Phylogenetic distances for neighbour dependent substitution processes

Mikael Falconnet

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. For a wide class of such nonreversible models, we show how to compute consistent, mathematically exact, estimators of the time elapsed between aligned sequences, for an ancestral sequence and a present one, and also for two present sequences. In both cases, we provide asymptotic confidence intervals, valid for nucleotidic sequences of finite length. We compute explicit formulas for the estimators and for their confidence intervals in the simplest nontrivial case, the Jukes-Cantor model with CpG influence.

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, 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 practical reasons, since some neighbour influences are well documented in the literature, and caused by well known biological mechanisms, and yield substitution rates which can be, in some cases, much larger than their independent counterparts. A class of mathematical models with neighbour influences was recently introduced by biologists, see [GGG96], and studied mathematically, see [BGP08], and through simulations, see [AH05] and [ABH03] for instance. The goal of the present paper is to show that one can compute exact formulas for consistent estimators of the distances between DNA sequences whose evolution is ruled by any model in this class.

As a proof of concept, we completely describe the construction in the simplest non trivial case, the Jukes-Cantor model with (symmetric) CpG influence, and we show that its evolution is ruled by finite sized linear systems. Note that for every model in this class one can write similar closed formulas.

Date: 10th December 2008.

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

Key words and phrases. Confidence intervals, DNA sequences, phylogenetic distances, CpG deficiency.

In section 1, we describe the class of manageable models introduced in [BGP08] 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. The other sections explain how we state 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 [BGP08], and we introduce some notations.

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

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 [JC69], [Kim80], [Fel81] and [HKY85] for instance. Even in codon evolution models, see [JTT92], one still assumes that different codons evolve independently. However, it is a well known experimental fact, see [DG00] 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 superposed mechanisms.

The first mechanism is an independent evolution of the sites as in the usual models. Hence it is characterized by a 4×4 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, possibly after a rescaling of the time, the rate of the substitutions of x by y is set to 1, for every nucleotides x and y in $\mathscr{A}$.

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 exponential time of 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 exponential time of 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 [BGP08] that the relevant class of models, called RN+YpR in this paper, is in fact larger than just described.

Firstly, RN stands for Rzhetsky-Nei and means that the 4×4 matrix of substitution rates which characterize the independent evolution of the sites must satisfy 4 equalities, summarized as follows: for every 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 case described above is the simplest non trivial (symmetric) one: the only YpR substitutions with positive rate are $CG\to CA$ and $CG\to TG$, and both happen at the same rate.

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

1.2. **Main properties.** We now recall some results of [BGP08], valid for every RN model with YpR influence. First, for every probability measure ν on $\mathscr{A}^{\mathbb{Z}}$, there exists a unique Markov process $(X(t))_{t\geqslant 0}$ on $\mathscr{A}^{\mathbb{Z}}$, with initial distribution ν, 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 ith coordinate $X_i(t)$ of $X(t)$ is the random value of the nucleotide at site i and time t. The process $(X(t))_{t\geqslant 0}$ is ergodic, its unique stationary distribution π on $\mathscr{A}^{\mathbb{Z}}$ is invariant and ergodic with respect to the translations of $\mathbb{Z}$, and π puts a positive mass on every cylinder of $\mathscr{A}^{\mathbb{Z}}$.

Thus, for every finite word $w=(w_i)_{0\leqslant i\leqslant \ell}$ written in the alphabet $\mathscr{A}$, $\pi(w)$ is positive. Furthermore, for every position i in $\mathbb{Z}$, $\mathbb{P}_\nu(X_{i:i+\ell}(t)=w)$ converges to $\pi(w)$ when $t\to+\infty$. (Here and later on, for every indices i and j in $\mathbb{Z}$ with $i\leqslant j$ and every symbol S, the shorthand $S_{i:j}$ denotes $(S_k)_{i\leqslant k\leqslant j}$.) Finally, if ξ in $\mathscr{A}^{\mathbb{Z}}$ is distributed along π, the empirical frequencies of any word w in ξ, observed along any increasing sequence of intervals of $\mathbb{Z}$, almost surely converge to $\pi(w)$.

