A Bayesian Compound Stochastic Process for Modeling Nonstationary and Nonhomogeneous Sequence Evolution

Samuel Blanquart and Nicolas Lartillot

Projet Méthodes et Algorithmes pour la Bioinformatique, LIRMM-CNRS, Montpellier, France

Variations of nucleotidic composition affect phylogenetic inference conducted under stationary models of evolution. In particular, they may cause unrelated taxa sharing similar base composition to be grouped together in the resulting phylogeny. To address this problem, we developed a nonstationary and nonhomogeneous model accounting for compositional biases. Unlike previous nonstationary models, which are branchwise, that is, assume that base composition only changes at the nodes of the tree, in our model, the process of compositional drift is totally uncoupled from the speciation events. In addition, the total number of events of compositional drift distributed across the tree is directly inferred from the data. We implemented the method in a Bayesian framework, relying on Markov Chain Monte Carlo algorithms, and applied it to several nucleotidic data sets. In most cases, the stationarity assumption was rejected in favor of our nonstationary model. In addition, we show that our method is able to resolve a well-known artifact. By Bayes factor evaluation, we compared our model with 2 previously developed nonstationary models. We show that the coupling between speciations and compositional shifts inherent to branchwise models may lead to an overparameterization, resulting in a lesser fit. In some cases, this leads to incorrect conclusions, concerning the nature of the compositional biases. In contrast, our compound model more flexibly adapts its effective number of parameters to the data sets under investigation. Altogether, our results show that accounting for nonstationary sequence evolution may require more elaborate and more flexible models than those currently used.

Introduction

Base composition has been shown to be highly variable among species (Jukes and Bhushan 1986; Montero et al. 1990; Bernardi 1993), a phenomenon generally denoted as compositional biases. When analyzing phylogenetic relationships between species using standard methods, a similar nucleotidic composition is often interpreted as phylogenetic signal, leading unrelated species to be grouped together in the resulting tree (Lockhart et al. 1992, 1994; Lake 1994; Galtier and Gouy 1995; Yang and Roberts 1995; Foster and Hickey 1999; Mooers and Holmes 2000; Foster 2004).

A first way to avoid this problem consists in recoding the character states into functional groups, so as to homogenize the composition between sequences. For instance, the RY coding (Woese et al. 1991) consists in replacing nucleotides A and G by R (purine) and C and T by Y (pyrimidine). In this way, only transversion events are considered, nucleotides A and G, C and T become synonymous and GC biases are removed. As transitions often occur more frequently than transversions (Brown et al. 1982), the RY coding also decreases saturation and this enhances ancient phylogenetic signal. It has been used for resolving deep divergences (Phillips and Penny 2002; Delsuc et al. 2003). An analogous coding system has been proposed for amino acid sequences (Dayhoff coding, Hrdy et al. 2004). More generally, one can accommodate the data by removing saturated sites from the analysis such as third codon positions (Swofford et al. 1996; Delsuc et al. 2002; Canbäck et al. 2004) or fast-evolving sites (Brinkmann and Philippe 1999; Philippe et al. 2000). These methods have not been specifically devised to deal with compositional biases, but assuming that biased sites are generally among the fast-evolving ones, they should nevertheless be efficient against them. Altogether, these filtering and recoding methods have proven quite effective, in particular, for resolving deep divergences (Philippe et al. 2000; Phillips and Penny 2002; Delsuc et al. 2003). On the other hand, they may also result in a loss of phylogenetic information. Moreover, the RY and the Dayhoff coding may not be efficient in all situations. For instance, the RY coding only alleviates GC biases and will not efficiently suppress more general compositional shifts in DNA composition.

All methods mentioned thus far aim at filtering away the compositional bias from the data set but do not question the underlying evolutionary model itself, which is still assumed to be stationary and homogeneous. In contrast, a series of methods based on nonstationary models have been proposed. Among them, there are distance-based methods, such as the "LogDet" (Lockhart et al. 1994), "paralinear" (Lake 1994), or modified "Tamura–Nei" (Tamura and Kumar 2002) distances, and full-likelihood approaches, based on maximum likelihood (Yang and Roberts 1995; Galtier and Gouy 1998) or Bayesian (Foster 2004) frameworks. Some of these methods have been successfully applied, in particular, to studies about ancestral GC contents (Galtier et al. 1999; Tarrio et al. 2001) or to phylogenetic inference from GC-biased sequences (Herbeck et al. 2004).

In the model proposed by Galtier and Gouy, a single parameter is associated to each branch of the phylogenetic tree. This parameter represents the branch-specific GC ratio. The model proposed by Yang and Robert accommodates more general biases as 3 free parameters, handling frequencies for each of the 4 nucleotides, are assigned to each branch. In both models, the values of the parameters are estimated from the data by maximum likelihood. However, associating different compositions to each branch may result in a large amount of free parameters, and problems of overparameterization may then be encountered (Foster 2004). In order to reduce such overparameterization effects, Foster (2004) proposed a model based on a predefined number of clusters of base frequencies, smaller than the number of branches. The base frequencies of each cluster, and the

Key words: phylogeny, MCMC, nonstationary, nonhomogeneous, compositional bias, compound stochastic process.

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

Mol. Biol. Evol. 23(11):2058–2071. 2006
doi:10.1093/molbev/msl091
Advance Access publication August 24, 2006

cluster each branch is associated to, are sampled from their joint posterior distribution by Markov chain Monte Carlo (MCMC, Neal 1993). Such a method may be effective in reducing the number of parameters. On the other hand, it still lacks flexibility as the number of different clusters is fixed to a prespecified number. Ideally, this number should be a free parameter of the phylogenetic inference. Moreover, it does not completely address the problems that are the original cause of the overparameterization of the branchwise models, such as those of Galtier and Gouy or Yang and Roberts. Fundamentally, this overparameterization is a consequence of the fact that the equilibrium frequencies of the substitution process are reassessed at the base of each branch of the tree, whereas in many practical situations, the equilibrium frequencies may have remained constant for time periods spanning several branches, sometimes entire groups. Moreover, in the branchwise models, changes of compositional bias are associated with speciation events, that is, with the nodes of the phylogenetic tree, which is not realistic; speciation events and changes of composition should ideally be uncoupled.

In this direction, an interesting solution was proposed by Huelsenbeck et al. (1999), allowing variations of the substitution rate along lineages, according to the so-called compound stochastic process. Specifically, substitutions between sequence states happen according to a classical first-order Markov process, whose instant rate is described by a second, piecewise constant, stochastic process. Realizations of the second stochastic process are sampled using MCMC methods. The number of rate change events is thus a free variable in the model of Huelsenbeck et al.

Here we propose a nonstationary model, also based on a compound stochastic process, generalizing the models of both Galtier and Gouy and Yang and Robert. As in the model of Huelsenbeck et al., we use an additional stochastic process operating along lineages, but this time, to model compositional shift events. Those events occur independently from speciations, according to a Poisson process, and are thus a free, Poisson distributed variable of our model. Importantly, this approach allows a flexible dimensionality, in contrast to previous proposed nonstationary models. We implemented this model in a Bayesian framework, using the MCMC approach to sample realizations of the compositional shift stochastic process, and applied it to several nucleotidic data sets. In particular, our results show that its free dimensionality leads our model to better fit the data, especially compared with the model proposed by Yang and Robert, which in contrast appears to be penalized by its lack of control of the dimensionality.

Methods

A set of homologous aligned sequences is available in the form of a data matrix of J sequences of K sites. Phylogenetic relationships between the J extant species are represented by a rooted binary tree, denoted as τ, whose nodes represent speciation events. A length is associated to each branch. Let $\mathbf{t} = \{t_1, \ldots, t_{2J-2}\}$, where $(1, \ldots, 2J-2)$ are branch indices, denote the set of branch lengths. Additionally, sites have their own rate of substitution, $\mathbf{r} = (r_1, \ldots, r_K)$, distributed according to a continuous gamma distribution.

Markovian Model of the Substitution Process

Probabilistic models in phylogenetics usually assume that sequence evolution can be seen as a Markovian process. This Markovian process is defined on a state space of size S ($S = 4$ for nucleotide, $S = 20$ for amino acid) and is characterized by a stochastic instantaneous rate matrix, $\mathbf{Q}$, of size $S \times S$. Given a stationary probability vector, π (or "profile"), of size S, and a matrix ρ of relative exchange rates between states, a stochastic matrix Q is obtained as follows:

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

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

At a site k of rate r_k, and along a branch j of length t_j, the evolutionary distance v_{jk} is

$$v_{jk} = t_j r_k.$$

A state l is substituted into a state m in an evolutionary distance v_{jk} with probability:

$$P_{lm}(v_{jk}) = \left[e^{v_{jk}Q}\right]_{lm}. \tag{3}$$

Nonstationary Model of Substitution

Homogeneous models of sequence evolution assume a single Markovian substitution process, defined by a single Q matrix operating along the whole tree. The Markovian process is assumed to be at equilibrium, and thus, the model is stationary. In this article, we design a nonstationary model of sequence evolution. We model shifts of sequence composition along lineages as a compound and piecewise constant stochastic process, defined as follows: discontinuous changes occur according to a Poisson process of rate ε. At each discontinuity point, the profile π of the substitution process (i.e., its stationary probability vector) switches to a new value π', directly drawn from a prespecified distribution G_0 on the simplex:

$$p(\pi'|\pi) = G_0(\pi').$$

We use by default a uniform distribution for G_0. Note that π' is independent from π, consequently the series of successive compositional shifts follows a 0th-order Markov process. The relative exchange rates ρ, which are here considered as free parameters, are kept constant in the whole tree, so that the substitution process is described by a stochastic matrix $Q = \pi\rho$ before, and $Q' = \pi'\rho$ after, the discontinuity point. Thanks to these compositional shift events, our nonstationary model can take into account substitution processes specific to each part of the tree. For example, high stationary probabilities for G and C found in a given lineage will drive an evolution toward a GC richer content.

