual substitutions $l \to l$ are allowed). In addition, we do not normalize the $\mathbf{Q}$ matrix (eq. 2). As a result, branch lengths are no more measured in terms of expected number of substitutions.

Site- and Time-Specific Nonparametric Devices

We model variations of the amino acid distributions, across sites and along time, using the 2 nonparametric devices initially implemented into 2 separate models, CAT (Lartillot and Philippe 2004) and BP (Blanquart and Lartillot 2006), respectively. In the following, superscripts ^{c} and ^{b} will be used as symbols denoting parameters related to the CAT or the BP component.

First, as in the CAT model, we assume that a mixture of K distinct categories, each defining its own Markovian process of substitution, is distributed among the I sites. Each category of the mixture, indexed by $k \in [1, \ldots, K]$, defines a normalized vector Π^{c}_{k}, of size S, and with positive entries, which will be called a "profile" in the following. Let $\Pi^{c} = (\Pi^{c}_{k})$, where $k \in [1, \ldots, K]$, denotes the set of profiles. The category a site i is associated to is specified by an allocation variable $z_i \in [1, \ldots, K]$. The vector $\mathbf{z} = (z_i)$, where $i \in [1, \ldots, I]$, is called the allocation vector (fig. 1). Note that in the original version of the CAT model, the mixture of profiles is described as a Dirichlet process (DP, Ferguson 1973; Antoniak 1974), which is implemented following the Neal's (2000) incremental algorithm. This implies that the number K of categories is a free parameter of the inference. In this work however, for computational reasons, we opt for a fixed number K. The present version of the CAT component is therefore akin to a classical mixture model.

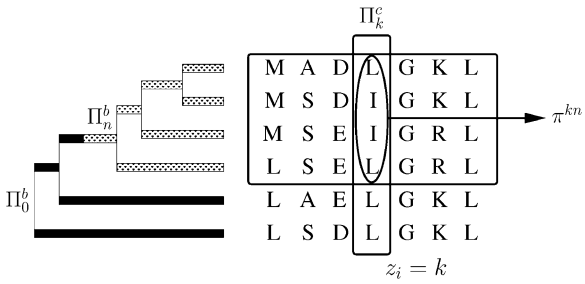

FIG. 1.—A realization of the nonstationary mixture process. A site i is allocated to category k, of profile Π^{c}_{k}. One break point has been created, splitting the tree into 2 areas under the influence of 2 distinct modulators, Π^{b}_{0} (black area) and Π^{b}_{n} (gray area, here $n = 1$). Profile Π^{c}_{k} and modulator Π^{b}_{n} are combined into a site- and lineage-specific set of stationary probabilities π^{kn}.

Second, the model allows for variations along lineages of the global amino acid composition. This is done using the compound stochastic process formalism (Huelsenbeck et al. 1999; Blanquart and Lartillot 2006). The idea is to define a piecewise constant stochastic process, running along the branches of the tree, and taking values in the space of positive normalized vectors of size S. This stochastic process emits events along the tree, each of which "modulates" the K substitution processes defined by the CAT mixture.

Specifically, a random number N of discontinuities, or break points, appear along the tree according to a Poisson process of rate λ. These break points define a partition of the tree into $N + 1$ distinct areas. A modulator, that is, a positive normalized vector Π^{b}_{n}, of size S, is then specified in each of the areas indexed by n (fig. 1). Let $\Pi^{b} = (\Pi^{b}_{n})$, where $n \in [0, \ldots, N]$, denote the set of the $N + 1$ modulation vectors. By convention, we will give each modulator the index of the break point n defining the lower boundary (i.e., closer to the root) of its operating area, and we set Π^{b}_{0} to be the modulator active immediately after the root of the tree. Two sets of parameters are used to describe break point coordinates along the tree: the set $\mathbf{y} = (y_n \in [1, \ldots, 2J - 2])_{n \in [0, \ldots, N]}$ yields the indices of branches supporting the $N + 1$ break points and $\mathbf{x} = (x_n \in [0, \ldots, 1])_{n \in [0, \ldots, N]}$ denotes the break points' relative coordinates on their respective branches. Hence, the overall modulation process

can be described as the combination of a Poisson process emitting events along the tree and at each event, a discrete stochastic shift of the modulation vector, from its current value Π^{b}_n to a new value Π^{b}_m. In particular, in the case where λ tends to 0, N also tends to 0 (i.e., no break point created along the tree), and one obtains a single area along which the default root modulator Π^{b}_0 is applied. The model therefore reduces itself to a time homogeneous one.

We then have, on one hand, site-specific profiles that are constant along the whole tree and on the other hand time specific modulators that are applied for all sites. Profiles and modulators now have to interact with each other in order to define site- and lineage-specific stationary probabilities. Specifically, the stationary probability vector π^{kn} of the substitution process of the mixture category k, in the area n of the tree, is obtained through a multiplicative combination of the profile Π^{c}_k with the modulator Π^{b}_n (fig. 1)

$$\pi^{kn}_s = \frac{1}{Z}\Pi^{\mathrm{c}}_{ks}\Pi^{\mathrm{b}}_{ns}, \forall s \in [1, \ldots, S], \tag{5}$$

where Z is a normalization factor

$$Z = \sum_{s=1}^{S} \Pi^{\mathrm{c}}_{ks}\Pi^{\mathrm{b}}_{ns}.$$

All the K profiles are therefore simultaneously modified by a given modulator n. This combination yields a total of $K(N+1)$ vectors of stationary probabilities $\boldsymbol{\pi} = (\pi^{kn})$, where $k \in [1, \ldots, K]$ and $n \in [0, \ldots, N]$, which are then used to compute the finite-time transition probabilities (eq. 4), along the relevant areas and for the relevant sites. The resulting model is thus heterogeneous both across sites and along lineages. Note that, unlike in the original versions of the CAT and BP models, the vectors associated to the K categories and to the $N+1$ modulators, although positive and normalized over the S states, are not themselves the stationary probabilities of any stochastic process.

The multiplicative rule used here (eq. 5) has several advantages. First, it allows one to create $K(N+1)$ processes out of only $K+N+1$ set of state frequencies, implying $19(K+N+1)$ free parameters. Second, a given category of the mixture will modulate its stationary probability vector while keeping a certain biochemical identity throughout the tree. And finally, upon traversing a break point, all the categories change in a coherent fashion: Each amino acid has its stationary probability either increased in all categories or decreased in all of them.

As explained in Galtier and Gouy (1998), because the stationary assumption does not hold, we can not assume that the stationary probabilities from which to draw the sequence at the root of the tree (i.e., $\pi^{k\infty}$, where ∞ stands for the infinite time elapsed before the root) should be equal to those of the substitution process starting at this point of the tree (i.e., π^{k0}). Therefore, we should normally define an additional modulator immediately upstream of the root, which would be denoted as Π^{b}_∞, and combine it with the profiles of the categories, so as to obtain the stationary probabilities $\pi^{k\infty}$ from which to draw the root sequence. On the other hand, the model, such as specified until now, is invariant by some particular transformations of the vertical Π^{c} and the horizontal Π^{b} components. Specifically, simultaneously multiplying the lth entry of the K profiles Π^{c}_k by a factor ζ, dividing the lth entry of the $N+1$ modulators Π^{b}_n by the same factor ζ, and renormalizing all profiles and modulators, will leave the $K(N+1)$ stationary probability vectors totally invariant. To alleviate this problem, we can fix one of the modulators to be "flat." Our choice is to consider Π^{b}_∞ as the flat modulator, which amounts to directly draw the sequence at the root from the profiles of the categories:

$$\pi^{k\infty} = \Pi^{\mathrm{c}}_k, \quad \forall k \in [1..K].$$

In summary, a realization of the model results into K categories and $N+1$ modulators, their associated profiles Π^{c} and modulation vectors Π^{b}, and their distribution, respectively, across sites $\mathbf{z}$ and over lineages $\mathbf{y}$ and $\mathbf{x}$.

The probability of the data, given particular values of the parameters mentioned previously (or likelihood), is computed using the dynamic programming pruning algorithm (Felsenstein 1981). This algorithm was adapted so as to propagate vectors of partial (conditional) likelihoods along each piecewise constant area of the tree, using the relevant π^{kn} probability vector (eq. 5).

Nonparametric Prior of the CAT Component

Profiles, and their mixture among sites, are drawn from the following hierarchical prior distribution. First, profiles Π^{c}_k are drawn independently and identically distributed (i.i.d.) from a base distribution $G^{\mathrm{c}}_0()$, which is a generalized Dirichlet parameterized by a center ϕ^{c} on the S-dimensional simplex and a concentration μ^{c} around that center:

$$\Pi^{\mathrm{c}}_k \sim \mathrm{Dirichlet}_{\mu^{\mathrm{c}}\phi^{\mathrm{c}}_1, \ldots, \mu^{\mathrm{c}}\phi^{\mathrm{c}}_S}().$$

In practice, this is done by drawing independent gamma variables and renormalizing them (Lartillot et al. 2007)

$$\Pi_s \sim \gamma_{\mu^{\mathrm{c}}\phi^{\mathrm{c}}_s, 1}(), \tag{6}$$

$$\Pi^{\mathrm{c}}_{k,s} = \frac{\Pi_s}{\sum_{s=1}^{S}\Pi_s}, \tag{7}$$

where $\gamma_{\mu^{\mathrm{c}}\phi^{\mathrm{c}}_s, 1}()$ is a standard gamma distribution of shape parameter $\mu^{\mathrm{c}}\phi^{\mathrm{c}}_s$. The prior density of a profile is then

$$p(\Pi^{\mathrm{c}}_k | \mu^{\mathrm{c}}, \phi^{\mathrm{c}}) = \frac{\Gamma\left(\sum_{s=1}^{S}\mu^{\mathrm{c}}\phi^{\mathrm{c}}_s\right)}{\sum_{s=1}^{S}\Gamma(\mu^{\mathrm{c}}\phi^{\mathrm{c}}_s)}\prod_{s=1}^{S}\left(\Pi^{\mathrm{c}}_{ks}\right)^{\mu^{\mathrm{c}}\phi^{\mathrm{c}}_s - 1}, \tag{8}$$

where $\Gamma()$ is the generalized factorial function.

