Accurate Estimation of Substitution Rates with Neighbor-Dependent Models in a Phylogenetic Context

JEAN BÉRARD¹ and LAURENT GUÉGUEN²,*

¹Institut Camille Jordan, UMR CNRS 5208, Université Lyon 1, Villeurbanne F-69622 Cedex, Université de Lyon, Lyon 69003, France; and ²Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS 5558, Université Lyon 1, Villeurbanne F-69622 Cedex, Université de Lyon, Lyon 69003, France;

*Correspondence to be sent to: Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS 5558, Université Claude Bernard - Lyon 1, 16, rue Raphaël Dubois, 69622 Villeurbanne cedex, France; E-mail: Laurent.Gueguen@univ-lyon1.fr.

Received 22 June 2010; reviews returned 7 December 2010; accepted 26 January 2012

Associate Editor: Olivier Gascuel

Abstract.—Most models and algorithms developed to perform statistical inference from DNA data make the assumption that substitution processes affecting distinct nucleotide sites are stochastically independent. This assumption ensures both mathematical and computational tractability but is in disagreement with observed data in many situations—one well-known example being CpG dinucleotide hypermutability in mammalian genomes. In this paper, we consider the class of RN95 + YpR substitution models, which allows neighbor-dependent effects—including CpG hypermutability—to be taken into account, through transitions between pyrimidine–purine dinucleotides. We show that it is possible to adapt inference methods originally developed under the assumption of independence between sites to RN95 + YpR models, using a mathematically rigorous framework provided by specific structural properties of this class of models. We assess how efficient this approach is at inferring the CpG hypermutability rate from aligned DNA sequences. The method is tested on simulated data and compared against several alternatives; the results suggest that it delivers a high degree of accuracy at a low computational cost. We then apply our method to an alignment of 10 DNA sequences from primate species. Model comparisons within the RN95 + YpR class show the importance of taking into account neighbor-dependent effects. An application of the method to the detection of hypomethylated islands is discussed. [CpG hypermutability; CpG islands; maximum likelihood phylogeny; neighbor-dependent substitution.]

Many stochastic models of nucleotide substitution have been developed since the early work of Jukes and Cantor (1969). They rely on increasingly elaborate Markov substitution processes as they incorporate more and more biological characteristics of actual nucleotide substitution processes (see Yang 1996; Lió and Goldman 1998; Felsenstein 2004 for an overview of evolutionary models).

The most common models are based on a 4 × 4 rate matrix describing a neighbor-independent Markov nucleotide substitution process. Such a matrix comprises 12 substitution rates, and biological hypotheses are usually built into the model to somehow constrain the parameters and have it reflect some relevant biological features. These models are mathematically tractable and allow for the development of efficient algorithms for phylogenetic inference, using, for example, maximum likelihood or Bayesian methods. Early algorithms made use of the assumptions that the sequences were at equilibrium and evolved under a reversible substitution model. To accommodate for situations where these assumptions are not met, more general maximum likelihood algorithms were developed, which can handle irreversible and nonstationary situations, and still are very efficient at performing inference, see Felsenstein (2004).

An important property shared by the models just described is neighbor independence: The matrix of substitution rates at a site is unaffected by the base composition at other sites in the sequence. As a consequence, the substitution processes affecting distinct sites in the sequence are stochastically independent, and, for example, the long-term frequency of a dinucleotide xy has to be equal to the product of the long-term frequencies of x and y. Statistical analyses have shown that, for many genomes, this is not true (Burge et al. 1992; Karlin and Burge 1995), even in noncoding regions. In mammalian genomes, the most important neighbor-dependent effect comes from the hypermutability of CpG dinucleotides: Because of cytosine methylation, the substitution rate of a CpG dinucleotide into TpG or CpA is about 10 times higher than the overall substitution rate of a C into a T or a G into an A (Bird 1980; Nachman and Crowell 2000; Zhang et al. 2007). The problem of extending available inference tools to include models that take into account such neighbor-dependent substitution rates is thus biologically relevant.

The main difficulty with neighbor-dependent substitution rates lies in the propagation of dependencies from site to site. Technically speaking, models with neighbor-dependent substitution rates lead to arbitrarily long-range dependencies, even though these dependencies are most often expected to decay exponentially fast along the sequence. The complexity of such processes is usually impossible to handle exactly, and either approximations or simulation-based methods have to be used.

One approach is to approximate the complicated dependence structure induced by the model, by a simpler structure leading to tractable computations. A surrogate for the original likelihood of the model can thus be computed and exploited within a maximum likelihood or a
Bayesian framework. Examples of this approach include using a Markov model representation of the nucleotide sequence (Duret and Galtier 2000; Arndt et al. 2003; Siepel and Haussler 2004) and working at the level of di- or trinucleotides only (Sved and Bird 1990; Arndt and Hwa 2005; Christensen et al. 2005; Christensen 2006; Peifer et al. 2008); see also Lunter and Hein (2004). This approach in general leads to computationally efficient algorithms, but the reliability of the corresponding approximations is usually difficult to assess.

Another approach consists in using data augmentation techniques, applying the Monte Carlo Markov Chain (MCMC) methodology to sample from the conditional distribution of the (unobserved) substitution histories of sites, given the observed data, and then using either maximum likelihood (such as Expectation-Maximization) or Bayesian strategies. See Pedersen and Jensen (2001), Hwang and Green (2004), Baele et al. (2008), Duret and Arndt (2008), Hobolth (2008), and Baele et al. (2010a, 2010b). This approach faithfully reflects the dynamics specified by neighbor-dependent substitution rates but leads in general to computationally intensive algorithms, for which convergence is an issue.

In this paper, we develop an original approach based on the use of the so-called RN95 + YpR neighbor-dependent substitution models. These models are obtained by combining a neighbor-independent RN95 model [named after Rzhetsky and Nei (1995), see also Schadt et al. (1998) for a thorough study of this model] to a set of transition rates between YpR dinucleotides, as illustrated in Fig. 1. The overall substitution dynamics of the RN95 + YpR model consists in the superposition of an RN95 model describing neighbor-independent nucleotide substitution rates and of a YpR model describing additional neighbor-dependent transition rates.