Ideally, the likelihood of particular values of the parameters (topology, branch lengths, etc.) has to be integrated over all realizations of the stochastic compositional shift process described above. However, computing this

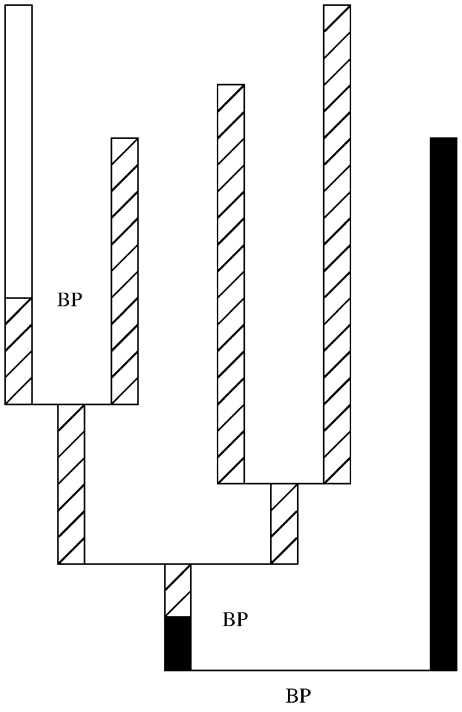

FIG. 1.—Splitting the tree into piecewise constant areas. Three BP are placed on the tree: a default break point placed at the root of the tree, defining the black area, and 2 other break points, defining the hatched and white areas. Each area involves a specific Markovian process of substitution.

integral directly is intractable. We, therefore, use the MCMC approach to sample realizations of this process. We call "break points," or BP, the points at which the discontinuous changes occur (fig. 1). A realization of the process is then defined by the number (N) of break points, each of them being specified by its position in the tree and its associated profile.

Data Structure for the Nonstationary Model

Conditional on a particular tree, of total length T, the number N of break points follows a Poisson process of rate ε. With N break points, plus a default root break point placed at the root of the tree (with index 0), we split the tree into $N + 1$ constant areas (i.e., where the substitution process is homogeneous).

A break point n is entirely characterized by its profile π_n, the lineage where it appears b_n (i.e., the branch onto the break point is placed), and the point along the branch at which it appears x_n (i.e., the relative position of the break point on its branch b_n: $0 < x_n < 1$), where $0 \leq n \leq N$. By extension, for the particular root break point, we defined $b_0 = 0$, $t_0 = 0$, and $x_0 = 0$. Let us denote by N_j the number of break point on the jth branch. Thus, N is equal to $\sum_{j=1}^{2J-2} N_j$, where J is the number of taxa and $2J - 2$ the number of branches.

Given the global set of relative exchange rates ρ and a break point profile π_n, one can compute the Q_n matrix (eqs. 1 and 2). Then, substitution probabilities between states (eq. 3) are computed, in each area of the split tree, using in each case the relevant rate matrix.

Finally, as the stationary assumption does not apply, we cannot assume that stationary probabilities at the root are equal to those of the process at this point of the tree (i.e., π_0). As in Galtier and Gouy, we therefore define an extraparameter π_∞, which represents the stationary probabilities at the root. Altogether, the parameter vector θ of the nonstationary model is written:

$$\theta = \{\tau, \mathbf{r}, \mathbf{t}, \rho, \varepsilon, N, (\pi_n), (b_n), (x_n), \pi_\infty\}.$$

Probability Densities of the Data Structure

The fact that we define break points with relative positions on branches induces a nontrivial prior distribution, which is explained here. Break points appear following a Poisson process of rate μ, as in Huelsenbeck et al. (1999). Thus, the number N_j of break points on branch j, of length t_j, follows a Poisson distribution of mean μt_j:

$$p(N_j) = \frac{e^{-\mu t_j}(\mu t_j)^{N_j}}{N_j!}.$$

Given N_j, all possible distributions of the N_j break points along the branch are equally likely. Denoting one such distribution by $X_j = \{x_1, x_2, \ldots, x_{N_j}\}$, such that $x_1 < x_2 < \ldots < x_{N_j}$ the density of X_j is $p(X_j) = \frac{1}{Z}$, where

$$Z = \int_{0<x_1} \int_{<x_2} \cdots \int_{<x_{N_j}<1} \mathrm{d}x_1 \mathrm{d}x_2 \cdots \mathrm{d}x_{N_j} = \frac{1}{N_j!}.$$

The joint probability density for N_j and X_j is thus:

$$p(N_j, X_j) = \frac{e^{-\mu t_j}(\mu t_j)^{N_j}}{N_j!} N_j! = e^{-\mu t_j}(\mu t_j)^{N_j}.$$

Taking the product over all branches and rearranging the factors yields the prior density of the overall break point distribution:

$$p(N, \mathbf{x}) = \prod_{j=1}^{2J-2} (e^{-\mu t_j}(\mu t_j)^{N_j}) = e^{\left(-\mu \sum_{j=1}^{2J-2} t_j\right)} \mu^{\left(\sum_{j=1}^{2J-2} N_j\right)} \prod_{j=1}^{2J-2} t_j^{N_j}$$

$$= e^{-\mu T} \mu^N \prod_{j=1}^{2J-2} t_j^{N_j} = \left(\frac{e^{-\mu T}(\mu T)^N}{N!}\right)\left(\frac{N!}{T^N} \prod_{j=1}^{2J-2} t_j^{N_j}\right).$$

This last formula can be decomposed into 2 factors: the first factor is the probability of observing N break points on the whole tree, given a Poisson distribution of rescaled rate $\varepsilon = \mu T$. The second is the probability density of a particular distribution of the N break points on the tree. We directly parameterize our model in terms of ε, rather than μ. This allows a more direct interpretation of the results: ε is simply the mean number of break points across the whole tree.

Canonical priors are used for all other model parameters. Specifically, we use a uniform prior over topologies (τ), a Gamma distribution of mean 1 and variance $\frac{1}{\alpha}$ for the rate across sites ($\mathbf{r}$), an exponential prior of mean $\frac{1}{\beta}$ for the branch lengths ($\mathbf{t}$), an exponential prior of mean 1 for relative exchange rate parameters (ρ), and finally, a uniform prior for the profiles (π). The hyperparameters

α, β, and ε are also free parameters of the model, and all follow exponential priors of mean 1.

Likelihood Computation

We can easily adapt the pruning algorithm of Felsenstein (1981) to compute the likelihood given the break points and the base frequencies. In effect, each break point is equivalent to a new node, so that a branch with N_j break points is subdivided into $N_j + 1$ segments. Along each segment, the vector of partial likelihoods is propagated using the relevant Q matrix, which is itself obtained by combining the global set of relative exchange rates with the local base frequencies.

MCMC Sampling

By Bayes theorem, the posterior probability is proportional to the prior times the likelihood:

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

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's algorithm. Applying MCMC to phylogenetic reconstruction problems has been developed by Larget and Simon (1999), Huelsenbeck and Ronquist (2001), and Huelsenbeck et al. (2002). We have implemented the nonstationary model into the software "PhyloBayes" (Lartillot and Philippe 2004), which provides a MCMC implementation in a stationary context. Briefly, at each step of a MCMC, one modifies the current value of parameter vector θ, according to a stochastic kernel $q(\theta, d\theta')$, obtaining a new value θ', which is accepted with probability:

$$p_{\text{accept}}(\theta') = \min\left(1, \frac{p(\theta'|D,M)}{p(\theta|D,M)}\frac{q(\theta', d\theta)}{q(\theta, d\theta')}\right),$$

where $\frac{p(\theta'|D,M)}{p(\theta|D,M)} = \frac{p(D|\theta',M)p(\theta'|M)}{p(D|\theta,M)p(\theta|M)}$ is the ratio of posterior densities, or Metropolis ratio, and $H = \frac{q(\theta',d\theta)}{q(\theta,d\theta')}$ is the Hastings ratio (Neal 1993), that is, the probability of proposing a backward modification on θ' that would exactly reverse the forward modification on θ, divided by the probability of the forward modification. Green (2003) provides a general formula for the Hastings ratio:

$$H = \frac{g'(w')}{g(w)}|J|, \qquad (5)$$

where w and w' are the set of random numbers picked with distribution g and g', when modifying θ into θ', or symmetrically θ' into θ. The second factor $|J| = |\det[\frac{\partial(\theta',w')}{\partial(\theta,w)}]|$ is the absolute value of the Jacobian determinant of the transformation from $\{\theta, w\}$ to $\{\theta', w'\}$. Originally, Green's formula was introduced for dealing with reversible MCMC moves, but as noted by Holder et al. (2005), this formula happens to be useful in much more general MCMC frameworks.

Update Mechanisms

Three stochastic kernels, or update mechanisms, were devised to update the break point structure mapped onto a given topology, allowing one to update the number, positions, and profiles of the break points. We also devised 3 topological update mechanisms that keep track of the break point structure and leave the total length of the tree and the number of break points unchanged: a "subtree pruning and regrafting" or SPR, as described by Swofford et al. (1996), a "node sliding" as described by Lartillot and Philippe (2004), and a topological move of the root's position. All these update mechanisms, and their corresponding Hastings ratios, are described in the Appendix. Rates across sites, relative exchange rates, branch lengths, and hyperparameters are updated as described by Lartillot and Philippe (2004).