Second, the allocations $\mathbf{z} = (z_i)_{i \in [1, \ldots, I]}$ of the sites to the K profiles are drawn i.i.d. from a multinomial of weight parameters $\mathbf{w} = (w_k)_{k \in [1, \ldots, K]}$; the weights are in turn considered as uniformly distributed:

$$z_i \sim \mathbf{w}$$

$$\mathbf{w} \sim \mathrm{Uniform}(). \tag{9}$$

In fact, the weight vector **w** can be integrated out analytically and the prior marginal probability of a particular allocation **z** is then

$$\begin{aligned} p(\mathbf{z}) &= \int_{\mathbf{w}} p(\mathbf{z}|\mathbf{w}) p(\mathbf{w}) \mathrm{d}\mathbf{w}, \\ &= (K-1)! \int_{\mathbf{w}} \prod_{k=1}^{K} w_k^{\eta_k} \mathrm{d}\mathbf{w}, \\ &= \frac{(K-1)!}{(K+I-1)!} \prod_{k=1}^{K} \eta_k!, \end{aligned} \tag{10}$$

where η_k is the number of sites the category k is allocated to $\left(\sum_{k=1}^{K} \eta_k = I\right)$.

Conversely, if needed, the weight vector **w** may be recovered, that is, resampled conditionally on a realization of **z**, by drawing it from a Dirichlet distribution of parameters $\eta_k + 1$ (eqs. 6 and 7). Conditional on this particular realization of **w**, the likelihood at a given site can then be summed over all possible values of z:

$$p(C_i|\theta, \mathbf{w}) = \sum_{k=1}^{K} w_k p(C_i|\theta, z_i = k), \tag{11}$$

where θ collectively denotes all the ever mentioned parameters. Note that in practice, we never use equation (11), except for computing the cross-validation scores (see below).

Nonparametric Prior of the BP Component

Each modulators Π_n^{b} are drawn i.i.d. from a base distribution $G_0^{\mathrm{b}}()$, which again is a generalized Dirichlet distribution (eq. 8), of concentration μ^{b} and center ϕ^{b} (distinct from μ^{c} and ϕ^{c}). The joint prior of a break point configuration is

$$p(N, \mathbf{x}, \mathbf{y}|\epsilon) = \left(\frac{\mathrm{e}^{-\epsilon}\epsilon^N}{N!}\right)\left(\frac{N!}{T^N}\prod_{j=1}^{2I-2} t_j^{N_j}\right),$$

where N_j denotes the number of break points on branch j, T the total length of the tree, and $\epsilon = \lambda T$ the rescaled rate of the Poisson process of break point creation (for details, see Blanquart and Lartillot 2006). This consists in the product of the prior probability of the number of modulator N (first factor), and given N, of the prior density of their distribution over the tree, as specified by **x** and **y** (second factor).

Priors on Parameters and Hyperparameters

All other model parameters and hyperparameters have canonical priors. We use uniform priors over topologies (τ) as well as for hyperparameters ϕ^{c} and ϕ^{b}, truncated and arbitrary bounded uniform priors for positive hyperparameters μ^{c}, μ^{b}, and ε, a hierarchical exponential prior of mean $\frac{1}{\beta}$ for branch lengths (**t**) and an exponential prior of mean 1 for hyperparameters α and β (respectively, the prior inverse variance of the distribution of rates across sites and the prior mean branch lengths).

Several alternative prior specifications to the one specified above were checked. All alternative priors on single parameters (i.e., the priors on N, α, β, and ε) yield results similar to those obtained under the current implementation (Supplementary Material online). In contrast, discarding hierarchical priors on population of parameters, such as branch lengths (hyperparameterized with β), or profiles and modulators (hyperparameterized with μ and φ), and instead describing each individuals by an uniform prior, has much more impact on the model inferences. In particular, discarding the 2 hierarchical nonparametric priors on Π^{c} and Π^{b} leads to a model lacking flexibility and reducing itself to homogeneous settings (i.e., all posterior probability concentrated on $N = 0$, for details, see Supplementary Material online).

MCMC Sampling

Let θ collectively denote the set of all the parameters of the model. By Bayes theorem, the posterior probability $p(\theta \mid D, M)$ is proportional to the prior $p(\theta \mid M)$ times the likelihood $p(D \mid \theta, M)$:

$$p(\theta|D, M) = \frac{p(D|\theta, M) p(\theta|M)}{p(D|M)},$$

where D denotes the data, M a given model, and θ a particular realization of its associated parameters.

In order to obtain a sample from the posterior distribution of θ, we use the MCMC sampling method, based on the Metropolis–Hasting algorithm. We had previously implemented the nonstationary model BP into the "PhyloBayes" software (Lartillot and Philippe 2004, http://www.lirmm.fr/mab/article.php3?id_article=329), which already provides a MCMC implementation of the CAT model. The MCMC kernels and Hastings ratios involved in these 2 models have been described previously (Lartillot and Philippe 2004; Blanquart and Lartillot 2006). The CAT–BP model is obtained by simply merging the CAT and the BP components. Essentially, we only need to modify our likelihood computation module, so as to use the stationary probability vectors defined by equation (5). Otherwise, no other MCMC update mechanism, in addition to those defined for either CAT or BP, is needed to sample from CAT–BP.

In their original versions, both CAT and BP involve transdimensional update mechanisms, allowing them to adjust their dimensionality according to the data. In the BP model, the number N of modulators is allowed to fluctuate, through an update mechanism that either creates a new modulator somewhere in the tree or deletes an already existing one. Likewise, the "Switch-Mode" update mechanism of CAT allows reallocation of each site either to one of the available categories of the mixture or to a newly created category. It thus simultaneously updates **z** and K.

In the CAT–BP combination introduced here, we keep the reversible jump component of the BP model, but concerning the mixture of the CAT component, we opt instead for a fixed dimensionality. This requires to specify K a priori, but this alleviates some difficulties of the MCMC sampling in variable-dimensional space. The CAT–BP model will be made more general in a subsequent version. To determine the value of K to be used in practice, we first run the CAT model on the data set of interest, recover the estimated

posterior mean number of components of the DP $\tilde{K}$, and fix K to this estimated value.

Accordingly, we have to simplify the original version of the Switch-Mode update mechanism, so as to keep K fixed and update the allocation vector $\mathbf{z}$ only. This is still done using a Gibbs sampler but, following equation (10) (and upon simplification of common factors), the distribution from which the allocation variable is drawn is now:

$$z_i' \sim (\eta_k + 1)p(C_i|\theta, z_i = k, \pi^{k*}), \quad \forall k = 1, \ldots, K,$$

where η_k is the occupation number of category k (the site being reallocated has first to be excluded from the mixture, thus, $\sum_{k=1}^{K} \eta_k = I - 1$, and not I), and π^{k*} stands for all the $N + 1$ modulations along the tree of the profile indexed by k.

Otherwise, Dirichlet resampling mechanisms, as described in Larget and Simon (1999), are used to update the shapes of all normalized vectors of parameters (Π^c, Π^b, ϕ^c and ϕ^b). The hyperparameters ϵ, μ^c, and μ^b are updated using multiplicative update mechanisms. Three different topological update mechanisms keeping track of the break point distribution were introduced (Blanquart and Lartillot 2006), a global SPR move, a local node-sliding move, and a move of the root position. All other parameters (i.e., branch lengths [$\mathbf{t}$] and its associated hyperparameter β, and the shape parameter α of the Gamma-distributed rates across sites) were updated as in Lartillot and Philippe (2004).

MCMC Settings

Chains were run for 5, 000 cycles, each cycle resulting in a saved sample, the first 3,000 samples were discarded as the "burn-in." A cycle is defined by several calls to all available update mechanisms, with tuning parameters empirically chosen such as to reach acceptance ratios from 30% to 90%. Each experiment was run twice in order to check MCMC convergence (see Supplementary Material online, part 2).

Cross-Validations

We compared the fit of alternative models by k-fold cross-validation (CV, Smyth 2000; Lartillot et al. 2007). The method consists in estimating the predictive power of a model M on part of the data, D_{test}, after having learnt the parameters on the other part D_{learn}, where D_{learn} and D_{test} result from a random split of the data set D. By definition, in k-fold CV, D_{test} amounts to a fraction of $1/k$ of the total number of aligned positions.

Formally, the predictive power is expressed as the probability of the test set D_{test} given the learning set D_{learn}:

$$p(D_{\text{test}}|D_{\text{learn}}, M) = \int_{\theta} p(D_{\text{test}}|\theta, M)p(\theta|D_{\text{learn}}, M)\mathrm{d}\theta.$$

This marginal probability is thus the expectation of the likelihood conditional on D_{test}, over the posterior under D_{learn}. In the following, we will call it the cross-validation likelihood. This expectation can be approximated by MCMC: one first gets A samples $(\theta^{(a)})_{a \neq \text{in } [1, \ldots, A]}$ from the posterior distribution under D_{learn}, $\theta^{(a)} \sim p(\theta \mid D_{\text{learn}}, M)$ and then averages the likelihood $p(D_{\text{test}} \mid \theta^{(a)}, M)$ over the A samples.