These properties stem from the following representation of the distribution π. There exists an i.i.d. sequence $(H_i)_{i\in\mathbb{Z}}$ of Poisson processes, and a measurable map

Ψ with values in $\mathscr{A}$, 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 π. 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 $|i-j| \geqslant 3$ 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.

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

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 $\mathscr{A}$, we use the shorthands $*x = \mathscr{A}\times\{x\}$, $x* = \{x\}\times\mathscr{A}$ and $\overline{x} = \mathscr{A}\setminus\{x\}$.

From now on and with the exception of the statement of theorem 6.1, $X(0)$ and $X^1(0) = X^2(0)$ are distributed according to π, so the system is stationary.

2. Summary of main results

Theorems 2.2 and 2.5 below provide asymptotic confidence intervals for the time elapsed between a present sequence and an ancestral one, for the Jukes-Cantor model with CpG influence of intensity r. These intervals are based on two consistent estimators of the elapsed time t. Our first estimator is based on the evolution of the frequency $(C,C)(t)$ when the time t varies and the other one on the evolution of $(A,A)(t)$.

Propositions 2.7, 2.8 and 2.9 allow to compare the convergence to equilibrium of $(C,C)(t)$ in models with influence and in independent models with corresponding rates of substitution.

Finally, theorem 2.11 provides an asymptotic confidence interval for the time since two present sequences issued from the same ancestral sequence diverged. Theorem 2.11 is the keystep for the creation of phylogenetic trees built by a distance-based method.

2.1. **Alignment of cytosines in an ancestral sequence and a present one.** Let $(C,C)_{\text{obs}}$ denote the observed value of (C,C) on two aligned sequences of length N, that is,

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

In figure 1.3 for instance, $N = 7$ and $(C,C)_{\text{obs}} = \frac{2}{7}$.

Definition 2.1. *Let T_C denote the estimator of the elapsed time t defined as the solution in t of the equation*

$$(C,C)(t) = (C,C)_{\text{obs}}.$$

Let κ_{obs}^C and v_{obs}^C denote observed quantities, defined as

$$\kappa_{\text{obs}}^C = r(C*,CG)_{\text{obs}} - 4(C,C)_{\text{obs}} - (C),$$

and

$$v_{\text{obs}}^C = (C,C)_{\text{obs}} - (C,C)_{\text{obs}}^2 + 2(CC,CC)_{\text{obs}} + 2(C*C,C*C)_{\text{obs}}.$$

As explained in section 6, the function $t \mapsto (C,C)(t)$ is decreasing from (C) to $(C)^2$, where (C) is the frequency of C sites at stationarity. Thus, T_C is unique and well defined for any pair of aligned sequences such that

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

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

We note that v_{obs}^C is always positive whereas κ_{obs}^C might be negative for some sequences of observations and lengthes N. However, from lemma 5.1 ini section 5, κ_{obs}^C is almost surely positive when N is large.

We now state our main result.

Theorem 2.2. *Assume that the ancestral sequence is at stationarity. Then, when $N \to +\infty$, $\kappa_{\text{obs}}^C\sqrt{N/v_{\text{obs}}^C}(T_C - t)$ converges in distribution to the standard normal law. An asymptotic confidence interval at level ε for t is*

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

where $z(\varepsilon)$ denotes any real number such that $\mathbb{P}(|Z| \geqslant z(\varepsilon)) \leqslant \varepsilon$ with Z a variable with standard normal law.

Proposition 2.3. *When t is large, the variations of T_C around t are of order $\mathrm{e}^{4t}/\sqrt{N}$.*

The meaning of proposition 2.3 is that one must observe a part of the sequence of length N at least of order e^{8t} to estimate t up to a given factor.

2.2. **Alignment of adenines in an ancestral sequence and a present one.** One can use (A,A) like (C,C) before. We skip the details and only state the results.

Let $(A,A)_{\text{obs}}$ denote the observed value of (A,A) on two aligned sequences of length N, that is,

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

In figure 1.3 for instance, $N = 7$ and $(A,A)_{\text{obs}} = \frac{2}{7}$.