Nonstationary Model Configurations

Several variants of our nonstationary model can be proposed. Instead of considering the general compositional shift process, where state frequencies of the π profiles are free parameters, one can constrain the model so that $\pi_C = \pi_G$ and $\pi_A = \pi_T$, according to a GC ratio parameter. Additionally, rather than considering the number of break points and their positions as free parameters, it is possible to constrain the model so that a break point is placed at the beginning of each branch. In this way, our nonstationary model reduces to the models proposed by Galtier and Gouy or Yang and Robert. More specifically, if one uses GC ratio parameters, one obtains the model proposed by Galtier and Gouy, denoted in the following as GG_{GC}, and otherwise, if profiles are left unconstrained, the settings are equivalent to the model proposed by Yang and Robert, denoted as YR_π. By homology, one denotes our model (considering the number of break points and their positions as free parameters) by BP_{GC} and BP_π, depending on whether GC or unconstrained profiles are used. Finally, when constraining the number of break points to $N = 0$, and setting $\pi_\infty = \pi_0$, the Markovian substitution process defined at the root is applied to the whole tree and our nonstationary model reduces to a stationary and homogeneous GTR + Gamma model, denoted in the following as STAT.

MCMC Settings

We define a MCMC cycle as the consecutive call to all relevant update mechanisms, given a model. Some update mechanisms are called several times, with different tuning parameters (see table S1, Supplementary Material online for details concerning a cycle). Continuous update mechanisms were tuned so as to reach an acceptance ratio of 30–70%. The transdimensional update (creating and deleting break point) cannot by tuned. Its acceptance ratio was highly variable, for example, of about 10% for a data set of 5 16S rRNAs, and seems to decrease as the lengths of sequences increase. We run chains for a total of 500,000 cycles, discarding a burn-in period of 100,000 cycles and saving 4,000 samples among the 400,000 remaining points. Some chains were performed under free topology, in which case we computed the majority-rule consensus

(Margush and McMorris 1981) from the topologies of the 4,000 resulting samples. When chains were run under fixed topology, we estimated for each branch j the mean number of break points:

$$\widehat{N}_j = \frac{1}{A}\sum_{a=1}^{A} N_j^a,$$

where A is the number of samples, and N_j^a is the number of break points placed on branch j at sample a. We also computed the mean profile per branch:

$$\widehat{\pi}_{ij} = \frac{1}{A}\sum_{a=1}^{A}\sum_{n=0}^{N_j^a}(x_{n+1}^a - x_n^a)\pi_{in}^a,\ 1 \le i \le S,$$

where $\widehat{\pi}_{ij}$ is the mean frequency, averaged over A samples, of state i along branch j, x_{n+1}^a and x_n^a are the relative positions of 2 consecutive break points (with $x_0^a = 0$ and $x_{N_j+1}^a = t_j^a$), t_j^a is the length of branch j, π_{in}^a is the frequency of state i defined at break point n, and a denotes a particular sample. We estimated standard errors (SE) on $\widehat{N}_j$ and $\widehat{\pi}_{ij}$ using the "window estimators" method described by Geyer (1992), discussed by Raftery and Lewis (1992), and used by Wilson et al. (2003). The method consists in estimating the effective size of the sample from its autocorrelation function. The SE is then equal to the standard deviation divided by the square root of the effective size. To check convergence of the MCMC, each experiment was run twice.

Model Comparison by Bayes Factor Evaluation

Given a data set D, the relative fit between 2 models M_0 and M_1 can be formulated by the ratio of their marginal likelihoods:

$$B = \frac{p(D|M_1)}{p(D|M_0)},$$

where

$$p(D|M_i) = \int_\Theta p(D|\theta, M_i)p(\theta|M_i)\mathrm{d}\theta.$$

A Bayes factor B greater (respectively, lower) than one indicates a support in favor of model M_1 (respectively, M_0). To numerically estimate the Bayes factor, we used the thermodynamic integration method (Ogata 1989; Gelman et al. 2004). An implementation of this method is provided in the *PhyloBayes* program (Lartillot and Philippe 2006). Specifically, we used the "model-switch" scheme as defined in that paper. We performed several types of thermodynamic integrations: 1) between one of the BP_π, BP_{GC}, YR_π, GG_{GC}, and STAT models and the stationary model already implemented in the *PhyloBayes* program, under free or fixed topology, and 2) between 2 topologies under a fixed model. The first type of thermodynamic integration allows us to compare fits of model configurations using the stationary model as a reference. The second type of thermodynamic integration is a way of determining the relative support of 2 candidate topologies under a given model. Although sampling the topology space already gives such an answer, problems in the chains' mixing behavior may be encountered, and thus evaluation of the following Bayes factor:

$$B = \frac{p(D|M, \tau_1)}{p(D|M, \tau_2)},$$

where τ_1 and τ_2 are the 2 candidate topologies, provides a confirmation of the results obtained under free topology.

For each experiment, we ran a 1,000,000 cycle long bidirectional thermodynamic integration. For each direction, we got a set of 1,000 samples. Sampling, thermic lag, and discretization errors are combined into a single 95% confidence interval for the Bayes factor approximation, as proposed by Lartillot and Philippe (2006).

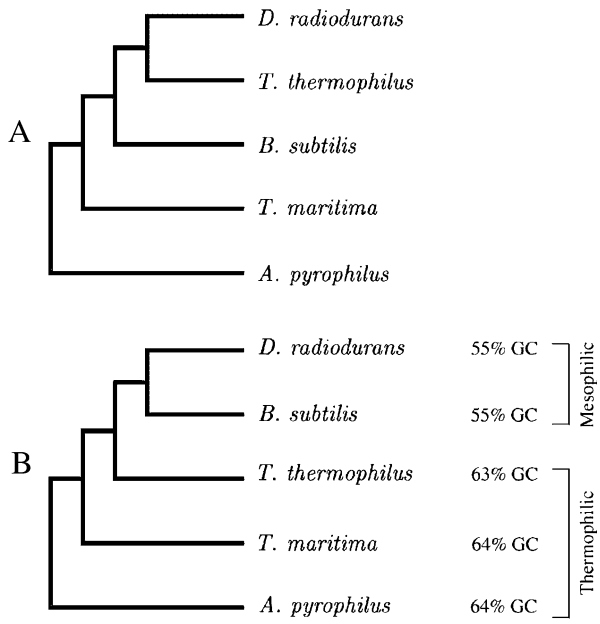

FIG. 2.—Candidate phylogenies for bacteria *Thermus thermophilus*, *Deinococcus radiodurans*, *Bacillus subtilis*, *Thermotoga maritima*, and *Aquifex pyrophilus*. (*A*) The assumed correct phylogeny. (*B*) A commonly obtained reconstruction artifact, where mesophilic bacteria with similar GC content attract together (as well as thermophilic GC richer bacteria, which also attract together). Percentages indicate the GC content of each species.

Material

We first applied the nonstationary model to a data set of 5 eubacterial (*Thermus thermophilus*, *Deinococcus radiodurans*, *Bacillus subtilis*, *Thermotoga maritima*, and *Aquifex pyrophilus*) 16S rRNAs, assembled by Embley et al. (1993). A topology τ_1, supported by much independent evidences (Murray 1991; Eisen 1995; Gupta 1998), groups *T. thermophilus* with *D. radiodurans*, to the exclusion of *B. subtilis*, *T. maritima*, and *A. pyrophilus* (fig. 2*A*). However, this set of sequences is known to be prone to phylogenetic reconstruction artifacts under stationary models due to the attraction of sequence of similar composition (Embley et al. 1993; Mooers and Holmes 2000; Foster 2004). The artifact leads to group together the mesophilic bacteria *D. radiodurans* and *B. subtilis*, leading to topology τ_2 (fig. 2*B*). In the following, we will call

τ_1 and τ_2, respectively, as the "correct" and the "artifact" topology.

We additionally compared the fits of the 4 nonstationary model configurations (BP_π, YR_π, BP_{GC}, and GG_{GC}, see Methods), on several sets of bacterial 16S rRNAs and of yeast genes. First, we analyzed 4 data sets of 5, 10, 15, and 20 Proteobacteria and Deinococci 16S rRNAs (species belonging to data sets 5, 10, and 20: *Pelobacter propionicus*, *Photobacterium profundum*, *Vitreoscilla stercoraria*, *Sulfurospirillum arcachonensis*, and *D. radiodurans*; to data sets 10, 15, and 20: *Thioploca ingrica*, *Burkholderia pseudomallei*, *Zymomonas mobilis*, *Alvinella pompejana epibiont*, and *T. thermophilus*; to data sets 15 and 20: *Syntrophus gentianae*, *Flavimonas oryzihabitans*, *Leptothrix cholodnii*, *Rhodospirillum molischianum*, and *Deinococcus murrayi*; to data set 15: *Desulfuromusa bakii*, *Zymobacter palmae*, *Rickettsia honei*, *Campylobacter rectus*, and *Thermus ruber*; and to data set 20: *Desulfovibrio fairfieldensis*, *Nitrosomonas europae*, *Acidosphaera rubrifaciens*, *Arcobacter cryaerophilus*, and *Thermus filiformis*).