The RN95 part of the model consists of a 4 × 4 rate matrix belonging to the RN95 class, named after Rzhetsky and Nei (1995) (see also Schadt et al. 1998 for an extensive study of this model). This means a matrix $S = (s_{x,y})_{x,y \in A}$, where $A = \{A, T, C, G\}$, that admits the form:

$$S = \begin{pmatrix} A & T & C & G \\ A & - & v_{T} & v_{C} & w_{G} \\ T & v_{A} & - & w_{C} & v_{G} \\ C & v_{A} & w_{T} & - & v_{G} \\ G & w_{A} & v_{T} & v_{C} & - \end{pmatrix}.$$

Such a matrix $S$ is characterized by eight parameters and reflects the assumption that the transversion rate $v_{x}$ depends only on the nucleotide $x$, resulting from the transversion and not on the nucleotide that has just been substituted.

The YpR part of the model is characterized by a list $r = (r_{CpG}, r_{CpA}, r_{CpA}, r_{CpG}, r_{TpG}, r_{TpG}, r_{CpG}, r_{TpG}, TpA, r_{TpA}, TpG, r_{TpA}, Cpa, r_{CpA}, TpA, r_{CpA}, Cpa)$ of eight additional substitution rates describing transitions between YpR dinucleotides, as illustrated in Fig. 1. The overall substitution dynamics of the RN95 + YpR model is specified through the family $R(x, y|u, v)$, $x, y, u, v \in A, x \neq y$, where $R(x, y|u, v)$ describes the rate of substitution of a nucleotide $x$ by a nucleotide $y$ when $x$ is surrounded by a left neighbor $u$ and a right neighbor $v$. Here, the rate of substitution has the usual meaning that,
during any infinitesimal time interval $dt$, a nucleotide $x$ with $u$ at its left and $v$ at its right has a probability $R(x, y|u, v)dt$ of being turned into a $y$. Given an RN95 rate matrix $S$ and a list of YpR substitution rates $r$, one sets

$$
R(x, y|u, v) := \begin{cases} 
  s_{x,y} + r_{u,X,Y}w_Y & \text{when } uX \text{ and } uY \text{ are YpR dinucleotides,} \\
  s_{x,v} + r_{X,Y}w_Y & \text{when } vX \text{ and } vY \text{ are YpR dinucleotides,} \\
  s_{x,y} & \text{otherwise.} 
\end{cases}
$$

(1)

Note that it is not possible for $ux$ and $xy$ to simultaneously be YpR dinucleotides since $x$ has to be either a purine or a pyrimidine. Hence, exactly one of the three possible cases listed in (1) occurs for any given $x, y, u, v$.

To illustrate the above definition, consider a piece of sequence of the form $\cdots$ ATCG $\cdots$. The overall substitution rate of the $C$ by a $T$ is then equal to $w_T + r_{CpG}$; $T\rightarrow C$ since the $C$ belongs to a $CpG$ dinucleotide. On the other hand, the substitution rate of the $C$ by an $A$ is just given by $v_A$ since the YpR part of the model does not contain additional rates for transversions. For the same reason, the substitution rate of the $G$ by an $A$ is $w_A + r_{CpG}$; $G\rightarrow A$, and the substitution rate of the $G$ by a $T$ is just $v_T$. Now, the overall substitution rate of the $T$ by a $C$ is equal to $w_C$ since the $T$ does not belong to a dinucleotide of the form YpR. As for the substitution rate of the $A$ to a $G$, we would need to know the nucleotide at the left of $A$ to determine its value.

In the sequel, we often consider restricted versions of the RN95 $+$ YpR model in which a model more specific than RN95 is used to describe neighbor-independent substitution rates. When model M is used instead of RN95, the resulting neighbor-dependent model is denoted by $M + YpR$ (e.g., T92 $+$ YpR, TN93 $+$ YpR, etc.). Also, most of the time, the only neighbor-dependent YpR substitutions we consider are $CpG\rightarrow TpG$ and $CpG\rightarrow CpA$, with the same rate for both to ensure strand symmetry. The resulting model is then denoted $M + CpGs$ (e.g., T92 $+$ CpGs, TN93 $+$ CpGs), where the “$s$” is for strand symmetry. Similarly, when we consider general strand-symmetric YpR substitutions, the resulting model is denoted by $M + YpRs$.

Note that RN95 $+$ YpR models are in general non-reversible, even when the RN95 rate matrix describing neighbor-independent substitutions is itself reversible (see online Supplementary Material available from http://datadryad.org, doi:10.5061/dryad.5vp21b10 for details).

\section*{Φ-encoding of polynucleotides}

Let $\rho$ denote the application, which fuses the two purines together, and $\eta$ the application, which fuses the two pyrimidines together, that is,

$$
\rho(A) := R = \rho(G), \quad \rho(C) := C, \quad \rho(T) := T
$$

and

$$
\eta(A) := A, \quad \eta(G) := G, \quad \eta(C) := Y = \eta(T).
$$

Given a word $W = x_1x_2 \cdots x_{m-1}x_m$ of length $m \geq 3$ written in the nucleotidic alphabet $A = \{A, T, C, G\}$, we define the $\Phi$-encoding of $W$ as

$$
\Phi(W) := \rho(x_1)x_2 \cdots x_{m-1}\eta(x_m).
$$

Now consider a nucleotide sequence $X_1(t) \cdots X_n(t)$ evolving according to an RN95 $+$ YpR model. The following two key properties were established in Bérand et al. (2008).

(P1) $\Phi$-encoded nonoverlapping polynucleotides evolve independently;

(P2) The time evolution of a $\Phi$-encoded polynucleotide of length $m$ forms a continuous-time Markov chain, characterized by a certain matrix of transition rates $Q(S, r, m)$, which depends only on, and can be computed explicitly from, $S, r$ and $m$.