Definition 2.4. *Let T_A denote the estimator of the elapsed time t defined as the solution in t of the equation*

$$(A,A)(t) = (A,A)_{\text{obs}}.$$

Let κ_{obs}^A and v_{obs}^A denote observed quantities, defined as

$$\kappa_{\text{obs}}^A = -4(A,A)_{\text{obs}} + r(*A,CG)_{\text{obs}} - (A),$$

where (A) is the frequency of sites occupied by an A at stationarity and

$$v_{\text{obs}}^A = (A,A)_{\text{obs}} + 2(AA,AA)_{\text{obs}} + 2(A*A,A*A)_{\text{obs}} - (A,A)_{\text{obs}}^2.$$

Theorem 2.5. *Assume that the ancestral sequence is at stationarity. Then, when $N \to +\infty$, $\kappa_{\text{obs}}^A\sqrt{N/v_{\text{obs}}^A}(T_A - t)$ converges in distribution to the standard normal law. An asymptotic confidence interval at level ε for t is*

$$\left[T_A - \frac{z(\varepsilon)}{\kappa_{\text{obs}}^A}\sqrt{\frac{v_{\text{obs}}^A}{N}}, T_A + \frac{z(\varepsilon)}{\kappa_{\text{obs}}^A}\sqrt{\frac{v_{\text{obs}}^A}{N}}\right],$$

where $z(\varepsilon)$ is such as in theorem 2.2.

Proposition 2.6. *When t is large, the variations of T_A around t are of order $\mathrm{e}^{4t}/\sqrt{N}$.*

2.3. **Comparisons with standard models.** First, we study the independent Jukes-Cantor model with the same overall rate of substitutions, then the independent Kimura model with the same transition and transversion overall rates, and finally the independent model with the same overall rate for each of the 12 possible substitutions.

Proposition 2.7. *The convergences of $(C,C)(t)$ and $(A,A)(t)$ to equilibrium when $t \to +\infty$ in the Jukes-Cantor model with CpG influence are slower than in the independent Jukes-Cantor model with the same global rate of substitution.*

In Kimura's model [Kim80], the rates of transitions ($A \leftrightarrow G$ and $T \leftrightarrow C$) and transversions ($A,G \leftrightarrow T,C$) may be different. Following the notations in [Yan06], these models describe independent evolutions of the sites with 4×4 infinitesimal

generators:

$$\begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{pmatrix} \cdot & \beta & \beta & \alpha \\ \beta & \cdot & \alpha & \beta \\ \beta & \alpha & \cdot & \beta \\ \alpha & \beta & \beta & \cdot \end{pmatrix} \end{array},$$

where the sum of the coefficients on every line is equal to 0.

Proposition 2.8. *The convergence of $(C,C)(t)$ and $(A,A)(t)$ to equilibrium when $t \to +\infty$ in the Jukes-Cantor model with CpG influence are the same than in the Kimura model with the same transversion and transition overall rates.*

We finally compare the Jukes-Cantor model with CpG influence to the independent RN model with the same overall rate for each of the 12 possible substitutions. The 4×4 infinitesimal generator is given by:

$$\begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{pmatrix} \cdot & 1 & 1 & 1 \\ 1 & \cdot & 1 & 1 \\ 1 & 1+\delta & \cdot & 1 \\ 1+\delta & 1 & 1 & \cdot \end{pmatrix} \end{array},$$

where the sum of the coefficients on every line is equal to 0 and $\delta = r(CG)/(C)$.

Proposition 2.9. *The convergence of $(C,C)(t)$ to equilibrium when $t \to +\infty$ in the Jukes-Cantor model with CpG influence is slower than in the model above whereas the convergence of $(A,A)(t)$ is the same.*

2.4. **Alignment of present sequences.** We build an estimator of the time since two present sequences issued from the same ancestral sequence diverged. We assume that the ancestral sequence is under the stationary regime. Mimicking subsections 2.1 and 2.2, one could build such an estimator on the observed value of $[A,A]$ but we only detail the estimator based on $[C,C]$.

