Syst. Biol. 57(5):675–692, 2008
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150802422324

A Model-Based Approach to Study Nearest-Neighbor Influences Reveals Complex Substitution Patterns in Non-coding Sequences

GUY BAELE,[1,2,3] YVES VAN DE PEER,[2,3] AND STIJN VANSTEELANDT[1]

[1]*Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 S9, B-9000 Ghent, Belgium*
[2]*Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Ghent, Belgium; E-mail: yves.vandepeer@psb.ugent.be (Y.V.d.P.)*
[3]*Bioinformatics and Evolutionary Genomics, Department of Molecular Genetics, Ghent University, B-9052 Ghent, Belgium*

Abstract.— In this article, we present a likelihood-based framework for modeling site dependencies. Our approach builds upon standard evolutionary models but incorporates site dependencies across the entire tree by letting the evolutionary parameters in these models depend upon the ancestral states at the neighboring sites. It thus avoids the need for introducing new and high-dimensional evolutionary models for site-dependent evolution. We propose a Markov chain Monte Carlo approach with data augmentation to infer the evolutionary parameters under our model. Although our approach allows for wide-ranging site dependencies, we illustrate its use, in two non-coding datasets, in the case of nearest-neighbor dependencies (i.e., evolution directly depending only upon the immediate flanking sites). The results reveal that the general time-reversible model with nearest-neighbor dependencies substantially improves the fit to the data as compared to the corresponding model with site independence. Using the parameter estimates from our model, we elaborate on the importance of the 5-methylcytosine deamination process (i.e., the CpG effect) and show that this process also depends upon the 5′ neighboring base identity. We hint at the possibility of a so-called TpA effect and show that the observed substitution behavior is very complex in the light of dinucleotide estimates. We also discuss the presence of CpG effects in a nuclear small subunit dataset and find significant evidence that evolutionary models incorporating context-dependent effects perform substantially better than independent-site models and in some cases even outperform models that incorporate varying rates across sites. [Bayes factor; context effect; context-dependent evolution; CpG effect; likelihood function; Markov chain Monte Carlo; nearest-neighbor influences; thermodynamic integration.]

The modeling of evolutionary processes has come a long way since the introduction of the first and simplest model of DNA evolution, the Jukes-Cantor model (Jukes and Cantor, 1969). The basic assumptions of this model express that each base or nucleotide in a sequence has an equal chance of replacement and that all nucleotides evolve independently with equal transition probabilities (and thus with equal nucleotide frequencies). Many extensions of this model have been proposed to relax its assumptions (in e.g., Kimura, 1980; Lanavé et al., 1984; Hasegawa et al., 1985), and the assumption of independent evolution of sites in nucleotide sequences remains in frequent use despite there being no real biological motivation. This is because this assumption proves mathematically convenient in composing likelihood functions in a maximum-likelihood framework (Felsenstein, 1981) and posterior densities in Bayesian phylogenetic inference (e.g., Rannala and Yang, 1996).

Over the past decade, a number of empirical studies have found indications that the assumption of independent evolution of sites is too restrictive. Morton (1995, 1997) and Morton and Clegg (1995) found that substitution bias, measured in terms of the transversion proportion, is significantly correlated with the composition of the two neighboring bases. They observed that transversions occur significantly more frequently than transitions in the rbcL gene, when both the 5′ and 3′ flanking nucleotides are an A or a T. When either or both neighbors are a C or a G, the opposite trend is found. In addition, when both flanking bases are A and/or T, a significant influence of nucleotides other than the immediate neighbors on substitution dynamics is observed (Morton, 1997; Morton et al., 1997). These findings suggest that the substitution process may be "context dependent," in the sense that neighboring base composition may influence the substitution bias at a particular site, such that substitution dynamics vary from site to site. Such context effects have been observed in pseudogenes, where large neighbor effects exist on transitions from C or G and smaller neighbor effects on transitions from A or G (Bulmer, 1986). Such context dependence may also result from influences of neighboring bases on the process of misincorporation (Mendelman et al., 1989).

A number of studies have focused on specific context effects, such as the CpG-methylation-deamination process. Using Monte Carlo simulations, Fryxell and Zuckerkandl (2000) have shown that the deamination of 5-methylcytosine causes underrepresentation of both the CpG and TpA dinucleotides. A systematic approach to this work has been introduced by Arndt et al. (2003). By allowing for mutations of both single nucleotides and pairs of neighboring nucleotides, these authors consider the evolution of an initial random sequence of nucleotides in discrete time steps, according to a set of update rules. From the stationary probability distribution of the considered mutation processes, the authors calculate the nucleotide and dinucleotide frequencies and use the computed dinucleotide odds ratios to measure whether a given dinucleotide pair is over- or underrepresented. They conclude that these correctly capture the strong underrepresentation of the CpG dinucleotides in the human chromosome 21. Recently, Bérard et al. (2008) have introduced a wide extension of the Tamura+CpG model, a class of models introduced and analyzed in the work of Duret and Galtier (2000), of neighbor-dependent substitution processes and have shown that these models are solvable. More precisely, the authors have proven that the frequencies of polynucleotides at equilibrium solve explicit finite-size linear systems and provide explicit and algebraic formulas for the stationary frequencies of non-degenerate neighbor-dependent models of DNA substitutions. Furthermore, their analysis provides some stringent independence properties of these models at equilibrium.

Given the available evidence, several attempts have been made to model context dependence in nucleotide models. Jensen and Pedersen (2000) considered the specific case of two sequences with a reversible substitution process and allowed for the instantaneous probability of change at any site to depend upon its nearest neighbors at the instant of the substitution. Their model consists of a first component that depends on the type of change, whereas the second component models the CpG-deamination process. Christensen et al. (2005) extended this to allow the first component to be an arbitrary reversible codon substitution model and allowed for further flexibility in modeling the CpG-deamination process. In view of the complexity of these approaches, inference is obtained using MCMC (Jensen and Pedersen, 2000) or EM-based pseudo-likelihood estimation (Christensen et al., 2005).

An alternative approach to modeling dependencies between neighboring sites is to reflect different selective constraints at different sites via correlation between the evolutionary rates at neighboring sites (Yang, 1996a). This has been considered through the use of a hidden Markov model (HMM), which assigns a rate of change to each site, according to a Markov process that depends on the rate of change at the previous neighboring site. The HMM approach thus models site dependencies through shared rate parameters but still assumes independent changes at the different sites, conditional on their rate of change. This approach is considered restrictive because the actual evolutionary events at the sites show a dependence that goes beyond their assignment to the same rate category (Felsenstein and Churchill, 1996). For instance, when observing the same base pair in different local sequence environments, Blake et al. (1992) found differences of varying magnitude in the rates of substitution of that base pair and determined an order of effect of both the 5′ and 3′ neighbor on the rates of substitution for transitions and transversions in a dataset comprising a large number of extant primate gene and pseudogene sequences.

A third approach to allowing for site dependence takes advantage of properties of the genetic code via codon-based models of protein change which group nucleotides into triplets that encode an amino acid (i.e., codons) (Goldman and Yang, 1994; Muse and Gaut, 1994). Such models acknowledge that the evolution of a base within a codon depends on whether this will cause a change in the encoding amino acid (i.e., whether the substitution is synonymous or non-synonymous). As such, the substitution of a nucleotide is allowed to depend on the two other nucleotides that form the codon, but different codons are still assumed to evolve independently.

The non-availability of a general and flexible approach to model site dependencies using nucleotide models has led to a number of higher order models of evolution. Such models make assumptions considering the co-evolution of a small number of adjacent sites but usually do not model these dependencies directly (Felsenstein, 2004). Schöniger and von Haeseler (1994) used a dinucleotide model of evolution to show the inadequate performance of models that assume site-independent evolution. Siepel and Haussler (2004) extended this approach by modeling dependencies with N adjacent nucleotides (to the left or right of the considered site) through a Markov model that allows for the evolution of a site to depend either on the identities of its predecessors or on the identities of its successors in the sequence. Although such a model does not preclude bidirectional influences, it may be difficult to interpret because such influences are not modeled explicitly. A further limiting aspect of this approach is the assumption of independent evolution for the ancestral sequences. Also, as the inferences are restricted to N adjacent nucleotides along each branch, context effects cannot cascade in both directions along a single branch of a phylogenetic tree.

Lunter and Hein (2004) calculate the joint likelihood of observing two sequences evolving under a dinucleotide model and use a Bayesian MCMC sampling approach to infer mutation rates. The authors ignore multiple substitutions involving four or more consecutive nucleotides. Given the dinucleotide model, such events comprise at least three independent overlapping substitutions, yet the mutation rates can still be faithfully recovered as shown using synthetic data.

Hwang and Green (2004) compare context-dependent rates across clades and find that these are broadly similar apart from lineage-specific multiplicative shifts in the baseline rate for certain context-dependent substitutions. The authors estimate separate context-dependent rate matrices for transcribed and untranscribed regions for six clades (or groups), leading to 12 rate matrices each consisting of 192 parameters (which is reduced to 96 parameters each by assuming equal rates of complementary events). Here, the context is defined by the two adjacent ancestral nucleotides, which makes this approach attractive because it models bidirectional influences directly. This approach does not use well-known evolutionary models (such as the general time-reversible model) and may drastically increase the number of unknown evolutionary parameters. Although it is obvious that extra parameters are needed to model site dependence, too many additional parameters will only add noise and imply a risk of overfitting, unless in combination with careful model-building strategies (Arndt and Hwa, 2005).

To accommodate this, we will adapt the ideas in Hwang and Green (2004) to study the influence of the composition of the neighboring bases on the substitution probabilities for a given site, using the well-known general time-reversible model. Our key result is an approach which allows evolution at a site to depend upon ancestral states at both the immediate flanking bases, along with an algorithm for constructing the likelihood under this model. We show how much additional information concerning evolutionary patterns can be obtained by allowing the flanking bases to evolve along a branch. We further examine a large non-coding vertebrate dataset for signs of context-dependent evolution, discuss the importance of modeling the CpG-deamination process, hint at the possible presence of a TpA effect, and show, using Bayes factors, that a context-dependent model may improve the model fit to a larger extent than among-site rate variation. We also examine context-dependent substitution behavior in a smaller plant dataset, for which we illustrate potential problems when adding extra parameters that are not sufficiently well supported by the data. Finally, we discuss the influence of such complex models on the estimation of branch lengths given the tree topology.

MATERIALS AND METHODS

The Likelihood Function

Since its introduction in phylogenetics, the likelihood function $L = \Pr(Y_{obs}|\tau)$ (for a phylogenetic tree τ) has had an enormous impact in the field. It expresses how likely the data Y_{obs} are obtained under a given set of assumptions about the process of evolution. In phylogenetics, the data usually consist of a set of nucleotide sequences, whereas the underlying assumptions include the evolutionary substitution model and the topology of the phylogenetic tree of those sequences, along with the corresponding branch lengths. Calculation of the likelihood function is enormously facilitated by the assumption that sites evolve independently as this allows for determining the likelihood function as a product of the likelihoods at each site separately (Felsenstein, 1981).