Second, we analyzed the BAS1 gene, chosen among the 106 genes of the data set assembled by Rokas and Carroll (2005). In the latter case, we investigate 2 versions of the data set, comprising 7 and 14 species (species belonging to data sets 7 and 14: *Saccharomyces paradoxus*, *Saccharomyces kudriavzevii*, *Saccharomyces castellii*, *Saccharomyces kluyveri*, *Kluyveromyces lactis*, *Debaryomyces hansenii*, and *Yarrowia lipolytica*; and to data set 14: *Saccharomyces cerevisiae*, *Saccharomyces mikatae*, *Saccharomyces bayanus*, *Candida glabrata*, *Candida albicans*, *Ashbya gossypii*, and *Kluyveromyces waltii*).

For each data set, we evaluated the Bayes factors using thermodynamic integration between the nonstationary models and the reference stationary model, under fixed topology (see Methods). For the 4 bacterial 16S rRNAs data sets, the topologies were obtained using the MrBayes software (Huelsenbeck and Ronquist 2001) under the default GTR + Gamma model, and for the 2 yeast gene data sets, we use the topology obtained using MrBayes, by Jeffroy et al. (2006) on amino acid sequences. In the latter case, MrBayes chains were run under the WAG + Gamma + Invariant model.

Results

Check of Model and Implementation

We performed several checks of our implementation. First, when the likelihood terms in the Metropolis–Hastings ratio are omitted, the MCMC should yield a sample from the prior distribution defined by our model, which we checked, marginally, on several parameters of interest (break point number and profiles, relative exchange rates, branch lengths, fig. S1, Supplementary Material online; and bipartitions, fig. S2, Supplementary Material online). Second, we compared the posterior mean values obtained under the default GTR + Gamma model of MrBayes, with those of our model configured as closely as possible to MrBayes (table S2, Supplementary Material online). All parameter posterior values determined under our model were close to those estimated by MrBayes: the largest relative difference is of 0.8%, obtained for the total tree length. This latter difference still represents 5 times the SE, but this could be explained by the fact that our stationary configuration is not strictly identical to that of MrBayes. In particular, the prior of the relative exchange rates is not the same. Finally, we performed simulations, and measured the hit probabilities, as was done by Wilson et al. (2003). The underlying idea is that, if the implementation is correct, the expected fraction of the simulations for which the true (simulation) value of a given parameter falls within the $p \times 100\%$ confidence interval should be equal to p. Our checks (table S3, Supplementary Material online) are consistent with these expected fractions.

Posterior Values of Model Parameters on Fixed Topologies

As a way of illustrating the behavior of our model, we performed a series of fixed topology analyses. We considered the data set of 5 16S rRNAs (*T. thermophilus*, *D. radiodurans*, *B. subtilis*, *T. maritima*, and *A. pyrophilus*) and run chains under the BP_π model, fixing the topology to its correct τ_1 or to its artifact τ_2 configuration. We rooted the tree as in Olsen et al. (1994) and Galtier and Gouy (1998), in the branch leading to *A. pyrophilus*. As the number of break points on each branch, as well as their respective profiles, may change during the MCMC, we propose to visualize the average effect of all these fluctuations by just looking at the mean posterior number of break points, and at the mean posterior profiles of stationary probabilities, on each branch (fig. 3 and table S4, Supplementary Material online).

On both the correct and the artifact topologies, AT rich profiles are favored along terminal branches leading to *B. subtilis* and *D. radiodurans* (fig. 3*A* and *B*, Supplementary Material online) and more specifically in the case of the artifact topology, also along the internal branch leading to the clade (*B. subtilis* and *D. radiodurans*) (fig. 3*A*). The model thus takes into account the compositional shift of *B. subtilis* and *D. radiodurans* toward an AT richer content. Moreover, one obtains mean posterior numbers of break point (1) of 1.030 on the *B. subtilis* and *D. radiodurans* ancestor branch, and of at most 0.074 on all other branches, for the artifact topology (fig. 3*C*), and (2) of 1.234 on the *B. subtilis* branch, of 1.269 on the *D. radiodurans* branch, and of at most 0.117 on all other branches, for the correct topology (fig. 3*D*). In other words, on average, the chains mainly sampled break points on branches leading to the most significantly biased sequences of the data set. Moreover, the model parsimoniously adapts to the correct, or to the artifact, topology and explains the compositional shift of *B. subtilis* and *D. radiodurans* toward AT richness, respectively, as a convergent evolution (2 independent events) or as a shared derived character (1 ancestral event).

Comparison of 2 Candidate Topologies

We then wanted to know which of the 2 candidate topologies is preferred, depending on the model used, that is, the nonstationary model BP_π or the stationary model STAT. As expected from previous analyzes (Foster 2004), under the STAT model *B. subtilis* groups with *D. radiodurans* (artifact topology τ_2), with a posterior probability of

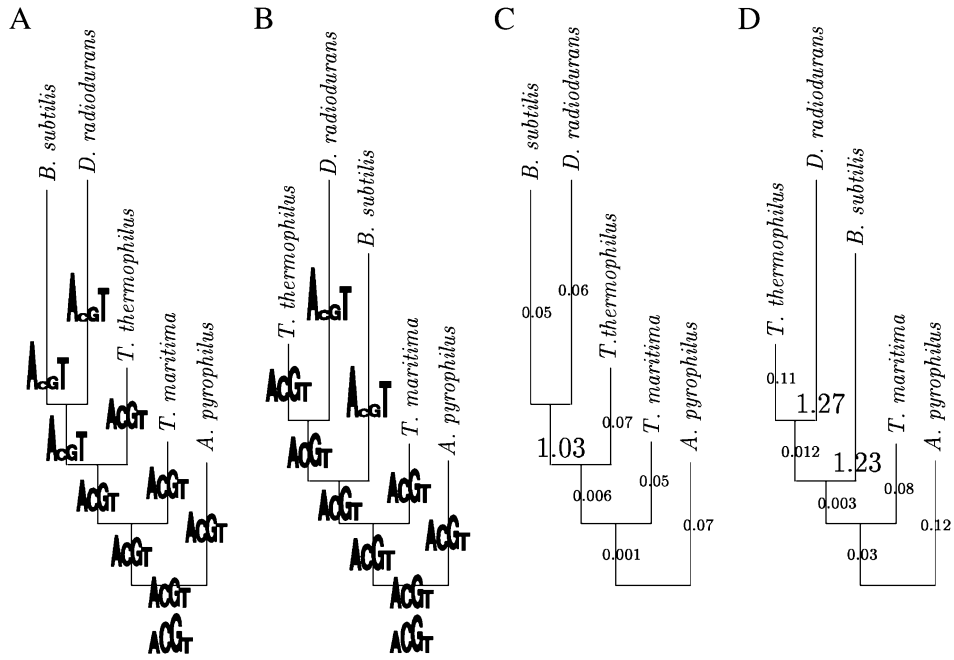

FIG. 3.—Posterior mean profiles and number of break points per branch obtained under the nonstationary model on the artifact (*A* and *C*) and the correct (*B* and *D*) topology. (*A*) and (*B*): posterior mean profiles, with the π_∞ profile placed at the very bottom of the figure. (*C*) and (*D*): posterior mean number of break points.

0.97. In contrast, under the BP_π model, we obtained the clade (*T. thermophilus* and *D. radiodurans*) with a posterior probability of 1 (correct topology τ_1).

As a confirmation, we evaluated the relative support of the τ_1 and τ_2 topologies, by thermodynamic integration under a fixed model, either BP_π or STAT (see Methods). A positive value of the logarithm of the Bayes factor (95% credibility interval [2.4, 15.7]) is obtained under the nonstationary model, whereas a negative value (in interval $[-9.5, -0.1]$) is obtained under the stationary model. This means that the correct τ_1 topology better fits the data under the BP_π model, and in contrast, that the artifact τ_2 topology is chosen by STAT. These results are consistent with our analyses under free topology and show that our model is able to recover the correct topology.

As previously suggested by Foster (2004), this disagreement between the 2 models may be explained by the fact that the stationary model tends to artifactually group together unrelated taxa sharing similar base composition. In contrast, the nonstationary model, handling compositional heterogeneities, would be able to discern the phylogenetic signal from the compositional bias. If this interpretation is correct, one would expect a better fit on this data set for the nonstationary model than for the stationary one. We, therefore, compared the 2 models by Bayes factor evaluation, using the thermodynamic integration method. Importantly, as the 2 models do not favor the same phylogeny, the integration was done under free topology (see Methods). We estimated the logarithm of the Bayes factor to lie in a 95% credibility interval of [59.3, 67.2] (table 1). This estimation strongly rejects the stationary model in favor of the BP_π model and thus retrospectively confirms previous results and assumptions considering the τ_1 topology to be closer than τ_2 to biological reality, and τ_2 to be an artifact caused by compositional bias (Murray 1991; Embley et al. 1993; Eisen 1995; Gupta 1998; Mooers and Holmes 2000; Foster 2004).