Let $[C,C]_{\text{obs}}$ denote the observed value of $[C,C]$, that is

$$[C,C]_{\text{obs}} = \frac{1}{N}\sum_{i=1}^{N} \widetilde{K}_i^C(t), \qquad \widetilde{K}_i^C(t) = \mathbf{1}\{X_i^1(t) = X_i^2(t) = C\}.$$

Definition 2.10. *Let $\widetilde{T}_C$ denote the estimator of the divergence time t defined as the solution in t of the equation*

$$[C,C](t) = [C,C]_{\text{obs}}.$$

Let $\widetilde{\kappa}_{\text{obs}}^C$ and $\widetilde{v}_{\text{obs}}^C$ denote observed quantities, defined as

$$\widetilde{\kappa}_{\text{obs}}^C = -8[C,C]_{\text{obs}} + 2r[C*,CG]_{\text{obs}} - 2(C),$$

and

$$\widetilde{v}_{\text{obs}}^C = [C,C]_{\text{obs}} + 2[CC,CC]_{\text{obs}} + 2[C*C,C*C]_{\text{obs}} - [C,C]_{\text{obs}}^2.$$

As in definition 2.1, we pay attention to the fact that $\widetilde{T}_C$ is unique and well defined for any pair of aligned sequences such that

$$(C)^2 < [C,C]_{\text{obs}} < (C).$$

As the observed quantity κ^C_{obs}, $\widetilde{\kappa}^C_{\text{obs}}$ is almost surely positive when N is large.

Theorem 2.11. *Assume that the common ancestral sequence is at stationarity. Then, when $N \to \infty$, $\widetilde{\kappa}^C_{\text{obs}}\sqrt{N/\widetilde{v}^C_{\text{obs}}}(\widetilde{T}_C - t)$ converges in distribution to the standard normal law. An asymptotic confidence interval at level ε for t is*

$$\left[\widetilde{T}_C - \frac{z(\varepsilon)}{\widetilde{\kappa}^C_{\text{obs}}}\sqrt{\frac{\widetilde{v}^C_{\text{obs}}}{N}}, \widetilde{T}_C + \frac{z(\varepsilon)}{\widetilde{\kappa}^C_{\text{obs}}}\sqrt{\frac{\widetilde{v}^C_{\text{obs}}}{N}}\right].$$

Proposition 2.12. *When t is large, the variations of $\widetilde{T}_C$ around t are of order $\mathrm{e}^{8t}/\sqrt{N}$.*

2.5. **Plan.** The rest of the paper is organized as follows. In section 3, we state the central limit theorems which the results in subsections 2.1 to 2.4 are based on. In section 4, we deal with $(C,C)_{\text{obs}}$ and $[C,C]_{\text{obs}}$ and we detail the proofs of central limit theorems for these quantities. In section 5, we show that the central limit theorems established in section 3 imply the theorems and the propositions of section 2. In sections 6 and 7, we characterize the evolutions of $(C,C)(t)$, $(A,A)(t)$ and $[C,C](t)$ and we state some monotony properties. Section 8 contains the proof of theorem 6.1, which rules the evolution of $(C,C)(t)$. Section 9 contains the proof of the monotony properties of $[C,C](t)$ and of some related functions.

3. Central limit theorems for the time estimators

3.1. **Central limit theorems for T_C and T_A.** We explicit the behaviour of T_C and T_A around t. To state our result, we first need central limit theorems for $(C,C)_{\text{obs}}$ and $(A,A)_{\text{obs}}$.

Proposition 3.1. *For $x \in \{C,A\}$, when $N \to +\infty$, $\sqrt{N}((x,x)_{\text{obs}} - (x,x)(t))$, converges in distribution to the centered normal distribution with variance $\sigma^2_x(t)$, where*

$$\sigma^2_x(t) = (x,x)(t) + 2(xx,xx)(t) + 2(x*x,x*x)(t) - 5(x,x)(t)^2.$$

We now deal with the variations of T_C and T_A around t.

For $x \in \{C,A\}$, let μ_x, denote the reciprocal function of $t \mapsto (x,x)(t)$, which is a diffeomorphism from proposition 6.3, and μ'_x. Then,

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

Besides, the derivative of μ_x with respect to t is

$$\mu'_x((x,x)(t)) = \frac{1}{(x,x)'(t)}$$

Using the delta method, see [vdV98], one gets the following result.