In our implementation, the likelihood of a column i is conditional on its allocation variable z_i. However, we have no prior knowledge about the allocation status of the sites of the test set D_{test}. Thus, the cross-validation likelihood needs to be integrated over $\mathbf{z}$. This in turn requires that, for each of the A samples, we draw a value for the prior weights of the K categories from a Dirichlet distribution of parameters $\eta_k + 1$ (eqs. 6 and 7). Given the weights $\mathbf{w} = (w_k)_{k \in [1, \ldots, K]}$, the cross-validation likelihood at column C_i of the test set D_{test} is given by equation (11). Taking the product over all sites of the test set, and averaging over the A samples, yields an estimate of the cross-validation likelihood score.

For a given random split, we then take as the CV score of a given model M the difference in cross-validation log likelihood between model M and a model of reference M_0:

$$\mathrm{CV}(M, D_{\text{test}}, D_{\text{learn}}|M_0) = \ln p(D_{\text{test}}|D_{\text{learn}}, M) - \ln p(D_{\text{test}}|D_{\text{learn}}, M_0).$$

Ideally, we want the expectation of this score over all possible random splits. In practice, the procedure is repeated on a series of N such random splits (here $N = 10$) and the score is averaged over these N replicates.

Posterior Predictive Tests

The posterior predictive method (Meng 1994; Gelman et al. 1996) allows one to investigate the ability of the models under study to capture some potentially important features of the observed data. Given a test statistic ω, the method consists in comparing the observed value $\omega^o = \omega(D)$ to the distribution $\omega^s = \omega(D^s)$ of ω under replicates D^s simulated from the posterior predictive distribution:

$$D^s \sim p(D^s|D, M),$$

or equivalently

$$D^s \sim p(D^s|\theta),$$

where $\theta \sim p(\theta \mid D, M)$ is a sample from the posterior distribution under model M, given the true data D. Following the MCMC paradigm, the posterior predictive distribution may be simulated by first drawing a set of A samples $(\theta^{(a)})_{a \in [1, \ldots, A]}$ from the posterior distribution and then use each sample $\theta^{(a)}$ to simulate a replicate, D^s_a.

The posterior predictive distribution of ω^s plays the role of the null distribution: If the observed value $\omega^o = \omega(D)$ falls within it, this means that the observed data do not reject the model. On the opposite, if the observed value is too extreme compared with the null distribution, the model is rejected. Posterior predictive testing depends on the chosen test statistic. The interpretation is that neither is the model able to reproduce the observed statistic ω nor will it correctly anticipate the impact that the features captured by ω may have on the inference.

The deviation of the observed value ω^o to the null distribution can be assessed by evaluating the P value (i.e., the number of replicates D^s leading to a value $\omega(D^s)$ more extreme than ω^o). However, in practice, P values obtained by MCMC

simulation have a low dynamical range and do not allow one to discriminate between strongly rejected models. As an alternative, the discrepancy between the observed value ω^o and the null distribution can also be expressed by a Z score Z_ω, a score of 1 representing a distance equal to the null distribution's standard deviation. As a rule of thumb, assuming the null distribution is normal, a Z score of 1.65 would correspond to a P value of 0.05 and a Z score of 3 to a P value of 0.001.

We used 4 different statistics. First, we devised 2 test statistics, respectively, accounting for site-specific biochemically restricted diversity (ω_d) and taxon-specific compositional biases (ω_b). The diversity at site i, denoted as n_i, is defined as the number of distinct observed amino acids. Averaging n_i over all sites yields the mean site diversity ω_d

$$\omega_d = \frac{1}{I}\sum_{i=1}^{I} n_i. \quad (12)$$

Concerning compositional biases, let $\mathcal{F}(\mathcal{D})$ and $\mathcal{F}(\mathcal{D}_j)$ denote the empirical amino acid distributions measured, respectively, over the whole data set and for a given taxon j. An estimate of the compositional bias of taxon j can be obtained by summing the square of the differences between $\mathcal{F}(\mathcal{D})$ and $\mathcal{F}(\mathcal{D}_j)$. Then, taking the maximum difference over all taxa yields the global bias ω_b

$$\omega_b = \mathrm{MAX}_{j\in[1..J]}\left(\sum_{s=1}^{S}\left(\mathcal{F}(D_j)_s - \mathcal{F}(D)_s\right)^2\right). \quad (13)$$

The 2 other test statistics depend on the sequences at some internal nodes of the tree, and thus cannot be directly deduced from observed data or replicates thereof but require first that we reconstruct a full substitutional history Ξ along the tree, which we did using stochastic mapping (Nielsen 2002). Specifically, given a parameter value θ drawn from the posterior distribution, $\theta \sim p(\theta \mid D)$, stochastic substitution mappings can be simulated conditional on the observed data D:

$$\Xi^o \sim p(\Xi|\theta, D).$$

The distribution of the statistic of interest over such simulated values of Ξ^o will yield the so-called "observed" distribution (strictly speaking, it is inferred, but based on the observed data). Alternatively, the algorithm can be run unconstrained, without trying to fit the data observed at the leaves, in which case the mapping is simulated conditional on the parameter value only:

$$\Xi^s \sim p(\Xi|\theta).$$

The distribution of the statistic over unconstrained mappings Ξ^s will yield the null (posterior predictive) distribution. In practice, given each of the sampled parameter value θ, one constrained Ξ^o and one unconstrained Ξ^s mappings were drawn.

The 2 test statistics based on mappings are conditional on a prespecified topology and focus on some clades of interest. They are defined as the number of homoplasies (convergences) and the number of synapomorphies (shared derived characters). Specifically, let $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ denote 4 internal branches, each defining a "clade" (in the unrooted sense of the term). Under the specified topology, clade $\mathcal{A}$ clusters with $\mathcal{B}$, and $\mathcal{C}$ with $\mathcal{D}$. In this context, for a given site i, if the states $\sigma_\mathcal{A}$ and $\sigma_\mathcal{B}$ inferred by the substitution mapping Ξ at the base of clades $\mathcal{A}$ and $\mathcal{B}$ are equal to some state X, and the states $\sigma_\mathcal{C}$ and $\sigma_\mathcal{D}$ at the base of clades $\mathcal{C}$ and $\mathcal{D}$ are equal to some state Y different from X, this defines an apparent synapomorphy (fig. 2*A*):

$$\mathcal{S} = ((\sigma_\mathcal{A} = \sigma_\mathcal{B}) \neq (\sigma_\mathcal{C} = \sigma_\mathcal{D})). \quad (14)$$

Conversely, if $\mathcal{A}$ and $\mathcal{C}$ (or $\mathcal{A}$ and $\mathcal{D}$) have a matching state X, and $\mathcal{B}$ and $\mathcal{D}$ (or $\mathcal{B}$ and $\mathcal{C}$) a same state Y, this defines a homoplasy (fig. 2*B*):

$$\mathcal{H} = ((\sigma_\mathcal{A} = \sigma_\mathcal{C}) \neq (\sigma_\mathcal{B} = \sigma_\mathcal{D})). \quad (15)$$

Given 4 clades of interest, the test statistics ω_s and ω_h, respectively, represent the overall number of apparent synapomorphies $\mathcal{S}$ and homoplasies $\mathcal{H}$ obtained over all sites under a given mapping.

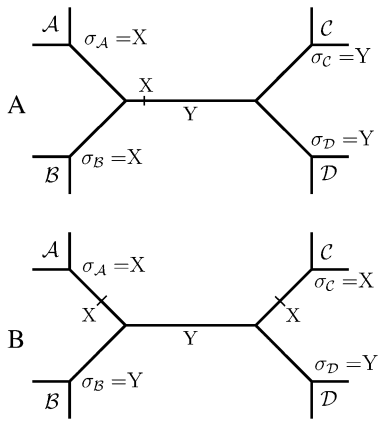

FIG. 2.—Illustration of synapomorphic (fig. 2*A*) and homoplasic (fig. 2*B*) site patterns. (*A*) A single substitution can explain the data observed at the leaves, which defines an apparent synapomorphy. (*B*) Two independent substitutions are needed to explain the site-pattern, which results in an unambiguous homoplasy.

Posterior Estimates Based on Substitution Mappings

Two additional estimations are extracted from posterior stochastic mappings. First, the stochastic **Q** matrices of the Markovian processes of substitution being obtained as described in equation (2), they are unnormalized and consequently branch lengths are not directly measured in expected number of substitution per site. In this situation, effective branch lengths are obtained a posteriori, using stochastic mappings sampled from the posterior distribution, simply by averaging over sites the number of substitutions H_{ij} mapped on branch j at site i:

$$t_j = \frac{1}{I}\sum_{i=1}^{I} H_{ij}.$$

Second, substitution mappings allow us to compute the saturation inferred by each model on the data set being investigated. Saturation can be defined as the number of

independent substitution events toward the same amino acid at a given site (thus, convergences or reversions). It can be visualized by plotting, for each position, the number n_i of distinct amino acid observed at site i as a function of the total number of inferred substitutions $H_i = \sum_{j=1}^{2J-2} H_{ij}$. Specifically, for site i, the number of substitution H_i inferred under a model is averaged over stochastic mappings drawn from the posterior distribution.

Material

We used a data set introduced in a previous work (Delsuc et al. 2005), consisting in an alignment of 4 mitochondrial proteins, COXI, COXII, COXIII, and CYTB, for 20 arthropod species (9 hexapods, 4 crustaceans, 2 myriapods, and 5 chelicerates). The species names, followed by the mitochondrial genomes accession numbers are: Hexapoda: (Diptera) *Anopheles gambiae* [NC_002084], *Drosophila yakuba* [NC_001322], *Chrysomya chloropyga* [NC_002697], (Lepidoptera) *Ostrinia furnacalis* [NC_003368], *Bombyx mandarina* [NC_003395], (Coleoptera) *Crioceris duodecimpunctata* [NC_003372], *Tribolium castaneum* [NC_003081], (Hymenoptera) *Apis mellifera* [NC_001566], and *Melipona bicolor* [NC_004529]; Crustacea: *Pagurus longicarpus* [NC_003058], *Penaeus monodon* [NC_002184], *Daphnia pulex* [NC_000844], and *Triops cancriformis* [NC_004465]; Myriapoda: *Narceus annularus* [NC_003343] and *Thyropygus* sp. [NC_003344]; Chelicerata: *Limulus polyphemus* [NC_003057], *Ixodes hexagonus* [NC_002010], *Ornithodoros moubata* [NC_004357], *Rhipicephalus sanguineus* [NC_002074], and *Varroa destructor* [NC_004454]. This data set was translated, yielding 1,243 aligned amino acid positions. Data, softwares, and "making of" are available on our Web site http://www.lirmm.fr/mab/article.php3?id_article=313.

