



**HAL**  
open science

## Phylogenetic distances for neighbour dependent substitution processes

Mikael Falconnet

► **To cite this version:**

Mikael Falconnet. Phylogenetic distances for neighbour dependent substitution processes. 2008. hal-00345897v3

**HAL Id: hal-00345897**

**<https://hal.science/hal-00345897v3>**

Preprint submitted on 20 Nov 2009 (v3), last revised 4 Jan 2010 (v4)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# PHYLOGENETIC DISTANCES FOR NEIGHBOUR DEPENDENT SUBSTITUTION PROCESSES

MIKAEL FALCONNET

ABSTRACT. We consider models of nucleotidic substitution processes where the rate of substitution at a given site depends on the state of its neighbours. Under these models, we investigate two situations : the first one is to estimate the time elapsed between an ancestral sequence at frequency stationarity and a present sequence, the second one is to estimate under the hypothesis that two sequences are issued from a common ancestral sequence also at frequency stationarity, the time since they have diverged. For the the simplest nontrivial case, the Jukes-Cantor model with CpG influence, we show with a mathematical proof how to compute consistent estimators for these two situations. We provide asymptotic confidence intervals, valid for nucleotidic sequences of finite length and we compute explicit formulas for the estimators and for their confidence intervals. In the general case of RN model plus YpR influence, we show that the results can be extended under one proviso : the equation defining the estimator does not fail to have a unique solution.

## INTRODUCTION

A crucial step in the computation of phylogenetic trees based on aligned DNA sequences is the estimation of the evolutionary times between these sequences. In most phylogenetic algorithms based on stochastic substitution models, one assumes that each site evolves independently from the others and, in general, according to a given Markovian kernel. This assumption is mainly due to the difficulty to work without the assumption of independence. To understand why, note that the distribution of the nucleotide at site  $i$  at a given time depends a priori on the values at previous times of the dinucleotides at sites  $i - 1$  and  $i + 1$ , whose joint distributions, in turn depend a priori on the values of some trinucleotides, and so on. It follows that we have to solve an infinite-dimensional linear system, generically hard to solve. Besides, there can be a large effect of the neighbours on the substitution rates. Since some neighbour influences are well documented in the literature, and caused by well known biological mechanisms, taking account

---

*Date:* November 20, 2009.

*1991 Mathematics Subject Classification.* Primary: 60J25. Secondary: 62P10; 62F25; 92D15; 92D20.

*Key words and phrases.* Markov processes, Confidence intervals, DNA sequences, Phylogenetic distances, CpG deficiency.

of the neighbour influences in substitution models has become a necessity, that is why a class of mathematical models with neighbour influences was recently introduced by biologists, see [4], and studied mathematically, see [1]. The goal of the present paper is to show that one can compute consistent estimators of the distances between DNA sequences whose evolution is ruled by a particular model in this class, and to detail the results that can be extended to the rest of the class without condition and those who need more assumption.

We completely describe the construction in the simplest non trivial case, the Jukes-Cantor model with (symmetric) CpG influence. We explain how the construction can be extended for every model in this class in the appendix.

In section 1, we describe the class of manageable models introduced in [1] and the main properties of the simplest one, the Jukes-Cantor model with CpG influence. In section 2, we summarize our main results on the estimation of the elapsed time between an old DNA sequence and a present one, and on the time since two present DNA sequences issued from the same ancestral sequence diverged. In appendix B, we give the extended results of section 2. The other sections explain how we prove our results. At the end of section 2, we give the plan of the rest of the paper.

## 1. MODELS WITH INFLUENCE

We first describe the class of models to which the results of this paper apply, and in particular the simplest one, called Jukes-Cantor model with CpG influence. Then, we mention its main mathematical properties, already established in [1], and we introduce some notations.

Recall that DNA sequences are encoded by the alphabet  $\mathcal{A} = \{A, T, C, G\}$ , where the letters stand for Adenine, Thymine, Cytosine and Guanine respectively. Thus, bi-infinite DNA sequences are encoded as elements of  $\mathcal{A}^{\mathbb{Z}}$  where  $\mathbb{Z}$  is the set of integers.

**1.1. Jukes-Cantor model with CpG influence.** In most models of DNA evolution, one assumes that each site evolves independently from the others and follows a given Markovian kernel, see [9], [10], [3] and [6] for instance. Even in codon evolution models, see [8], one often assumes that different codons evolve independently with some exceptions as [7]. However, it is a well known experimental fact, see [2] by example, that the nature of the close neighbours of a site can modify, notably in some cases, the substitution rates observed at this site. To take account of these observations, we consider models, in continuous time, where the sequence evolves under the combined effect of two superimposed mechanisms.

The first mechanism is an independent evolution of the sites as in the usual models. Hence it is characterized by a  $4 \times 4$  matrix of substitution rates,

each rate being the mean number of substitutions per unit of time. The simplest case is the Jukes-Cantor model, where each substitution happens at the same rate. Hence, the rate of the substitutions of  $x$  by  $y$  is set to 1, for every nucleotides  $x$  and  $y$  in  $\mathcal{A}$ .

A second mechanism is superimposed, which describes the substitutions due to the influence of the neighborhood: the most noticeable case is based on experimentally observed CpG-methylation-deamination processes, whose biochemical causes are well known. Hence we assume that the substitution rates of cytosine by thymine and of guanine by adenine in CpG dinucleotides are both increased by an additional nonnegative rate  $r$ .

This means for example that any  $C$  site whose right neighbour is not occupied by a  $G$ , changes at global rate 3, hence after an exponentially random time with mean  $1/3$ , and when it does, it becomes an  $A$ , a  $G$  or a  $T$  with probability  $1/3$  each. On the contrary, any  $C$  site whose right neighbour is occupied by a  $G$ , changes at global rate  $s = 3 + r$ , hence after an exponentially random time with mean  $1/s$ , and when it does, it becomes an  $A$ , a  $G$  or a  $T$  with unequal probabilities  $1/s$ ,  $1/s$ , and  $(1 + r)/s$  respectively.

The case  $r = 0$  corresponds to the usual Jukes-Cantor model. As soon as  $r \neq 0$ , the evolution of a site is not independent of the rest of the sequence. Hence the evolution of the complete sequence is Markovian (on a huge state space), but not the evolution of a given site, nor of any given finite set of sites.

Recall from [1] that the relevant class of models, called RN+YpR in this paper, is in fact larger than just described.

As already mentioned, the results of this paper about Jukes-Cantor models with CpG influence are adapted to every RN model with YpR influence in the appendix.

**1.2. Main properties.** We work on the space  $\mathcal{A}^{\mathbb{Z}}$  with the topology product and the cylindric  $\sigma$ -algebra defined as the coarsest  $\sigma$ -algebra (i.e. the one with the fewest measurable sets) such that every projection on  $\mathcal{A}^{\mathbb{Z}}$  is a measurable function.

We now recall some results of [1], valid for every RN model with YpR influence. First, for every probability measure  $\nu$  on  $\mathcal{A}^{\mathbb{Z}}$ , there exists a unique Markov process  $(X(t))_{t \geq 0}$  on  $\mathcal{A}^{\mathbb{Z}}$ , with initial distribution  $\nu$ , associated to the transition rates above. Thus, for every time  $t$ ,  $X(t)$  describes the whole sequence and, for every  $i$  in  $\mathbb{Z}$ , the  $i$ th coordinate  $X_i(t)$  of  $X(t)$  is the random value of the nucleotide at site  $i$  and time  $t$ . Under a non-degeneracy condition on the rates of a RN model with YpR influence, the process  $(X(t))_{t \geq 0}$  is ergodic, its unique stationary distribution  $\pi$  on  $\mathcal{A}^{\mathbb{Z}}$  is invariant and ergodic with respect to the translations of  $\mathbb{Z}$ , and  $\pi$  puts a positive mass on every finite word  $w = (w_i)_{0 \leq i \leq \ell}$  written in the alphabet  $\mathcal{A}$ . The notation  $\pi(w)$  is abusive because  $\pi$  is a measure on  $\mathcal{A}^{\mathbb{Z}}$  but it is

a shorthand for  $\pi(\Pi_{0,\ell}^{-1}(\{w\}))$ , where  $\Pi_{0,\ell}$  is such that for every  $x \in \mathcal{A}^{\mathbb{Z}}$ ,  $\Pi_{0,\ell}(x) = (x_i)_{0 \leq i \leq \ell}$ .