Proposition 3.2. *For $x \in \{C,A\}$, when $N \to +\infty$, $\sqrt{N}(T_x - t)$ converges in distribution to the centered normal distribution with variance $\sigma_x^2(t)/(x,x)'(t)^2$.*

3.2. **Central limit theorem for $\widetilde{T}_C$.** As for $(C,C)_{\mathrm{obs}}$, we prove a central limit theorem for $[C,C]_{\mathrm{obs}}$.

Proposition 3.3. *When $N \to \infty$, $\sqrt{N}([C,C]_{\mathrm{obs}} - [C,C](t))$ converges in distribution to the central normal distribution with variance $\widetilde{\sigma}_C^2(t)$, where*

$$\widetilde{\sigma}_C^2(t) = [C,C](t) + 2[CC,CC](t) + 2[C*C,C*C](t) - 5[C,C](t)^2.$$

Let $\widetilde{\mu}_C$ denote the reciprocal function of $t \mapsto [C,C](t)$, which is a diffeomorphism from proposition 7.3, and $\widetilde{\mu}_C'$ its derivative, hence

$$\widetilde{T}_C = \widetilde{\mu}_C([C,C]_{\mathrm{obs}}), \quad t = \widetilde{\mu}_C([C,C](t)) \quad \text{and} \quad \widetilde{\mu}_C'([C,C](t)) = \frac{1}{[C,C]'(t)},$$

for every nonnegative t. The delta method yields the following result.

Proposition 3.4. *When $N \to \infty$, $\sqrt{N}(\widetilde{T}_C - t)$ converges in distribution to the centered normal distribution with variance $\widetilde{\sigma}_C^2(t)/[C,C]'(t)^2$.*

4. Properties of the observed quantities

We detail the properties of $(C,C)_{\mathrm{obs}}$, $(A,A)_{\mathrm{obs}}$ and $[C,C]_{\mathrm{obs}}$. We assume that $N \geqslant 2$.

4.1. **Description of $(C,C)_{\mathrm{obs}}$ and $(A,A)_{\mathrm{obs}}$.**

Lemma 4.1. *For $x \in \{C,A\}$, the mean of $(x,x)_{\mathrm{obs}}$ with respect to π is $(x,x)(t)$. The variance of $(x,x)_{\mathrm{obs}}$ with respect to π is $\sigma_x^2(N,t)$, where*

$$\begin{aligned} N\sigma_x^2(N,t) &= (x,x)(t) - (x,x)(t)^2 + 2(1-1/N)\big((xx,xx)(t) - (x,x)(t)^2\big) + \\ &\quad + 2(1-2/N)\big((x*x,x*x)(t) - (x,x)(t)^2\big). \end{aligned}$$

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

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

The variance of each $K_i^x(t)$ is $\mathrm{var}(K_1^x(t)) = (x,x)(t) - (x,x)(t)^2$. The independence at distance 3 implies that each covariance for $|i-j| \geqslant 3$ is zero. The invariance by translation of π shows that each of the $2(N-1)$ covariances such that $i = j \pm 1$ is

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

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

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

This concludes the proof. □

Corollary 4.2. *For every positive u,*

$$\mathbb{P}_\pi(|(C,C)_{\mathrm{obs}} - (C,C)(t)| > u) \leqslant 15/(16Nu^2),$$

and

$$\mathbb{P}_\pi(|(A,A)_{\mathrm{obs}} - (A,A)(t)| > u) \leqslant 105/(100Nu^2).$$

Proof. Since $(xx,xx)(t) \leqslant (x,x)(t)$ and $(x*x,x*x)(t) \leqslant (x,x)(t)$,

$$N\sigma_x^2(N,t) \leqslant 5(x,x)(t)\big(1-(x,x)(t)\big).$$

Thanks to proposition 6.3, the function $t \mapsto (x,x)(t)$ is decreasing, so $(x,x)(t) \leqslant (x,x)(0)$ for every $t \geqslant 0$. In the stationary regime, $(x,x)(0) = (x)$ and, for every $r \geqslant 0$,

$$(C) = \frac{4+r}{16+5r} \leqslant \frac{1}{4} \quad \text{and} \quad (A) = \frac{4+3/2r}{16+5r} \leqslant \frac{3}{10}.$$