For example, (P1) implies that the three $\Phi$-encoded polynucleotides

$$
\Phi(X_1(t)X_2(t)X_3(t)), \quad \Phi(X_4(t)X_5(t)X_6(t)), \quad \text{and} \quad \Phi(X_7(t)X_8(t)X_9(t)X_{10}(t))
$$

evolve independently since these are nonoverlapping polynucleotides. On the other hand, (P1) cannot be applied, for example, to the pair

$$
\Phi(X_2(t)X_3(t)X_4(t)) \quad \text{and} \quad \Phi(X_4(t)X_5(t)X_6(t)),
$$

since these are overlapping nucleotides. As an illustration of (P2), note that the time evolution of $\Phi(X_1(t)X_2(t)X_3(t))$ corresponds to a continuous-time Markov chain with rate matrix $Q(S, r, 3)$. See online Supplementary Material for an explicit definition of $Q(S, r, 3)$. Note that neither (P1) nor (P2) holds for a general neighbor-dependent substitution process. These properties are a specific feature of RN95 $+$ YpR models, allowing for the development of the specific inference approach described in the next section.

\section*{Position of RN95 $+$ YpR models}

Several classical neighbor-independent models fall in the RN95 class. The simplest one is the Jukes and Cantor model (Jukes and Cantor 1969), denoted JC69, where all substitution rates are equal. Then, Kimura’s K80 model (Kimura 1980) has a single parameter, which accounts for the ratio between transition and transversion rates, whereas in Felsenstein’s F81 model (Felsenstein 1981), three parameters reflect the stationary distribution of the nucleotides. Both sets of parameters are used in the models described in Felsenstein (1984) and Hasegawa et al. (1985), denoted HKY85. A special case of the HKY85 model is the T92 model, described in Tamura (1992),
which explicitly accounts for the GC versus AT equilibrium frequencies. The TN93 model, described in Tamura and Nei (1993), is a generalization of the HKY85 model with two distinct transition rates. So far, all the models that have been described are reversible, and, in fact, TN93 is the most general reversible model falling in the RN95 class. Thus, the RN95 class does not cover all the GTR model class (Lanave et al. 1984) and not all the RN95 models are reversible either. Besides reversibility, DNA strand symmetry is an important assumption. As is the case for reversibility, the RN95 class does not cover all the general strand-symmetric model (which we call the L95 model after Lobry 1995), and, conversely, not every RN95 model is strand symmetric. The most general strand-symmetric model falling in the RN95 class comprises three parameters, and we denote it RN95s.

Neighbor dependence is introduced in our model through additional transition rates within YpR dinucleotides. From a modeling perspective, this corresponds to a rather parsimonious choice, obtained by applying a simple perturbation scheme to an existing site-independent model. This kind of approach is also found in Pedersen and Jensen (2001), Arndt et al. (2003), Arndt and Hwa (2005), Duret and Arndt (2008), Peifer et al. (2008), and Hobolth (2008). On the other hand, more complex models have been considered, in which the substitution rate at a site is allowed to fully depend upon its left and right neighbors. This kind of approach is found, for example, in Lunter and Hein (2004), Hwang and Green (2004), Siepel and Haussler (2004), Christensen et al. (2005), Christensen (2006), and Baele et al. (2008, 2010a, 2010b) and leads to parameter-rich models (e.g., the general non-reversible context-dependent model considered in Baele et al. 2010b has 192 substitution parameters). Broadly speaking, more complex models are better able to reflect elaborate features of actual biological substitution processes. However, a more parsimonious model may be desirable when one is interested in accurate parameter estimation on relatively short sequences.

Inference Based of $\Phi$-Encodings for RN95 + YpR Models

The core idea of our approach to inference is the following. Consider a division of the sequence of interest into short nonoverlapping dinucleotides of given length. Then, according to (P1) and (P2), the $\Phi$-encodings of these dinucleotides can be treated rigorously as independently evolving “sites,” using the methods already developed for site-independent substitution models. Specifically, we use a division into nonoverlapping dinucleotides. For convenience, assume that the length $n$ of the sequence is a multiple of 3, that is, $n = 3q$. Then, let

$$Y_k^0(t) := \Phi(X_{3k-2}(t)X_{3k-1}(t)X_{3k}(t)); k = 1, \ldots, q.$$  

Properties (P1) and (P2) allow us to treat $Y_1^0(t), \ldots, Y_q^0(t)$ as a sequence of independently evolving “sites,” with each $Y_k^0(t)$ belonging to the alphabet

$$A_3 = \{C, T, R\} \times \{A, T, C, G\} \times \{A, G, Y\},$$

and evolving according to the substitution process on $A_3$ specified by the rate matrix $Q(S, r, 3)$.

Using a maximum likelihood approach, we can thus infer the parameters of the underlying RN95 + YpR model $(S$ and $r)$, as well as the branch lengths and topology of the phylogenetic tree, in very much the same way as for models assuming independence between sites. On the practical side, it is possible to recycle existing inference algorithms by having them use the alphabet $A_3$ instead of $A$ and the matrix $Q(S, r, 3)$ instead of a usual $4 \times 4$ rate matrix. The alphabet size is then $|A_3| = 3 \times 3 = 3$ instead of just $|A| = 4$, hence an additional but manageable computational cost. On the theoretical side, this approach automatically inherits the properties of maximum likelihood inference based on sequences of independent and identically distributed observations, such as asymptotic consistency and efficiency (see, e.g., van der Vaart 1998).

Averaged log-likelihood.—Due to the use of $\Phi$-encodings in the definition of the sequence $Y_k^0(t)$, the approach described above ignores a substantial part of the information contained in the sequence alignment data: No distinction is made between As and Gs at sites $i = 3k - 2$, nor between Cs and Ts at sites $i = 3k$. To improve upon this approach, we take into account not only one but all three reading frames associated with the division of the sequence into dinucleotides. More precisely, extend the definition of $Y_k^0(t)$ by letting, for $\ell = 1, 2$,

$$Y_k^\ell(t) := \Phi(X_{3k-2+\ell}(t)X_{3k-1+\ell}(t)X_{3k+\ell}(t)).$$

Each $\ell = 0, 1, 2$ gives rise to a log-likelihood of the observed data given the model, $\hat{L}_\ell$, and we define the averaged log-likelihood as

$$\hat{L}(S, r) := \frac{1}{3} \left( \hat{L}_0(S, r) + \hat{L}_1(S, r) + \hat{L}_2(S, r) \right).$$

The approach we adopt throughout the paper is to perform maximum likelihood inference with respect to the averaged log-likelihood $\hat{L}(S, r)$.