Furthermore, for every position  $i$  in  $\mathbb{Z}$ ,  $\mathbb{P}_\nu(X_{i:i+\ell}(t) = w)$  converges to  $\pi(w)$  when  $t \rightarrow +\infty$ , where  $\mathbb{P}_\nu$  stands for the probability under the initial measure  $\nu$ . Here and later on, for every indices  $i$  and  $j$  in  $\mathbb{Z}$  with  $i \leq j$  and every symbol  $S$ , the shorthand  $S_{i:j}$  denotes  $(S_k)_{i \leq k \leq j}$ . Finally, if  $\xi$  in  $\mathcal{A}^{\mathbb{Z}}$  is distributed according to  $\pi$ , the empirical frequencies of any word  $w$  in  $\xi$ , observed along any increasing sequence of intervals of  $\mathbb{Z}$ , almost surely converge to  $\pi(w)$ .

All of the above properties stem from the following representation of the distribution  $\pi$ . There exists an i.i.d. sequence  $(H_i)_{i \in \mathbb{Z}}$  of Poisson processes, and a measurable map  $\Psi$  with values in  $\mathcal{A}$ , such that if one sets

$$\xi_i = \Psi(H_{i-1}, H_i, H_{i+1})$$

for every site  $i$  in  $\mathbb{Z}$ , then the distribution of  $(\xi_i)_{i \in \mathbb{Z}}$  is  $\pi$ . In particular, any collections  $(\xi_i)_{i \in I}$  and  $(\xi_i)_{i \in J}$  are independent as soon as the subsets  $I$  and  $J$  of  $\mathbb{Z}$  are such that  $|i - j| \geq 3$  for every sites  $i$  in  $I$  and  $j$  in  $J$ . We call this property independence at distance 3.

**1.3. Notations.** Our estimators are based on various quantities provided by the alignment of the two sequences.

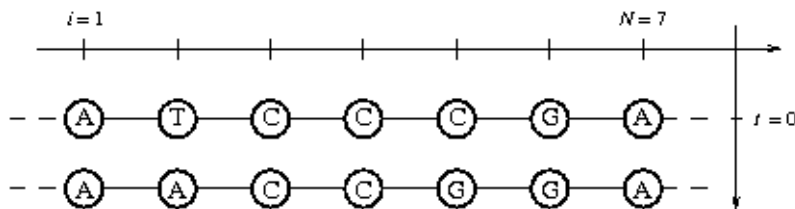


FIGURE 1. Alignment of an ancestral sequence and a present one

For every  $\ell \geq 0$  and every word  $w$  of length  $\ell + 1$  written in the alphabet  $\mathcal{A}$ , say that site  $i$  is occupied at time  $t$  by  $w$  if  $X_{i:i+\ell}(t) = w$ . For every triple of subsets  $W$ ,  $W'$  and  $W''$  of words and every couple of times  $t$  and  $s$ ,  $(W)(t)$  denotes the frequency of sites occupied by any of the words in  $W$  at time  $t$ , that is

$$W(t) = \lim_N \frac{1}{N} \sum_{i=0}^N \sum_{w \in W} H_i(t, w), \quad \text{where} \quad H_i(t, w) = \mathbf{1}\{X_{i:i+\ell}(t) = w\},$$

$(W, W')(t)$  the frequency of sites occupied by any of the words in  $W$  at time 0 and any of the words in  $W'$  at time  $t$ , that is

$$(W, W')(t) = \lim_N \frac{1}{N} \sum_{i=0}^N \sum_{w \in W, w' \in W'} H_i(0, w) H_i(t, w').$$

The limits above exist thanks to the ergodicity of  $\pi$  with respect to translations.

When comparing two present sequences, we use the following notations. For every sets  $W$  and  $W'$  of words and every time  $t$ ,  $[W, W'](t)$  denotes the frequency of sites occupied by a word of  $W$  in the left sequence (denoted by  $X^1$ ) and by a word of  $W'$  in the right sequence (denoted by  $X^2$ ).

We identify a word  $w$  and the set of words  $\{w\}$ . For every letter  $x$  in the alphabet  $\mathcal{A}$ , we use the shorthands  $*x = \mathcal{A} \times \{x\}$ ,  $x* = \{x\} \times \mathcal{A}$ ,  $x*x = \mathcal{A} \times \{x\} \times \mathcal{A}$  and  $\bar{x} = \mathcal{A} \setminus \{x\}$ .

## 2. SUMMARY OF MAIN RESULTS

Theorem 2.3 below provides asymptotic confidence intervals for estimation procedures of the time elapsed between a present sequence and an ancestral one and for the time since two present sequences issued from the same ancestral sequence diverged, for the Jukes-Cantor model with CpG influence of intensity  $r$ . These intervals are based on two consistent estimators of the elapsed time  $t$  and two consistent estimators of the time of divergence. Our first estimator is based on the evolution of the frequency  $(C, C)(t)$  when the time  $t$  varies and the second on the evolution of  $(A, A)(t)$ . These estimators match the classic ones used for the original Jukes-Cantor model when  $r$  is equal to zero. The symmetry of the roles played by  $A$  and  $G$ , or by  $C$  and  $T$  in the Jukes-Cantor model with CpG influence gives immediately  $(A, A)(t) = (T, T)(t)$  and  $(G, G)(t) = (C, C)(t)$ . The third estimator is based on the evolution of the frequency  $[C, C](t)$  when the time  $t$  varies and the last one on the evolution of  $[A, A](t)$ . Even if the results are given in the same theorem, there is a main difference between  $[C, C]$  and  $[A, A]$ . Indeed, as we explained in sections 5 and 6, we have

**Theorem 2.1.** *Under the Jukes-Cantor model with CpG effect, for every positive  $t$ ,*

$$[C, C](t) = (C, C)(2t), \quad [A, A](t) \neq (A, A)(2t).$$

In appendix B, theorem B.1 provides an asymptotic confidence interval for the time elapsed between a present sequence and an ancestral one, for the RN model with YpR influences, under the condition that the estimator is well-defined in the general case.

Theorem 2.3 is the keystone for the creation of phylogenetic trees built by a distance-based method. At the moment, a prior knowledge of the parameter  $r$  is needed to apply the method. We prospect in a future work to perform simulations and find a method to estimate parameter  $r$ .

Let  $(x, x)_{\text{obs}}$  and  $[x, x]_{\text{obs}}$  denote for every  $x \in \{A, C\}$  the observed value of  $(x, x)$  and  $[x, x]$  on two aligned sequences of length  $N$ , that is,

$$(x, x)_{\text{obs}} = \frac{1}{N} \sum_{i=1}^N K_i^x(t), \quad \text{with} \quad K_i^x(t) = \mathbf{1}\{X_i(0) = X_i(t) = x\},$$

and

$$[x, x]_{\text{obs}} = \frac{1}{N} \sum_{i=1}^N \tilde{K}_i^x(t), \quad \text{with} \quad \tilde{K}_i^x(t) = \mathbf{1}\{X_i^1(t) = X_i^2(t) = x\}.$$

In figure 1 for instance,  $N = 7$  and  $(C, C)_{\text{obs}} = \frac{2}{7}$ .