We accommodate site dependence by proposing an evolutionary model which explicitly incorporates dependencies on neighboring sites in both directions of each considered site, unlike Siepel and Haussler (2004). Construction of the likelihood is more difficult under our model because standard factorization of a distribution is one-sided, either conditioning on neighboring sites to the left of the considered site or to the right (Cowell et al., 1999). In order to construct the likelihood, we represent a phylogenetic tree as a directed acyclic graph (DAG). Here, the vertices represent the ancestral nucleotides, the edges represent direct dependencies between sites, and the direction of the arrow expresses the direction of evolution (see Fig. 1; Cowell et al., 1999). Hence, each node present in a phylogenetic tree is represented by n nodes in the DAG, one for each site in the (observed or ancestral) sequence. The assumption that substitutions

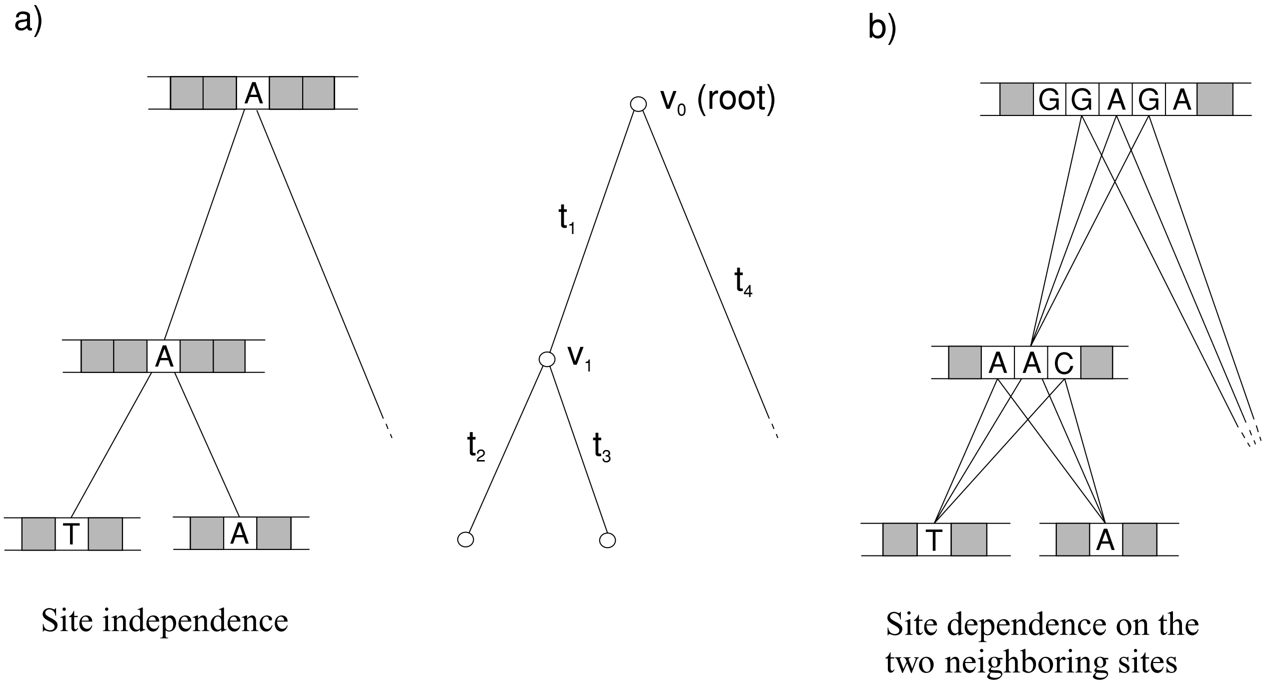

FIGURE 1. Likelihood construction under different assumptions of the evolutionary process for a site at position i in the same phylogenetic tree. (a) Under the assumption of independence, a site only depends on the site at the same position of its evolutionary ancestor. Given the ancestral states in this figure, site i adds the following term to the overall joint density for the given tree: $p(A)p(A|A,t_1)p(T|A,t_2)p(A|A,t_3)$. Each probability in this expression originates from the same evolutionary model. (b) Under the assumption that a nucleotide not only depends on its ancestor at the same position but also on the two immediate flanking nucleotides in its ancestral sequence. Under this structure an increasing, but weakening, range of dependencies (although not explicitly modeled) can be seen across the tree towards the root. Site i adds the following term to the overall joint density for the given tree: $p(A)p(A|GAG,t_1)p(T|AAC,t_2)p(A|AAC,t_3)$. These probabilities use two different evolutionary models: (1) a model, which is used for all sites whose neighbors in the parental sequence are both G, to calculate the contribution of site i over branch t_1; and (2) another model, which is used for all sites whose left neighbor is A and right neighbor is C in the parental sequence, to calculate the contribution of site i over branches t_2 and t_3.

only depend on both flanking bases can now be represented by mapping each branch in the phylogenetic tree onto $3n$ edges in the DAG, as shown in Figure 1b. Although the displayed DAG merely allows for direct dependence on the immediate neighboring sites, note that the underlying model implies (decaying) longer range dependencies by the fact that direct dependencies may propagate over the different edges. For comparison, the stronger assumption of independence is shown in Figure 1a.

Under the assumptions displayed by the DAG, the likelihood can be easily built using probabilistic modeling techniques (Cowell et al., 1999). Let $V = \{v_0, v_1, \ldots\}$ be the set of nodes in the DAG; i.e., V contains all the observed and ancestral nucleotides in the phylogenetic tree. To determine the likelihood function of those nodes under the assumptions of the DAG, we specify the conditional distributions of the state X_v at each node $v \in V$ given its "parents" $X_{pa(v)}$ under the DAG (i.e., given the states of the ancestors that are directly connected to v). This distribution $p(x_v|x_{pa(v)})$ will be obtained from our context-dependent model. Then the likelihood function under our model can be written as

$$L = \prod_{v \in V} p(x_v \mid x_{pa(v)}).$$

Under the assumption that evolution has been proceeding for a very long time (according to the particular model of independent base substitution used), it is reasonable to take $p(x_v|x_{pa(v)})$ for the nodes at the root of the tree (which have no parent(s)) to be the equilibrium probability of base x_v under that model (Felsenstein, 2004).

Even though we model site-dependent evolution, there is no need for new evolutionary models in the sense that any known model can be reused, as illustrated in Figure 1. Indeed, standard evolutionary models, such as Kimura's two-parameter model (Kimura, 1980), the Tamura-Nei model (Tamura and Nei, 1993), and the general time-reversible model (Lanavé et al., 1984), amongst others, can be used for evaluating $p(x_v|x_{pa(v)})$, but with evolutionary parameters depending on the two immediate flanking bases $pa(v)$ in the parental sequence of a given site. As each flanking base has four possible identities, this implies that each site evolves according to 1 out of 16 possible models of evolution along the branch to its descendant sequence.

The example shown in Figure 1 assumes that the nucleotides at the interior nodes of the tree are observed and known. In principle, one may correct for this by summing the likelihood function over all four nucleotides for each missing nucleotide in the ancestral sequences (assuming that there are no missing nucleotides in the observed sequences; Felsenstein, 1981). Under the assumption of site independence, this summation may be done in a computationally economical way using the "pruning" approach (Felsenstein, 1973, 1981). Unfortunately, this approach is no longer feasible when evolution is allowed to depend on the adjacent neighbors (unless the dependencies are constrained at some point), making calculation computationally cumbersome. In the next section, we handle the increased complexity through Bayesian Markov chain Monte Carlo simulation with data augmentation.

Bayesian Markov Chain Monte Carlo

Bayesian inference of phylogeny is based on a quantity called the posterior probability function of a tree, in the same way as maximum-likelihood inference is based on the likelihood function. Although the posterior probability is generally tedious to calculate, simulating from it is relatively easy through the use of Markov chain Monte Carlo (MCMC) methods (Gilks et al., 1996; Huelsenbeck et al., 2001). As the previous section illustrates, relaxing the assumption of independent evolution leads to computational difficulties. In this article, we handle the increased computational complexity via a data augmentation scheme (Tanner and Wong, 1987).

Data Augmentation

Let θ be the collection of unknown parameters indexing the evolutionary model of interest, Y_{obs} the observed nucleotide sequences (i.e., the observed data), and Y_{mis} the unknown ancestral sequences (i.e., the missing data). The observed-data posterior $f(\theta|\, Y_{obs}) = \frac{f(Y_{obs}|\theta) f(\theta)}{f(Y_{obs})}$ is intractable under our model because it involves the likelihood of the observed data, which is computationally cumbersome. However, when Y_{obs} is "augmented" by a random draw for Y_{mis} from the distribution $f(Y_{mis}|\, Y_{obs}, \theta)$ of the ancestral sequences, the resulting complete-data posterior $f(\theta|\, Y_{obs}, Y_{mis})$ becomes tractable. In Tanner and Wong's data augmentation algorithm (1987), the missing nucleotides $Y_{mis}^{(t)}$ of the ancestral sequences in the t^{th} step of the algorithm are drawn from the conditional predictive distribution of Y_{mis}, given a current guess $\theta^{(t-1)}$ of the parameter(s):

$$Y_{mis}^{(t)} \sim f(Y_{mis}|\, Y_{obs}, \theta^{(t-1)}).$$

This step is often referred to as the Imputation-step, or I-step. To acknowledge the uncertainty on this guess $\theta^{(t-1)}$, this is followed by the Posterior-step or P-step, in which (a) new value(s) of $\theta^{(t)}$ is drawn, conditional on $Y_{mis}^{(t)}$, from its complete-data posterior:

$$\theta^{(t)} \sim f\left(\theta|\, Y_{obs}, Y_{mis}^{(t)}\right).$$

Upon convergence of the algorithm, the iterates $\theta^{(t)}$ are random draws from $f(\theta|\, Y_{obs})$.

Prior Distributions

We have used the general time-reversible model (GTR; Lanavé et al., 1984) to study site interdependencies, with the following substitution probabilities:

$$\begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} - & \pi_G r\,AG & \pi_C r\,AC & \pi_T r\,AT \\ \pi_A r\,AG & - & \pi_C rCG & \pi_T rGT \\ \pi_A r\,AC & \pi_G rCG & - & \pi_T rCT \\ \pi_A r\,AT & \pi_G rGT & \pi_C rCT & - \end{pmatrix} \end{array}$$

Let $\pi = \{\pi_A, \pi_C, \pi_G, \pi_T\}$ be the set of base frequencies and θ the evolutionary parameters, with $\theta = \{2\pi_A\pi_G r\,AG, 2\pi_A\pi_C r\,AC, 2\pi_A\pi_T r\,AT, 2\pi_G\pi_C rCG, 2\pi_G\pi_T rGT, 2\pi_C\pi_T rCT\}$ the terms of the scaling formula that binds the parameters of the model. Let T be the set of branch lengths with $t_b (t_b \geq 0)$ one arbitrary branch length and μ a hyperparameter in the prior for t_b in T. The following prior distributions $q(\cdot)$ were chosen for our analysis, with $\Gamma(.)$ the Gamma function:

$$\begin{aligned} \pi &\sim \text{Dirichlet}(1,1,1,1), q(\pi) = \Gamma(4) \\ &\quad \text{on } 0 \leq \pi_m \leq \sum_m \pi_m = 1, \\ \theta &\sim \text{Dirichlet}(1,1,1,1,1,1), q(\theta) = \Gamma\,(6) \\ &\quad \text{on } 0 \leq \theta_i \leq \sum_i \theta_i = 1, \\ t_b|\mu &\sim \text{ Exponential }(\mu), q\,(t_b|\mu) = \frac{1}{\mu} e^{-(1/\mu)t_b} \\ &\quad \text{for each } t_b \text{in } T, \text{ and} \\ \mu &\sim \text{ Inv - gamma }(2.1,1.1), \\ q\,(\mu) &= \frac{(1.1)^{(2.1)}}{\Gamma(2.1)} \mu^{-(2.1+1)} e^{-1.1/\mu}, \mu > 0. \end{aligned}$$

Branch lengths are assumed i.i.d. given μ. When the model allows for the presence of multiple contexts of evolution, each context is assumed to have its own prior, independently of other contexts.