One practical advantage of this approach is that it can still be carried out by recycling algorithms originally devised for models that assume independence between sites, even though overlapping $\Phi$-encoded dinucleotides generally do not evolve independently under an RN95 + YpR model. More specifically, maximum likelihood inference with respect to $\hat{L}(S, r)$ can be achieved by applying a maximum likelihood algorithm that treats the whole family of $\Phi$-encoded dinucleotides

$$(Y_k^\ell(t), \ell = 0, 1, 2; 1 \leq k \leq q)$$

as independently evolving sites governed by the rate matrix $Q(S, r, 3)$ (see online Supplementary Material).
From a theoretical point of view, maximum-likelihood inference with respect to $\hat{L}(S, r)$ shares some similarities with pseudolikelihood inference (see, e.g., Christensen et al. 2005; Christensen 2006 for an illustration of this approach in the context of neighbor-dependent substitution models) since $\hat{L}(S, r)$ is obtained by summing the individual log-likelihood values computed for each $\Phi$-encoded trinucleotide $Y^\ell_k(t)$ (up to the factor $1/3$, which is irrelevant as far as maximization is concerned). However, one major difference is that, for a given phase $\ell$, $\hat{L}^\ell(S, r)$ is the exact value of the log-likelihood associated with the family of $\Phi$-encoded trinucleotides $(Y^\ell_k(t), 1 \leq k \leq q)$, so that it is only the dependence between distinct phases, not within them, which makes $\hat{L}(S, r)$ distinct from the exact value of the log-likelihood of the whole alignment.

A theoretical justification.—Let us now sketch a theoretical justification for the use of $\hat{L}(S, r)$, based on asymptotic theory. Denote by $n$ the number of nonoverlapping trinucleotides in the alignment used for the inference. Under generic assumptions, standard asymptotic theory for independent identically distributed sequences (see, e.g., van der Vaart 1998) yields that, for each $\ell = 0, 1, 2$, as $n$ goes to infinity, $n^{-1/2}(\hat{L}^\ell(S, r) - E_{\langle S_0, r_0 \rangle}(L(S, r)))_{\ell=0,1,2}$ is that of a tridimensional centered normal vector $Z = (Z_0, Z_1, Z_2)$ whose covariance matrix is of the form

$$\text{cov}(Z) = \begin{pmatrix} \sigma & c & c \\ c & \sigma & c \\ c & c & \sigma \end{pmatrix}.$$ 

As a consequence, the definition of the averaged log-likelihood $\hat{L}$ through (2) provides (in the sense of asymptotically unbiased with minimum variance) the optimal way of combining the three log-likelihood values $\hat{L}^\ell(S, r)$, $\ell = 0, 1, 2$ to produce an estimate of $E_{\langle S_0, r_0 \rangle}(L(S, r))$.

Dealing with the root distribution.—Since RN95 + YpR models are nonreversible, the probability distribution of the sequence at the root plays a role in the inference, and the likelihood of the observed alignment of sequences has to be averaged with respect to this distribution. To comply with our approach that treats nonoverlapping $\Phi$-encoded trinucleotides as independently evolving “sites,” this distribution must be such that nonoverlapping $\Phi$-encoded trinucleotides are independent. This property is not true in general; for example, it rules out models for the root sequence distribution based on Markov chains (see, e.g., Hwang and Green 2004; Baele et al. 2010a). A typical case where it holds is the stationary distribution of an RN95 + YpR model (that may be distinct from the one used to describe the subsequent evolution along the tree). In this paper, we consider two different options to deal with the root sequence distribution. The first one, denoted Fstat, consists in using the stationary distribution of the RN95 + YpR model describing the substitution process along the tree. The second one, denoted Fest, consists in including the root distribution of $\Phi$-encoded trinucleotides as an additional set of parameters to be estimated in the overall likelihood maximization procedure. Clearly, Fstat is best suited to situations where the root is assumed to be at stationarity, whereas Fest can be used in general nonstationary situations.

Extension to longer polynucleotides.—So far, we have discussed our approach in the case where a division of the sequence into trinucleotides is used. A division into longer polynucleotides can also be considered; the size of the corresponding alphabet is then equal to $N = 3 \times 4^{m-2} \times 3$, where $m$ is the polynucleotide length. The main obstruction to using values of $m$ much larger than 3 is that the computational cost of our method quickly becomes prohibiting. For instance, computing the spectral decomposition of the rate matrix has a cost of order $N^4$, which is about $10^{11}$ with as small a value as $m = 5$ (see, e.g., de Koning et al. 2010 for a general discussion of computational cost associated with phylogenetic likelihood calculations).

DATA
Simulated Data Set
Simulated data provide a framework in which there is a well-defined and known stochastic substitution model governing the evolution of sequences, so that estimates can be directly compared with the true value. Our simulations mimic an alignment of Human, Chimpanzee, and Macaque sequences. The tree used in our simulations is depicted in Fig. 2, and evolution along the branches of this tree is governed by the T92 + CpGs model. This model comprises three parameters: $\kappa$ is the ratio between the transition and transversion rates, $\theta$ is the equilibrium G + C frequency of the T92 model alone, and $\rho = \frac{r_{CpG, CpA}}{s_{C, T}} = \frac{r_{CpG, TpG}}{s_{C, T}}$ is the ratio between the symmetric CpG neighbor-dependent transition rate and the independent transition rate from C to T. The distribution at the root is the stationary distribution of a T92 + CpGs model with $\rho = \rho_2$, whereas the subsequent evolution along the tree uses a T92 + CpGs model with $\rho = \rho_1$. 


the values $\kappa = 3$ and $\theta = 0.4$ being fixed throughout. In the sequel, we use the name Sim_Dat to refer to artificial data simulated according to this scheme.

Note that we deliberately chose to run simulations on a shallow tree with few leaves so as to produce a data set in which the sequences carry very little information about the substitution process. We think of the resulting data set as a worst-case scenario for parameter estimation, compared with real situations.