**Definition 2.2.** Let  $T_x$  and  $\tilde{T}_x$  denote the estimators of the elapsed time  $t$  and the divergence time  $t$ , defined for every  $x \in \{A, C\}$ , as the solution in  $t$  of the equations

$$(x, x)(t) = (x, x)_{\text{obs}} \quad \text{and} \quad [x, x](t) = [x, x]_{\text{obs}}.$$

Let  $\kappa_{\text{obs}}^x$ ,  $\tilde{\kappa}_{\text{obs}}^x$ ,  $\nu_{\text{obs}}^x$  and  $\tilde{\nu}_{\text{obs}}^x$  denote observed quantities, defined for every  $x \in \{A, C\}$  as

$$\begin{aligned} \kappa_{\text{obs}}^C &= 4(C, C)_{\text{obs}} + r(C*, CG)_{\text{obs}} - (C), \\ \kappa_{\text{obs}}^A &= 4(A, A)_{\text{obs}} - r(*A, CG)_{\text{obs}} - (A), \\ \nu_{\text{obs}}^x &= (x, x)_{\text{obs}} - 5(x, x)_{\text{obs}}^2 + 2(xx, xx)_{\text{obs}} + 2(x * x, x * x)_{\text{obs}}, \end{aligned}$$

and up to turning  $(, )$ s into  $[, ]$ s,

$$\tilde{\kappa}_{\text{obs}}^x = 2\kappa_{\text{obs}}^x \quad \text{and} \quad \tilde{\nu}_{\text{obs}}^x = \nu_{\text{obs}}^x.$$

We note that  $\kappa_{\text{obs}}^x$ ,  $\tilde{\kappa}_{\text{obs}}^x$ ,  $\nu_{\text{obs}}^x$  and  $\tilde{\nu}_{\text{obs}}^x$  might be negative for some sequences of observations and lengths  $N$ . However, from lemma 4.1 in section 4,  $\kappa_{\text{obs}}^x$ ,  $\tilde{\kappa}_{\text{obs}}^x$ ,  $\nu_{\text{obs}}^x$  and  $\tilde{\nu}_{\text{obs}}^x$  are almost surely positive when  $N$  is large.

As explained in sections 5 and 6, in the Jukes-Cantor model with CpG influence, for every  $x \in \{A, C\}$ , the functions  $t \mapsto (x, x)(t)$  and  $t \mapsto [x, x](t)$  are decreasing on  $[0, +\infty)$  from  $(x)$  to  $(x)^2$ , where  $(x)$  is the frequency of  $x$  sites at stationarity. Thus,  $T_x$  and  $\tilde{T}_x$  are unique and well defined for any pair of aligned sequences such that

$$(x)^2 < (x, x)_{\text{obs}} < (x).$$

Thanks to the ergodicity of the model, this condition is almost surely satisfied when  $N$  is large enough because  $(x, x)_{\text{obs}} \rightarrow (x, x)(t)$  and  $[x, x]_{\text{obs}} \rightarrow [x, x](t)$  almost surely when  $N \rightarrow \infty$ .

However, even if  $T_x$  and  $\tilde{T}_x$  are unique and well defined, we don't have explicit formulas to compute them. It is necessary to use numerical methods to solve equation  $(x, x)(t) = (x, x)_{\text{obs}}$ , for example. Fortunately, we have explicit formulas for  $(x, x)(t)$  and  $[x, x](t)$  in the Jukes-Cantor model with CpG influence.

We now state our main result.

**Theorem 2.3.** *Assume that the ancestral sequence is at stationarity. Then, in the Jukes-Cantor model with CpG influence, for every  $x \in \{A, C\}$ , when  $N \rightarrow +\infty$ ,  $\kappa_{\text{obs}}^x \sqrt{N/\nu_{\text{obs}}^x} (T_x - t)$ , respectively  $\tilde{\kappa}_{\text{obs}}^C \sqrt{N/\tilde{\nu}_{\text{obs}}^C} (\tilde{T}_C - t)$ , converges in distribution to the standard normal law. An asymptotic confidence interval at level  $\varepsilon$  for the elapsed time  $t$ , respectively for the time of divergence  $t$ , is*

$$\left[ T_x - \frac{z(\varepsilon)}{\kappa_{\text{obs}}^x} \sqrt{\frac{\nu_{\text{obs}}^x}{N}}, T_x + \frac{z(\varepsilon)}{\kappa_{\text{obs}}^x} \sqrt{\frac{\nu_{\text{obs}}^x}{N}} \right],$$

respectively

$$\left[ \tilde{T}_x - \frac{z(\varepsilon)}{\tilde{\kappa}_{\text{obs}}^x} \sqrt{\frac{\tilde{\nu}_{\text{obs}}^x}{N}}, \tilde{T}_x + \frac{z(\varepsilon)}{\tilde{\kappa}_{\text{obs}}^x} \sqrt{\frac{\tilde{\nu}_{\text{obs}}^x}{N}} \right],$$

where  $z(\varepsilon)$  denotes the unique real number such that  $\mathbb{P}(|Z| \geq z(\varepsilon)) = \varepsilon$  with  $Z$  a variable with standard normal law.

Theorem 2.3 implies that, for large  $N$ , the width of the confidence interval scales as  $N^{-1/2}$  times a function of  $t$ , and for large  $t$ , this function scales as  $e^{4t}$  for the time elapsed and  $e^{8t}$  for the time of divergence according to formulas given in corollaries 5.2 and 6.2. Heuristically, we can say that one must observe a part of the sequence of length  $N$  at least of order  $e^{8t}$  for the time elapsed,  $e^{16t}$  for the time of divergence, to estimate  $t$  up to a given factor.

The rest of the paper is organized as follows. In section 3, we state central limit theorems for the time estimators for the Jukes-Cantor model with CpG influence and for the general model under conjecture 3.4. In section 4, we show that the central limit theorems established in section 3 imply theorem 2.3 of section 2. In section 5, and 6, we characterize the evolutions of  $(C, C)(t)$  and  $[C, C](t)$ , and in section 6 the evolutions of  $(A, A)(t)$  and  $[A, A](t)$ . We state some monotonicity properties in these two sections.

In appendix A, we give a short description of the general RN model with YpR influence. In appendix B, we give an extension of theorem 2.3 to the general model under conjecture 3.4, and in appendix C the justification of this extension. In appendix D, we describe some simulations performed to reinforce conjecture 3.4.

### 3. CENTRAL LIMIT THEOREMS FOR THE TIME ESTIMATORS

We give here central limit theorems for the time estimators in the general model. The strategy is the following. We first deal with  $(x, x)_{\text{obs}}$  and  $[x, x]_{\text{obs}}$ . We compute exactly the variance of this quantities thanks to the independence at distance 3. Then, we establish mixing properties of the sequence due to the independence at distance 3. Then, we use a central limit



theorem for mixing sequences. To state central limit theorem for the time estimators, we use the delta method, and in order to do that, we need to know that  $t \mapsto (x, x)(t)$  and  $t \mapsto [x, x](t)$  are diffeomorphisms. This will be a conjecture for the general model whereas we prove it for the Jukes-Cantor model with CpG influence.

**3.1. Variance computations.** We detail the properties of  $(C, C)_{\text{obs}}$ ,  $(A, A)_{\text{obs}}$ ,  $[C, C]_{\text{obs}}$  and  $[A, A]_{\text{obs}}$ . We assume that  $N \geq 2$ .

**Lemma 3.1.** *In the general RN model plus YpR influence, for  $x \in \{C, A\}$ , the mean of  $(x, x)_{\text{obs}}$ , respectively  $[x, x]_{\text{obs}}$ , with respect to  $\pi$  is  $(x, x)(t)$ , respectively  $[x, x](t)$ . The variance of  $(x, x)_{\text{obs}}$ , respectively  $[x, x]_{\text{obs}}$  with respect to  $\pi$  is  $\sigma_x^2(N, t)$ , respectively  $\tilde{\sigma}_C^2(N, t)$ , where*

$$N\sigma_x^2(N, t) = (x, x)(t) - (x, x)(t)^2 + 2(1 - 1/N)((xx, xx)(t) - (x, x)(t)^2) + 2(1 - 2/N)((x * x, x * x)(t) - (x, x)(t)^2),$$

and up to turning  $(, )$ s into  $[, ]$ s

$$\tilde{\sigma}_x^2(N, t) = \sigma_x^2(N, t).$$

*Proof.* The random variables  $(K_i^x(t))_{i \in \mathbb{Z}}$ , respectively  $(\tilde{K}_i^x(t))_{i \in \mathbb{Z}}$ , are Bernoulli random variables identically distributed with respect to  $\pi$ , their common mean is  $(x, x)(t)$ , respectively  $[x, x](t)$ , and  $(x, x)_{\text{obs}}$ , respectively  $[x, x]_{\text{obs}}$ , is the empirical mean of the  $N$  values  $K_i^x(t)$ , respectively  $\tilde{K}_i^x(t)$ , for  $i$  from 1 to  $N$ . Thus, we obtain the value of  $\mathbb{E}((x, x)_{\text{obs}})$ , respectively  $\mathbb{E}([x, x]_{\text{obs}})$ , as  $(x, x)(t)$ , respectively  $[x, x](t)$ . Furthermore,

$$N^2\sigma_x^2(N, t) = \sum_{i=1}^N \text{var}(K_i^x(t)) + 2 \sum_{1 \leq i < j \leq N} \text{cov}(K_i^x(t), K_j^x(t)).$$