We make a few remarks relating to the above choices. First, the family of Dirichlet distributions is the obvious prior choice for both base frequencies and model parameters of the general time-reversible model as it exploits the fact that both sets of parameters sum to a constant. Second, by specifying a prior distribution for the base frequencies, we do not fix their values in our MCMC run but sample values for them (Larget and Simon, 1999; Suchard et al., 2001; Huelsenbeck et al., 2002; Zwickl and Holder, 2004). In contrast, some MCMC methods fix the values of the base frequencies at either the empirical estimate from the observed data (Li et al., 2000) or at values determined by preliminary MCMC sampling (Mau et al., 1999). Such empirical estimates may be biased when taxon selection oversamples certain subgroups since these estimates give equal weight to all taxa and fixing these parameters can also lead to overestimation of the precision of other parameters (Suchard et al., 2001). Finally, the use of a hyperparameter for the branch-length priors is to reduce sensitivity of the posterior to the prior, in line with Yang and Rannala (2005).

The MCMC Update Process

Each new state of the Markov chain is proposed via a Gibbs cycle. In each step of the cycle, a (set of) parameter(s) is updated conditional on the remaining parameter(s) (sets) using a Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970). Specifically, in each cycle, the following update process is used:

$$\begin{gathered} Y_{mis} \mid T, \tau, \pi, \theta, \mu, Y_{obs} \\ \mu \mid \tau, T, \pi, \theta, Y_{obs}, Y_{mis} \\ \pi \mid \tau, T, \theta, \mu, Y_{obs}, Y_{mis} \\ \theta \mid \tau, T, \pi, \mu, Y_{obs}, Y_{mis} \\ T \mid \tau, \pi, \theta, \mu, Y_{obs}, Y_{mis} \end{gathered}$$

Under the data augmentation approach explained above, the ancestral sequences need to be updated during each update cycle of the chain, after which the remaining parameters (e.g., base frequencies, ...) are updated based on these most recent estimates of the ancestral nucleotides. The ancestral sequences are updated sequentially, one site at a time, from top to bottom in the considered tree and each ancestral site is updated during each update cycle. Starting with the root and descending down to the leaf nodes of the tree, the missing nucleotides in each ancestral sequence are augmented from the first site moving along the sequence up to the last site. Each ancestral "augmented" site is imputed by a draw from a multinomial distribution with probability

$$\begin{aligned} Y_{mis,i} &\sim P(Y_{mis,i} = \mathrm{X} \mid T, \tau, \pi, \theta, \mu, Y_{obs}, Y_{mis,-i}) \\ &= \frac{L_X}{L_A + L_C + L_G + L_T}, X \in \{A, C, G, T\}. \end{aligned}$$

where $Y_{mis,i}$ represents the state of the ancestor that is being augmented at site i, $Y_{mis,-i}$ represents the set of states for all remaining ancestors, and L_X is the complete data likelihood when $X \in \{A, C, G, T\}$ is the value augmented for the considered ancestral site.

The hyperparameter for the branch lengths is updated using a normal driver, which reflects around 0. In particular, a new value μ' for the hyperparameter is proposed based upon its current value μ as $\mu' = |\mu + \varepsilon|$, with $\varepsilon \sim N(0, \sigma^2 = 0.1)$ (as in Suchard et al., 2001). During each update cycle, new frequency values $(\pi'_A, \pi'_C, \pi'_G, \pi'_T)$ are proposed by sampling, based on their current values $(\pi_A, \pi_C, \pi_G, \pi_T)$, from a Dirichlet distribution with density

$$\frac{\Gamma(\alpha_0)}{\prod_{i=1}^{n} \Gamma(\pi_i \alpha_0)} \prod_{i=1}^{n} (\pi'_i)^{\pi_i \alpha_0 - 1},$$

where α_0 (set to 1000 in our analysis) controls the degree to which the new values are free to vary (Huelsenbeck, 2000). We have used an analogous approach to propose updates of the parameters of the general time-reversible model (Larget and Simon, 1999; Huelsenbeck, 2000). In the case of context-dependence, each context-specific evolutionary model is sequentially updated once per generation of the chain and independently of the other models. Finally, each update cycle is completed by updating each branch length in the tree using a normal driver centered on the current value of the branch length. Let $t_b \in T$ be a branch of the tree topology, then a new proposal t'_b for the branch length is $t'_b = |t_b + \varepsilon|$, with $\varepsilon \sim N(0,\sigma^2 = 0.1)$ (Suchard et al., 2001).

Because it is seldom necessary to spend much effort in choosing starting values (Gilks et al., 1996), we draw starting values for the hyperparameter μ, the branch lengths $T \mid \mu$, the base frequencies π, and model parameters θ directly from their prior distribution and carefully monitor the mixing rate of each parameter. We equate the starting ancestral nucleotide Y_{mis} for each site with the nucleotide of its left child. Because we observed that the imputation of the ancestral nodes converges quickly in the beginning stages of the burn-in (data not shown), this approach seems reasonable. In addition, we have repeatedly run the chain with different starting parameter values drawn from their prior distributions and confirmed that convergence to the same posterior distribution occurred.

To the best of our knowledge, no tree transition kernels for a Bayesian MCMC framework have been devised for phylogenetic trees that carry along the ancestral states. Determining such transition kernels is a peculiar problem because at each point in the chain, all ancestral states must be known in order to calculate the complete-data likelihood. Indeed, when proposing a move to a new tree, no estimates are available for the ancestral node which appears where the disconnected node or clade is reinserted. Existing (RJ)MCMC approaches (Green, 1995) are currently prohibiting to perform moves to a new tree. In this article, we have therefore developed inference for nearest-neighbor dependent evolutionary models under a fixed posterior consensus tree. How to infer phylogenetic trees under our model is the topic of ongoing research.

Context-Dependence Using Data Augmentation

Under the assumption of context-dependent evolution, the evolutionary process at a given site may depend upon the evolution of its neighbors. One may acknowledge this by calculating the probability of three given bases (i.e., the site under evolution and its two neighbors) at a given time t along a branch via differential equations (Arndt and Hwa, 2005). Using these probabilities, an approximation to the (log) likelihood may be obtained but at a high computational cost. In view of this, a more attractive approach is to partition branches into several discrete time units (Hwang and Green, 2004).

In this article, we make the weak assumption that the identities of the immediate flanking neighbors remain constant across a single branch partition of the tree so as to allow for increased computational flexibility in model choice. Because this assumption is most likely violated for longer branches (Felsenstein, 2004), lacks biological realism, and given the fact that a pruning approach, as introduced by Felsenstein (1981), is not computationally feasible, alternate measures are necessary to improve on this approximation.

Several approaches to accurately model such situations have already been used to study protein evolution under the assumption of dependent change among codons with the aim of understanding both secondary and tertiary structure (e.g., Parisi and Echave, 2001). As the approach of Parisi and Echave (2001) is computationally inefficient, MCMC approaches which sample substitution histories either between two observed sequences or along a phylogenetic tree have been devised (Robinson et al., 2003; Rodrigue et al., 2005; Yu and Thorne, 2006).

Although these approaches could be applied to nearest-neighbor dependence models, they are computationally cumbersome. We therefore employ a branch-partitioning approach similar to Hwang and Green (2004) as it logically extends our data augmentation approach. This approach allows us to use as many partitions as necessary, with more partitions allowing for a more accurate reflection of reality but also increasing computational requirements. We believe, however, that the largest improvements in our approach will occur when initially partitioning a branch into a small number of partitions and that further partitioning will only increase the computational burden, without causing further meaningful changes to the results. To gain insight whether a sufficient number of partitions was used, we examined whether dividing the branches effectively yields different estimates for the evolutionary substitution parameters. As this approach will generate short branch partitions, we conjecture that a model which is reversible per context will also be (approximately) reversible overall.

Bayes Factor Calculation

By allowing for context-dependent evolution, evolutionary models become more parameter-rich. As previously discussed (Steel 2005), consistency problems may arise with such high-dimensional models, along with potential computational burdens. In view of this, a model-selection approach should be used that penalizes the addition of extra parameters unless there is a sufficiently impressive improvement in fit between model and data (Steel, 2005). One such objective criterion is the Bayes factor (Kass and Raftery, 1995). This is a ratio of two marginal likelihoods obtained under the two models to be compared. Because the harmonic mean estimator of the Bayes factor systematically favors parameter-rich models, we have chosen to calculate Bayes factors using thermodynamic integration (Lartillot and Philippe, 2006). For each model comparison, we have calculated the Bayes factor using the model-switch integration method (Lartillot and Philippe, 2006) and we have performed a bidirectional check; i.e., we have calculated both annealing and melting integrations under various settings to obtain very similar runs, as suggested in the work of Rodrigue et al. (2006). When comparing different models, we report Bayes factor estimates for both annealing and melting integrations, as well as their mean.

Data

To evaluate the presence of context effects we analyzed two datasets. A first dataset consists of 10 vertebrate species (Human, Chimpanzee, Gorilla, Orangutan, Baboon, Macaque, Vervet, Marmoset, Dusky Titi, and Squirrel Monkey). This dataset is a subset of the alignment analyzed in the work of Margulies et al. (2006). The original dataset consists of 31 sequences, all orthologous to a ~1.9-Mb region on human chromosome 7q31.3 (chr7:115 404 472 to 117 281 897 [May 2004 freeze] on the UCSC Genome Browser; Karolchik et al. 2003; Kent et al. 2002) and were generated by the NISC Comparative Sequencing Program (Thomas et al. 2003). This genomic region contains the cystic fibrosis transmembrane conductance regulator gene (CFTR). The original dataset was aligned using TBA (Blanchette et al. 2004), of which we have taken a subset using *maf_order* (Human as reference sequence, this avoids rerunning TBA with this subset of 10 sequences; Blanchette et al. 2003, http://bio.cse.psu.edu/). Ancestral repeats in the human sequence were detected using RepeatMasker (Smit et al. 1996–2004) with the RepBase Update libraries (September 2007; Jurka 2000). Simple repeats, low complexity regions, members of the Alu family, RNA elements that diverged less than 25%, and L1 elements that diverged less than 20% from the reconstructed ancestral sequence were removed using an adaptation of the script by Elliot Margulies (Margulies et al., 2003). The resulting coordinates were mapped onto our alignment, after which the ancestral repeat sequences were extracted. Only continuous stretches of at least 15 bases without gaps were retained from this subset. We have refrained from any further post-processing of this resulting alignment, as Siepel and Haussler (2003) and Hwang and Green (2004) found that their results, which were estimated on an alignment also containing the CFTR gene, did not appear to be very sensitive to their post-processing methods. Our retained ancestral repeat dataset consists of 114,726 sites for each of the 10 sequences and we refer to it as the "Ancestral Repeats" dataset.

A second (smaller) dataset consists of 20 small subunit (SSU) rRNA genes (nuclear) obtained from the alignment of Karol et al. (2001) after removing the gaps in our subset. We have used the following sequences: *Cyanophora paradoxa*, *Nephroselmis olivacea*, *Chlamydomonas moewusii*, *Volvox carteri*, *Paulschulzia pseudovolvox*, *Coleochaete orbicularis 2651*, *Coleochaete solute 32d1*, *Coleochaete irregularis 3d2*, *Coleochaete sieminskiana 10d1*, *Zygnema peliosporum*, *Mougeotia* sp758, *Gonatozygon monotaenium 1253*, *Onychonema* sp832, *Cosmocladium perissum 2447*, *Lychnothamnus barbatus 159*, *Nitellopsis obtusa F131B*, *Chara connivens F140*, *Lamprothamnium macropogon X695*, *Arabidopsis thaliana*, and*Taxus mairei*. We used the 50% majority-rule posterior consensus tree under the general time-reversible model with site independence in our analysis. We refer to this dataset as the "Nuclear SSU rRNA" dataset.