Results

Free Topology Analyses

Phylogenetic analyses of arthropod complete mitochondrial genomes have yielded surprising results (Nardi et al. 2003), subsequently argued to be likely artifactual (Delsuc et al. 2003, 2005). In the 35 arthropod species phylogeny displayed in the original paper (Nardi et al. 2003), obtained by a ML analysis using a fairly simple model (uniform rates across sites), not only were the hexapods (insects and collembolans) found polyphyletic but also 2 species of bees, *A. mellifera* and *Heterodoxus macropus*, clustered with ticks, among chelicerates. In their subsequent reanalysis, Delsuc et al. (2003) investigated the phylogenetic relationships of the same arthropod species, but this time using a nucleotidic data set gathering the 4 most conserved mitochondrial genes COXI, COXII, COXIII, and CYTB. The first and third codon positions were RY coded (Woese et al. 1991), and the data set was analyzed with a partitioned model, attributing 3 independent substitution models to each codon positions, and accounting for varying rates across sites. This analysis succeeded to recover both insect and chelicerate monophyly. Similar results were later obtained by Delsuc et al. (2005) on the same 4 mitochondrial genes, with a reduced taxon resampling of 20 arthropod species. Investigating this second data set, Delsuc et al. (2005) showed that the ML phylogenetic reconstruction clusters 4 chelicerates among insects, on the branch leading to hymenopterans *A. mellifera* and *M. bicolor*, while the ML analysis of the RY-coded data set succeeds in recovering both insect and chelicerate monophyly.

We focused our experiments on the latter 20 species data set, translated into amino acid sequences (1,243 aligned positions). We first performed phylogenetic analyses of this data set using 4 alternative models: GTR (with both the MrBayes and PhyloBayes implementations), BP, CAT, and CAT–BP. Among them, GTR is time and sequence homogeneous, BP accounts for variations of the stationary probabilities along time, CAT for variations along the sequences, and CAT–BP for both.

The GTR topology reconstructed by MrBayes, as well as other parameter estimates, are identical to the ones obtained by our implementation. As expected, considering previous results (Delsuc et al. 2003; Nardi et al. 2003; Delsuc et al. 2005), the topology estimated by GTR displays an artifactual attraction between chelicerates and hymenopterans. More specifically, GTR produces an erroneous and strongly supported clustering (posterior support of 1) of the 2 bees *Apis* and *Melipona* with the tick *Varroa destructor* (fig. 3*A*). As suggested previously (Delsuc et al. 2005), this result may be attributed to the significantly faster evolution and the similar compositional biases of the artifactually clustered sequences.

The first observation consistent with this is that the 2 branches leading to the bees *Apis* and *Melipona* and to *Varroa* are the longest ones in the estimated tree: respectively 0.63 and 0.39, for a total tree length of 5.48. The second observation bears on the nucleotidic content of the 3 species, measured on the nucleotidic data set to be the most AT rich: 81% AT for *Melipona*, 79% AT for *Apis*, and 75% AT for *Varroa*. The AT richness of *Apis* was shown to induce a bias also at the amino acid level (Foster et al. 1997), which in turn was suggested to have an impact on the phylogenetic estimation accuracy (Foster and Hickey 1999). One may thus expect the GTR reconstruction artifact to be due to a combination of saturation and compositional biases.

However, the BP model also produced the artifact (fig. 3*B*), even though it has been specifically designed to deal with, and shown efficient against, the negative impact of compositional biases on phylogenetic reconstructions (Blanquart and Lartillot 2006). At first, this suggests that the similar compositional biases displayed by the 3 species is not a sufficient stand-alone explanation of the observed artifact.

Similarly, the CAT model, although designed in order to accurately handle site saturation and able to avoid LBA in some cases (Lartillot et al. 2007) also produced the artifactual clade, albeit with a slightly lower posterior probability support of 0.99 (fig. 3*C*). Topological analyses under models CAT and BP thus suggest that the artifactual attraction of bees and ticks cannot be overcome by exclusively accounting for either site-specific features or independent evolution toward similar compositional biases.

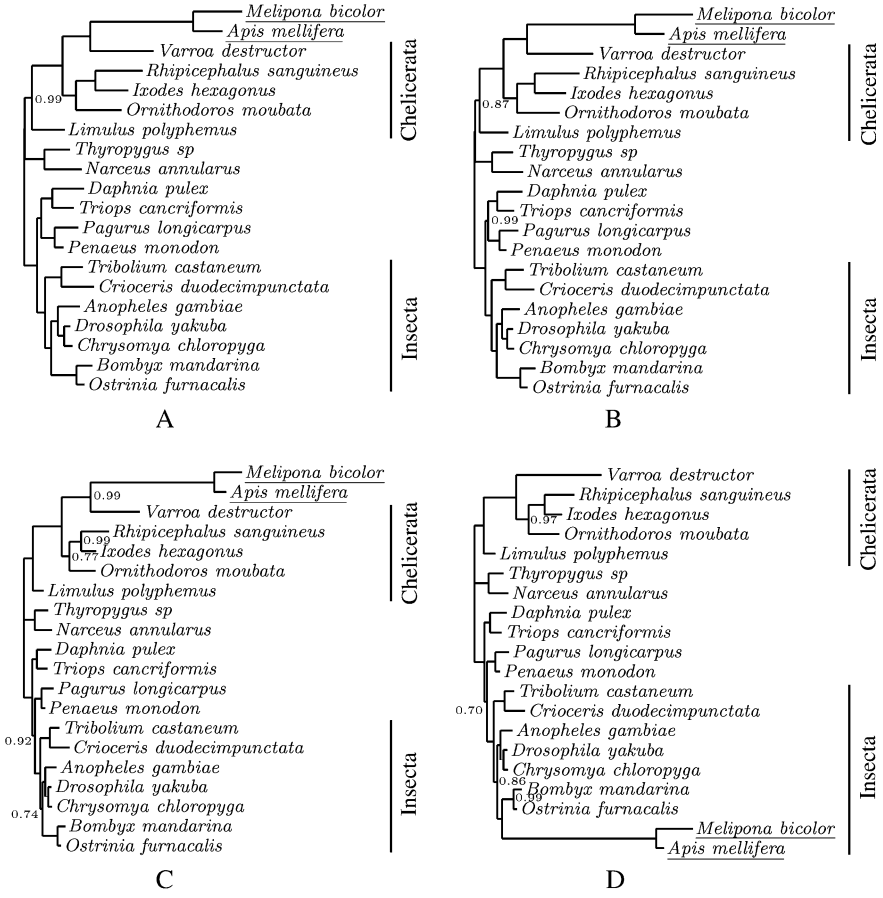

FIG. 3.—Posterior consensus trees obtained under GTR (*A*), BP (*B*), CAT (*C*), and CAT–BP (*D*) (only posterior supports <1 are indicated).

We run the CAT–BP model on this data set, setting $K = 50$, which is approximately the posterior mean number of categories found under the CAT model (i.e., $\tilde{K}=60.1\pm6.4$). The dimensionality of the modulation process is free and is determined by a trade-off between the prior and the data. In the present case, the number of break points converged to $\tilde{N}=6.92\pm0.14$ (Supplementary Material online, table 1).

In contrast to the other 3 models, CAT–BP converged to a much more reasonable phylogenetic tree (fig. 3*D*), recovering the monophyly of insects by clustering the bees with lepidopterans (posterior probability, pp = 1), and the monophyly of ticks (pp = 1). Note that the tree found by Delsuc et al. (2005) slightly differed from ours, in that hymenopterans were found sister group of coleopterans, rather than of lepidopterans, as we found here. Hence, by combining together the respective features of the CAT and the BP models, CAT–BP appears to be able to overcome an artifact that is not correctly detected by either CAT or BP.

Interestingly, the lengths of the 2 branches leading to bees and to *Varroa*, as well as the total tree length, increase as one goes from GTR to BP, CAT, and CAT–BP (table 1), indicating a progressively better detection of saturation. This is confirmed by the saturation graphs (see Methods): GTR yields the largest, and CAT–BP the smallest amount of saturation with BP and CAT in-between (fig. 4). Importantly, compared with GTR, one observes that CAT infers a much higher saturation than BP and also that the increase in saturation observed between GTR and CAT–BP is higher than the sum of the net increases observed between GTR and CAT and between GTR and BP. The latter point suggests a synergistic behavior in the anticipation of saturation resulting from the coupling between CAT and BP.