The variance of each  $K_i^x(t)$  is  $\text{var}(K_1^x(t)) = (x, x)(t) - (x, x)(t)^2$ . The independence at distance 3, valid for any RN+YpR model, implies that each covariance for  $|i - j| \geq 3$  is zero. The invariance by translation of  $\pi$ , valid for any RN+YpR model, shows that each of the  $(N - 1)$  covariances such that  $i = j - 1$  is

$$\text{cov}(K_1^x(t), K_2^x(t)) = (xx, xx)(t) - (x, x)(t)^2.$$

Finally, each of the  $(N - 2)$  covariances such that  $i = j - 2$  is

$$\text{cov}(K_1^x(t), K_3^x(t)) = (x * x, x * x)(t) - (x, x)(t)^2.$$

The same arguments hold for  $\tilde{\sigma}_C^2(N, t)$ . This concludes the proof.  $\square$

**3.2. Central limit theorems for  $(x, x)_{\text{obs}}$  and  $[x, x]_{\text{obs}}$ .** To prove the convergence in distribution to the normal law, we use the following result.

**Theorem 3.2** (Hall and Heyde [5]). *Let  $(V_i)_{i \in \mathbb{Z}}$  denote a stationary, ergodic, centered, square integrable sequence. Let  $\mathcal{F}_0 = \sigma(V_i; i \leq 0)$  denote the  $\sigma$ -algebra generated by the random variables  $V_i$  for  $i \leq 0$ . For every positive integer  $n$ , introduce*

$$U_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n V_i.$$

Assume that

- (i) for every positive  $n$ , the series  $\sum_{k \geq 1} \mathbb{E}(V_k \mathbb{E}(V_n | \mathcal{F}_0))$  converges,
- (ii) the series  $\sum_{k \geq K} |\mathbb{E}(V_k \mathbb{E}(V_n | \mathcal{F}_0))|$  converges to zero when  $n \rightarrow +\infty$ , uniformly with respect to  $K$ .

Then  $\mathbb{E}(U_n^2)$  converges to a real number  $\sigma^2 \geq 0$  when  $n \rightarrow +\infty$ . Furthermore, if  $\sigma^2 > 0$ , then  $U_n / \sqrt{\sigma^2}$  converges in distribution to the standard normal distribution.

**Proposition 3.3.** *In the general RN model plus YpR influence, for  $x \in \{C, A\}$ , when  $N \rightarrow +\infty$ ,  $\sqrt{N}((x, x)_{\text{obs}} - (x, x)(t))$ , respectively  $\sqrt{N}([x, x]_{\text{obs}} - [x, x](t))$ , converges in distribution to the centered normal distribution with variance  $\sigma_x^2(t)$ , respectively  $\tilde{\sigma}_x^2(t)$ , where*

$$\sigma_x^2(t) = (x, x)(t) + 2(xx, xx)(t) + 2(x * x, x * x)(t) - 5(x, x)(t)^2,$$

and up to turning  $(, )$ s into  $[, ]$ s

$$\tilde{\sigma}_x^2(t) = \sigma_x^2(t).$$

*Proof.* For any RN+YpR model, for  $x \in \{C, A\}$ , the sequence  $(K_i^x(t))_{i \in \mathbb{Z}}$ , respectively  $(\tilde{K}_i^x)_{i \in \mathbb{Z}}$ , is stationary and ergodic. Let  $V_i^x = K_i^x(t) - (x, x)(t)$ , respectively  $\tilde{V}_i^x = \tilde{K}_i^x - [x, x](t)$ . This defines a sequence  $(V_i^x)_{i \in \mathbb{Z}}$ , respectively  $(\tilde{V}_i^x)_{i \in \mathbb{Z}}$ , such that the first hypothesis of theorem 3.2 holds. We now check conditions (i) et (ii). The independence at distance 3, valid for any RN+YpR model, implies that, for every  $n \geq 3$ ,  $\mathbb{E}(V_n^x | \mathcal{F}_0^x) = \mathbb{E}(V_n^x) = 0$ , respectively  $\mathbb{E}(\tilde{V}_n^x | \tilde{\mathcal{F}}_0^x) = \mathbb{E}(V_n^x) = 0$ . Hence we only have to check the cases  $n = 1$  and  $n = 2$ .

For every  $k \geq 3$ ,  $V_k^x$ , respectively  $\tilde{V}_k^x$ , is independent of  $\mathcal{F}_0^x$ , respectively  $\tilde{\mathcal{F}}_0^x$ , and  $\mathbb{E}(V_n^x | \mathcal{F}_0^x)$ , respectively  $\mathbb{E}(\tilde{V}_n^x | \tilde{\mathcal{F}}_0^x)$ , is  $\mathcal{F}_0^x$ -measurable, respectively  $\tilde{\mathcal{F}}_0^x$ -measurable, hence

$$\mathbb{E}(V_k^x \mathbb{E}(V_n^x | \mathcal{F}_0^x)) = \mathbb{E}(V_k^x) \mathbb{E}(\mathbb{E}(V_n^x | \mathcal{F}_0^x)) = 0,$$

and

$$\mathbb{E}(\tilde{V}_k^x \mathbb{E}(\tilde{V}_n^x | \tilde{\mathcal{F}}_0^x)) = \mathbb{E}(\tilde{V}_k^x) \mathbb{E}(\mathbb{E}(\tilde{V}_n^x | \tilde{\mathcal{F}}_0^x)) = 0.$$

This implies (i) and (ii), hence theorem 3.2 applies. To compute the asymptotic variance in the theorem, we note that the variance of  $\sqrt{N}((x, x)_{\text{obs}} - (x, x)(t))$ , respectively  $\sqrt{N}([x, x]_{\text{obs}} - [x, x](t))$ , is  $N\sigma_x^2(N, t)$ , respectively  $N\tilde{\sigma}_x^2(N, t)$ , which converges to  $\sigma_x^2(t)$ , respectively  $\tilde{\sigma}_x^2(t)$ , when  $N \rightarrow +\infty$ .  $\square$

**3.3. Central limit theorems for  $T_x$  and  $\tilde{T}_x$ .** We describe explicitly the behaviour of  $T_x - t$  and  $\tilde{T}_x - t$ . To state our result, we use the central limit theorems given in proposition 3.3, but we now need to separate the Jukes-Cantor model with CpG influence and the general RN model plus YpR influence.

For  $x \in \{C, A\}$ , let  $\mu_x$ , respectively  $\tilde{\mu}_x$ , denote the reciprocal function of  $t \mapsto (x, x)(t)$ , respectively  $t \mapsto [x, x](t)$ , (both defined on  $[0, +\infty)$  and decreasing from  $(x)^2$  to  $(x)$ ).

We are sure from propositions 5.3, 6.3 and 6.5, that  $t \mapsto (x, x)(t)$  and  $t \mapsto [x, x](t)$  are diffeomorphisms in the Jukes-Cantor model with CpG influence. However in the general RN model plus YpR influence, it is only a conjecture. We performed simulations described in appendix D, to test the relevance of the decrease of  $t \mapsto (C, C)(t)$ .

**Conjecture 3.4.** *In the RN model plus YpR influence, for  $x \in \{C, A\}$ , the functions  $t \mapsto (x, x)(t)$  and  $t \mapsto [x, x](t)$  defined on  $[0, +\infty)$  are diffeomorphisms from  $(x)$  to  $(x)^2$ .*

Then,

$$T_x = \mu_x((x, x)_{\text{obs}}) \quad \text{and} \quad t = \mu_x((x, x)(t)),$$

and

$$\tilde{T}_x = \tilde{\mu}_x([x, x]_{\text{obs}}) \quad \text{and} \quad t = \tilde{\mu}_x([x, x](t)).$$

Besides, the derivatives of  $\mu_x$  and  $\tilde{\mu}_x$ , with respect to  $t$  are

$$\mu'_x((x, x)(t)) = \frac{1}{(x, x)'(t)} \quad \text{and} \quad \tilde{\mu}'_x([x, x](t)) = \frac{1}{[x, x]'(t)}.$$

Using the delta method, see [11], one gets the following result.