Results

MCMC Software Program

We have developed a software program to perform a Bayesian MCMC analysis given a user-defined tree, a dataset of aligned sequences (excluding gaps) and a general time-reversible model with behavior possibly dependent on the composition of the immediate flanking sites. Specifically, our program allows for this general time-reversible model to be assigned to a site based upon its two nearest neighbors. This way, up to 16 different models can be used in an MCMC analysis, one for each context. To avoid overparameterization, parameters can be shared between contexts when differences between those contexts are weak. For the remainder of this section, we will use the following notation to clarify the influence of the neighboring bases: AXG represents a site X under consideration, flanking with a base A as the 5′ neighbor and a base G as the 3′ neighbor.

Ancestral Repeats

Our analysis of the Ancestral Repeats dataset is based upon the phylogenetic tree shown in Figure 2. The context-dependent analysis was run for 100,000 update cycles (comprising about 103.3 billion parameter or missing ancestor updates), of which the first 20,000 were discarded as the burn-in. Analysis of the parameter estimates for each of the parameters (rAC, rAG, rAT, rCG, rCT, and rGT) in each of the 16 possible contexts of evolution yields a large diversity in substitution behavior, which we explain further in this section. We observe large variations in the transition parameters (rAG and rCT) depending on the identities of the immediate flanking bases as opposed to much smaller variations in the transversion parameters (rAC, rAT, rCG, and rGT). Note that due to the structure of the general time-reversible model, an increase in one parameter for a given neighboring base composition implies a decrease in at least one other parameter for that same neighboring base composition.

The substitution behavior for the transition parameters can be seen in Figure 3 for a tree with all branches divided into three to allow for evolving neighboring bases along each branch. We have compared these parameter estimates with those estimated when the branches are left undivided but found only small differences (see Fig. 1 of the online Supplementary Material; available at www.systematicbiology.org), suggesting that fixing the identities of the neighboring bases along a branch does not prevent our algorithm from reliably detecting substitution patterns. Further divisions of the branches did not

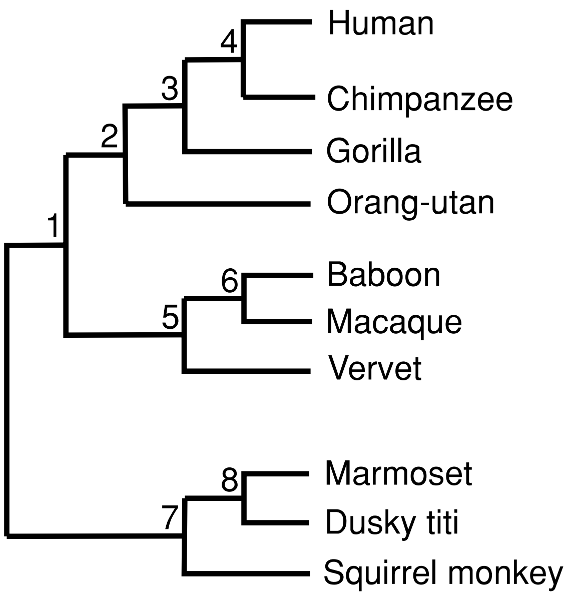

FIGURE 2. Rooted posterior tree of the 114,726-bp Ancestral Repeats dataset, following the phylogenetic tree reported in Margulies et al. (2006); branch lengths not drawn to scale.

lead to noticeable differences in substitution parameter estimates (data not shown).

The transversion parameters are mostly unaffected by the base composition of the two immediate flanking bases, except possibly for the rCG substitution parameter which shows a weak dependence on the neighboring bases (see Fig. 2 of the online Supplementary Material). Information for this parameter is weak, however, as suggested by its large posterior variance. This is most likely due to the underrepresentation of the CpG dinucleotide in mammalian DNA.

The transition substitution parameters (rAG and rCT) are most heavily affected by the neighboring base composition. We were unable to find a positive correlation between the A+T content of the flanking bases and each of the substitution parameters, in contrast to previous studies (Morton, 1995, 1997; Morton et al. 1997). As an example, the four contexts with the highest A+T content (i.e., AXA, AXT, TXA, and TXT) result in rCT parameter estimates that vary greatly between contexts, being large in context AXA and small in TXT. The inverted pattern for the rAG substitution parameter depending on A+T content, as opposed to the rCT substitution parameter, might be explained by considering Watson-Crick pairs. An example of this is shown for the CpG deamination process in Figure 4. In the presence of guanine at the 3′ side, cytosine is more likely to mutate into thymine; i.e., the CpG effect causes a C to mutate more likely to a T in the presence of G as the 3′ neighbor. The opposing base, which has C as its 5′ neighbor, correspondingly mutates to an A. We thus expect the rAG parameter to be elevated when C is the 5′ neighbor. This is supported by the results in Figure 3 upon averaging the estimates over the identity of the preceding base. Note that although we have modeled a dependence on both neighboring bases, the dependence on only the preceding site is of interest here to aid interpretation and because unidirectional dependencies have been suggested in previous studies. For example, Zhang and Gerstein (2003) show that the four transitional substitutions C → T, G → A, A → G, T → C and the transversion T → A are significantly affected by the 5′ neighboring base. Note from our results in Figure 3 that both the 5′ and 3′ neighboring bases affect the transitional substitutions. In other words, the rCT parameter increases when the 3′ neighboring base is a G (i.e., the CpG effect) and the level of the increase clearly depends upon the 5′ neighboring base.

CpG effects.—The presence of the CpG-deamination process is pronounced in Figure 3, with four of the six highest rCT estimates being obtained when the 3′ neighbor is guanine. The importance of the 5-methylation deamination process (i.e., the CpG effect) in mammals has been recognized and lies at the basis of an underrepresentation of the CpG dinucleotide in vertebrates (Jabbari and Bernardi, 1998). We also observe a decrease in rCT substitution parameter estimates depending on the succeeding base, in the following order: G > A > C > T, within contexts with the same preceding base. It thus follows that both the preceding base and the succeeding base determine the substitution pattern of a given site.

As far as transversions are concerned, the main differences depending on the neighboring base composition are observed in the rCG substitution parameter. Indeed, the rCG parameter shows variation when either the preceding base is a C or the succeeding base is a G, which indicates the presence of another context-dependent process (i.e., CpG transversion). This process has been found important in modeling evolution, albeit less than the CpG transition process (Arndt and Hwa, 2005). Arndt and Hwa (2005) motivate this process by the existence of another CpG-based process, which is likely also triggered by the methylation of cytosine, although Blake et al. (1992) and Siepel and Haussler (2004) believe the mechanism behind CpG transversions to be different from that behind CpG transitions. The rAT substitution parameter estimates seem almost entirely unaffected by the composition of the two immediate flanking bases, with the highest estimates when the neighboring context is TXA. It is unclear whether this can be attributed to the instability of the TT dimer, which has been shown to mainly trigger 3′ T → A transversions both in single-stranded and hairpin-containing vectors (Banerjee et al. 1988; Arndt and Hwa 2005).

TpA effects.—Whereas four of the six highest rCT estimates are observed when the 3′ neighbor is guanine, the other two elevated estimates are those found in the AXA and GXA contexts, suggesting higher estimates for the rCT parameter when A is the 3′ neighbor, depending on the identity of the preceding base. This finding, i.e., that an A as the succeeding base increases the rCT substitution parameter, supports previous observations (Blake et al., 1992; Hess et al., 1994). Just like the well-known CpG process of evolution, there thus appears to

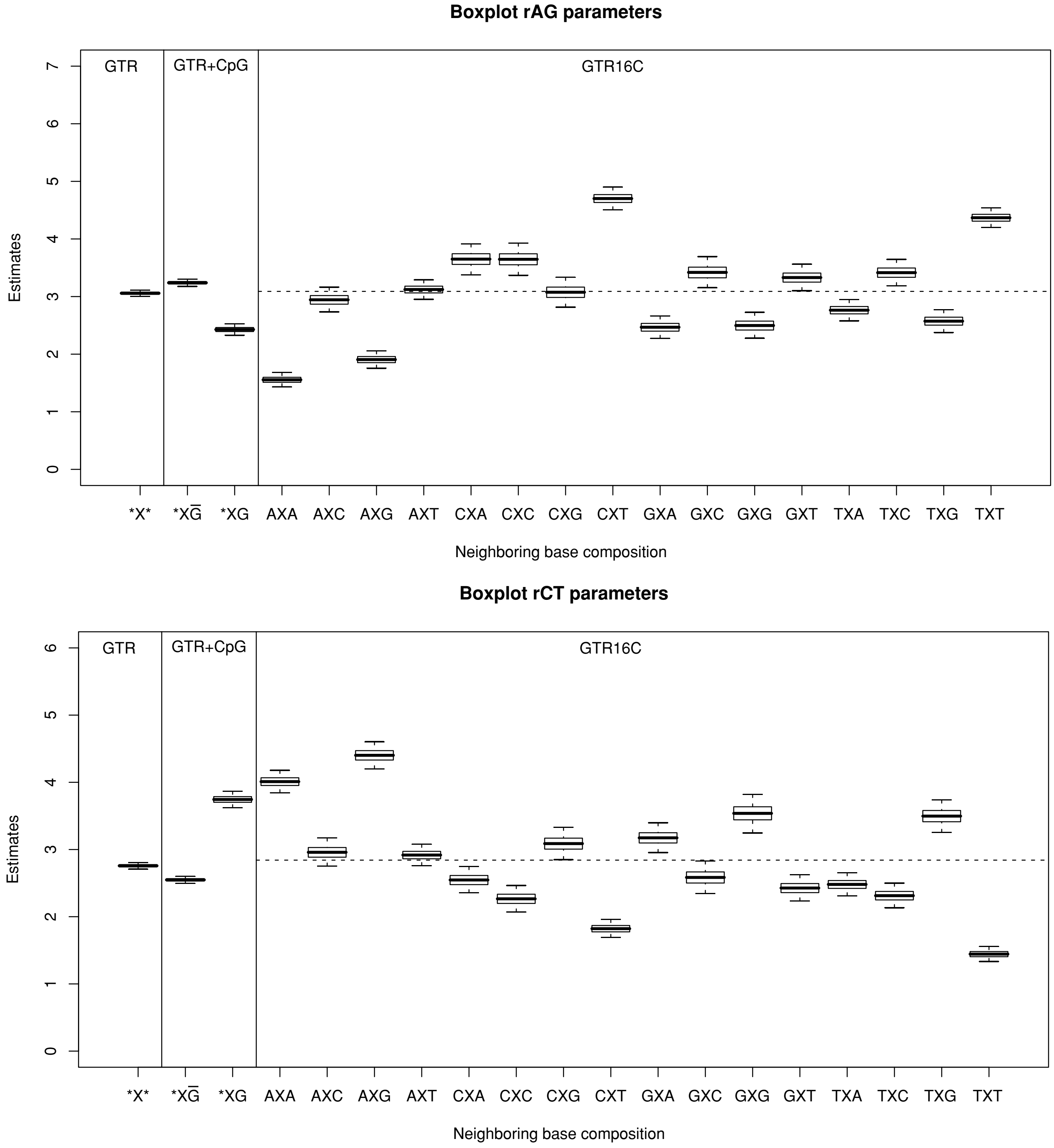

FIGURE 3. Ancestral Repeats dataset: 95% credibility intervals versus context for the rAG and rCT parameters in our independence model (GTR), our CpG model (GTR+CpG), and our 16-context model (GTR16C), with all branches divided into three pieces to allow for evolving neighboring sites. The independence model only estimates one rCT parameter (indicated by *X*), whereas the CpG model estimates two rCT parameters: one rCT parameter when the 3′ neighbor is guanine and one rCT parameter when the 3′ neighbor is not guanine (indicated respectively by *XG and $*X\overline{G}$) and the 16-context model estimates 16 rCT parameters (from AXA to TXT). For both transition parameters, a strong dependence on the evolutionary context can be observed. The rCT estimates reveal the inverse pattern of the rAG estimates due to the lack of influence on the neighboring base composition of the transversion parameters (see the online supplementary material; www.systematicbiology.org).