Primate DNA Data Set

We applied our method on a data set consisting of a human genome sequence aligned with genomes from nine other primate species (Chimpanzee, Gorilla, Orangutan, Macaque, Baboon, Marmoset, Tarsier, Gray mouse lemur, and Galago), see Fig. 3. Specifically, we retrieved the sequence of the ENm001 region of the ENCODE Pilot Project (Consortium 2007) (position 115,810,521 to position 117,687,946 on chromosome 7) from the hg19 version of the human genome, together with aligned primate sequences, as available from the Galaxy web tool (http://galaxy.psu.edu/). This alignment is denoted ENm001 in the sequel.

Among the 1,877,425 bases in the human sequence of ENm001, 810,623 belong to repeated elements, according to the RepeatMasker (Smit et al. 1996) annotations on the human sequence from the University of California–Santa Cruz Table Browser (http://genome.ucsc.edu/cgi-bin/hgTables). The corresponding portion of ENm001 was extracted, and, to obtain a data set similar to the “Ancestral Repeats” data set (Baele et al. 2008), pieces of the alignment corresponding to simple repeats, low complexity regions, members of the Alu family, and RNA elements that diverged less than 25% and L1 elements that diverged less than 2% from the reference RepBase sequence, were removed. The resulting alignment is 460,402 sites long (428,042 bp in the human sequence) and is denoted ENm001_AR in the sequel.

Finally, Enm001_sw denotes a division of Enm001 into nonoverlapping sliding windows of length 2000 bases.

RESULTS

Assessment on Simulated Data

Accuracy of parameter estimation.—To assess the accuracy of the estimates of the CpG hypermutability rate delivered by our method, we performed tests on the simulated data set Sim_Dat. The value of the hypermutability parameters was fixed to $\rho = \rho_1 = \rho_2 = 10$. Figure 4 shows the results of the inference of the hypermutability rate $\rho$ produced by maximum average likelihood with T92 + CpGs + Fstat, for a sequence length varying from 1000 to 20,000 bases (each estimation was repeated 100 times). As expected from the theoretical considerations presented in the Materials and Methods section, the results do not show any substantial bias. Moreover, the estimation is reasonably accurate for sequences as short as 2000 bases long.

Let us point out that, since the depth of the tree is only 0.029 substitution per site, a sequence of length 2000 has accumulated on average 58 substitutions from the root sequence. Moreover, since the root sequence is at the equilibrium of the T92 + CpGs model with a rather large ($\rho = 10$) CpG hypermutability ratio, only a tiny fraction of these substitutions consists of CpG substitutions. As a result, the data contain few hints of CpG hypermutability, and one expects the estimation of $\rho$ to be difficult.

Estimations on the same data with T92 + CpGs + Fest instead of T92 + CpGs + Fstat yield similar results, although with a larger variance, as expected due to the larger number of parameters to be estimated (see online Fig. 1).

Comparison against alternative methods.—We performed comparisons between the estimates of the CpG hypermutability rate delivered by several methods, using the simulated data set Sim_Dat, with sequences of length 10,000. The value of the hypermutability parameters was fixed to $\rho = \rho_1 = \rho_2 = 10$. Each estimation was repeated 100 times. The results are shown on Fig. 5. The
name DA08 refers to the method developed in Duret and Arndt (2008), which is based on MCMC data augmentation. The name BVV10 refers to the method developed in Baele et al. (2008) and improved in Baele et al. (2010a), also based on MCMC data augmentation. Finally, H08 refers to the method developed in Hobolth (2008), where the special form of the model allows for a direct maximum likelihood estimation with respect to the stationary distribution. Note that a value of \( \rho \) cannot always be straightforwardly calculated from the estimated model parameters produced by the various methods. The details are given in the online Appendix.

The computer program corresponding to the DA08 method uses the general single-site strand-symmetric model (L95) plus strand-symmetric CpG effect. Note that the T92 + CpGs model is included in the L95 + CpGs class, so that our simulated data set Sim.Dat is indeed a legitimate test case for the DA08 method. Since the L95 class is not included in RN95, it is not possible to perform a comparison between DA08 and our maximum average likelihood method under exactly the same set of assumptions on the underlying substitution model. To ensure fairness, we chose to work with a subclass of RN95 containing T92 and admitting the same number of parameters (six) as L95, namely the TN93 class. Moreover, since the root distribution is not prescribed in DA08, we performed estimations with the TN93 + CpGs + Fest method instead of TN93 + CpGs + Fstat. As regards the BVV10 method, the substitution model consists in 16 distinct GTR models describing the rate matrix at a site given the 16 possible combinations of nucleotides at its left and right neighbors. Due to its lack of reversibility (even when the flanking bases are fixed, see online Supplementary Material), the T92 + CpGs model used to produce the data set does not fit into this framework. The model underlying the H08 postulates a special form for CpG hypermutability in which T92 + CpGs does not meet either.

With these precisions kept in mind, one sees that our approaches provide the most accurate estimates. The DA08 method does not show substantial bias but achieves a much larger variance than TN93 + CpGs + Fest. Finally, the BVV10 and H08 methods deliver estimates with low variance but sizeable bias, reflecting the inadequacy between the model used to produce the simulations and the one assumed in the inference method.

**Estimation of a recently modified CpG hypermutability rate.**—We assessed the accuracy of the estimation of the CpG hypermutability rate on the simulated data set Sim.Dat with two distinct rates \( \rho_1 \) and \( \rho_2 \) and sequences 5000 bases long. One may think of this data set as representing a piece of DNA having evolved for an infinitely long time prior to the root in a methylated background—associated with a large value for \( \rho_2 \), and then, starting from the root, evolving in a hypomethylated environment—associated with a low value for \( \rho_1 \) or the other way around. The results obtained by maximum averaged likelihood estimation with T92 + CpGs + Fest are shown on Fig. 6 (each estimation was repeated 100 times). One can see that the value of \( \rho_1 \) is accurately estimated over a wide range of values of \( \rho_2 \).