**Proposition 3.5.** *In the Jukes-Cantor model with CpG influence, for  $x \in \{C, A\}$ , when  $N \rightarrow +\infty$ ,  $\sqrt{N}(T_x - t)$ , respectively  $\sqrt{N}(\tilde{T}_x - t)$ , converges in distribution to the centered normal distribution with variance  $\sigma_x^2(t)/(x, x)'(t)^2$ , respectively  $\tilde{\sigma}_x^2(t)/[x, x]'(t)^2$ .*

*Under conjecture 3.4, the same results hold for the RN model plus YpR influence.*

#### 4. PROOFS OF THE RESULTS OF SECTION 2 IN JC + CPG MODEL

Proposition 3.5 yields the variation of  $T_x$  and  $\tilde{T}_x$  around  $t$  for  $x \in \{C, A\}$ . A priori, to build a confidence interval for  $t$  from this proposition requires to

know the value of  $(x, x)'(t)$  and of  $\sigma_x^2(t)$ , respectively  $[x, x]'(t)$  and of  $\tilde{\sigma}_x^2(t)$ , which both depend on the quantity  $t$  to be estimated.

As is customary, Slutsky's lemma (see [11]) allows to bypass this difficulty through the observed quantities  $\kappa_{\text{obs}}^x$  and  $\nu_{\text{obs}}^x$ , respectively  $\tilde{\kappa}_{\text{obs}}^x$  and  $\tilde{\nu}_{\text{obs}}^x$ , defined in section 2. Indeed, Slutsky's lemma states that if two sequences of random variables  $(X_N)_N$  and  $(Y_N)_N$  are such that  $(X_N)_N$  converges in distribution to a random variable  $X$  and  $(Y_N)_N$  converges in probability to a constant  $c$ , then the sequence  $(X_N Y_N)_N$  converges in distribution to the random variable  $cX$ .

**Lemma 4.1.** *In the Jukes Cantor model with CpG influence, for  $x \in \{C, A\}$ , when  $N \rightarrow +\infty$ ,  $\kappa_{\text{obs}}^x \rightarrow -(x, x)'(t)$  and  $\nu_{\text{obs}}^x \rightarrow \sigma_x^2(t)$  almost surely.*

*Proof.* The equalities

$$\begin{aligned} (C, C)'(t) &= -4(C, C)(t) - r(C*, CG)(t) + (C), \\ (A, A)'(t) &= -4(A, A)(t) + r(*A, CG)(t) + (A), \\ [C, C]'(t) &= -8[C, C](t) - 2r[C*, CG](t) + 2(C), \\ [A, A]'(t) &= -8[A, A](t) + 2r[*A, CG](t) + 2(A), \end{aligned}$$

given in sections 5 and 6, and the almost sure convergence of the observed quantities  $(C, C)_{\text{obs}}$ ,  $(C*, CG)_{\text{obs}}$ ,  $(CC, CC)_{\text{obs}}$ ,  $(C*C, C*C)_{\text{obs}}$ ,  $(A, A)_{\text{obs}}$ ,  $(*A, CG)_{\text{obs}}$ ,  $(AA, AA)_{\text{obs}}$  and  $(A*A, A*A)_{\text{obs}}$ , and the same turning  $(\cdot)$ s into  $[\cdot]$ s, to the corresponding limiting values, when  $N \rightarrow +\infty$ , imply the desired convergences.  $\square$

We apply Slutsky's lemma to the sequence  $(X_N) = (\sqrt{N}(T_x - t))$ , respectively  $(\tilde{X}_N) = (\sqrt{N}(\tilde{T}_x - t))$ , which converges in distribution to the centered normal law with variance  $\sigma_x^2(t)/(x, x)'(t)^2$ , respectively  $\tilde{\sigma}_x^2(t)/[x, x]'(t)^2$ , from proposition 3.5, and the sequence  $(Y_N) = (\kappa_{\text{obs}}^x/\sqrt{\nu_{\text{obs}}^x})$ , respectively  $(\tilde{Y}_N) = (\tilde{\kappa}_{\text{obs}}^x/\sqrt{\tilde{\nu}_{\text{obs}}^x})$ , which converges in probability to  $-(x, x)'(t)/\sigma_x(t)$ , respectively  $-[x, x]'(t)/\tilde{\sigma}_x(t)$ , from lemma 4.1. This implies theorem 2.3.

## 5. EVOLUTIONS OF $(C, C)(t)$ AND $[C, C](t)$ IN JC+CPG MODEL

Under the Jukes-Cantor model with CpG effect, dinucleotides coded as  $\{C, \bar{C}\} \times \{G, \bar{G}\}$  have autonomous evolution with the following  $4 \times 4$  rate matrix  $Q$ :

$$\begin{array}{c} \begin{array}{cccc} & CG & \bar{C}G & \bar{C}\bar{G} & C\bar{G} \\ \begin{array}{l} CG \\ \bar{C}G \\ \bar{C}\bar{G} \\ C\bar{G} \end{array} & \begin{pmatrix} -(6+2r) & 3+r & 0 & 3+r \\ 1 & -4 & 3 & 0 \\ 0 & 1 & -2 & 1 \\ 1 & 0 & 3 & -4 \end{pmatrix} \end{array} \end{array}$$

The dynamics of the dinucleotides can be represented with the graph given in figure 2.

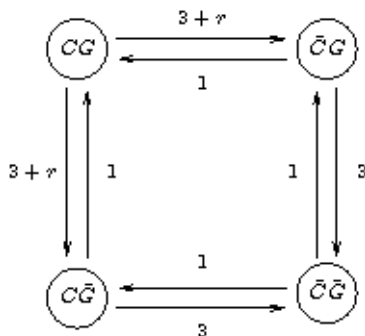


FIGURE 2. Dynamics of dinucleotides encoded as  $\{C, \bar{C}\} \times \{G, \bar{G}\}$

The exponential of the corresponding matrix can be explicitly computed. One can also compute explicitly the stationary frequencies of dinucleotides coded as  $\{C, \bar{C}\} \times \{G, \bar{G}\}$  using the same matrix. That is

$$\begin{aligned} (CG) &= \frac{1}{16 + 5r}, & (C\bar{G}) &= \frac{3 + r}{16 + 5r}, \\ (\bar{C}\bar{G}) &= \frac{9 + 3r}{16 + 5r}, & (\bar{C}G) &= \frac{3 + r}{16 + 5r}. \end{aligned}$$

These stationary frequencies have already been derived in [1] by Bérard, Gouéré and Piau [1].

We observe that  $(C, C)(t)$  can be expressed as a linear combination of terms of the form  $(XY, ZT)(t)$  where  $(X, Y)$  and  $(Z, T)$  belong to  $\{C, \bar{C}\} \times \{G, \bar{G}\}$ .

It is then clear that an explicit expression for  $(C, C)(t)$  can be obtained, and that an expression of  $(C, C)'(t)$  in terms of dinucleotide frequencies holds.

**Proposition 5.1.** *The evolution of  $(C, C)(t)$  satisfies the linear differential equation*

$$(C, C)'(t) = -4(C, C)(t) - r(C^*, CG)(t) + (C)(0).$$

Proposition 5.1 is valid out of equilibrium. We use it at stationarity hence, in particular, for the initial values

$$(C)(0) = (C) = \frac{4 + r}{16 + 5r}, \quad (CG)(0) = (CG) = \frac{1}{16 + 5r}.$$

The equation in proposition 5.1 yields expressions of  $(C, C)(t)$ . Consider the positive real numbers  $u$ ,  $u_+$  and  $u_-$  defined as

$$u = \sqrt{4 + 2r + r^2}, \quad u_+ = 6 + r + u, \quad u_- = 6 + r - u.$$

**Corollary 5.2.** *In the stationary regime,*

$$(C, C)(t) = c_0 e^{-4t} + c_+ e^{-u_+ t} + c_- e^{-u_- t} + (C)^2,$$

with

$$c_0 = \frac{3+r}{2(16+5r)} \quad \text{and} \quad c_{\pm} = \frac{3+r}{4u(16+5r)^2} (u(16+3r) \mp (32+14r+3r^2)).$$

As expected,

$$c_+ + c_- + c_0 = (C) - (C)^2.$$

Furthermore, for every positive  $r$ ,

$$4 < u_- < 5 < 2r + 7 < u_+ < 2r + 8.$$

Although the Jukes-Cantor model with CpG effect is not reversible, the dynamics of dinucleotides encoded as  $\{C, \bar{C}\} \times \{G, \bar{G}\}$  with respect to this model is reversible. This can easily be checked by looking at the cycles in figure 2.