Table 1
Lengths in Posterior Number of Substitutions per Sites Inferred on the GTR (upper quadrant) and the CAT–BP (lower quadrant) Topologies

	Apis mellifera, Melipona bicolor	Varroa destructor	Tree length
GTR	0.59	0.37	5.09
BP	0.60	0.47	5.38
CAT	1.62	0.71	9.93
CAT–BP	2.05	1.23	10.49
GTR	0.82	0.54	5.40
BP	0.78	0.57	5.41
CAT	2.07	1.13	9.40
CAT–BP	2.23	1.34	11.03

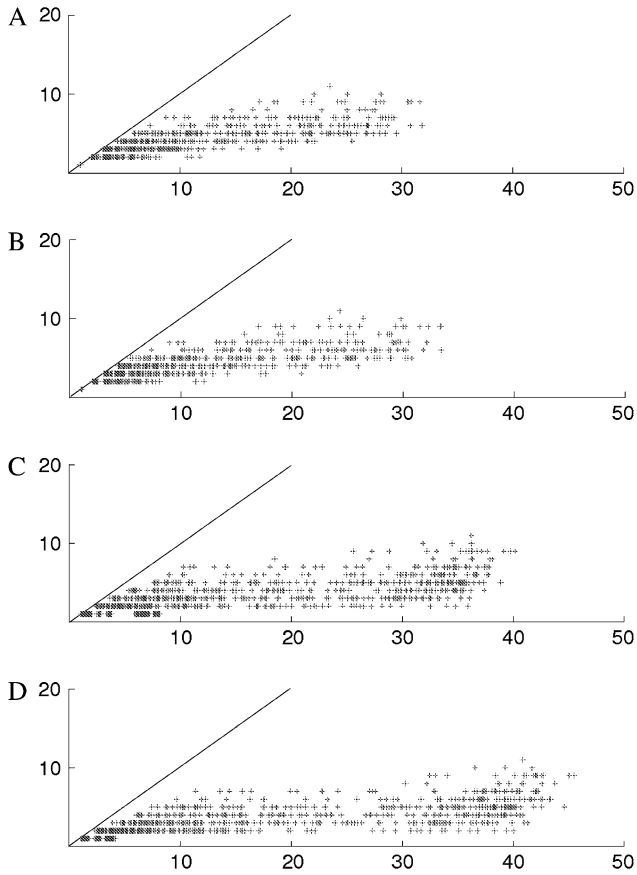

FIG. 4.—Saturation graph obtained under GTR (*A*), BP (*B*), CAT (*C*), and CAT–BP (*D*). For each site, the diversity (i.e., the number of distinct states observed at the corresponding column) is plotted against the posterior mean number of substitution inferred under each model.

Posterior Predictive Assessment of Site- and Taxon-Specific Amino Acid Distributions

The arthropod data set investigated here thus confronts the models with highly saturated sequences and 2 clades (bees and *Varroa*) evolving fast and toward strong and similar compositional biases. The phylogenetic signal present in the sequences may thus be significantly blurred by noise, which the 4 models do not seem to handle in the same way.

In order to understand why they behave differently when confronted to saturated sequences, the 4 models were checked for their ability to account for the site- and taxon-specific distributions of amino acids observed in the data. We used the posterior predictive method, with 2 test statistics, respectively, denoted as ω_d, for site diversity, and ω_b, for compositional biases (see Methods, eqs. 12 and 13). The posterior predictive experiments were performed on the fixed topologies found by CAT–BP and GTR. The results are quite similar, whichever topology is used, and we thus focus in the following on results obtained on the CAT–BP topology.

A protein site usually does not display all the 20 possible amino acids, but rather a small subset generally characterized by similar biochemical properties (Hasegawa and Fitch 1996; Crooks and Brenner 2005). Standard empirical matrices, such as used in the GTR or the BP models, try to implicitly capture site-specific selective effects through the relative exchange rates ρ_{lm} of the substitution process: The exchange rates between biochemically similar amino acids are higher than between unrelated amino acids. However, upon repeated amino acid replacements, the process encoded by the 20×20 matrix rapidly reaches the stationary equilibrium, which in particular implies that saturated sites should display a relatively wide spectrum (high diversity) of amino acids.

In contrast, mixture models such as CAT or CAT–BP explicitly account for site-specific biochemical constraints through a set of *K* F81 processes, each of which is supposed to represent a particular biochemical environment. Importantly, unlike standard models based on one single empirical matrix, this way of encoding the biochemical constraint implies that sites may be saturated and still display very few amino acids, namely, those that have a significant weight according to the relevant profile of the mixture. If the profiles are sparse, which they are most of the time, the site-specific substitution processes are in effect operating over small subsets of the full amino acid alphabet (Lartillot et al. 2007).

Thus, in practice, the 4 models investigated in this work lead to different predictions concerning the mean site-specific amino acid diversity likely to be observed. In particular, on saturated data, CAT and CAT–BP would tend to predict a lower diversity than GTR and BP. Thanks to the posterior predictive formalism, it is possible to check which of these 4 models better predicts the diversity actually observed.

On the arthropod mitochondrial data, the observed mean diversity ω_d^o is of 2.74 distinct amino acids per site. Upon posterior predictive simulations, the GTR and BP models produce a mean diversity of, respectively, 3.35 and 3.42, which corresponds to *Z* scores Z_d of 6.8 and 7.4, respectively. Thus, GTR and BP are strongly rejected by the data concerning their ability to correctly anticipate the observed site-specific biochemical restrictions. In contrast, CAT and CAT–BP both simulate a mean diversity of 2.72, which is very close to the observed statistic $\omega_d^o = 2.74$. The *Z* scores are of 0.3 and 0.2, respectively, and the *P* values (0.3 and 0.4) indicate that CAT and CAT–BP are not rejected (fig. 5). Thus, among the 4 models, CAT and CAT–BP provide a better explanation of the observed site-specific amino acid distributions.

We next checked the aptitude of the models to account for the amino acid compositional variations between taxa. The most compositionally diverged sequence, compared with the overall data set empirical distribution, is *Melipona*, which displays an observed bias statistic ω_b^o of 0.007. The models CAT and GTR involve stationary Markovian processes of substitution assumed to be at equilibrium all along the tree. The probability of observing a given state anywhere in the tree is thus equal to its stationary probability, and those models are therefore expected to simulate homogeneous sequence compositions. Consistent with this, under CAT and GTR, the posterior predictive distribution of the test statistic ω_b^s displays a mean of respectively 0.0006 and 0.0005, which is about 10 times lower than the observed value. The *Z* scores (39.6 and 44.0, respectively)

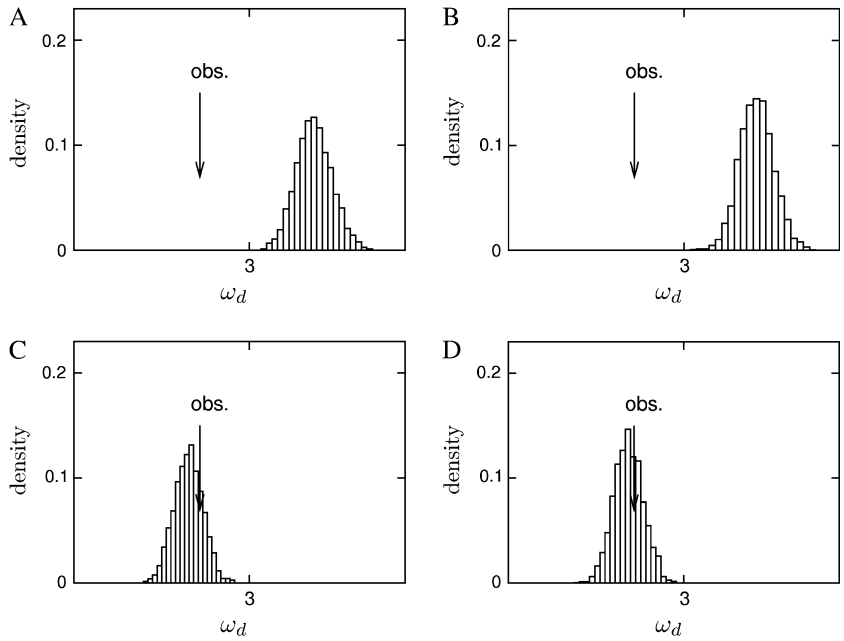

FIG. 5.—Observed value (ω_d^o=2.74, arrow) and posterior predictive distribution of the diversity test statistic ω_d under GTR (*A*), BP (*B*), CAT (*C*), and CAT–BP (*D*).

indicate that both stationary models are strongly rejected concerning their ability to anticipate the compositional biases observed in the data (fig. 6).

In contrast, BP and CAT–BP are explicitly nonstationary and nonhomogeneous. They account for lineages-specific compositional shifts, which allows them to simulate much more heterogeneous sequence compositions compared with stationary models. Accordingly, BP and CAT–BP predict a mean value of 0.005 and 0.004 for the bias statistic, thus much closer to the observed value 0.007. However, the *Z* scores of 2.2 and 3.6 still indicate that BP and CAT–BP are also rejected at the 0.05 level ($Z_\omega > 1.65$, see Methods), albeit far less strongly than are CAT and GTR.

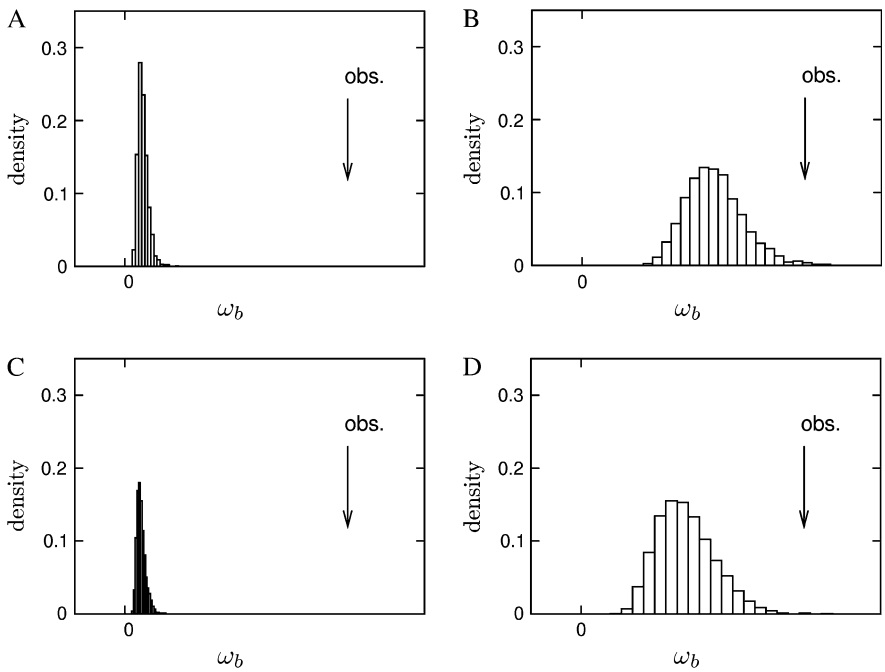

FIG. 6.—Observed value (ω_b^o=0.007, arrow) and posterior predictive distribution of the compositional test statistic ω_b under GTR (*A*), BP (*B*), CAT (*C*), and CAT–BP (*D*).