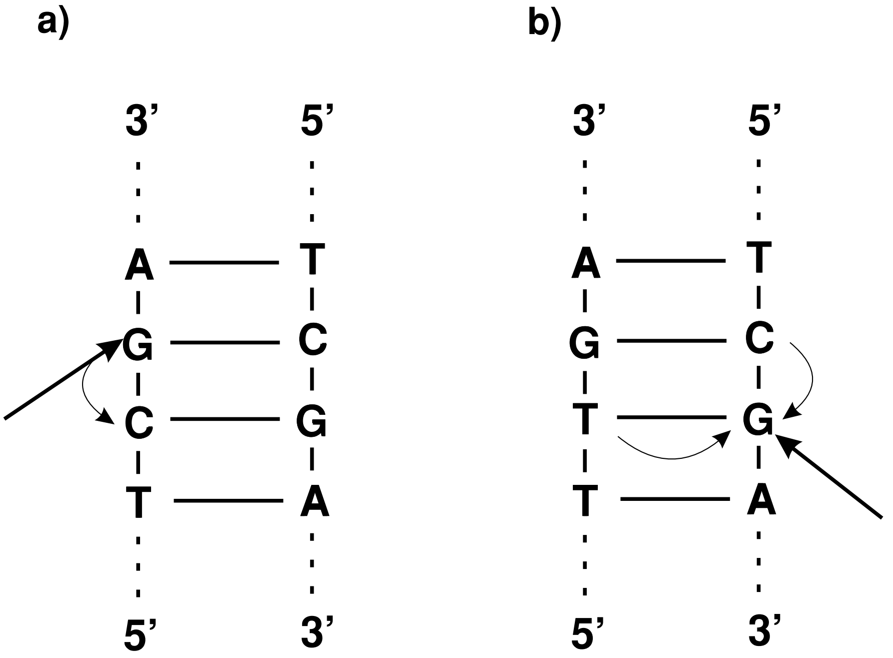

FIGURE 4. Influence of neighboring bases on the substitution behavior of a given site in the case of a CpG effect. (a) Base C is influenced by the presence of a G to its right and the CpG-effect (with an elevated rCT substitution parameter) comes into play: the C mutates to a T, as can be seen in (b) where the Watson-Crick pair CG has mutated to a TG pair. To maintain stability in this stem region, the base at the opposing side of the stem must mutate as well. This is not a CpG effect but its inverse effect; i.e., the mutation of a G with a C preceding it. The base G mutates to an A, as can be seen in (c), which explains the higher parameter estimates for the rAG parameter estimates when the base is preceded by a C. A similar effect (a TpA-effect) can be described for the rCT substitution parameter in the TpA base pair, albeit somewhat weaker.

be a "TpA effect," albeit only conditional on the preceding base. This is also perceivable in the rAG parameter when the preceding base is a T or an A and in the rCT parameter when the succeeding base is a T.

Like the CpG dinucleotide, the TpA dinucleotide has been found to be underrepresented in vertebrate DNA sequences, with Fryxell and Zuckerkandl (2000) claiming that the CpG effect is responsible for both. The CpG underrepresentation is said to be compensated by the overrepresentation of CpA and TpG and these two dinucleotides are also just one mutation away from the TpA dinucleotide (Duret and Galtier, 2000; Arndt and Hwa, 2003). Zhang and Gerstein (2003) claim that the underrepresentation of TpA is caused by an increased substitution probability of A in the presence of T, to make TpA less conserved than other dinucleotides such as ApA, GpA, and CpA.

Several possible explanations for the underrepresentation of TpA have been put forward, such as an increased substitution probability of T to C in TpA dinucleotides (i.e., a "TpA effect" where T mutates to C in the presence of A at the 3′ side; Blake et al., 1992; Hess et al., 1994) and non-CpG methylation, a contentious issue in mammalian DNA (Ramsahoye et al., 2000). The deamination of 5-methylcytosine occurring (in the CpG dinucleotide) in the reverse strand of DNA can lead to CpA dinucleotides, which further mutate to TpA (Yang et al., 2004). Because the TpA dinucleotide is underrepresented in mammalian DNA, however, another process must be responsible for (further) altering of TpA.

Dinucleotide frequencies.—We have also examined the single and dinucleotide frequencies of this dataset. The single nucleotide frequencies were obtained from our software program, whereas we used a simple counting method to establish the dinucleotide frequencies from the observed sequences. The results can be seen in Table 1. Although such frequencies are mainly informative of the genetic composition of the analyzed data (which comprises repeats with different features), overall these frequencies can be helpful in identifying important substitution processes. Based on the estimated single-nucleotide frequencies, there is an overabundance of both A and T, at the expense of C and G. The single-nucleotide frequencies can be used to calculate dinucleotide odds ratios, which indicate whether a

TABLE 1. Ancestral Repeats dataset: estimates of the single nucleotide frequencies, as obtained from the MCMC analysis, and estimates of the dinucleotide frequencies, obtained using a simple counting method.

α	f_α	$f_{\alpha A}$	$f_{\alpha C}$	$f_{\alpha G}$	$f_{\alpha T}$
A	0.300	0.093	0.050	0.069	0.085
C	0.192	0.071	0.046	0.005	0.072
G	0.192	0.059	0.037	0.046	0.052
T	0.317	0.074	0.061	0.074	0.104

TABLE 2. Ancestral Repeats dataset: estimates of the dinucleotide odds ratios [$\rho_{\alpha\beta} = f_{\alpha\beta}/(f_\alpha f_\beta)$], based on the estimates of the dinucleotide frequencies, obtained using a simple counting method. These odds ratios indicate whether a specific dinucleotide pair is underrepresented ($\rho_{\alpha\beta} < 1$) or overrepresented ($\rho_{\alpha\beta} > 1$). As can be seen, the CpG dinucleotide is heavily underrepresented, which will be compensated by the overrepresentation of CpA and TpG. The underrepresentation of TpA, although less drastic, is also apparent from this table.

$\rho_{\alpha\beta}$	$\beta = A$	C	G	T
$\alpha = A$	1.039	0.877	1.199	0.894
C	1.235	1.253	0.147	1.191
G	1.030	1.006	1.249	0.858
T	0.781	1.011	1.306	1.035

certain dinucleotide is under- or overrepresented. As can be seen from Table 2, the CpG dinucleotide is heavily underrepresented, which is compensated (in part) by an increased presence of CpA and TpG dinucleotides. Apart from these dinucleotides, which are directly influenced by the CpG process, an overrepresentation of both CpC and GpG dinucleotides and an underrepresentation of the TpA dinucleotides are the most obvious conclusions. These estimates show roughly the same pattern as results by Arndt et al. (2003).

The results from Table 2 illustrate the complexity of evolutionary patterns. Judging by our rCT parameter estimates, the AXA context leads to an increased probability of substitution between C and T. Blake et al. (1992) observed an increased probability of substituting a T with a C in the presence of A as the 3′ neighbor. This suggests that triplets such as ATA (TAT on the reverse strand) are likely to mutate into ACA (TGT). This concurs with our rCT (rAG) estimates but has implications for the presence of the dinucleotides involved. The presence of both the ApT and TpA dinucleotide will decrease, as observed in Table 2, and the ApC and CpA dinucleotide counts will increase. The latter is difficult to confirm, given that CpA is already affected by the CpG process. On the reverse strand, however, the TpA and ApT counts increase again (thus canceling the previous increases) and the TpG and GpT counts increase as well. The increase of the GpT count, however, could not be observed in our estimates, which could indicate that many of the GpT dinucleotides have mutated to another dinucleotide.

Bayes factors.—Finally, note that Bayes factor calculations using thermodynamic integration (Lartillot and Philippe, 2006) are computationally cumbersome, given the magnitude of this dataset. Bayes factors are generally divided into four categories depending on their value: from 1 to 3, indicating nothing worth reporting; from 3 to 20, indicating positive evidence of one model over another; from 20 to 150, indicating strong evidence of one model over another; and larger than 150, indicating significant (or very strong) evidence of one model over another (Kass and Raftery, 1995). Given the computational demands, we have focused our efforts on comparing plausible situations of site-dependencies, reflected in three dependence models of which the improvement over the independence model is given in Table 3.

TABLE 3. Ancestral Repeats dataset: due to the computational requirements, only a few selected complex models could be tested for this large dataset. A model containing all 16 contexts (GTR16C), a model which incorporates CpG effects (GTR+CpG), a model aimed at incorporating CpG effects dependent on the previous base (GTR+XpCpG), and a model incorporating varying rates across sites (GTR+Γ4; discrete gamma distribution with four rate classes) were evaluated against the independent general time-reversible model (GTR) by calculating the appropriate Bayes factors. All these models significantly outperform the independence model, and the GTR16C model even outperforms a model incorporating varying rates across sites (GTR+Γ4). For each model (first column), the number of evolutionary contexts is reported (second column; corresponding number of parameters between brackets) along with the log Bayes factor for both annealing (third column) and melting (fourth column) schemes of the model-switch integration method (Lartillot and Philippe, 2006). The mean value of these two schemes (i.e., a bidirectional check; fifth column) is then used to calculate the actual Bayes factor (sixth column).

Model	Contexts	Annealing	Melting	Bidirectional	Bayes factor
GTR16C	16 (96)	630.63	653.67	642.15	76.26×10^{277}
GTR+Γ4	1 (6)	342.75	369.38	356.06	43.27×10^{153}
GTR+ XpCpG	5 (30)	148.86	167.07	157.96	40.10×10^{67}
GTR+CpG	2(12)	134.00	141.72	137.86	74.52×10^{58}
GTR	1 (6)	0	0	0	1

First, we have tested a two-context model (which uses one set of parameters for those sites whose right neighbor is a guanine, and one set of parameters for all other sites) aimed at incorporating CpG effects. Although this model performs significantly better (shown by a Bayes factor of 74.52×10^{58}) than the independence model, there is an indication that the CpG-deamination process depends on the preceding base (see Fig. 3; Zhang and Gerstein, 2003). We have thus tested a CpG model dependent upon the preceding base, consisting of a general independent site category and four site-dependent categories; i.e., one for each possible preceding base for the CpG-deamination process. This model outperforms the simple CpG model, as can be seen in Table 3. Given the wide range of parameter estimates detected, apart from those related to the CpG-deamination process, we have tested a model containing all 96 parameters and found that it outperforms any other model tested, yielding a Bayes factor of 76.26×10^{277} over the independence model.

As the evolutionary rate can vary enormously among sites, modeling among-site rate variation (or rates across sites) will in many cases significantly outperform a standard independence model. We have tested this assumption using a discrete gamma approximation with four rate classes to model the varying rates across sites (Yang, 1994). With a Bayes factor of 43.27×10^{153}, modeling among-site rate variation yields a significant improvement in model fit over the independence model and also outperforms both CpG models tested. The mean estimate for the shape parameter of the gamma distribution equals 1.156. However, our most complex site-dependent model clearly outperforms the model with among-site rate variation, indicating that the Ancestral Repeats dataset offers tremendous support for the context effects studied. Given the computational burden, we have refrained from testing models that superimpose among-site rate variation upon site-dependence but conjecture that such models will further improve model fit.