Reversibility means that the dynamics will look the same whether time runs forward or backward. As a result, given two sequences at stationarity, the probability of data in a state is the same whether one sequence is ancestral to the other or both are descendants of an ancestral sequence at stationarity. Roughly speaking, for every  $(X, Y)$  and  $(Z, T)$  that belong to  $\{C, \bar{C}\} \times \{G, \bar{G}\}$ , going from a  $XY$  at time  $t$  to 0 then back to a  $ZT$  at time  $t$  on another branch, is equivalent to going from a  $XY$  to at time 0 to a  $ZT$  at time  $2t$ .

In consequence, for every positive  $t$ , we have

$$[C, C](t) = (C, C)(2t).$$

For every positive  $r$ , the parameters  $c_{\pm}$  and  $c_0$ , are positive. This proves the following proposition.

**Proposition 5.3.** *In the Jukes-Cantor model with CpG influence, the functions  $t \mapsto (C, C)(t)$  and  $t \mapsto [C, C](t)$  defined on  $[0, +\infty)$ , are decreasing diffeomorphisms from  $(C)$  to  $(C)^2$ .*

## 6. EVOLUTIONS OF $(A, A)(t)$ AND $[A, A](t)$ IN JC+CPG

Like  $(C, C)$ , it is possible to encode dinucleotides such that under the Jukes-Cantor model with CpG effect,  $(A, A)$  is a linear combination of terms involved in an autonomous evolution. It suffices to encode the dinucleotides as  $\{C, \bar{C}\} \times \{A, G, Y\}$ , and the dynamics can be represented with the graph given in figure 3.

However, we don't use this encoding to compute  $(A, A)(t)$ . Indeed, the evolution matrix associated to this encoding is a  $6 \times 6$  matrix whereas it is possible to deal with the  $4 \times 4$  matrix  $Q$ , defined in section 5, to state the evolution of  $(A, A)$  as explained below.

We choose to present this encoding because it is a way to understand the difference between the role of  $C$  and  $A$  in the Jukes Cantor model with CpG

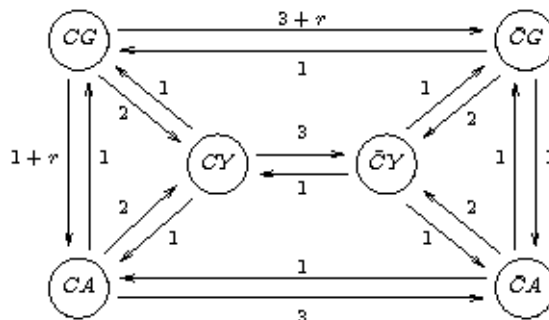


FIGURE 3. Dynamics of dinucleotides encoded as  $\{C, \bar{C}\} \times \{A, G, Y\}$

effect. Indeed, the dynamics of dinucleotides encoded as  $\{C, \bar{C}\} \times \{A, G, Y\}$  is not reversible. This can be checked by looking at the cycle  $CA \rightarrow CY \rightarrow CG \rightarrow CA$  in figure 3. In consequence, unlike  $[C, C]$ , we can't state a relation between  $[A, A](t)$  and  $(A, A)(2t)$ , and in fact, as soon as  $r > 0$ , the equality

$$[A, A](t) = (A, A)(2t),$$

is false.

We describe a way to state the expression of  $(A, A)(t)$ . We state easily the following proposition.

**Proposition 6.1.** *The evolution of  $(A, A)(t)$  satisfies the linear differential equation*

$$(A, A)'(t) = -4(A, A)(t) + r(*A, CG)(t) + (A)(0).$$

Let  $U(t)$  denote the time dependent vector defined as

$$\begin{pmatrix} (*A, CG)(t) \\ (*A, \bar{C}G)(t) \\ (*A, \bar{C}\bar{G})(t) \\ (*A, C\bar{G})(t) \end{pmatrix},$$

then we have

$$U'(t) = {}^tQU(t).$$

We can now compute  $(*A, CG)(t)$ , infer the value of  $(A)(0)$  at stationarity and finally state the expression of  $(A, A)(t)$ .

**Corollary 6.2.** *In the stationary regime,*

$$(A, A)(t) = a_0 e^{-4t} + a_+ e^{-u_+ t} + a_- e^{-u_- t} + (A)^2,$$

with

$$a_0 = \frac{80 + 31r}{32(16 + 5r)},$$

and

$$a_{\pm} = \frac{512 + 384r + 106r^2 + 13r^3 \mp u(256 + 18r + 13r^2)}{64u(16 + 5r)^2}.$$

For every positive  $r$ , the parameters  $a_{\pm}$  and  $a_0$ , are positive. This proves the following proposition.

**Proposition 6.3.** *In the Jukes-Cantor model with CpG influence, the function  $t \mapsto (A, A)(t)$  defined on  $[0, +\infty)$ , is a decreasing diffeomorphism from  $(A)$  to  $(A)^2$ .*

We deal now with the evolution of  $[A, A](t)$ . Introduce the constant matrices

$$M = \begin{pmatrix} -8 & 2r & 0 & 0 \\ 0 & -(12 + 2r) & 1 & 1 \\ 0 & -r & -8 & 0 \\ 0 & -r & 0 & -8 \end{pmatrix}, \quad N = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & r & 0 \\ 0 & r & 0 & 0 \\ 0 & r & 0 & 0 \end{pmatrix},$$

$$O = \begin{pmatrix} -8 & -2r & 0 & 0 \\ 1 & -(12 + 2r) & -r & 1 \\ 0 & 4 & -(16 + 4r) & 0 \\ 0 & -2r & 0 & -8 \end{pmatrix}, \quad B = \begin{pmatrix} 2(A) \\ (CG) \\ (C) + (A) \\ (A) + (G) \end{pmatrix},$$

$$D = \begin{pmatrix} 2(C) \\ (CG) \\ 0 \\ 2(C) \end{pmatrix},$$

and the time-dependent vectors

$$V(t) = \begin{pmatrix} [A, A](t) \\ [*A, CG](t) \\ [*A, C*](t) \\ [A, G](t) \end{pmatrix}, \quad W(t) = \begin{pmatrix} [C, C](t) \\ [C*, CG](t) \\ [CG, CG](t) \\ [C*, *G](t) \end{pmatrix}.$$

**Proposition 6.4.** *The evolution of  $V(t)$  and  $W(t)$  is ruled by the linear differential systems*

$$V'(t) = MV(t) + NW(t) + B, \quad W'(t) = OW(t) + D.$$

The exponential of the block matrix  $\begin{pmatrix} M & N \\ 0 & O \end{pmatrix}$  can be explicitly computed.

The explicit spectrum of this matrix is

$$\{-8, -8, -8, -v_+, -v_-, -2u_+, -2u_-, -(12 + 2r)\}, \quad v_{\pm} = 10 + r \pm u.$$

Thus, it is possible to compute an explicit expression for  $[A, A](t)$ . The computation under Maple shows that  $[A, A](t)$  possesses nonzero coefficients in front of  $e^{-v_+t}$  and  $e^{-v_-t}$ , and this fact proves that  $[A, A](t)$  can't be equal to  $(A, A)(2t)$ . However, this two quantities are very closed as we can see on figure 4.