Comparison between Nonstationary Model Configurations

Several nonstationary models have already been proposed (Yang and Roberts 1995; Galtier and Gouy 1998; Foster 2004), which differ mostly by the kind of bias that they consider (i.e., GC or general biases). In addition all these models are branchwise (i.e., the stationary probabilities of the substitution process are reassessed at the base of each branch). We wanted to provide a comprehensive analysis of the relative merits of some of these models, in particular those of Yang and Roberts (1995) and Galtier and Gouy (1998), whose configurations can be reproduced in our implementation (i.e., YR_π and GG_{GC}, see Methods). We, therefore, performed a general Bayes factor analysis of the bacterial 16S rRNA data set introduced in the

Table 1
Logarithm of the Bayes Factor Estimated for Several Alternative Models on 16S rRNAs (*Thermus thermophilus*, *Deinococcus radiodurans*, *Bacillus subtilis*, *Thermotoga maritima*, and *Aquifex pyrophilus*). The Stationary Model Is Used as a Reference (the comparison to the STAT configuration is a control, which is expected to be close to 0). 95% Credibility Interval Is Shown

BP_π	YR_π	BP_{GC}	GG_{GC}	STAT
[59.3, 67.2]	[50.4, 65.6]	[59.3, 65.5]	[59.0, 65.5]	[−4.5, 5.3]

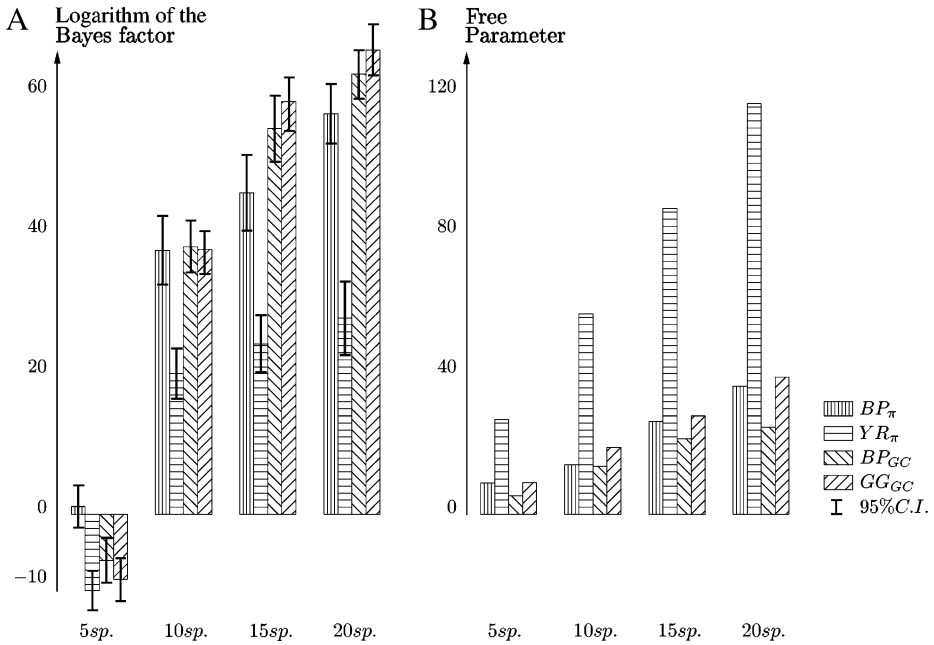

FIG. 4.—(*A*) Bayes factor estimations, obtained by thermodynamic integration under fixed topology, on 4 16S rRNA data sets. Error bars stand for 95% CI. (*B*) Mean posterior number of free parameters for each of the considered data sets (inferred from table 2).

previous section, under all these model configurations and using thermodynamic integration under free topology.

According to our results, all nonstationary models better fit the data than the stationary model, confirming that the stationary model is strongly rejected on this data set (table 1). In addition, among the nonstationary models under consideration, BP_π obtained the best Bayes factor and YR_π the worst. However, the estimated Bayes factors are very close to each other, which suggests that the differences are not significant in this case.

In order to get a more informative view of the relative merits of the nonstationary models, we performed additional Bayes factor evaluations on 4 other data sets of 5, 10, 15, and 20 Proteobacteria and Deinococci 16S rRNAs (see Material). On the 10, 15, and 20 species data sets, the estimated Bayes factors are all positive, rejecting the stationary model in favor of the nonstationary ones (fig. 4*A*). Only the 5 species data set behaves differently. However, it displays weak compositional heterogeneity and is close to stationarity (a χ^2 test yields a minimum *P* value of 0.22 over the 5 taxa). Accordingly, on this data set, all models except BP_π are rejected in favor of the stationary model. Interestingly, the mean posterior number of break points sampled under the BP_π model is close to 0 in this case (95% CI = [0, 2], table 2), indicating that, when the analyzed data display no significant compositional bias, the compound process model reduces itself to the stationary model.

As shown in table 2, the mean posterior number of break points inferred by BP_π and BP_{GC} models remains smaller than the number of branches, which implies that these models always use fewer free parameters than their homologous branchwise versions. This could indicate that not all the stationary profiles assumed by the YR_π and GG_{GC} models (i.e., one per branch) are useful and, thus, that some of them represent no real compositional shift events. Consistent with this observation, we note that the YR_π model systematically obtained the worst Bayes factor and, at the same time, handles the largest number of free parameters (table 2 and fig. 4*B*). These correlation between Bayes factors and model dimensionality can be explained as follows: for *n* successive branches along which there are no compositional shift events, the YR_π has to infer *n* times anew the same profile; this is a highly unlikely configuration a priori, which penalizes the model by lowering its marginal likelihood. Note that this interpretation does not totally fit all the observations. In particular, the BP_{GC} model handles the smallest number of free parameters, yet it obtains a smaller Bayes factor than its homologous model, GG_{GC}. However, this may be due to the conservative prior we chose for the ε parameter, which leads the mean number of break points ε and, consequently, the posterior number of break points to tend toward one. This prior seems to penalize the BP_{GC} model, compared with its homologous GG_{GC} having a fixed number of break points, especially when many break points must be fitted, that is, when the number of analyzed biased sequences increases, as is the case here (table 2).

Table 2
Posterior Mean Number of Break Points Sampled under Nonstationary Models, on Data Sets of 5, 10, 15, and 20 *Proteobacteria* and *Deinococci* 16S rRNAs. 95% CI Are Shown within Brackets

	BP_π	BP_{GC}	YR_π and GG_{GC}
5 Species	0.7 [0, 2]	1.6 [0, 4]	8
10 Species	2.0 [1, 3]	5.8 [2, 12]	18
15 Species	5.1 [3, 7]	9.7 [5, 16]	28
20 Species	7.6 [3, 11]	11.4 [7, 16]	38

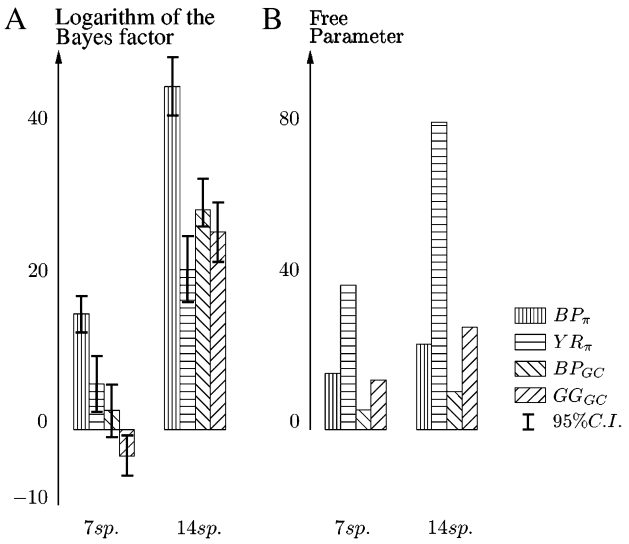

FIG. 5.—(*A*) Bayes factor estimations, obtained by thermodynamic integration under fixed topology, on 2 yeast gene data sets. Error bars stand for 95% CI. (*B*) Mean posterior number of free parameters for the 2 data sets (inferred from table 3).

Finally, the 2 GC models (BP_{GC} and GG_{GC}) obtained the best fit on the data sets of 15 and 20 species. This may indicate that these data sets do not contain significant biases other than GC biases. Consistent with this, rRNA stems have similar proportions of G and C nucleotides (Higgs 2000), and thus, the observed compositional biases should be well described by GC ratio parameters. In this case, GC bias–based models avoid another overparameterization effect as they do not have to repeatedly infer similar proportions for A and T and for G and C.

However, not all biases are GC, and therefore, to offer a more complete spectrum of model comparisons, we analyzed data sets displaying more unequal composition in G and C nucleotides. We evaluated the Bayes factors of the nonstationary models for 2 data sets of 7 and 14 yeast species, for the BAS1 gene (see Material). In these 2 cases, the BP_{GC} and GG_{GC} were penalized and did not obtain the best Bayes factor (fig. 5*A*). Importantly, on the 14 species data set, the YR_π model again obtained the worst Bayes factor and, at the same time, was the model involving the highest number of free parameters (table 3 and fig. 5*B*). In contrast, on both data sets, the BP_π model obtained the best fit. Note that only 2.2 (95% CI = [1, 4]), and 4.1 (95% CI = [2, 6]), break points are inferred on average, respectively, on the 7 and 14 species data sets (table 3), which results in a considerably smaller number of free parameters, compared with YR_π. These observations reinforce the interpretation proposed above for the lack of fit of YR_π, that is, that it is fundamentally an overparameterization problem.

Importantly, in the present case, this overparameterization phenomenon causes YR_π to obtain a worse fit than GG_{GC}. Thus, relying on branchwise models only, one would conclude that the bias of the 14 species data set is a pure GC bias, rather than a more general one. However, the BAS1 gene displays very unequal proportions in G and C (35% of A, 16% of C, 28% of G, and 21% of T). Consistent with this, the break point models show a better fit in favor of general biases (BP_π), compared with GC biases (BP_{GC}). Hence, we are here in a case where the lack of control of the number of parameters inherent to branchwise models would have resulted in a wrong biological interpretation. In contrast, our compound process model, which is able to control its dimensionality according to the data, provides a more reliable conclusion.