Altogether, our posterior predictive experiments indicate that a mixture of substitution processes is needed in order to anticipate the observed site-specific biochemical restrictions and, concomitantly, that nonstationarity and nonhomogeneity is required to better anticipate the observed compositional biases. Among the 4 models investigated here, CAT–BP is the only one simultaneously implementing those 2 requirements.

Probability of Homoplasies

The arthropod data set presumably displays conflicting convergent signals between hymenopterans and the chelicerate *Varroa*, which are apparently strong enough to compete with the true phylogenetic signal. In such a context, models that do not correctly anticipate how frequent homoplasies (i.e., convergences and reversions) can be may fail at detecting them and may instead misinterpret them as synapomorphies (i.e., shared derived characters). This could be the reason why GTR, BP, and CAT artifactually cluster bees with ticks. In contrast, the ability of CAT–BP to correctly cluster bees within insects may reflect its better anticipation of the level of homoplasy in the data set.

To investigate this in more detail, we performed a posterior predictive analysis, using as the test statistic the number of convergences between the ancestor of hymenopterans and *Varroa*. Specifically, we constrained the monophyly of insects, by assuming the topology found by CAT–BP and defined the 4 clades ($\mathcal{A}$) hymenopterans, ($\mathcal{B}$) lepidopterans, ($\mathcal{C}$) *Varroa*, and ($\mathcal{D}$) other chelicerates (*Ixodes*, *Ornithodoros*, and *Rhipicephalus*). Under these settings, we simulated substitution mappings, either conditional on the data at the leaves (observed) or not (posterior predictive), and in each case, we recorded the number of substitution mappings for which the states at the base of the 4 clades $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ correspond to a convergent pattern between the bees and *Varroa* (ω_h, eq. 15). For a model to predict significantly less convergences than what is actually observed means that it does not offer any good explanation for these observed convergences and is therefore biased in favor of the artifact.

The observed distribution of the statistic ω_h^o, obtained from mappings constrained by the observed data is very similar across the 4 models, with a mean of approximately 30 convergences (fig. 7*A*). In contrast, the posterior predictive distribution ω_h^s differs significantly between the models. In all 4 cases, it is shifted toward lower values compared with the observed distribution, indicating that the 4 models anticipate less homoplasies in favor of the artifact than they are forced to infer on the true data. Among the 4 models, GTR displays the strongest under-anticipation $(\omega_h^s = 12.42)$. Compared with GTR, the BP model performs slightly better $(\omega_h^s = 15.13)$. The anticipation of the CAT model are much better $(\omega_h^s = 27.53)$. In fact, CAT anticipates nearly twice as many homoplasies than does either GTR or BP, suggesting that the predominant model violation is caused by the nonconsideration of site-specific effects, rather than by compositional effects. Finally, among the 4 models, CAT–BP displays the best anticipation of the occurrence of convergences $(\omega_h^s = 28.40)$, which is consistent with the fact that it recovers a more reasonable topology.

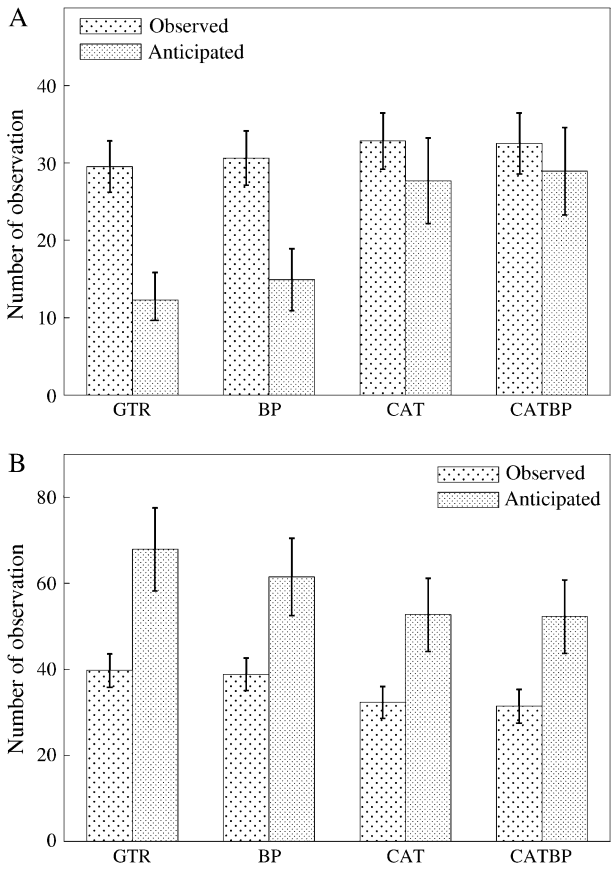

FIG. 7.—Observed and predictive number of homoplasies (*A*) and of apparent synapomorphies (*B*) under GTR, BP, CAT, and CAT–BP.

Note that the observed and predictive distributions of ω_h are not only significantly different under CAT–BP but also under CAT. Yet, compared with CAT–BP, CAT still obtains the artifact, although with a lesser support (pp = 0.99) compared with BP and GTR (pp = 1). In this respect, it should be stressed that the present experiment is just meant for illustrative purposes and does certainly not capture all possible sources of signals in favor of the artifact. Other site patterns apart from those that we have considered here probably have an influence on the phylogenetic estimates obtained under each model.

A similar experiment using as a test statistic, not the number of homoplasies ω_h, but the number of apparent synapomorphies ω_s (equation 14), yields an inverted picture: the expected number of synapomorphies under the posterior predictive distribution, successively under GTR, BP, CAT, and CAT–BP, forms a decreasing series, consistent with the fact that models anticipating more noisy data will expect not only more homoplasies but also less synapomorphies (fig. 7*B*). However, in contrast to what is observed in the case of the number of homoplasies ω_h, even CAT and CAT–BP are rejected for the ω_s statistic, as the posterior predictive mean number of synapomorphies are significantly higher than the observed means. In other words, all models observe much less phylogenetic signal in favor of the correct phylogenetic relationships than what they

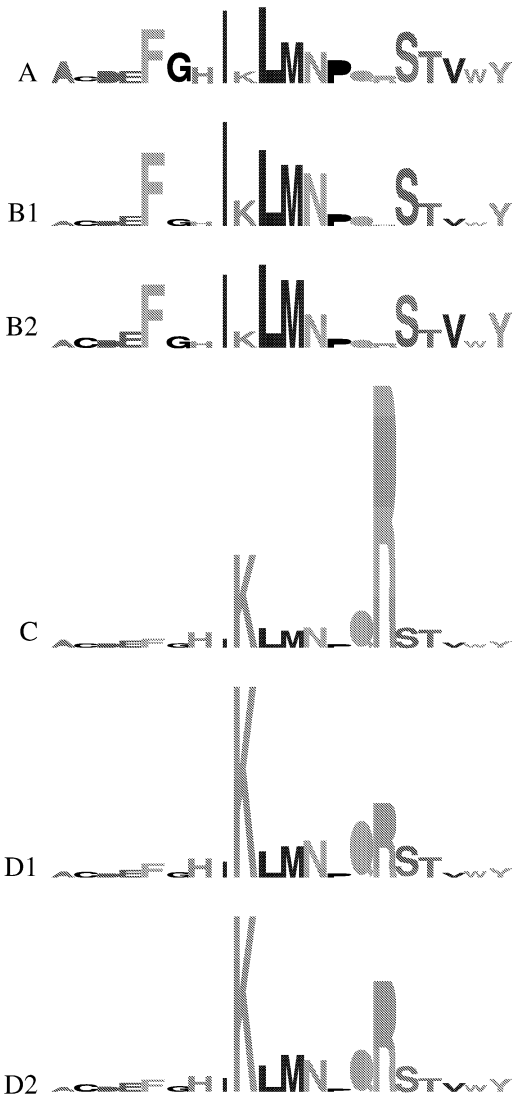

FIG. 8.—Mean stationary probabilities inferred under GTR (*A*) and CAT (*C*), BP (*B*), and CAT–BP (*D*), along the branches leading to the hymenopteran clade (1) and to *Varroa destructor* (2).

would have expected, which is in itself a good indicator of remaining model violations.

Cross-Validations

Finally, the fits of the 4 models were compared by cross-validation (see Methods). The cross-validation scores under each model, and using either the CAT–BP or the GTR topology, were evaluated on 10 random replicated splits of the data set. The GTR model was used as a reference.

The best fit is obtained by the BP model, followed by CAT–BP, GTR, and finally CAT (table 2). The fact that BP is better than GTR, and similarly, that CAT–BP is better than CAT, indicates an improvement brought by the addition of nonstationarity. On the other hand, the worse fit of CAT compared with GTR, and similarly, of CAT–BP compared with BP, seems to indicate that the mixture of Poisson processes is not adequate for this data set. A reason could be that the data set is too small for CAT to be able to infer statistically robust categories. Note also that the order between the 4 models was not stable across the 10 replicates. For instance, BP was found less good than GTR on 1 replicate, CAT better than GTR on 2, and CAT–BP better than BP on 2 out of the 10 replicates. This further indicates that the data set is too small to draw definitive conclusions about the relative merits of the models investigated here.