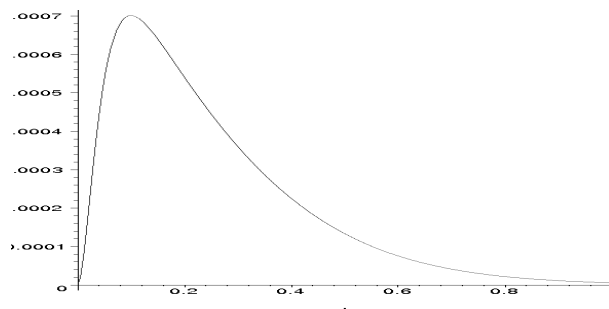


FIGURE 4. Representation of  $t \mapsto [A, A](t) - (A, A)(2t)$ , when  $r = 10$

We don't give here the explicit expression of  $[A, A](t)$ , however it seems that the following proposition occurs.

**Proposition 6.5.** *In the Jukes-Cantor model with CpG influence, the function  $t \mapsto [A, A](t)$  defined on  $[0, +\infty)$ , is a decreasing diffeomorphism from  $(A)$  to  $(A)^2$ .*

#### APPENDIX A. SHORT DESCRIPTION OF THE RN MODEL WITH YpR INFLUENCE AND NOTATIONS

Firstly, RN stands for Rzhetsky-Nei and means that the  $4 \times 4$  matrix of substitution rates which characterize the independent evolution of the sites must satisfy 4 equalities, summarized as follows: for every pair of nucleotides  $x$  and  $y \neq x$ , the substitution rate from  $x$  to  $y$  may depend on  $x$  but only through the fact that  $x$  is a purine ( $A$  or  $G$ , symbol  $R$ ) or a pyrimidine ( $C$  or  $T$ , symbol  $Y$ ). For instance, the substitution rates from  $C$  to  $A$  and from  $T$  to  $A$  must coincide, likewise for the substitution rates from  $A$  to  $C$  and from  $G$  to  $C$ , from  $C$  to  $G$  and from  $T$  to  $G$ , and finally from  $A$  to  $T$  and from  $G$  to  $T$ . The 4 remaining rates, corresponding to purine-purine and to pyrimidine-pyrimidine substitutions, are free.

Secondly, the influence mechanism is called YpR, which stands for the fact that one allows any specific substitution rates between any two YpR dinucleotides ( $CG$ ,  $CA$ ,  $TG$  and  $TA$ ) which differ by one position only, for a total of 8 independent parameters. The Jukes-Cantor model with CpG effect is the simplest non trivial one: the only YpR substitutions with positive rate are  $CG \rightarrow CA$  and  $CG \rightarrow TG$ , and both happen at the same rate.

Recall that  $Y$  denote the set of pyrimidines defined as  $Y = \{T, C\}$ , and  $R$  the set of purines defined as  $= \{A, G\}$ .

The  $4 \times 4$  matrix of substitution rates which characterize the independent evolution of the sites in RN model is given by

$$\begin{matrix} & A & T & C & G \\ \begin{matrix} A \\ T \\ C \\ G \end{matrix} & \begin{pmatrix} \cdot & v_T & v_C & w_G \\ v_A & \cdot & w_C & v_G \\ v_A & w_T & \cdot & v_G \\ w_A & v_T & v_C & \cdot \end{pmatrix} \end{matrix}.$$

The influence mechanism called YpR adds specific rates of substitutions from each YpR dinucleotide as follows.

- Every dinucleotide  $CG$  moves to  $CA$  at rate  $r_A^C$  and to  $TG$  at rate  $r_T^G$ .
- Every dinucleotide  $TA$  moves to  $CA$  at rate  $r_C^A$  and to  $TG$  at rate  $r_G^T$ .
- Every dinucleotide  $CA$  moves to  $CG$  at rate  $r_G^C$  and to  $TA$  at rate  $r_T^A$ .
- Every dinucleotide  $TG$  moves to  $CG$  at rate  $r_C^G$  and to  $TA$  at rate  $r_A^T$ .

#### APPENDIX B. EXTENSION OF THEOREM 2.3 TO THE RN MODEL WITH YPR INFLUENCE

Under conjecture 3.4, it is possible to generalize theorem 2.3 by suitably generalizing the definitions of  $\kappa$  and  $\nu$  given in section 2.

Let  $\kappa_{\text{obs}}^{RN}$  and  $\nu_{\text{obs}}^{RN}$  defined as

$$\begin{aligned} \kappa_{\text{obs}}^{RN} &= -v_C(C, A)_{\text{obs}} - w_C(C, T)_{\text{obs}} + u_C(C, C)_{\text{obs}} - v_C(C, G)_{\text{obs}} \\ &\quad - r_C^A(C^*, TA)_{\text{obs}} - r_C^G(C^*, TG)_{\text{obs}} + r_T^A(C^*, CA)_{\text{obs}} + r_T^G(C^*, CG)_{\text{obs}}. \\ \nu_{\text{obs}}^{RN} &= \nu_{\text{obs}}^C. \end{aligned}$$

In the case where  $v_C = w_C = 1$ ,  $u_C = 3$ ,  $r_C^A = r_C^G = r_T^A = 0$  and  $r_T^G = r$ , which is the case in the Jukes Cantor model with CpG influence, we note that  $\kappa_{\text{obs}}^{RN} = \kappa_{\text{obs}}^C$ .

On the other side, the observed quantity  $\nu_{\text{obs}}^C$  is unchanged between the Jukes Cantor model with CpG influence and the RN model with YpR influence because lemma 3.1 occurs for the general model.

Once again, Slutsky's lemma, through the observed quantities  $\kappa_{\text{obs}}^{RN}$  and  $\nu_{\text{obs}}^{RN}$  is the key to state theorem B.1 below from proposition 3.5.

**Theorem B.1.** *Assume that the ancestral sequence is at stationarity. Under conjecture 3.4, when  $N \rightarrow +\infty$ ,  $\kappa_{\text{obs}}^{RN} \sqrt{N/\nu_{\text{obs}}^{RN}}(T_C - t)$  converges in*

distribution to the standard normal law. An asymptotic confidence interval at level  $\varepsilon$  for  $t$  is

$$\left[ T_C - \frac{z(\varepsilon)}{\kappa_{\text{obs}}^{RN}} \sqrt{\frac{\nu_{\text{obs}}^{RN}}{N}}, T_C + \frac{z(\varepsilon)}{\kappa_{\text{obs}}^{RN}} \sqrt{\frac{\nu_{\text{obs}}^{RN}}{N}} \right],$$

where  $z(\varepsilon)$  denotes the unique real number such that  $\mathbb{P}(|Z| \geq z(\varepsilon)) = \varepsilon$  with  $Z$  a variable with standard normal law.

As in the Jukes-Cantor model with CpG influence, the estimator  $T_C$  is defined implicitly for the RN model with YpR influence. We don't have an explicit formula for  $(C, C)(t)$  in the general model, but there are numerical methods to compute a closed solution of the theoretical solution of the differential linear system, and consequently it is possible to solve equation  $(C, C)(t) = (C, C)_{\text{obs}}$  with numerical methods.

#### APPENDIX C. EVOLUTION OF $(C, C)(t)$ IN RN + YpR

We base our description of the method in the general RN model with YpR influence on the encoding of dinucleotides as  $\{R, T, C\} \times \{Y, G, A\}$ , which has autonomous evolution.

The detailed description of the corresponding  $9 \times 9$  matrix is given below as  $m(uv, xy)$ , where  $uv$  and  $xy$  are generic elements of the alphabet.