One of the ideas developed in this paper is that an estimation of the CpG hypermutability rate based on aligned sequence data may lead to a more reliable detection of a hypomethylated environment than the traditional CpGo/e-based criteria (see below). To illustrate this point, we computed the value of the CpGo/e ratio in the artificial “human” sequences simulated in Sim.Dat. The results are shown on Fig. 6 (each estimation was repeated 100 times). One can see that the value of CpGo/e is quite sensitive to the value of \( \rho_2 \), as opposed to our estimation of \( \rho_1 \).

### Results on the Primate DNA Data Set

**Comparison of models within the RN95 + YpR class.**—Using the averaged log-likelihood computed on \( \Phi \)-encodings, we performed a series of goodness-of-fit comparisons on the ENm001 and ENm001_AR alignment data sets, for a variety of neighbor-dependent substitution models. Since the validity of the approach based on \( \Phi \)-encodings is limited to the RN95 + YpR class, the only comparisons allowed are between models that belong to this class. Still, within this limited framework, we can assess the impact of including/neglecting neighbor dependence in the substitution model and of a variety of assumptions about the neighbor-independent part of the model.

We considered seven neighbor-independent models (JC69, K80, T92, HKY85, TN93, RN95s, and RN95), in combination with no (denoted +0), strand-symmetric CpG only (denoted +CpGs), and general strand-symmetric YpR (denoted +YpRs), neighbor-dependent substitution rates. For each of the resulting models, maximum likelihood estimation was performed. Even for neighbor-independent (i.e., +0) models, the likelihood was computed using \( \Phi \)-encoded trinucleotides, so as to make comparisons possible. Table 1 shows the resulting log-likelihood values, together with the estimation of the CpG hypermutability rate.

Several conclusions are consistently supported by both data sets. First, it appears that the two most important factors affecting goodness-of-fit are the inclusion of CpG neighbor dependency and the consideration of distinct transition and transversion rates.
Indeed, a model as simple as $K80 + CpGs$ leads to a better fit than all other models with no neighbor dependency. On the other hand, $JC69 + CpGs$ is inferior to the more complicated +0 models, even though it outperforms $K80 + 0$. Moreover, we note that the improvement of the fit obtained by using +YpRs instead of +CpGs is much smaller than the one resulting from using +CpGs instead of +0.

We also see that, for the neighbor-independent part of the model, the assumption of reversibility gets less support than strand symmetry, even when nonreversible YpR neighbor-dependent substitutions are included. Indeed, $TN93$ is the most general reversible model in the RN95 class (with five parameters) but always does worse than RN95s (with only three parameters). The quality of the fit of $TN93$ is greatly improved when one includes CpG or more generally YpR neighbor dependency, but even then the model gets less support than RN95s.

Finally, we observe that the RN95s and RN95 models yield quite comparable results in terms of goodness-of-fit and estimation of the hypermutability rate, even though the RN95s model has only three free parameters. Note that these results are consistent with those of Siepel and Haussler (2004) and Baele et al. (2010b) and confirm the relevance of the strand symmetry assumption.

As regards the estimation of the hypermutability rate $\rho$, our method produces values that are of the same order of magnitude as those obtained in previous studies, except for the JC69 model. In Arndt et al. (2003), on human intergenic DNA, Arndt and collaborators estimated the neighbor-dependent CpG→TpG substitution rate to be 9.51 times the neighbor-independent nucleotide C→T substitution rate. In Arndt and Hwa (2005), on human AluSx SINEs, this ratio was estimated to be 10.45. In Lunter and Hein (2004), the estimated

---

### Table 1. Gain of maximum log-likelihoods for ENm001 and ENm001_AR data sets, compared with JC69 + 0 model, and estimation of the CpG hypermutability rate, with various reconstruction models (np: number of free parameters)

<table>
<thead>
<tr>
<th>Model</th>
<th>np</th>
<th>ENm001_AR Log-likelihood</th>
<th>$\rho$</th>
<th>ENm001 Log-likelihood</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>JC69 + CpGs</td>
<td>1</td>
<td>31,384.1</td>
<td>30.242</td>
<td>132,534.1</td>
<td>25.078</td>
</tr>
<tr>
<td>JC69 + YpRs</td>
<td>4</td>
<td>40,812.0</td>
<td>48.381</td>
<td>173,431.3</td>
<td>38.855</td>
</tr>
<tr>
<td>K80 + 0</td>
<td>1</td>
<td>28,161.6</td>
<td>118,040.5</td>
<td>12.719</td>
<td></td>
</tr>
<tr>
<td>K80 + CpGs</td>
<td>2</td>
<td>54,888.2</td>
<td>15.774</td>
<td>226,295.3</td>
<td>12.719</td>
</tr>
<tr>
<td>K80 + YpRs</td>
<td>5</td>
<td>54,848.2</td>
<td>17.548</td>
<td>228,052.1</td>
<td>14.521</td>
</tr>
<tr>
<td>T92 + 0</td>
<td>2</td>
<td>37,678.4</td>
<td>12.915</td>
<td>164,839.6</td>
<td>9.16</td>
</tr>
<tr>
<td>T92 + CpGs</td>
<td>3</td>
<td>55,882.1</td>
<td>12.915</td>
<td>234,796.9</td>
<td>9.16</td>
</tr>
<tr>
<td>T92 + YpRs</td>
<td>6</td>
<td>56,076.7</td>
<td>13.286</td>
<td>236,027.8</td>
<td>10.283</td>
</tr>
<tr>
<td>HKY85 + 0</td>
<td>4</td>
<td>37,701.8</td>
<td>164,874.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HKY85 + CpGs</td>
<td>5</td>
<td>55,905.0</td>
<td>12.915</td>
<td>234,829.6</td>
<td>9.16</td>
</tr>
<tr>
<td>HKY85 + YpRs</td>
<td>8</td>
<td>56,100.2</td>
<td>13.288</td>
<td>236,062.2</td>
<td>10.284</td>
</tr>
<tr>
<td>TN93 + 0</td>
<td>5</td>
<td>37,707.8</td>
<td>164,901.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TN93 + CpGs</td>
<td>6</td>
<td>55,913.0</td>
<td>12.916</td>
<td>234,861.6</td>
<td>9.17</td>
</tr>
<tr>
<td>TN93 + YpRs</td>
<td>9</td>
<td>56,106.0</td>
<td>13.287</td>
<td>236,088.8</td>
<td>10.284</td>
</tr>
<tr>
<td>RN95s + 0</td>
<td>3</td>
<td>38,661.3</td>
<td>169,535.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RN95s + CpGs</td>
<td>4</td>
<td>56,416.9</td>
<td>12.094</td>
<td>237,311.3</td>
<td>9.242</td>
</tr>
<tr>
<td>RN95s + YpRs</td>
<td>7</td>
<td>56,754.1</td>
<td>11.978</td>
<td>239,298.5</td>
<td>9.158</td>
</tr>
<tr>
<td>RN95 + 0</td>
<td>7</td>
<td>38,690.2</td>
<td>169,596.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RN95 + CpGs</td>
<td>8</td>
<td>56,447.4</td>
<td>12.092</td>
<td>237,375.7</td>
<td>9.242</td>
</tr>
<tr>
<td>RN95 + YpRs</td>
<td>11</td>
<td>56,782.8</td>
<td>11.979</td>
<td>239,358.8</td>
<td>9.158</td>
</tr>
</tbody>
</table>
value is 14.1 for human chromosome 21 and 7.6 for human chromosome 10.