Since $\theta(1-\theta) \leqslant 3/16$ for every $\theta \leqslant 1/4$, and $\theta(1-\theta) \leqslant 21/100$ for every $\theta \leqslant 3/10$, $N\sigma_C^2(N,t) \leqslant 15/16$ and $N\sigma_A^2(N,t) \leqslant 105/100$. Chebychev's inequality concludes the proof. □

4.2. **Proof of proposition 3.1.** To prove the convergence in distribution to the normal law, we use the following result.

Theorem 4.3 (Hall and Heyde [HH80])**.** *Let $(V_i)_{i\in\mathbb{Z}}$ denote a stationary, ergodic, centered, square integrable sequence. Let $\mathscr{F}_0 = \sigma(V_i\,;\, i \leqslant 0)$ denote the σ-algebra generated by the random variables V_i for $i \leqslant 0$. 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 $\sum_{k\geqslant 1} \mathbb{E}(V_k\mathbb{E}(V_n|\mathscr{F}_0))$ converges,*

(ii) *the series $\sum_{k\geqslant K} |\mathbb{E}(V_k\mathbb{E}(V_n|\mathscr{F}_0))|$ converges to zero when $n \to +\infty$, uniformly with respect to K.*

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

For $x \in \{C,A\}$, the sequence $(K_i^x(t))_{i\in\mathbb{Z}}$ is stationary and ergodic. Let $V_i^x = K_i^x(t) - (x,x)(t)$. This defines a sequence $(V_i^x)_{i\in\mathbb{Z}}$ such that the first hypothesis of theorem 4.3 holds. We now check conditions (i) et (ii). The independence at distance 3 implies that, for every $n \geqslant 3$, $\mathbb{E}(V_n^x|\mathscr{F}_0^x) = \mathbb{E}(V_n^x) = 0$. Hence we only have to check the cases $n = 1$ and $n = 2$.

For every $k \geqslant 3$, V_k^x is independent of $\mathscr{F}_0^x$ and $\mathbb{E}(V_n^x|\mathscr{F}_0^x)$ is $\mathscr{F}_0^x$-measurable, hence

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

This implies (i) and (ii), hence theorem 4.3 applies. To compute the asymptotic variance in the theorem, we note that the variance of $\sqrt{N}((x,x)_{\text{obs}} - (x,x)(t))$ is $N\sigma_x^2(N,t)$, which converges to $\sigma_x^2(t)$ when $N \to +\infty$.

4.3. **Case of $[C,C]_{\text{obs}}$.** As for $(C,C)_{\text{obs}}$, we have a description of $[C,C]_{\text{obs}}$.

Lemma 4.4. *The mean of $[C,C]_{\text{obs}}$ with respect to π is $\widetilde{m}_C(t) = [C,C](t)$. The variance of $[C,C]_{\text{obs}}$ with respect to π is $\widetilde{\sigma}_C^2(N,t)$, where*

$$\begin{aligned} N\widetilde{\sigma}_C^2(N,t) &= [C,C](t) - [C,C](t)^2 + 2(1-1/N)\big([CC,CC](t) - [C,C](t)^2\big) + \\ &\quad + 2(1-2/N)\big([C*C,C*C](t) - [C,C](t)^2\big). \end{aligned}$$

The proof of lemma 4.4 is similar to the proof of lemma 4.1. Chebychev's inequality implies the following estimate.

Corollary 4.5. *For every positive u and integer $N \geqslant 2$,*

$$\mathbb{P}_\pi(|[C,C]_{\text{obs}} - \widetilde{m}_C(t)| > u) \leqslant 1/(Nu^2).$$

Proof. The inequalities used in corollary 4.2 hold with $[\cdot,\cdot]$ functions instead of $(\cdot,\cdot)$ functions. □

To prove proposition 3.3, we apply theorem 4.3 to the sequence $(\widetilde{W}_i)_{i\in\mathbb{Z}}$ defined by $\widetilde{W}_i = \widetilde{K}_i^C - \widetilde{m}_C(t)$ for every i.