Table 2
Cross-Validation Scores Using GTR as the Reference under the Fixed CAT–BP and GTR Topologies

	CAT–BP topology	GTR topology
BP	28.6 ± 10.6	25.7 ± 9.3
CAT	–8.8 ± 14.7	–8.5 ± 16.5
CAT–BP	15.8 ± 17.1	17.6 ± 17.2

In conclusion, more experiments on various data sets are needed in order to determine whether or not the CAT–BP model yields a good fit in general, as suggested by all other results obtained in the present study.

Discussion

The Importance of Combining Model Features

We have introduced a new model for protein sequence evolution, which gathers the advantages of both mixture and nonstationary approaches. Models behaviors were investigated using posterior predictive tests, based on 2 statistics, meant to measure how the models account for site- or taxon-specific amino acid distributions. Importantly, the GTR, BP, and CAT models were all rejected by either one or both of these 2 statistics. Concomitantly, all 3 models produced a phylogenetic reconstruction artifact, clustering the most AT biased and fastest evolving species in our arthropod case study. In contrast, the CAT–BP combination performed much better for the 2 posterior predictive tests. It was also the only model able to avoid the phylogenetic artifact. The more sensible phylogenetic tree estimated by CAT–BP, as well as its more accurate anticipations of heterogeneities along time and along the sequence, suggest that simultaneously accounting for site-specific and compositional effects results in a synergistic improvement of the phylogenetic inference.

Interestingly, this synergistic effect implies that, in itself, a phylogenetic artifact cannot be unequivocally attributed to the violation of a particular assumption of the model. For instance, in the present case, it is difficult to say that the attraction of bees and ticks is "caused by" the similar compositional biases of these 2 taxa or that it is "due to" the high level of convergence mechanically implied by the fact that the effective amino acid alphabet at each site is very small.

A more satisfactory interpretation of this problem, as indicated by our posterior predictive experiments, is that CAT and BP tend to produce the same artifact because they both fail at correctly explaining why so many sites support a sister group relationship between bees and ticks; yet, this

failure has a different cause in each case. CAT does not understand that similar compositional biases may cause spurious convergences, whereas BP does not understand that selective restrictions of the amino acid alphabet at a given site imply higher levels of convergent substitution toward the same amino acid at that site. Conversely, this means that both problems have to be correctly dealt with in the frame of one and the same model, if one wants to recover a reasonable phylogeny.

To illustrate more clearly how this synergy works, we focused on the particularly striking case of the columns displaying only K and R. We roughly estimated how each model evaluates the probability of convergent evolution at those columns, in the context of the present artifact. The stationary probabilities of the substitution process inferred by each model, averaged over all K/R sites of the alignment, and on the 2 branches leading to bees and Varroa are shown on figure 8. Compared with the global equilibrium frequency profiles estimated by GTR (fig. 8*A*), the profiles estimated by BP along the 2 branches of interest (fig. 8*B*) are enriched in amino acids such as Isoleucine (I) and Methionine (M), and depleted in Alanine (A) and Glycine (G), which is consistent with the hypothesis that biases at the amino acid level are essentially driven by the GC bias in the present case (Foster et al. 1997). On the other hand, the overall shape of the profile is not fundamentally different between these 2 models, and more importantly, both profiles are flat. In contrast, high stationary probabilities for states K and R are inferred under the mixture models CAT and CAT–BP (fig. 8*C* and *D*), thus accommodating for the underlying selective constraints at those K/R sites. Interestingly, whereas the overall frequency of K over lineages is globally smaller than R, as displayed under CAT, the relation between them is inverted on the 2 branches leading to the AT-rich clades bees and *Varroa*, as displayed specifically under CAT–BP. This is consistent with the fact that K, but not R, is encoded with AT-rich codons. Yet, such an inversion in the relative stationary probability of the 2 amino acids is not observed when comparing the profiles under GTR and GTR-BP (fig. 8*A* and *B*).

From these profiles, one can get a crude estimate of how each model estimates the probability for a K/R site to undergo convergent evolution toward K along the 2 lineages, simply by the square of the stationary probability of K (Lartillot et al. 2007). This amounts to assuming that the 2 branches are infinitely long and that processes are at equilibrium. The resulting probability is about only 0.0004 under GTR, 0.001 under BP, and goes up to 0.03 under CAT, to reach more than 0.1 under CAT–BP. Seen from CAT–BP, this clearly illustrates why only the combination works: If one does not account for nonstationarity, one underevaluates how likely convergent evolution toward K will be by a factor 3 (comparing CAT–BP and CAT). And if one does not model heterogeneities along the sequence, one misses the point by a factor 100 (comparing CAT–BP and BP). In both cases, although for different reasons, the 2 models cannot "understand" why there should be so many convergences toward state K between bees and ticks, which leads them to take this apparent signal as a true signal, and thus, to produce the same artifact.

More generally, in asymmetrical situations such as the present one, which is a typical case of LBA, the apparent signal in favor of the artifactual clustering of the 2 long branches is often stronger in the absolute than the true phylogenetic signal (authentic synapomorphies). In such situations, and provided that the data set is big enough, any under-anticipation of the probability of homoplasies, for whichever reason, will cause the same systematic artifact, where the 2 long branches will be put together. In practice, this implies that models jointly combining all the essential features leading to a better anticipation of phylogenetic noise will be the only one able to correctly deal with such challenging phylogenetic problems.

Comments on the Reconstructed Phylogeny

As already mentioned, RY recoding had been thus far the only method able to recover the monophyly of insects using the nucleotidic version of the data set investigated here (Delsuc et al. 2003, 2005). This may be due to the fact that the RY recoding may jointly alleviate AT bias effects, by recoding purines A and G into the state R and pyrimidine C and T into Y, and the overall saturation, by only considering transversions, whereas eliminating the much more frequent transitions. The resulting effects would then be similar to those described for CAT–BP. In this direction, we tried several recoding schemes at the level of the amino acid alphabet (Rodriguez-Ezpeleta et al. 2007). However, we still obtained the artifactual tree in all cases (not shown). Amino acid recoding thus seems inefficient here. More generally, the divide between phylogenetic noise and phylogenetic signal may be more subtle than what is assumed by amino acid recoding schemes, which may therefore be either inefficient or result in a too extreme loss of phylogenetic information. Accordingly, one should probably prefer the use of CAT–BP, or more elaborate models in the same spirit, entailing the possibility of much more refined interpretations of the data in terms of noise and signal.

Although the phylogeny of *Hexapoda* remains controversial, the 9 insect species investigated here are thought to group monophyletically into the so-called Holometabola phylum. Within this phylum, many studies have argued for the monophyly of Mecopterida, including Diptera and Lepidoptera to the exclusion of Hymenoptera and Coleoptera (Wheeler et al. 2001; Whiting 2002; Castro and Dowton 2005; Delsuc et al. 2005; Savard et al. 2006). Concerning the placement of hymenopterans within Holometabola, no clear consensus has emerged yet. They may be found first emerging species (Castro and Dowton 2005; Savard et al. 2006), sister group of Mecopterida (Wheeler et al. 2001; Whiting 2002; Castro and Dowton 2005) or of Coleoptera (Delsuc et al. 2003, 2005). We indeed recover Diptera, Lepidoptera, Hymenoptera, and Coleoptera as strongly supported monophyletic clades. However, our topology displays very low supports for the positions of those 4 clades among Holometabola and further breaks the assumed Mecopterida monophyly. It should thus be considered with caution.

Future Directions

Intuitively, one would like to interpret the posterior distribution of the modulators along the tree in historical terms. However, the patterns found in the present analysis (see supplementary fig. S1, Supplementary Material online) do not lend themselves to an easy cladistic interpretation, as what was found previously on a simpler case (Blanquart and Lartillot 2006). This observation suggests that, at least under the present version of the model, the break point distribution is more an ad hoc device providing the model with some nonparametric flexibility with respect to compositional heterogeneities than a reliable account of the history of events of compositional drifts along the lineages. Note, however, that this does not prevent the posterior mean modulators along each branch to provide potentially good estimates of the average compositional propensities along the corresponding lineages.

Apart from that, both our posterior predictive experiments and cross-validations show that the CAT–BP model has still quite a few weaknesses. For instance, although BP and CAT–BP have been specifically devised to account for compositional biases, both are slightly rejected by the compositional test ($P < 0.05$), indicating that the stochastic process we have proposed to describe the modulations of the stationary probabilities over time is not sufficiently sophisticated to accurately catch the dynamics of compositional shifts along lineages. Several of its features can be improved. First, modulators are drawn i.i.d. from the G_0^b distribution. As proposed previously (Blanquart and Lartillot 2006), it would be interesting to test first order Markov processes instead, which would allow more frequent but less dramatic changes along with time. The overall dimensionality/granularity of the process could be tuned through the hyperparameters controlling the amplitude of the discrete shifts or the rate of break point creation. In another direction, the Dirichlet prior on modulators itself is questionable and could also be a limiting factor for describing modulator shapes and likely transformations along time. As an alternative parameterization, one could rely on modulators defined in a log-scale, and endowed with a multivariate Gaussian prior. Compared with a Dirichlet, a multivariate Gaussian can be hyperparameterized with much more flexibility (200 degrees of freedom [df] available, compared with the 20 df of the Dirichlet). Finally, instead of using a piecewise constant process, one could use a continuous autocorrelated diffusion process, similar to those used for clock relaxation, but transposed in a multivariate framework. All those alternative specifications are currently under investigation.