We note that the estimated values of $\rho$ obtained for the ENm001 and ENm001_AR data sets are comparable, although consistently larger on repeated elements (data set ENm001_AR), as expected from the results of Meunier et al. (2005) and Meissner et al. (2008). Also, the estimated values tend to be larger when the neighbor-independent part of the model is assumed to be reversible. Finally, RN95s and the more general RN95 produce exactly the same estimates.

In view of the above results, we chose to work with the RN95s model for the subsequent analysis of the ENm001 data set.

Impact on phylogenetic reconstruction.—We assessed the impact of taking into account neighbor dependence on the estimation of phylogenetic tree branch lengths. Specifically, we compared the results obtained with the RN95s + CpGs and RN95s + 0 models on each window of the ENm001_sw data set for the estimation of the ratio $l_i/l_0$, where $l_i$ denotes the length between the leaf corresponding to the human sequence and node $i$ in Fig. 3. As a result, the estimates are on average equal on short distances (less than 0.05 substitution per site), but on longer distances, the estimates obtained with RN95s + 0 are smaller than those obtained with RN95s + CpGs. For instance, between node 0 and node 5, the average distance estimated with model RN95s + CpGs is 0.0963, and the ratio is 1.09 higher on average than with RN95s + 0 (see Fig. 2 in the online Appendix).

Variation of CpG Hypermutability

Estimating CpG hypermutability along the sequence.—We applied our inference method to the primate DNA data set, using maximum averaged likelihood with the RN95s + CpGs model combined with Fest. Since the results obtained on the simulated data set SIM_DAT suggest that our method is accurate enough to probe the variability of the CpG hypermutability rate $\rho$ at a scale of a few kilobases, we computed estimates of $\rho$ for each of the sliding windows contained in ENm001_sw. The resulting map of CpG hypermutability is shown in Fig. 7. One sees that our estimates of $\rho$ lie most often between 7 and 16, with occasional peaks well below or above these values.

Observed and stationary CpGo/e.—Our method delivers an estimate of hypermutability based on the inferred value of the hypermutability rate on an alignment of

![Graph showing CpG hypermutability](http://sysbio.oxfordjournals.org/)

**Figure 7.** Prediction of hypomethylated islands in ENm001. Top: Estimation of $\rho$ on sliding windows of length 2000. The gray line is the 2.25 threshold to discriminate hypomethylated islands, and the dots beneath locate the hypomethylated islands predicted with this criterion. Middle: The RefSeq genes and the CpG islands predicted by other methods. Bottom: CpGo/e of the human sequence under study, on the same windows.
sequences. A more usual measure of CpG hypermutability is the \(\frac{\text{CpG}}{\text{e}}\) ratio, defined as the ratio between the observed frequency of CpG dinucleotides and the frequency we would expect if nucleotides at distinct sites were stochastically independent,

\[
\frac{\text{CpG}}{\text{e}} = \frac{\text{frequency of CpG}}{\text{frequency of C} \times \text{frequency of G}}.
\]

In the long run, CpG hypermutability leads to a depletion of the number of CpG dinucleotides, relative to the abundance that would be observed in the absence of neighbor-dependent effects. As a consequence, large values of the hypermutability rate are expected to correspond to low values of \(\frac{\text{CpG}}{\text{e}}\), and, conversely, low values of the hypermutability rate are expected to correspond to values of \(\frac{\text{CpG}}{\text{e}}\) close to one. Such a correspondence relies on the long-term effect of CpG hypermutability and may be somewhat imperfectly verified in situations where the substitution process has not been given enough time to substantially alter the sequence composition. For example, one may imagine a transposable element recently inserted in a background whose methylation characteristics differ from those in which it had previously evolved. The simulations shown in Fig. 6 provide an artificial example of this phenomenon.

To illustrate this point on the primate DNA data set, we compare the value of the actually observed \(\frac{\text{CpG}}{\text{e}}\) ratio with the long-term stationary value \(\frac{\text{CpG}}{\text{e}}^*\) predicted from our estimate of the hypermutability rate (the formula for \(\frac{\text{CpG}}{\text{e}}^*\) is given in the online Supplementary Material), for each of the sliding windows of length 2000 of the ENm001_sw alignment. The results are shown on Fig. 8. Each point corresponds to a pair (\(\frac{\text{CpG}}{\text{e}}, \frac{\text{CpG}}{\text{e}}^*\)) for one of the sliding windows of ENm001_sw. While there is broad agreement between the two statistics, substantial differences are observed for windows whose (\(\frac{\text{CpG}}{\text{e}}, \frac{\text{CpG}}{\text{e}}^*\)) lies around (0.4, 0.25). We interpret this discrepancy as a sign that, for these windows, the effect of the CpG hypermutability rate is not yet fully reflected in the base composition.

Note that the distribution of \(\frac{\text{CpG}}{\text{e}}^*\) shows a clear division between a large cluster of values around \(\frac{\text{CpG}}{\text{e}}^* = 0.2\) and a more scattered set of values lying above \(\frac{\text{CpG}}{\text{e}}^* = 0.4\). The same division is observed on the distribution of \(\frac{\text{CpG}}{\text{e}}\), but the distinction between the two groups of values is less well-marked.