Table 3
Posterior Mean Number of Break Points Sampled under Nonstationary Models, on Data Sets of 7 and 14 BAS1 Genes of Yeast Species. 95% CI Are Shown within Brackets

	BP_π	BP_{GC}	YR_π and GG_{GC}
7 Species	2.2 [1, 4]	1.6 [0, 3]	12
14 Species	4.1 [2, 6]	4.0 [2, 7]	26

Discussion

The nonstationary model introduced here differs from previous full-likelihood–based models handling compositional bias phenomena (Yang and Roberts 1995; Galtier and Gouy 1998; Foster 2004) by allowing one to infer a free number of compositional shift events along lineages. This was done using the compound stochastic process method, inspired from Huelsenbeck et al. (1999), to model variations of substitution rates along lineages. To deal with the implied variations in the model dimensionality, we used the Green (2003) formalism. Compared with the models proposed by Galtier and Gouy and Yang and Robert, and as we were able to show by Bayes factor evaluations, our model is less subject to overparameterization effects, especially when many species are analyzed.

The overparameterization issue seems to be highly important and is particularly conspicuous in the case of the branchwise general compositional shift model (the Yang and Robert–like settings). In our experiments, this model always involved the greatest number of parameters and, at the same time, obtained the worst fit. One would expect this overparameterization problem to loose its importance as the length of the alignment increases as the branchwise versions of the nonstationary model are consistent in the limit of infinite sequence length (Chang 1996). But in practice, it should be remembered that many phylogenetic studies are conducted with rRNA, using a large number of taxa (Maidak et al. 1996; Cole et al. 2003). As was demonstrated previously by Hasegawa and Hashimoto (1993), rRNAs are often compositionally biased and should therefore be investigated using adequate nonstationary models. Yet, in such cases, branchwise models will probably not be so reliable because of overparameterization. In contrast, our break point version should behave more reliably. Branchwise models may also be problematic when applied to the amino acid sequences, as amino acid alphabet implies 19 free parameters per branch (instead of 3 for nucleotides), which would have overwhelming deleterious consequences on the fit of the model and maybe also on the estimated phylogeny. Because amino acid sequences can be biased (Foster et al. 1997; Foster and Hickey 1999), an efficient nonstationary

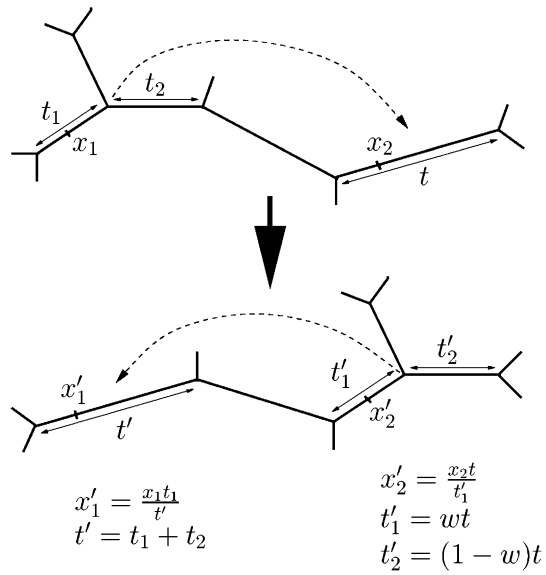

FIG. 6.—Computing the Jacobian in the case of the SPR move. Variables indicated on the drawing are as in the text (Appendix). Two break points with relative positions x_1 and x_2 are shown. Branch lengths and break point's relative positions are calculated given the equations shown at the bottom of the figure.

model is also needed in this case. Normally, the break point model should be able to handle this correctly.

It is not clear whether the branchwise model of Foster is sensitive to such overparameterization effects. It was, in fact, explicitly designed to avoid these problems. However, the equilibrium frequencies still need to be rechosen among the available ones at the base of each branch, which also has a cost, in principle, and possibly a significant one in a context where there are few compositional shift events, compared with the number of nodes. In addition, at least in the current version of this model, the number of profiles needs to be fixed a priori, which lacks flexibility. In this respect, it would be informative to modify this model such as it handles a free number of profiles, using reversible-jump Monte Carlo (e.g., as in Green 2003) and then to compare this modified version to our model. Or conversely, to draw the profile of each break point of our compound process model from a predefined set of profiles, as in Foster's model.

Apart from this, some improvements of the realism of our model can be considered. First, the exponential prior on the apparition rate of compositional shift events is conservative and penalizes the model when many events have to be fitted. To avoid this problem, one could instead use another prior (e.g., uniform). Second, one could change the uniform distribution G_0, from which profiles are a priori created and evaluated, to a generalized Dirichlet, whose hyperparameters can also be inferred. Third, modeling the compositional shifts as piecewise constant processes is not realistic and should be considered as a convenient mathematical device. In this respect, another improvement of the model would be to use a first order, rather than a 0th order, Markov process, when creating the profile of a new break point. Each break point profile depending on the profile of the previous break point, more break points may be created, which may allow to model quasi continuous compositional shift, although at the cost of an increasing computational time (the complexity of our modified pruning algorithm depending linearly on the number of break points). All these elaborations of our model could improve the quality of our reconstruction of the history of compositional trends of the substitution process. In each case, Bayes factor evaluation can be performed to see which of these configurations actually improve the resulting fit.

In a completely opposite direction, one could try to simplify the current model and keep only its most essential aspects. In particular, if one is not so much interested in the detailed reconstruction of the history of the compositional shifts, but only in the phylogenetic reconstruction, it might make sense to consider the branch of the phylogenetic tree as the fundamental unit of resolution. In this context, one could go back to the usual practice consisting in constraining the stochastic events to appear at the tree nodes, although now, not systematically, but with a probability that could itself be estimated from the data. This proposition remains different from the models proposed by Foster, Galtier and Gouy, and Yang and Robert, as (1) the equilibrium frequencies are not systematically reassessed for all branches, and as (2) each event leads to an independent compositional drift. Such a simplified version of the compound stochastic process presented here may have the same statistical properties than more complex versions, while avoiding the overparameterization pitfalls.

Supplementary Material

Supplementary figures S1 and S2 and tables S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We wish to thank Olivier Gascuel and Nicolas Galtier for their participation to the discussions over the model and its implementation. We are also grateful to Hervé Philippe, Nicolas Rodrigue, as well as 3 referees, for their helpful comments on this manuscript. This work was financially supported by the "60^{eme} commission franco-québécoise de coopération scientifique" and by the french Centre National de la Recherche Scientifique, through the ACI-IMPBIO Model-Phylo funding program.

Appendix

Creating and Deleting Break Points

Note that, in the following, w will denote a random number picked from a uniform distribution.

The number of break points N is a free parameter of the nonstationary model. The "create/delete" MCMC update mechanism creates, or deletes, a break point with probability $p_{\text{create}} = \frac{1}{2}$, or $p_{\text{delete}} = \frac{1}{2}$, if $N > 0$. If $N = 0$, then there is no break point to delete and $p_{\text{create}} = 1$. When deleted, a break point is uniformally picked among the N extant ones (except the default root break point, which cannot be

destroyed for obvious reasons). The probability to delete one of the extant break points is thus:

$$p_d = p_{\text{delete}} \frac{1}{N}.$$

When created, the position of the new break point on the tree, that is, its branch b and its relative position x on its branch, are picked uniformally. The probability of creating a break point, at any relative position, on a given branch j is thus $p(b = j) = \frac{t_j}{T}$. The profile π of the newly created break point is randomly picked following a uniform Dirichlet distribution: $p(\pi) = \text{Dir}(\pi)$. Taking the product yields the overall probability:

$$p_c = p_{\text{create}} \frac{t_j}{T} \text{Dir}(\pi).$$

By definition, the Hastings ratio, H, is the probability of the backward move divided by the probability of the forward move. Thus, in the case of break point creation, $H_c = \frac{p_d}{p_c}$, and reciprocally, in the case of break point deletion, $H_d = \frac{p_c}{p_d}$:

$$H_c = \frac{p_{\text{delete}} \frac{1}{N+1}}{p_{\text{create}} \frac{t_j}{T} \text{Dir}(\pi)}, \quad (6)$$

$$H_d = \frac{p_{\text{create}} \frac{t_j}{T} \text{Dir}(\pi)}{p_{\text{delete}} \frac{1}{N}}. \quad (7)$$

Note that some factors involved in these expressions will cancel out with the ratio of prior probabilities, $\frac{p(\theta')}{p(\theta)}$. According to equation (4), this ratio is:

$$\frac{p(\theta')}{p(\theta)} = \frac{\varepsilon\, t_j\, \text{Dir}(\pi)}{T}, \quad (8)$$

in the case of a creation, and

$$\frac{p(\theta')}{p(\theta)} = \frac{T}{\varepsilon\, t_j\, \text{Dir}(\pi)}, \quad (9)$$

in the case of a deletion. Combining equation (6) with equation (8) and equation (7) with equation (9) yields factored Hastings-prior ratios, HP:

$$\text{HP}_c = H_c \frac{p(\theta')}{p(\theta)} = \frac{\varepsilon\, p_{\text{delete}}}{(N+1) p_{\text{create}}}, \text{HP}_d = H_d \frac{p(\theta')}{p(\theta)} = \frac{N\, p_{\text{create}}}{\varepsilon\, p_{\text{delete}}},$$

that is, in the case of a creation: $\text{HP}_c = \frac{\varepsilon}{2}$ if $N = 0$ and $\text{HP}_c = \frac{\varepsilon}{N+1}$ if $N > 0$, and in the case of a deletion: $\text{HP}_d = \frac{2}{\varepsilon}$ if $N = 1$ and $\text{HP}_d = \frac{N}{\varepsilon}$ if $N > 1$.