Nuclear SSU rRNA

Our analysis for the nuclear SSU data is based upon the posterior consensus tree as shown in Figure 5. The tree was rooted using the outgroup sequence *Cyanophora paradoxa*. Convergence of the clade posterior probabilities was confirmed using AWTY (Nylander et al., 2004). The analysis was run during 200,000 update cycles (comprising about 6.2 billion parameter or missing ancestor updates), of which the first 20,000 were discarded as the burn-in. In order to be consistent with the analysis of the CFTR dataset, we have also split the branches of the posterior tree for this dataset into multiple pieces. As this alignment is smaller (containing only 1619 sites compared to 114,726 sites for the CFTR dataset), broader credibility intervals will be obtained. This makes it more difficult to determine the influence of splitting a branch. We have thus extrapolated the settings for the analysis of the CFTR dataset where each branch section has a mean length of 4.86×10^{-3} substitutions per site. This required us to split each branch into four equal parts, which only led to small differences compared to when the branches are left undivided (see Fig. 3 of the online Supplementary Material).

CpG effects.—The presence of CpG effects can be seen in Figure 6 where the rCT substitution parameter estimates for each context are shown. Those contexts consisting of a G as the 3′ neighbor of the given site clearly yield the highest estimates. Apart from the CpG-effects, the CXA and CXT evolutionary contexts stand out with the lowest estimates for the rCT parameter. The estimates for the transversion parameters in the different neighboring base compositions now seem to be more widespread and vary considerably around their mean values (Fig. 6; mean represented by the dotted line) as opposed to the transversion estimates in the CFTR dataset. This variation, however, makes it difficult to determine trends in the transversion estimates depending on the neighboring base composition. As in the CFTR dataset, it makes sense to interpret the estimates conditional on the identity of the preceding base. Conditional on the 5′ neighbor, the estimates for the rAC transversion are generally higher with increasing A+T content. This can be seen by considering the four contexts that have thymine as the 5′ neighbor: TXA, TXC, TXG, and TXT. The contexts TXA

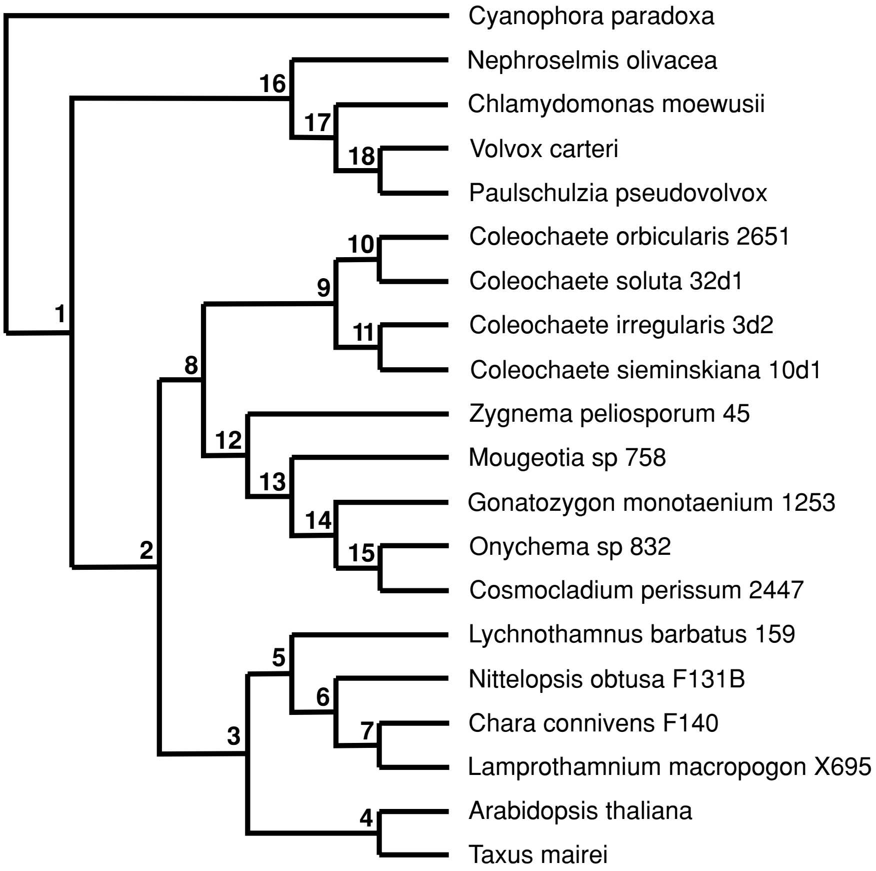

FIGURE 5. Rooted tree of the 1619-bp Nuclear SSU rRNA dataset, with branch lengths not drawn to scale, as determined using MrBayes (Ronquist and Huelsenbeck, 2003; GTR, equal rates, 2×10^6 iterations). Given the low percentage (±3%) of gaps present in the original alignment, these could be removed without affecting the assumed dependencies.

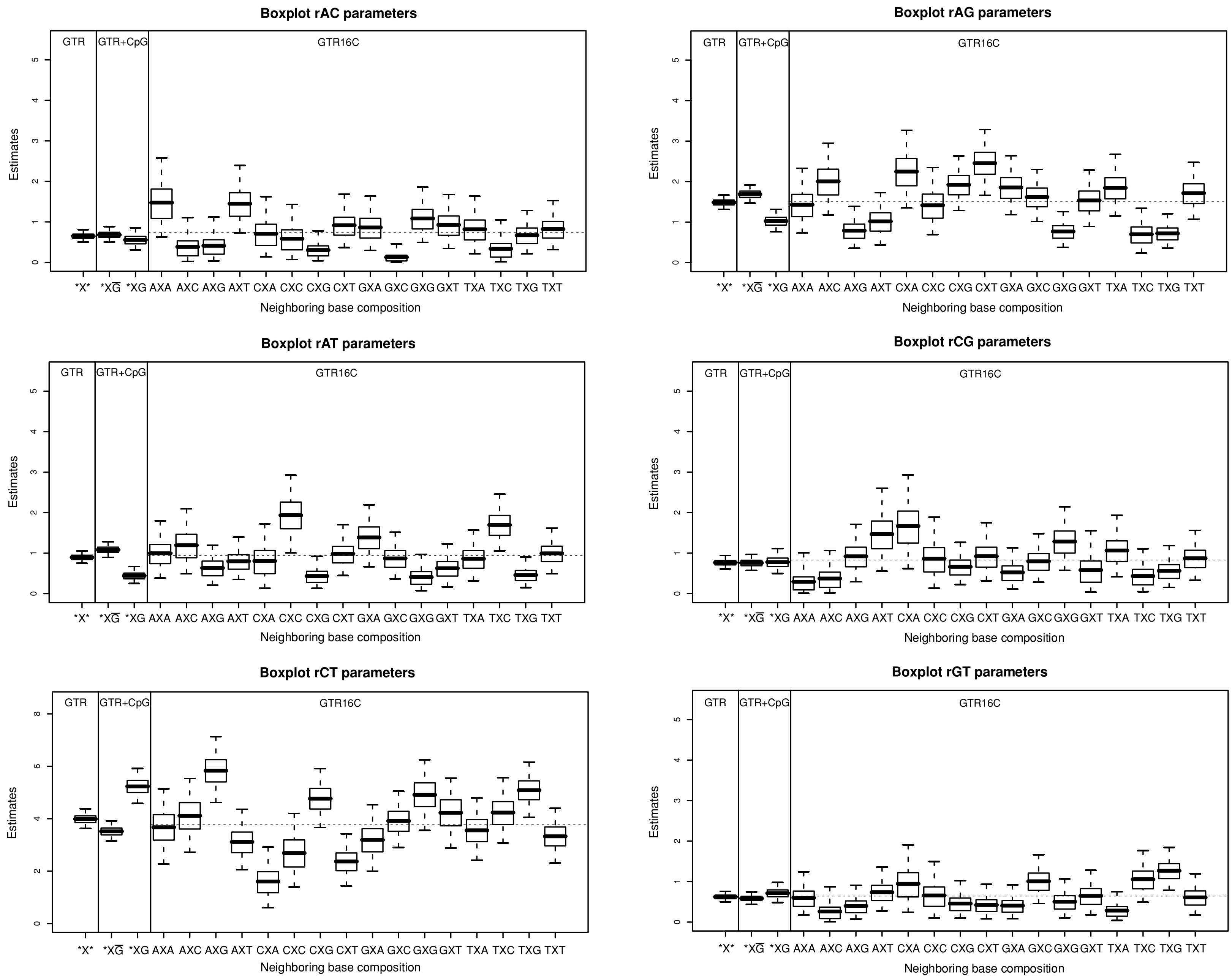

FIGURE 6. Nuclear SSU rRNA dataset: comparative plots for all 96 evolutionary parameters present in our independence model (GTR), our CpG model (GTR+CpG), and our 16-context model, with all branches divided into four pieces to allow for evolving neighboring sites. The independence model only estimates one rCT parameter (indicated by *X*), whereas the CpG model estimates two rCT parameters: one rCT parameter when the 3′ neighbor is guanine and one rCT parameter when the 3′ neighbor is not guanine (indicated respectively by *XG and $*X\overline{G}$) and the 16-context model estimates 16 rCT parameters (from AXA to TXT).

and TXT lead to the highest rAC estimates of those four contexts but cannot be claimed as leading to higher estimates than, for example, the GXA and GXT contexts. This observation is most outspoken when the 5′ neighbor is adenine. A similar observation can be made for the rCG transversion parameter, where a C or T as the 5′ neighbor leads to the higher estimates with increased A+T content, whereas this pattern is not observed with either an A or a G as the 5′ neighbor. The rAT transversion parameter does not show any relationship regarding the presence of higher estimates and the A+T content of the immediate neighbors. Finally, only when the 5′ neighbor is an A does an increased A+T content lead to higher rGT estimates, whereas a T as the 5′ neighbour leads to lower rGT estimates with increasing A+T content. No clear pattern could be established when the 5′ neighbor is either a C or a G.

Bayes factors.—Given the smaller number of sites in this alignment, Bayes factor calculations using thermodynamic integration (Lartillot and Philippe, 2006) are more computationally feasible, so that we were able to test a larger number of models. We have tested many possible assumptions of site dependencies to determine which neighbor combinations are informative and which mainly add noise. Starting from the standard general time-reversible model, we have constructed and evaluated all 16 possible 12-parameter models, consisting of a general independent site category and one site-dependent category that models the evolution of the neighboring base combination under consideration. This approach is reminiscent of the work by Arndt and Hwa (2005), although we try to determine which evolutionary contexts are informative instead of trying to identify single evolutionary processes which maximize the potential of the model. As for the Ancestral Repeats dataset, we have tested whether using a separate evolutionary model for each evolutionary context (i.e., using every possible parameter) improves fit and whether the two CpG scenarios (i.e., independent and dependent on the preceding base) yield a significant improvement over the independence model. The resulting Bayes factors are reported in Table 4. These Bayes factors indicate that there is significant support for a CpG effect that is dependent upon the preceding base, but that in this plant dataset the modeling of a CpG effect independent of the preceding base is preferred. Other than these two models, only one other model (CXT) offers a significant improvement (Bayes factor larger than 150) over the independent sites approach using a general time-reversible model.