Let  $v_R$  and  $v_Y$  denote the quantities defined as

$$v_R = v_A + v_G, \quad v_Y = v_T + v_C.$$

We have

$$\begin{aligned} m(uv, xy) &= 0, & \text{if } u \neq x \text{ and } v \neq y; \\ m(Rx, ux) &= v_u, & \text{if } x \in \{Y, G, A\} \text{ and } u \in \{C, T\}; \\ m(ux, Rx) &= v_R, & \text{if } x \in \{Y, G, A\} \text{ and } u \in \{C, T\}; \\ m(Ru, Rv) &= w_v, & \text{if } \{u, v\} = \{A, G\}; \\ m(xY, xu) &= v_u, & \text{if } x \in \{R, C, T\} \text{ and } u \in \{A, G\}; \\ m(xu, xY) &= v_R, & \text{if } x \in \{R, C, T\} \text{ and } u \in \{A, G\}; \\ m(uY, vY) &= w_v, & \text{if } \{u, v\} = \{T, C\}; \\ m(xu, xv) &= w_v + r_v^x, & \text{if } \{u, v\} = \{A, G\} \text{ and } x \in \{T, C\}; \\ m(ux, vx) &= w_v + r_v^x, & \text{if } \{u, v\} = \{C, T\} \text{ and } x \in \{A, G\}. \end{aligned}$$

It is then clear that quantities such as  $(C, C)(t)$  can be computed provided that we can find the exponential of the rate-matrix, and that quantities such as  $(C, C)'(t)$  have computable explicit expressions in terms of frequencies expressed in the coded dinucleotide-alphabet  $\{R, T, C\} \times \{Y, G, A\}$ .

APPENDIX D. SIMULATIONS PERFORMED

As a support to the conjecture that  $t \mapsto (C, C)(t)$  always defines a diffeomorphism in the general RN + YpR case, we refer to simulations having been performed. We give the range of parameter values that we explored and one example of program for one set of parameters, here a Kimura model with CpG influence.

D.1. Range of parameter values explored.

$v_A$	1	1	1	1	1	1	1	$w_A$	1	3	0.3	0.3	3	3	3
$v_T$	1	1	1	1	1	2	0.3	$w_T$	1	3	0.3	0.3	3	6	1
$v_C$	1	1	1	1	1	1	2	$w_C$	1	3	0.3	0.3	3	3	1
$v_G$	1	1	1	1	1	2	10	$w_G$	1	3	0.3	0.3	3	6	0.1
$r_A^C$	10	10	10	10	0.3	10	10	$r_T^G$	10	10	10	10	0.3	10	5
$r_C^A$	0	0	0	10	0.3	5	1	$r_G^T$	0	0	0	10	0.3	5	0.5
$r_G^C$	0	0	0	10	0.3	3	20	$r_T^A$	0	0	0	10	0.3	3	3
$r_C^G$	0	0	0	10	0.3	1	0.3	$r_A^T$	0	0	0	10	0.3	1	0.1

D.2. One example of simulation performed on Maple. The following program illustrate a simulation performed with the parameter values:

$$v_A = v_T = v_C = v_G = 1, \quad w_A = w_T = w_C = w_G = 3,$$

$$r_A^C = r_T^G = 10, \quad r_C^A = r_G^T = r_G^C = r_T^A = r_C^G = r_A^T = 0.$$

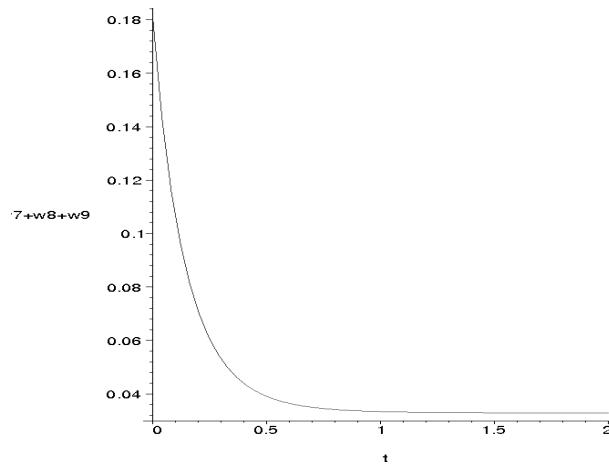
It is a Kimura model with CpG influence.

```
restart;
with(LinearAlgebra): with(plots):
vA := 1: wA := 3: vT := vA: wT := wA: vC := vA: wC := wA: vG :=
vA: wG := wA: vR := vA + vG: vY := vT + vC: rCA := 10: rGT := rCA:
rAC := 0: rTG := 0: rCG := 0: rAT := 0: rGC := 0: rTA := 0:
CAs := CAs: CYs := CYs: CGs := CGs:
Q := << -(vY + wG + vT + vC) | vA | wA | vR | 0 | 0 | vR | 0 |
0 >, < vY | -(vA + vG + vT + vC) | vY | 0 | vR | 0 | 0 | vR | 0
>, < wG | vG | -(wA + vY + vT + vC) | 0 | 0 | vR | 0 | 0 | vR >,
< vT | 0 | 0 | -(vR + vY + wG + rTG + wC + rAC) | vA | wA + rTA
| wT + rAT | 0 | 0 >, < 0 | vT | 0 | vY | -(vR + vA + vG + vC) |
vY | 0 | wT | 0 >, < 0 | 0 | vT | wG + rTG | vG | -(vR + wA + rTA
+ vY + wC + rGC) | 0 | 0 | wT + rGT >, < vC | 0 | 0 | wC + rAC |
0 | 0 | -(vR + wT + rAT + vY + wG + rCG) | vA | wA + rCA >, < 0
| vC | 0 | 0 | vC | 0 | vY | -(vR + wT + vA + vG) | vY >, < 0
| 0 | vC | 0 | 0 | wC + rGC | wG + rCG | vG | -(vR + wT + rGT +
wA + rCA + vY) > >:
u := < u1, u2, u3, u4, u5, u6, u7, u8, u9 >:
```

```

v := Multiply(Q,u):
eqns := {seq( v[i] = 0, i=1..9)}:
sols := solve(eqns):
assign(sols);
eqn := {sum('u[i]', 'i'=1..9)=1}:
sol := solve(eqn):
assign(sol);
CAs := u[7]: CYs := u[8]: CGs := u[9]:
w := < w1(t), w2(t), w3(t), w4(t), w5(t), w6(t), w7(t), w8(t), w9(t)
>:
wdiff := < diff(w1(t),t), diff(w2(t),t), diff(w3(t),t), diff(w4(t),t),
diff(w5(t),t), diff(w6(t),t), diff(w7(t),t), diff(w8(t),t), diff(w9(t),t)
>:
wtemp := wdiff - Multiply(Q,w):
dsys := seq( wtemp[i] = 0, i=1..9):
init := w1(0) = 0, w2(0) = 0, w3(0) = 0, w4(0) = 0, w5(0) = 0, w6(0)
= 0, w7(0) = CAs, w8(0) = CYs, w9(0) = CGs:
fcns := {seq(w[i], i=1..9)}:
dsol := dsolve({dsys,init}, fcns, numeric ):
odeplot(dsol, [t,sum( 'w[i]', 'i'=7..9)],0..2);

```



## REFERENCES

- [1] J. Bérard, J.-B. Gouéré, and D. Piau. Solvable models of neighbor-dependent nucleotide substitution processes. *Mathematical Biosciences*, 211:56–88, 2008.

- [2] L. Duret and N. Galtier. The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact. *Molecular biology and evolution*, 17:1620–1625, 2000.
- [3] J. Felsenstein. Evolutionary trees from DNA sequences : A maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.
- [4] N. Galtier, M. Gouy, and C. Gautier. Seaview and phylo\_win, two graphic tools for sequence alignment and molecular phylogeny. *Comput. Applic. Biosci.*, 12:543–548, 1996.
- [5] P. Hall and C. C. Heyde. *Martingale limit theory and its applications*. Academic Press, New York, 1980.
- [6] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, 22:160–174, 1985.
- [7] J. L. Jensen and A. Pedersen. Probabilistic models of DNA sequence evolution with context dependent rates substitution. *Adv. Appl. Prob.*, 32:459–517, 2000.
- [8] D. Jones, W. Taylor, and J. Thornton. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, 8:275–282, 1992.
- [9] T.H. Jukes and C.R. Cantor. *Mammalian protein metabolism*, chapter Evolution of Protein Molecules, pages 21–132. Academic Press, New York, 1969.
- [10] M. Kimura. A Simple Method for Estimating Evolutionary Rates of Base Substitutions Through Comparative Studies of Nucleotide Sequences. *J. Mol. Evol.*, 10:111–120, 1980.
- [11] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

UNIVERSITÉ JOSEPH FOURIER GRENOBLE 1, INSTITUT FOURIER UMR 5582 UJF-CNRS,  
100 RUE DES MATHS, BP 74, 38402 SAINT MARTIN D'HÈRES, FRANCE