Updating Break Point Positions

To update the relative positions of break points, we randomly pick one of them, except the default root break point, and set its new relative position as $x' = x + \lambda(w - 0.5)$, where λ is the tuning parameter, and w is a uniform [0, 1] number. If $x' \leq 0$ or $x' \geq 1$, we reflect x' back into [0, 1]. The corresponding Hastings ratio is 1. Note that this update mechanism may swap the relative positions of 2 break points on their branch, and as a result, modify the effect areas.

Updating Break Point Profiles

To update break point profiles, we uniformally pick one of them, including the profile of the default root break point. The new profile π' is picked from a Dirichlet distribution centered on the current π profile value: $\pi' \sim \text{Dirichlet}(\lambda\pi_1, \lambda\pi_2, \ldots, \lambda\pi_S)$, where λ is the tuning parameter specifying the amplitude of the update mechanism. The Hastings ratio is given by Larget and Simon (1999):

$$H = \prod_{m=1}^{S} \frac{\pi_m^{\lambda\pi'_m - 1} \Gamma(\lambda\pi_m)}{\pi'^{\lambda\pi_m - 1}_m \Gamma(\lambda\pi'_m)},$$

where $\Gamma(\,)$ is the Gamma function. We also use this MCMC move to update the π_∞ profile.

Subtree Pruning and Regrafting

The SPR, described in Swofford et al. (1996), is a global topological move, in which a subtree is pruned and reconnected elsewhere in the remaining tree. In this move, the total length of the tree is left unchanged, and branches behave like solid "sticks." Taking advantage of this property, one can easily generalize the SPR update mechanism to the present nonstationary context, essentially, by tracking the positions of the break points during the topological change as if they were clipped at a given position along a given stick. However, when a break point is on one of the branches that will be split, or merged into another branch, its "relative" position will change. This will induce a nontrivial Hastings ratio, which we compute using Green's formula (eq. 5).

Specifically, let t_1 and t_2 denote the lengths of the branches to be merged in a branch of length $t' = t_1 + t_2$, and symmetrically t will stand for the length of the branch split into 2 branches of lengths t'_1 and t'_2, $t = t'_1 + t'_2$ (fig. 6). Let $w \in [0, 1]$ be the uniform random number that was used to decide the position of the regrafting. Then we have

$$t'_1 = tw,$$

$$t'_2 = t(1 - w),$$

$$t' = t_1 + t_2.$$

The reverse move would have been performed upon drawing a random number:

$$w' = \frac{t_1}{t_1 + t_2}.$$

The Jacobian of the SPR move, without taking break points into account, is thus:

$$J_0 = \frac{\partial(t', t'_1, t'_2, w')}{\partial(t, t_1, t_2, w)} = \begin{vmatrix} 0 & w & 1-w & 0 \\ 1 & 0 & 0 & \frac{t_2}{(t_1+t_2)^2} \\ 1 & 0 & 0 & \frac{-t_1}{(t_1+t_2)^2} \\ 0 & t & -t & 0 \end{vmatrix},$$

whose determinant's absolute value is

$$|\det(J_0)| = \frac{t'_1 + t'_2}{t_1 + t_2}. \quad (10)$$

Now, let $x_k \in [0, 1]$ denote the relative position of break point k before, and x'_k after, the topological move. Suppose that break point k was placed on branch of length t_1. Its new X coordinate on the new branch of length $t' = t_1 + t_2$ will be $x'_k = \frac{x_k t_1}{t_1+t_2}$. More generally, we have:

$$x'_k = \begin{cases} \frac{x_k t_1}{t_1+t_2} & \text{if break point } k \text{is on } t_1, \\ \frac{x_k t_2 + t_1}{t_1+t_2} & \text{if break point } k \text{ is on } t_2, \\ \frac{x_k}{w} & \text{if break point } k \text{ is on t and ends up on } t'_1, \\ \frac{x_k - w}{1-w} & \text{if break point } k \text{ is on } t \text{ and ends up on } t'_2. \end{cases} \tag{11}$$

The Jacobian of the SPR move, taking break points into account, is

$$J = \begin{vmatrix} \frac{\partial(x'_k)}{\partial(x_l)} & \frac{\partial(t',t'_1,t'_2,w')}{\partial(x_l)} \\ \frac{\partial(x'_k)}{\partial(t,t_1,t_2,w)} & J_0 \end{vmatrix}.$$

As $\{t', t'_1, t'_2, w'\}$ are not functions of any x_l, $\frac{\partial(t',t'_1,t'_2,w')}{\partial(x_l)} = 0$, so that the cross derivative $\frac{\partial(x'_k)}{\partial(t,t_1,t_2,w)}$ cancel out and the determinant factors into:

$$\det(J) = \det(J_0)\det\left(\frac{\partial(x'_k)}{\partial(x_l)}\right).$$

Moreover, terms $\frac{\partial(x'_k)}{\partial(x_l)} = 0$ for $k \neq l$, so that $\frac{\partial(x'_k)}{\partial(x_l)}$ is a diagonal matrix whose determinant is

$$\det\left(\frac{\partial(x'_k)}{\partial(x_l)}\right) = \prod_{k=1}^{K} \frac{\partial(x'_k)}{\partial(x_k)}, \tag{12}$$

where K is the number of break points whose relative positions have changed during the SPR. Given equation (11), the derivatives involved in this product would be:

$$x'_k = \begin{cases} \frac{t_1}{t_1+t_2} & \text{if break point } k \text{ is on } t_1, \\ \frac{t_2}{t_1+t_2} & \text{if break point } k \text{ is on } t_2, \\ \frac{1}{w} & \text{if break point } k \text{ is on } t \text{ and ends up on } t'_1, \\ \frac{1}{1-w} & \text{if break point } k \text{ is on } t \text{ and ends up on } t'_2. \end{cases} \tag{13}$$

Let N_1, N_2, and N denote the number of break points initially placed on the branches of length t_1, t_2, and t, respectively, and N'_1, N'_2, and N', the number of break points finally placed on the branches of length t'_1 and t'_2 and on the merged branch of length t', respectively (we then have $N' = N_1+N_2$, $N=N'_1+N'_2$, and $K=N+N_1+N_2= N'+N'_1+N'_2$). With the derivatives given by equation (13), we now reformulate equation (12):

$$\det\left(\frac{\partial(x'_k)}{\partial(x_l)}\right) = \left(\frac{t_1}{t_1+t_2}\right)^{N_1}\left(\frac{t_2}{t_1+t_2}\right)^{N_2}\left(\frac{1}{w_c}\right)^{N'_1}\left(\frac{1}{1-w_c}\right)^{N'_2}. \tag{14}$$

Finally, we have $\frac{g'(w')}{g(w)} = 1$ (as the random numbers w and w' are picked uniformally), and thus, gathering equation (10) and equation (14) yields the Hastings ratio of the SPR move:

$$\begin{aligned} H &= \frac{g'(w')}{g(w)}|\det(J)| \\ &= \left(\frac{t'_1+t'_2}{t_1+t_2}\right)\left(\frac{t_1^{N_1}\, t_2^{N_2}\,(t'_1+t'_2)^{N'_1+N'_2}}{(t_1+t_2)^{N_1+N_2}\, t'^{N'_1}_1\, t'^{N'_2}_2}\right). \end{aligned} \tag{15}$$

More generally, during a topological move, each time a break point will swap from a branch of length t to a branch of length t', its relative position will change and a new term will appear in the Jacobian $\frac{\partial(x'_k)}{\partial(x_l)}$:

$$\frac{\partial(x'_k)}{\partial(x_k)} = \frac{t}{t'}. \tag{16}$$

Node Sliding

The node sliding, described in Lartillot and Philippe (2004), is a local topological move inspired from the LOCAL move (Larget and Simon 1999). The difference with the LOCAL move is simply that the tree length is left unchanged. Let a, b, c, u, and v denote 5 nodes in the tree, topologically associated into branches $u \to c$, $u \to v$, $v \to a$, and $v \to b$, of lengths t_1, t_2, t_3, and t_4, respectively. We randomly choose between the 2 paths $c - u - v - a$ and $c - u - v - b$, a path C, of origin c and of length $t_C = t_1 + t_2 + t_3$ or $t_C = t_1 + t_2 + t_4$. Node u is moved along path C at a new position $B = t_1 + \lambda\left(w - \frac{1}{2}\right)$, where λ is a tuning parameter and w is a uniform random [0, 1] number. This results in 2 cases, depending on how B compares with $t_1 + t_2$: 1) $B > t_1 + t_2$, the topology is modified and node u swaps on branch $v \to a$ or $v \to b$, which is split, and branches $u \to c$ and $u \to v$ merge together, or: 2) $B < t_1 + t_2$, the topology is not modified and only the branch lengths t_1 and t_2 change.