Three contexts with a G on the 3′ side of the considered site (TXG, GXG, and AXG) offer only a slight to mild advantage over the independence model, probably due to the lack of data given the large variation in posterior means across context for the different evolutionary

TABLE 4. Nuclear SSU rRNA dataset: the performance of the 16 possible two-context models (i.e., models treating one neighbor combination consistently different from all other neighbor combinations, which are grouped into one category), a model containing all 16 contexts (GTR16C), a model that incorporates CpG effects (GTR+CpG), a model aimed at incorporating CpG effects dependent on the previous base (GTR+XpCpG), and a model incorporating varying rates across sites (GTR+Γ4; discrete gamma distribution with four rate classes), was calculated and ordered performance-wise (starting with the best performing model). The GTR+CpG model only considers two contexts: one for those sites flanked with a G at the 3′ side (so the CpG effects are considered unconditional of the 5′ neighboring site) and one for all other sites. The rates across sites model offers the highest improvement, with other significant improvement provided by those models incorporating CpG effects and a two-context model (CXT). The performance of the GTR+Γ4 model is surpassed only by combining rates across sites with models describing CpG effects. Many models perform worse than the independence model, due to the low amount of data. For each model (first column), the number of evolutionary contexts is reported (second column; corresponding number of parameters between brackets) along with the log Bayes factor for both annealing (third column) and melting (fourth column) schemes of the model-switch integration method (Lartillot and Philippe, 2006). The mean value of these two schemes (i.e., a bidirectional check; fifth column) is then used to calculate the actual Bayes factor (sixth column).

Model	Contexts	Annealing	Melting	Bidirectional	Bayes factor
GTR+CpG+Γ4	2 (12)	500.0532	506.0389	503.0461	2.95×10^{218}
GTR+XpCpG+Γ4	5 (30)	500.3109	504.7276	502.5193	1.74×10^{218}
GTR+Γ4	1 (6)	494.5248	503.6716	499.0982	5.69×10^{216}
GTR+CpG	2 (12)	10.6526	13.5526	12.1026	1.80×10^{5}
GTR+XpCpG	5 (30)	6.3761	4.3131	5.3446	2.09×10^{2}
GTR+CT	2 (12)	4.0705	5.9777	5.0241	1.52×10^{2}
GTR+TG	2 (12)	4.0812	4.6931	4.3872	80.41
GTR+TC	2 (12)	0.9263	3.0526	1.9895	7.31
GTR+GG	2 (12)	0.7634	1.5995	1.1815	3.26
GTR+CA	2 (12)	0.2511	1.3585	0.8048	2.24
GTR+AG	2 (12)	1.6107	−1.2555	0.1776	1.19
GTR	1 (6)	0	0	0	1
GTR+CG	2 (12)	−0.4311	−0.0559	−0.2435	7.38×10^{-1}
GTR+GC	2 (12)	−2.6343	0.8306	−0.9019	4.06×10^{-1}
GTR+CC	2 (12)	−3.6232	−1.0649	−2.3441	9.59×10^{-2}
GTR+AT	2 (12)	−2.5154	−2.4718	−2.4936	8.26×10^{-2}
GTR+AC	2 (12)	−4.1737	−1.3632	−2.7685	6.28×10^{-2}
GTR+AA	2 (12)	−3.4107	−3.8217	−3.6162	2.69×10^{-2}
GTR+TA	2 (12)	−4.1968	−3.0893	−3.6431	2.62×10^{-2}
GTR+GA	2 (12)	−6.6161	−1.7411	−4.1786	1.53×10^{-2}
GTR+TT	2 (12)	−4.1336	−4.4931	−4.3134	1.34×10^{-2}
GTR+GT	2 (12)	−5.9059	−6.4557	−6.1808	2.07×10^{-3}
GTR16C	16 (96)	−18.9466	−16.8655	−17.9061	1.67×10^{-8}

parameters. Two other models (TXC and CXA) were slightly better than the standard independence model. All other 12-parameter models performed worse than the independence model, indicating a lack of support in the data for the additional parameters. Even though we have found a wide range of substitution patterns (see Fig. 6), modeling every possible context of evolution based on the two immediate flanking bases performs worst. This can be attributed to overfitting of the data (Rosenkranz and Raftery, 1994).

In this plant dataset, we have also tested the impact of modeling among-site rate variation (or rates across sites) using a discrete gamma-approximation using four rate classes (Yang, 1994). The mean estimate for the shape parameter of the gamma distribution equals 0.203, indicating that most sites have very low substitution rates, whereas few sites have very high rates (Yang, 1996a). We found that modeling varying rates across sites substantially increases the model fit over the independence model and over each dependent model we have tested. This can be attributed to strong purifying selection in rRNA (see, e.g., Rooney, 2004), which can be captured by assuming varying rates across sites. Because only a few dozen cases of ancestral repeats that have come under purifying selection are known (Kamal et al., 2006), assuming varying rates across sites in the ancestral repeats dataset did not yield higher Bayes factors than the assumption of context-dependent substitution probabilities. As can be seen in Table 4, this small dataset makes it feasible to superimpose among-site rate variation on the assumption of context-dependence, which leads to an even (slightly) better model fit.

Branch Length Estimation

Our analysis of the ancestral repeats dataset with 16 neighboring base compositions revealed a wide variation in substitution patterns, specifically for the transition parameters. Given this variation, the independence model may fail to capture the observed effects correctly We have compared the branch-length estimates between a 16-context-dependent model, a rates-across-sites model, and the independence model (see Table 1 in the online Supplementary Material). Ignoring the branch connecting ancestral sequences 7 and 8 (because of its short length), the differences in branch lengths under the context-dependent model ranged from an increase of 1.27% to an increase of 5.60%, with the largest differences in branch-length estimates appearing in the ((Marmoset, Dusky Titi), Squirrel Monkey) clade. The differences in branch lengths when assuming varying rates across sites ranged from a decrease of 2.75% to an increase of 4.21%.

It has been shown that incorporating varying rates across sites allows for more accurate distance calculations (Yang, 1996b). We have found for this dataset that modeling varying rates across sites substantially increases the model fit, but that the 16-contexts model improves model fit to a much larger extent. This is also reflected in the branch-length estimates under these two assumptions: in all but two branches (i.e., the branch connecting the outgroup to the ingroup and the branch leading to Dusky Titi), the context-dependent model caused larger differences in branch-length estimates than the rates-across-sites model. Under our context-dependent model, the branches leading to the observed sequences of Marmoset and Squirrel Monkey show differences that are respectively 5.60% and 5.41% in length. The branch length between ancestral sequences 7 and 8 shows a large decrease under the context-dependent model but a dramatic decrease when assuming varying rates across sites. It should, however, be noted that the absolute difference is modest and that mixing of this branch length is poorer, given its short length.

We repeated the previous comparison for the nuclear SSU rRNA dataset, now using the best-fitting CpG model since it yields a better model fit than the 16-contexts model. As can be seen from Table 2 in the online Supplementary Material, differences in branch lengths occur when modeling context-dependent evolution. The differences in branch-length estimates ranged from 0.1% up to 6.6%, with the largest difference appearing on the branch leading to Lychnothamnus and the smallest difference occurring at the branch leading to Arabidopsis. Assuming varying rates across sites, however, yields larger differences in branch lengths from 2.5% up to 62.4%, except for the branch connecting ancestral sequences 1 and 2. The branch length estimates now increase in all but two cases, versus half of the cases for the context-dependent model. The latter thus often suggests a decrease in branch length whereas the among-site rate variation model indicates a (large) increase in branch length.

Discussion and Future Work

In this paper, we have introduced a general framework to model site dependencies. Although this paper focuses on the influences of the immediate flanking sites on the evolution of a site, our approach can accommodate fairly general dependence structures. Phylogenetic inference now requires more computational resources but will not be seriously hindered by the use of data augmentation. We have shown that adding extra parameters to model site dependencies may result in both an increase and decrease in performance. Although our results show that modeling all nearest-neighbor combinations leads to a drastic improvement in model fit in a large dataset of non-coding sequences, further improvements may be possible because the evolutionary process may be very similar for some neighbor combinations. Separate models for these combinations are then bound to have a negative impact on performance and thus these contexts should either be omitted or be grouped together to reduce sampling variation. The general time-reversible model that underlies our model might also be restrictive, as certain substitution processes appear to be non-reversible (see, e.g., Blake et al., 1992; Hwang and Green, 2004). The reported substitution probabilities we have reported thus summarize the estimates as if all these substitution processes are symmetric. For some substitution processes, this assumption is reasonable. For example, the transition between C and T has been reported to be fairly symmetric in contexts CXA, GXT, GXA, and AXA (Hwang and Green, 2004). In other contexts, mainly those corresponding with CpG dinucleotides, the symmetry of transition might be oversimplistic as the substitution probability in one direction may differ drastically from the other. This is probably the reason why our transition parameter estimates between C and T are lower than those reported by Hwang and Green (2004), as the transition probability from T to C is generally lower than the transition probability from C to T. Modeling such asymmetric processes will introduce additional parameters but might further improve the model fit to the data.

Empirical studies have shown that neighboring bases have an influence on the occurrence of a base and that, both in coding and non-coding sequences, at least a second-order Markov chain is required for their representation (Blaisdell, 1985). Erickson and Altman (1979) found only a slight first-order Markov chain dependence but a significant second-order dependence in their analysis of the virus MS2 sequence. These authors also reported a dependency of the third codon site identity on the identity of both the preceding and succeeding site in some genes. A motivation for a first-order Markov chain along the root sequence in addition to a dependency on the neighboring bases can be found in the work of Jensen and Pedersen (2000). Hwang and Green (2004) modeled a second-order Markov chain along the root sequence, although not allowing for such a dependency structure in the other internal sequences. The modeling of such dependencies along the sequence results in a drastic increase in the number of parameters as well as computation time. To the best of our knowledge, the ensuing increase in model fit has not been evaluated in terms of a Bayes factor (or some other type of model selection), although differences in the nucleotide and dinucleotide frequencies suggest its relevance (Arndt et al., 2003).

We anticipate that modeling site-specific dependencies may have an important impact on several aspects of phylogenetic inference, as in some cases these dependence models even outperform models that incorporate varying rates across sites. As we have shown in this article, the estimation of branch lengths is influenced by using context-dependent models. We thus anticipate that clade posterior probabilities and even the tree topology might be estimated more accurately using these context-dependent models. Standard (independence) models, which ignore such dependencies, are prone to incorrectly state the precision of the results and might also yield biased phylogenetic inferences by misestimating the "distance" between observed nucleotide sequences. It remains an open question how to infer phylogenetic trees under our model because the tree space is discrete and new tree proposals in the MCMC algorithm typically require the introduction of new ancestral states that were not previously considered in the data augmentation algorithm. This is the topic of ongoing research.

As we are mainly interested in the effect of the neighboring bases on the substitutions at a particular site, we did not account for varying rates across sites in our estimation of the substitution parameters. In our estimations, all sites are thus assumed to have an equal rate of change per unit branch length. Note that varying rates across sites and context dependence are entirely different evolutionary processes. In particular, current rates-across-sites models do not allow the evolutionary parameters to change at each internal node (e.g., when the ancestral neighboring states of a site are estimated to differ across the tree). We have therefore tested the impact of context effects as well as assuming varying rates across sites in both datasets. Testing the combined performance of these two processes was only feasible on a small dataset, but we suspect that additional gains (and thus better models, in terms of Bayes factors) can be obtained by combining these two processes.