**Hypomethylated islands.**—In some stretches of DNA, CpG dinucleotides are less methylated. These are often associated with promoter regions (Ponger et al. 2001) and are typically 300–3000 bases long. We call such stretches hypomethylated islands in the sequel. In the human genome, the average \(\frac{\text{CpG}}{\text{e}}\) is about 0.25 (as we have seen also in Fig. 8), but it is higher in some of these stretches called CpG-islands. Basically (Gardiner-Garden and Frommer 1987), CpG-islands are characterized as segments at least 200 bases long, with a G + C proportion at least 0.5 and \(\frac{\text{CpG}}{\text{e}} \geq 0.6\).

To our knowledge, no method aimed at detecting hypomethylated islands exploits the information contained in an alignment. Instead, all the methods we know rely solely on the composition of the sequence of interest. Some are based on the \(\frac{\text{CpG}}{\text{e}}\) (Gardiner-Garden and Frommer 1987; Ponger and Mouchiroud 2001), others on the CpG density (Hackenberg et al. 2006), or on a Hidden Markov Model for the sequence (Durbin et al. 1998; Wu et al. 2010). The approach we develop consists in using the estimation of the hypermutability rate \(\rho\) inferred from an aligned set of sequences, to locate hypomethylated islands.

As explained earlier, statistics based exclusively on the composition of the sequence may be misleading indicators of the current methylation level. In order to prevent the detection of false positives, methods that look for CpG-islands are most often used on data from which recently inserted elements have been masked (with the exception of Wu et al. 2010). On the other hand, our method exploits phylogenetic information about the sequence of interest and should be less affected by this kind of problem. In fact, the estimates of the hypermutability rate presented in this section were all obtained from nonmasked data.

To single out hypomethylated islands, a precise limit on the value of \(\rho\) has to be fixed. We based our approach on estimations of \(\rho\) on alignments around the Transcription Start Sites of the human genome, leading to a threshold value of 2.25. The details of the method are given in the online Appendix.

We see in Fig. 7 that our predictions of hypomethylated islands match perfectly with the CpG-islands found by other methods, with the exception of the results of Wu et al. (2010). Note that the additional CpG-islands found by Wu et al. (2010) all lie inside recently inserted repeated elements (less than 8.5% divergence from the reference sequence).


**DISCUSSION**

For the RN95 + YpR class of neighbor-dependent nucleotide substitution models, we have described a method that maps the substitution process affecting a sequence onto a process in which sites evolve independently. This allows the usual inference algorithms to be exploited without introducing additional approximations. Likelihood computations using Felsenstein's recursion algorithm (Felsenstein 2004) with our model are roughly 81 times slower than with the usual neighbor-independent DNA substitution models since we have to use an alphabet that is 9 times larger. This makes our method practical for the study of large data sets.

Using a simulated data set mimicking aligned human–chimpanzee–macaque sequences, we have shown that the CpG-hypermutability rate is accurately estimated even on alignments containing little information on the substitution process (few sequences, shallow trees, and sequences as short as 2000 bases long). Extending the method to polynucleotides of length larger than three may further improve the accuracy, although it would substantially increase the computational requirements.

On a primate DNA alignment comprising sequences from 10 species, we performed goodness-of-fit comparisons between models within the RN95 + YpR class. We found that the two most important factors affecting goodness-of-fit are the inclusion of CpG neighbor dependency and the consideration of distinct transition and transversion rates. Also, for the neighbor-independent part of the model, the reversibility assumption is less supported than strand symmetry. Following Yang (1994), an interesting perspective would be to incorporate rate variability in the model, taking care of the fact that the rate of evolution of a $\Phi$-encoded trinucleotide depends on the respective rates of evolution of the three corresponding nucleotides.

We also performed comparisons of estimated branch lengths on the primate DNA alignment. Taking into account, CpG hypermutability in the substitution model leads to no effect on short branches but to a noticeable though not major effect on branches as long as 0.1 substitution per site. It would be interesting to study this phenomenon on trees with longer branches, such as the whole mammal family tree, where more substitutions due to the CpG influence are expected.

On the primate DNA data set, we computed estimates of the CpG hypermutability rate. The overall rate is found to be approximately 9.2. Thanks to the accuracy of the method, we were able to probe the variability of this rate at a scale of a few kilobases. This leads to a new approach to the detection of hypomethylated islands, based on the comparison of the estimated hypermutability rate with a threshold value. Unlike other methods that rely solely on the composition of the sequence under study, our method takes advantage of the available phylogenetic information. The results suggest that this method has the ability to accurately detect hypomethylated islands and can lead to a more thorough analysis of CpG-hypermutability in the whole human genome. However, our method does not yield a precise localization of the boundaries of hypomethylated islands, and it would be interesting to combine it with more elaborate segmentation methods, using, for example, hidden Markov models. Another perspective is to use branch-dependent hypermutability rates to look for the evolution of methylation in hypomethylated islands.

Finally, following Pedersen et al. (1998) and Christensen et al. (2005), it would be interesting to extend our approach to codon evolution models, to study the effect of CpG hypermutability on the estimation of selection. Such an extension should take care of the fact that $\Phi$-encoding of codons is not fully compatible with the genetic code.

**SUPPLEMENTARY MATERIAL**

Supplementary material, including data files and/or online-only appendices, can be found at http://www.sysbio.oxfordjournals.org/ and in the Dryad data repository (DOI:10.5061/dryad.5vp21b10).

**ACKNOWLEDGMENTS**

We thank Laurent Duret and Manolo Gouy for their useful help, as well as the associate editor and three anonymous reviewers for their very constructive suggestions. Our method has been implemented in the Bio++ suite (Dutheil et al. 2006). The maximum likelihood program is called bppml, which sources, and standalone executables are available at the address http://home.gna.org/bppsuite. Usage files are also available on this site. All alignment simulations have been performed with the Python suite alfacinha available at the address http://pbil.univ-lyon1.fr/software/alfacinha/.

**REFERENCES**