Relative positions of break points placed on branches whose lengths are modified will change, inducing a nontrivial Hastings ratio. Calculation of this ratio, using Green's formula, is very close to that performed for the SPR move and will, therefore, not be fully explained here. Briefly, the absolute value of the Jacobian determinant, without taking break points into account, is equal to 1 in both cases (i.e., with or without the node-sliding proposal results in a topological change). Additionally, each time a break point swaps from one branch, of length t, to another, of length t', a new term equal to $\frac{t}{t'}$ appears in the Jacobian $\frac{\partial(x'_k)}{\partial(x_l)}$ (eq. 16), yielding the following Hastings ratio when the topology is modified:

$$H = \frac{g'(w')}{g(w)}|\det(J)| = \frac{t_1^{N_1}\, t_2^{N_2}\,(t'_1+t'_2)^{N'_1+N'_2}}{(t_1+t_2)^{N_1+N_2}\, t'^{N'_1}_1\, t'^{N'_2}_2},$$

where t'_1 and t'_2 are the lengths of the split branches, and symmetrically t_1 and t_2 are the lengths of the 2 merged branches. In the other case, when the topology is not modified:

$$H = \frac{g'(w')}{g(w)}|\det(J)| = \frac{t_1^{N_1}\, t_2^{N_2}}{t'^{N'_1}_1\, t'^{N'_2}_2},$$

where t'_1 and t'_2 are new lengths of the 2 branches of initial lengths t_1 and t_2. In both cases, N_1, N_2, N'_1, and N'_2 denote

the number of break points placed on branches of lengths t_1, t_2, t'_1, and t'_2, respectively, and $\frac{g'(w')}{g(w)} = 1$, as w and w' are uniform random numbers.

Updating the Root Position

Because the model is nonstationary, and thus not reversible, the pulley principle of Felsenstein (1981) no longer applies and the likelihood now depends on the position of the root (Yang and Roberts 1995; Galtier and Gouy 1998). We thus implemented a topological move of the root position. During this move, we simply disconnect the root node from the tree (the root's sibling branches are merged together) and reconnect it at position $B = wt$, where w is a uniform random [0, 1] number on a randomly chosen branch of length t. The branch onto which the root is reconnected is split into 2 branches. Using Green's formula, we obtain the same Hastings ratio as for the SPR move, and this using exactly the same derivation (eq. 15), where t_1 and t_2 are lengths of branches to be merged, and t'_1 and t'_2 are lengths of the 2 branches resulting from the split, N_1, N_2, N'_1, and N'_2 are the numbers of break points placed on branches of lengths t_1, t_2, t'_1, and t'_2, respectively.

Update Hyperparameter ε

Finally, the prior mean number of break points, ε, is a free parameter of the model and has therefore to be updated. To do this, we pick w uniformally in [0, 1], we set $\varepsilon' = \varepsilon e^{\left(w-\frac{1}{2}\right)\lambda}$, where λ is a tuning parameter. Only uniform random numbers are involved, so that the Hastings ratio is simply equal to the Jacobian: $H = |\det(J)| = \frac{\partial(\varepsilon')}{\partial(\varepsilon)} = e^{\left(w-\frac{1}{2}\right)\lambda}$.

Literature Cited

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

Brinkmann H, Philippe H. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol 16:817–25.

Brown WM, Prager EM, Wang A, Wilson AC. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol 18:225–39.

Canbäck B, Tamas I, Andersson SG. 2004. A phylogenomic study of endosymbiotic bacteria. Mol Phylogenet Evol 21:1110–22.

Chang JT. 1996. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. Math Biosci 137:51–73.

Cole JR, Chai B, Marsh TL, et al. (11 co-authors). 2003. The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. Nucleic Acids Res 31:442–3.

Delsuc F, Phillips MJ, Penny D. 2003. Comment on "Hexapod origins: monophyletic or paraphyletic?". Science 301:1482.

Delsuc F, Scally M, Madsen O, Stanhope MJ, de Jong WW, F.M.C., Springer MS, Douzery EJ. 2002. Molecular phylogeny of living Xenarthrans and the impact of character and taxon sampling on the placental tree rooting. Mol Biol Evol 19:1656–71.

Eisen JA. 1995. The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. Mol Biol Evol 41:1105–23.

Embley TM, Thomas RH, Williams RAD. 1993. Reduced thermophilic bias in the 16S rDNA sequence from *Thermus ruber* provides further support for a relationship between *Thermus* and *Deinococcus*. Syst Appl Microbiol 16:25–9.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximun likelihood approach. Mol Evol 17:368–76.

Foster PG. 2004. Modeling compositional heterogeneity. Syst Biol 53:485–95.

Foster PG, Hickey DA. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J Mol Evol 48:284–90.

Foster PG, Jermiin LS, Hickey DA. 1997. Nucleotide composition bias affects amino acid content in protein coded by animal mitochondria. J Mol Evol 44:282–8.

Galtier N, Gouy M. 1995. Inferring phylogenies from DNA sequences of unequal base composition. Evolution 92:11317–21.

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 15:871–9.

Galtier N, Tourasse N, Gouy M. 1999. A nonhyperthermophilic common ancestor to extant life forms. Science 283:220–1.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2004. Bayesian data analysis. Chapman and Hall/CRC.

Geyer CJ. 1992. Practical Markov chain Monte Carlo. Stat Sci 7:473–83.

Green PJ. 2003. Trans-dimensional Markov Chain Monte Carlo. In: Green PJ, Hjort NL, Richardson S, editors. Highly structured stochastic systems. Oxford University Press. p 179–98.

Gupta RS. 1998. Protein phylogenies and signature sequences: a reappraisal of evolutionary relationship among Archaebacteria, Eubacteria, and Eukaryotes. Microbiol Mol Biol Rev 62:1435–91.

Hasegawa M, Hashimoto T. 1993. Ribosomal RNA trees misleading? Nature 361:23.

Herbeck JT, Degnan PH, Wernegreen JJ. 2004. Nonhomogeneous model of sequence evolution indicates independent origins of primary endosymbionts within the enterobacteriales (γ-Proteobacteria). Mol Biol Evol 22:520–32.

Higgs PG. 2000. RNA secondary structure: physical and computational aspects. Q Rev Biophys 33:199–253.

Holder MT, Lewis PO, Swofford DL, Larget B. 2005. Hastings ratio of the LOCAL proposal used in Bayesian phylogenetics. Syst Biol 54:961–5.

Hrdy I, Hirt R, Dolezal P, Bardonova L, Foster P, Tachezy J, Embley T. 2004. Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature 432:618–22.

Huelsenbeck JP, Larget B, Miller RE, Ronquist F. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst Biol 51:673–88.

Huelsenbeck JP, Larget B, Swofford D. 1999. A compound poisson process for relaxing the molecular clock. Genetics 154:1879–92.

Huelsenbeck JP, Ronquist F. 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–5.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet 22:225–31.

Jukes TH, Bhushan V. 1986. Silent nucleotide substitutions and G + C content of some mitochondrial and bacterial genes. J Mol Evol 24:39–44.

Lake JA. 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Evolution 91:1455–9.

Larget B, Simon DL. 1999. Markov Chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol Biol Evol 16:750–9.
Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095–109.
Lartillot N, Philippe H. 2006. Computing Bayes factors using thermodynamic integration. Syst Biol 55:195–207.
Lockhart PJ, Howe CJ, Bryant DA, Beanland TJ, Larkum AW. 1992. Substitutional bias confounds inference of Cyanelle origin from sequence data. J Mol Evol 34:153–62.
Lockhart PJ, Steel MA, Hendy MD, Penny D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol 11:605–12.
Maidak BL, Olsen GJ, Larsen N, Overbeek R, McCaughey MJ, Woese CR. 1996. The Ribosomal Database Project (RDP). Nucleic Acids Res 24:82–5.
Margush T, McMorris FR. 1981. Consensus n-trees. Bull Math Biol 43:239–44.
Montero LM, Salinas J, Matassi G, Bernardi G. 1990. Gene distribution and isochore organization in the nuclear genome of plants. Nucleic Acids Res 18:1859–67.
Mooers AO, Holmes EC. 2000. The evolution of base composition and phylogenetic inference. Trends Ecol Evol 15:365–9.
Murray RGE. 1991. The family Deinococcaceae. In: Balows A, Trüper HG, Dworkin M, Harder W, Schleifer KH, editors. The prokaryotes. Volume 4. London: Springer. p 3733–44.
Neal RM. 1993. Probabilistic inference using Markov Chain Monte Carlo methods. Technical Report CRG-TR-93-1.
Ogata Y. 1989. A Monte Carlo method for high dimensional integration. Numerishe Mathemetik 55:137–57.
Olsen GJ, Woese CR, Overbeek R. 1994. The winds of (evolutionary) change: breathing new life into microbiology. J Bacteriol 176:1–6.
Philippe H, Germot A, Moreira D. 2000. The new phylogeny of eukaryotes. Curr Opin Genet Dev 10:596–601.
Phillips MJ, Penny D. 2002. The root of the mammalian tree inferred from whole mitochondrial genomes. Mol Phylogenet Evol 28:171–85.
Raftery AE, Lewis SM. 1992. [Practical Markov chain Monte Carlo]: comment: one long run with diagnostics: implementation strategies for Markov Chain Monte Carlo. Stat Sci 7:493–7.
Rokas A, Carroll SB. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol Biol Evol 22:1337–44.
Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. 1996. Phylogenetic inference. In: Hillis DM, Moritz G, Mable BK, editors. Molecular systematics. Volume 11. Sinauer Associates.
Tamura K, Kumar S. 2002. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. Mol Biol Evol 19:1727–36.
Tarrio R, Rodriguez-Trelles F, Ayala FJ. 2001. Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae. Mol Phylogenet Evol 18:1464–73.
Wilson IJ, Weale ME, Balding DJ. 2003. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. Royal Stat Soc 166:155–201.
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 14:364–71.
Yang Z, Roberts D. 1995. On the use of nucleic acid sequences to infer branchings in the tree of life. Mol Biol Evol 12: 451–8.

Ziheng Yang, Associate Editor

Accepted July 26, 2006