Finally, when inferring trees under the assumption of site independence, assembling a dataset does not require specific precautions. Should a position or region of a sequence be difficult to align, it can easily be deleted since it has no effect on other positions of regions within that same sequence. When considering the assumption of site dependence, this is no longer true. When gaps are removed in the alignment, certain nucleotides will be lost and dependencies between the remaining nucleotides will be weakened. This will reduce the power to detect context effects and could bias the results to those from the independent model.

ACKNOWLEDGMENTS

We would like to thank Joseph P. Bielawski, Alexei Drummond, and Arndt von Haeseler for helpful comments. Editor in Chief Jack Sullivan, Associate Editor Olivier Gascuel, Nicolas Lartillot, and two anonymous referees provided useful comments on the first version of this article. S.V. acknowledges support from IAP research network grant nr. P06/03 from the Belgian government (Belgian Science Policy).

REFERENCES

Arndt, P. F., C. B. Burge, and T. Hwa. 2003. DNA sequence evolution with neighbor-dependent mutation. J. Comp. Biol. 10:313–322.

Arndt, P. F., and T. Hwa. 2005. Identification and measurement of neighbor-dependent nucleotide substitution processes. Bioinformatics 21:2322–2328.

Banerjee, S. K., R. B. Christensen, C. W. Lawrence, and J. E. Leclerc. 1988. Frequency and spectrum of mutations produced by a single cis-syn thymine-thymine cyclobutane dimmer in a single-stranded vector. Proc. Natl. Acad. Sci. USA 85:8141–8145.

Bérard, J., J. B. Gouéré, and D. Piau. 2008. Solvable models of neighbor-dependent substitution processes. Math. Biosci. 211:56–88.

Blaisdell, B. E. 1985. Markov chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eukaryotic nuclear DNA sequences both protein-coding and noncoding. J. Mol. Evol. 21:278–288.

Blake, R. D., S. T. Hess, and J. Nicholson-Tuell. 1992. The influence of nearest neighbours on the rate and pattern of spontaneous points mutations. J. Mol. Evol. 34:189–200.

Blanchette, M., W. J. Kent, C. Riemer, L. Elnitski, A. F. A. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. D. Green, D. Haussler, and W. Miller. 2004. Aligning multiple genomic sequences with the Threaded Blockset Aligner. Genome Res. 14:708–715.

Bulmer, M. 1986. Neighbouring base effects on substitution rates in pseudogenes. Mol. Biol. Evol. 3:322–329.

Christensen, O. F., A. Hobolth, and J. L. Jensen. 2005. Pseudo-likelihood analysis of codon substitution models with neighbour-dependent rates. J. Comp. Biol. 12:1166–1182.

Cowell, R., A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. 1999. Probabilistic networks and expert systems. Springer-Verlag, New York.

Duret, L., and N. Galtier. 2000. The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochors is due to a mathematical artefact. Mol. Biol. Evol. 17:1620–1625.

Erickson, J. W., and G. Altman. 1979. A search for patterns in the nucleotide sequence of the MS2 genome. J. Math. Biol. 7:219–230.

Felsenstein, J. 1973. Maximum likelihood and minimum steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. 22:240–249.

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

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

Felsenstein, J., and G. A. Churchill. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. Mol. Biol. Evol. 13:93–104.

Fryxell, K. J., and E. Zuckerkandl. 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. Mol. Biol. Evol. 17:1371–1383.

Gelman, A., and D. B. Rubin. 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7:457–472.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1996. Markov chain Monte Carlo in practice. Chapman & Hall, London.

Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11:725–736.

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

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

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

Hess, S. T., J. D. Blake, and R. D. Blake. 1994. Wide variations in neighbour-dependent substitution rates. J. Mol. Biol. 236:1022–1033.

Huelsenbeck, J. P. 2000. Likelihood-based inference of phylogeny. Marine Biological Laboratory Workshop on Molecular Evolution: Lectures. July 30–August 11, 2000.

Huelsenbeck, J. P., J. P. Bollback, and A. M. Levine. 2002. Inferring the root of a phylogenetic tree. Syst. Biol. 51:32–43.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314.

Hwang, D. G., and P. Green. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc. Natl. Acad. Sci. USA 101:13994–14001.

Jensen, J. L., and A.-M. K. Pedersen. 2000. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. Adv. Appl. Prob. 32:499–517.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism, Vol. III (M. N. Munro, ed). Academic Press, New York.

Jurka, J. 2000. RepBase Update: A database and an electronic journal of repetitive elements. Trends Genet. 9:418–420.

Kamal, M., X. Xie, and E. S. Lander. 2006. A large family of ancient repeat elements in the human genome is under strong selection. Proc. Natl. Acad. Sci. USA 103:2740–2745.

Karol, K. G., R. M. McCourt, M. T. Cimino, and C. F. Delwiche. 2001. The closest living relatives of land plants. Science 294:2351–2353.

Karolchik, D., R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, and W. J. Kent. 2003. The UCSC genome browser database. Nucleic Acids Res. 31:51–54.

Kass, R. E., and A. E. Raftery. 1995. Bayes factors. J. Am. Stat. Assoc. 90:773–795.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. 2002. The human genome browser at UCSC. Genome Res. 12:996–1006.

Kimura, M. 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.

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

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

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

Li, S., D. K. Pearl, and H. Doss. 2000. Phylogenetic tree construction using Markov chain Monte Carlo. J. Am. Stat. Assoc. 95:493–508.

Lunter, G., and J. Hein. 2004. A nucleotide substitution model with nearest-neighbour interactions. Bioinformatics 20(Suppl. 1):i216–i223.

Margulies, E. H., M. Blanchette, NISC Comparative Sequencing Program, D. Haussler, and E. D. Green. 2003. Identification and characterization of multi-species conserved sequences. Genome Res. 13:2507–2518.

Margulies, E. H., C. W. Chen, and E. D. Green. 2006. Differences between pair-wise and multi-sequence alignment methods affect vertebrate genome comparisons. Trends Genet. 22:187–193.

Mau, B., M. A. Newton, and B. Larget. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics 55:1–12.

Mendelman, L. V., M. S. Boosalis, J. Petruska, and M. F. Goodman. 1989. Nearest neighbour influences on DNA polymerase insertion fidelity. J. Biol. Chem. 264:14415–14423.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equations of state calculations by fast computing machines. J. Chem. Phys. 21:1087–1092.

Morton, B. R. 1995. Neighbouring base composition and transversion/transition bias in a comparison of rice and maize chloroplast noncoding regions. Proc. Natl. Acad. Sci. USA 92:9717–9721.

Morton, B. R. 1997. The influence of neighbouring base composition on substitutions in plant chloroplast coding sequences. Mol. Biol. Evol. 14:189–194.

Morton, B. R., and M. T. Clegg. 1995. Neighbouring base composition is strongly correlated with base substitution bias in a region of the chloroplast genome. J. Mol. Evol. 41:597–603.

Morton, B. R., V. M. Oberholzer, and M. T. Clegg. 1997. The influence of specific neighbouring bases on substitution bias in noncoding regions of the plant chloroplast genome. J. Mol. Evol. 45:227–231.

Muse, S. V., and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. 11:715–724.

Nylander, J. A. A., J. C. Wilgenbusch, D. L. Warren, and D. L. Swofford. 2008. AWTY (are we there yet?): A system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 24:581–583.

Parisi, G., and J. Echave. 2001. Structural constraints and emergence of sequence patterns in protein evolution. Mol. Biol. Evol. 18:750–756.

Ramsahoye, B. H., D. Bininskiewicz, F. Lyko, V. Clark, A. P. Bird, and R. Jaenisch. 2000. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proc. Natl. Acad. Sci. USA 97:5237–5242.

Rannala, B., and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J. Mol. Evol. 43:304–311.

Robinson, D. M., D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol. Biol. Evol. 20:1692–1704.

Rodrigue, N., N. Lartillot, D. Bryant, and H. Philippe. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene 347:207–217.

Rodrigue, N., H. Philippe, and N. Lartillot. 2006. Assessing site-interdependent phylogenetic models of sequence evolution. Mol. Biol. Evol. 23:1762–1775.

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

Rooney, A. P. 2004. Mechanisms underlying the evolution and maintenance of functionally heterogeneous 18S rRNA genes in Apicomplexans. Mol. Biol. Evol. 21:1704–1711.

Rosenkranz, S. L., and A. E. Raftery. 1994. Covariate selection in hierarchical models of hospital admission counts: A Bayes factor approach. Technical Report no. 268.

Schöniger, M., and A. von Haeseler. 1994. A stochastic model for the evolution of autocorrelated DNA sequences. Mol. Phylogenet. Evol. 3:240–247.

Siepel, A., and D. Haussler. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol. Biol. Evol. 21:468–488.

Smit, A. F. A., R. Hubley, and P. Green. 1996–2004. RepeatMasker Open-3.0. http://www.repeatmasker.org

Steel, M. 2005. Should phylogenetic models be trying to "fit an elephant"? Trends Genet. 21:307–309.

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

Tamura, K., and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10:512–526.

Tanner, M. A., and W. H. Wong. 1987. The calculation of posterior distributions by data augmentation. J. Am. Stat. Assoc. 82:528–540.

Thomas, J. W., J. W. Touchman, R. W. Blakesley, G. G. Bouffard, S. M. Beckstrom-Sternberg, E. H. Margulies, M. Blanchette, A. C. Siepel, P. J. Thomas, J. C. McDowell, B. Maskeri, N. F. Hansen, M. S. Schwartz, R. J. Weber, W. J. Kent, D. Karolchik, T. C. Bruen, R. Bevan, D. J. Cutler, S. Schwartz, L. Elnitski, J. R. Idol, A. B. Prasad, S.-Q. Lee-Lin, V. V. B. Maduro, T. J. Summers, M. E. Portnoy, N. L. Dietrich, N. Akhter, K. Ayele, B. Benjamin, K. Carlaga, C. P. Brinkley, S. Y. Brooks, S. Granite, X. Guan, J. Gupta, P. Haghighi, S.-L. Ho, M. C. Huang, E. Karlins, P. L. Laric, R. Legaspi, M. J. Lim, Q. L. Maduro, C. A. Masiello, S. D. Mastrian, J. C. McCloskey, R. Pearson, S. Stantripop, E. E. Tiongson, J. T. Tran, C. Tsurgeon, J. L. Vogt, M. A. Walker, K. D. Wetherby, L. S. Wiggins, A. C. Young, L.-H. Zhang, K. Osoegawa, B. Zhu, B. Zhao, C. L. Shu, P. J. De Jong, C. E. Lawrence, A. F. Smit, A. Chakravarti, D. Haussler, P. Green, W. Miller, and E. D. Green. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424:788–793.

Yang, A. S., M. R. H. Estécio, K. Doshi, Y. Kondo, E. H. Tajara, and J.-P. J. Issa. 2004. A simple method for estimating global DNA methylation using bisulfite PCR of repetitive DNA elements. Nucleic Acids Res. 32:e38.

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

Yang, Z. 1996b. Phylogenetic analysis using parsimony and likelihood methods. J. Mol. Evol. 42:294–307.

Yang, Z., and B. Rannala. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. Syst. Biol. 54:455–470.

Yu, J., and J. L. Thorne. 2006. Dependence among sites in RNA evolution. Mol. Biol. Evol. 23:1525–1537.

Zhang, Z., and M. Gerstein. 2003. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. Nucleic Acids Res. 31:5338–5348.

Zwickl, D. J., and M. T. Holder. 2004. Model parameterization, prior distributions, and the general time-reversible model in Bayesian phylogenetics. Syst. Biol. 53:877–888.

First submitted 4 January 2008; reviews returned 31 March 2008; final acceptance 17 June 2008

Associate Editor: Olivier Gascuel