We also note that, in spite of its good performances in the homoplasy test (ω_h), the CAT–BP model is rejected by the test measuring the number of apparent synapomorphies (ω_s). More experiments on various data sets are obviously required to confirm this tendency, but this suggests remaining model violations that are likely to have a significant impact on the phylogenetic accuracy, as they directly bear on the model's anticipations concerning the intensity and the structure of primary phylogenetic signal. Many candidate misspecifications can be suggested, which could be responsible for this weakness, among which the possible insufficiencies of the modulation process mentioned above, but also of the mixture model across sites, or to yet other features, like heterotachy, or even variations with time of the site-specific biochemical preferences. The latter 2 phenomena can be modeled using Markov modulated models (Galtier 2001; Holmes and Rubin 2002). Importantly, as was illustrated by this study, such model improvements, all of which have already been proposed in the recent phylogenetic literature, should now be tested in combination with each other, rather than separately, so as to take advantage of potential synergistic effects such as the one observed here between CAT and BP.

Supplementary Material

Supplementary material, figures S1–S11, and tables 1–5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We wish to thank Olivier Gascuel, Hervé Philippe, and Nicolas Galtier for their participation to the discussions over the model and its implementation. We are also grateful to Frédéric Delsuc for having provided the arthropod data set and also to 2 referees for their helpful comments on this manuscript. This work was financially supported by the "60eme commission franco-québécoise de coopération scientifique," by the French Centre National de la Recherche Scientifique, through the ACI-IMPBIO Model-Phylo funding program, and by the French Agence Nationale pour la Recherche, MITOSYS.

Literature Cited

Antoniak CE. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Statistics. 2:1152–1174.

Barry D, Hartigan JA. 1987. Asynchronous distance between homologous DNA sequences. Biometrics. 43:261–276.

Bernardi G. 1993. The vertebrate genome: isochores and evolution. Mol Biol Evol. 10:186–204.

Blanquart S, Lartillot N. 2006. A Bayesian compound stochastic process for modeling non-stationary and nonhomogeneous sequence evolution. Mol Biol Evol. 23:2058–2071.

Bogatyreva NS, Finkelstein AV, Galzitskaya OV. 2006. Trend of amino acid composition of proteins of different taxa. J Bioinform Comput Biol. 4:597–608.

Bollback JP. 2002. Bayesian model adequacy and choice in phylogenetics. Mol Biol Evol. 19:1171–1180.

Boussau B, Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. Syst Biol. 55:756–768.

Bruno WJ. 1996. Modeling residue usage in aligned protein sequence via maximum likelihood. Mol Biol Evol. 13:1368–1374.

Castro LR, Dowton M. 2005. The position of the Hymenoptera within the Holometabola as inferred from the mitochondrial genome of Perga condei (Hymenoptera: Symphyta: Pergidae). Mol Phylogenet Evol. 34:469–479.

Crooks GE, Brenner SE. 2005. An alternative model of amino acid replacement. Bioinformatics. 21:975–980.

Das S, Paul S, Bag SK, Dutta C. 2006. Analysis of Nanoarchaeum equitans genome and proteome composition: indications for hyperthermophilic and parasitic adaptation. BMC Genomics. 7:1–16.
Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 6:361–375.
Delsuc F, Phillips MJ, Penny D. 2003. Comment on "Hexapod origins: monophyletic or paraphyletic?". Science. 301:1482.
Dimmic MW, Mindell DP, Goldstein RA. 2000. Modeling evolution at the protein level using an adjustable amino acid fitness model. Pac Symp Biocomput. 5:18–29.
Felsenstein J. 1978. Cases in which parsimony or compatibility method will be positively misleading. Syst Zool. 27:401–410.
Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 17:368–376.
Ferguson T. 1973. A Bayesian analysis of some nonparametric problems. Statistics. 1:209–230.
Foster PG. 2004. Modeling compositional heterogeneity. Syst Biol. 53:485–495.
Foster PG, Hickey DA. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J Mol Evol. 48:284–290.
Foster PG, Jermiin LS, Hickey DA. 1997. Nucleotide composition bias affects amino acid content in protein coded by animal mitochondria. J Mol Evol. 44:282–288.
Fukuchi S, Yoshimune K, Wakayama M, Moriguchi M, Nishikawa K. 2003. Unique amino acid composition of proteins in halophilic bacteria. J Mol Evol. 327:347–357.
Galtier N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol. 18:866–873.
Galtier N, Gouy M. 1995. Inferring phylogenies from DNA sequences of unequal base composition. Evolution. 92:11317–11321.
Galtier N, Gouy M. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol Biol Evol. 15:871–879.
Gelman A, Meng XL, Stern H. 1996. Posterior predictive assessment of model fitness via realized discrepancies. Stat Sin. 6:733–807.
Gibson A, Gowri-Shankar V, Higgs PG, Rattray M. 2005. A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods. Mol Biol Evol. 22:251–264.
Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 11:725–736.
Gowri-Shankar V, Rattray M. 2007. A reversible jump method for Bayesian phylogenetic inference with a nonhomogeneous substitution model. Mol Biol Evol. 24:1286–1299.
Halpern AL, Bruno WJ. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol. 15:910–917.
Hasegawa M, Fitch WM. 1996. Dating the cenancester of organisms. Science. 274:1750.
Holmes I, Rubin GM. 2002. An expectation maximization algorithm for training hidden substitution models. J Mol Biol. 317:753–764.
Hudelot C, Gowri-Shankar V, Jow H, Rattray M, Higgs PG. 2003. RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences. Mol Phylogenet Evol. 28:241–252.
Huelsenbeck JP, Larget B, Swofford D. 1999. A compound poisson process for relaxing the molecular clock. Genetics. 154:1879–1892.
Jones DT, Taylor WR, Thomton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Bioinformatics. 8:275–282.
Jow H, Hudelot C, Rattray M, Higgs PG. 2002. Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. Mol Biol Evol. 19:1591–1601.
Jukes TH, Bhushan V. 1986. Silent nucleotide substitutions and G + C content of some mitochondrial and bacterial genes. J Mol Evol. 24:39–44.
Kennedy SP, Ng WV, Salzberg SL, Hood L, DasSarma S. 2001. Understanding the adaptation of Halobacterium species NRC-1 to its extreme environment through computational analysis of its genome sequence. Genome Res. 11:1641–1650.
Lake JA. 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Evolution. 91: 1455–1459.
Lanave C, Preparata G, Saccone C, Serio G. 1984. A new method for calculating evolutionary substitution rates. J Mol Evol. 20:86–93.
Larget B, Simon DL. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol Biol Evol. 16:750–759.
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. 7:S4.
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.
Lobry JR, Chessel D. 2003. Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. J Appl Genet. 44:235–261.
Lobry JR, Necsulea A. 2006. Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes. Gene. 30:128–136.
Lockhart PJ, Howe CJ, Bryant DA, Beanland TJ, Larkum AW. 1992. Substitutional bias confounds inference of cyanelle origin from sequence data. J Mol Evol. 34:153–162.
Lockhart PJ, Steel MA, Hendy MD, Penny D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol. 11:605–612.
Meng XL. 1994. Posterior predictive p-values. Ann Stat. 22:1142–1160.
Montero LM, Salinas J, Matassi G, Bernardi G. 1990. Gene distribution and isochore organization in the nuclear genome of plants. Nucleic Acids Res. 18:1859–1867.
Mooers AO, Holmes EC. 2000. The evolution of base composition and phylogenetic inference. Trends Ecol Evol. 15:365–369.
Nardi F, Spinsanti G, Boore J, Carapelli A, Dallai R, Frati F. 2003. Hexapod origins: monophyletic or paraphyletic? Science. 299:1887–1889.
Neal RM. 2000. Markov chain sampling methods for Dirichlet process mixture models. J Comput Graph Stat. 9:249–265.
Nielsen R. 2002. Mapping mutations on phylogenies. Syst Biol. 51:729–739.
Nielsen R, Huelsenbeck JP. 2002. Detecting positively selected amino acid sites using posterior predictive P-values. Pac Symp Biocomput. 7:576–588.
Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol. 20:1692–1704.
Rodrigue N, Lartillot N, Bryant D, Philippe H. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene. 347:207–217.
Rodriguez F, Oliver JL, Marin A, Medina JR. 1990. The general stochastic model of nucleotide substitution. J Theor Biol. 142:485–501.

Rodriguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst Biol. 56: 389–399.

Savard J, Tautz D, Richards S, Weinstock GM, Gibbs RA, Werren JH, Tettelin H, Lercher MJ. 2006. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. Genome Res. 16:1334–1338.

Singer GA, Hickey DA. 2000. Nucleotide bias causes a genome-wide bias in the amino acid composition of proteins. Mol Biol Evol. 17:1581–1588.

Singer GA, Hickey DA. 2003. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. Gene. 317:39–47.

Smyth P. 2000. Model selection for probabilistic clustering using cross-validated likelihood. Stat Comput. 9:63–72.

Tavaré S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lect Math Life Sci. 17:57–86.

Tekaia F, Yeramian E. 2006. Evolution of proteomes: fundamental signatures and global trends in amino acid compositions. BMC Genomics. 7:1–11.

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

Wheeler WC, Whiting MF, Wheeler QD, Carpenter JM. 2001. The phylogeny of the extant Hexapod orders. Cladistics. 17:113–169.

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.

Whiting MF. 2002. Phylogeny of the Holometabolous insect orders: molecular evidence. Zool Scr. 31:69–83.

Woese CR, Achenbach L, Rouviere P, Mandelco L. 1991. Archaeal phylogeny: reexamination of the phylogenetic position of Archaeoglobus fulgidus in light of certain composition-induced artifacts. Syst Appl Microbiol. 14:364–371.

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

Yang Z, Roberts D. 1995. On the use of nucleic acid sequences to infer branchings in the tree of life. Mol Biol Evol. 12:451–458.

Andrew Roger, Associate Editor

Accepted January 20, 2008