5. Proofs of the results of section 2

5.1. **Proof of theorems 2.2, 2.5 and 2.11.** Proposition 3.2 yields the variation of 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 $\sigma_C^2(t)$, which both depend on the quantity t to be estimated.

As is customary, Slutsky's lemma allows to bypass this difficulty throught the observed quantities κ_{obs}^x and v_{obs}^x, defined in section 2.

Lemma 5.1. *For $x \in \{C,A\}$, when $N \to +\infty$, $\kappa_{\text{obs}}^x \to -(x,x)'(t)$ and $v_{\text{obs}}^x \to \sigma_x^2(t)$ almost surely.*

Proof. The equalities

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

and

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

given by theorem 6.1, 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 $(A*A,A*A)_{\text{obs}}$ to the corresponding theoretical values, when $N \to +\infty$, imply the desired convergences. □

Proposition 3.1, lemma 5.1 and Slutsky's lemma imply theorems 2.2 and 2.5.

Lemma 5.2. *When $N \to +\infty$, $\widetilde{\kappa}^C_{\text{obs}} \to -[C,C]'(t)$ and $\widetilde{\nu}^C_{\text{obs}} \to \widetilde{\sigma}^2_C(t)$ almost surely.*

Proof. The equality

$$[C,C]'(t) = -8[C,C](t) - 2r[C*,CG](t) + 2(C),$$

given by theorem 7.1, and the almost sure convergence of the observed quantities $[C,C]_{\text{obs}}$, $[C*,CG]_{\text{obs}}$, $[CC,CC]_{\text{obs}}$ and $[C*C,C*C]_{\text{obs}}$ to the corresponding theoretical values, when $N \to +\infty$, imply the desired convergences. □

Proposition 3.4, lemma 5.2 and Slutsky's lemma imply theorem 2.11.

5.2. **Proofs of propositions 2.7, 2.8 and 2.9.** Firstly, from corollary 6.2, for every value of r, the convergence of $(C,C)(t)$ and $(A,A)(t)$ to equilibrium when $t \to +\infty$ in the Jukes-Cantor model with CpG influence are like e^{-4t}.

Secondly, in Jukes-Cantor models with CpG influence, every nucleotide changes at rate 3 due to unconditional substitution rates, plus every dinucleotide CpG changes at rate $2r$. Hence the global rate of substitution is

$$3 + 2r(CG) = 3 + \frac{2r}{16+5r}.$$

On the other hand, in the independent Jukes-Cantor model of parameter λ, the global rate of substitution is 3λ. Hence one should set

$$\lambda = 1 + \frac{2r/3}{16+5r}.$$

For independent Jukes-Cantor models, [Yan06] computes $(C,C)(t) = (A,A)(t) = \frac{1}{16} + \frac{3}{16}\mathrm{e}^{-4\lambda t}$. Since $\lambda > 1$ for every $r > 0$, the comparison with the independent Jukes-Cantor model is done.

Thirdly, in Jukes-Cantor models with CpG influence, every transversion occurs at rate 1, and every nucleotide may have two possible transversions. Hence the global rate of transversion is 2. The transitions $G \to A$ and $C \to G$ occur at rate $1 + r(CG)$, the transitions $A \to G$ and $G \to C$ occur at rate 1. Hence the global rate of transition is $1 + 2r(CG) = 1 + 2r/(16+5r)$. On the other hand, in Kimura models, the global rate of transition is α and the global rate of transversion is 2β. Hence one should set

$$\alpha = 1 + \frac{2r}{16+5r} \quad \text{and} \quad \beta = 1.$$

For independent Kimura models, [Yan06] computes

$$(C,C)(t) = (A,A)(t) = \frac{1}{16} + \frac{1}{16}\mathrm{e}^{-4\beta t} + \frac{1}{8}\mathrm{e}^{-(2\alpha+2\beta)t},$$

so the convergence of $(C,C)(t)$ and $(A,A)(t)$ to equilibrium when $t \to +\infty$ in Kimura models is like $\mathrm{e}^{-4\beta t}$, when one assumes $\alpha > \beta$. Since $\beta = 1$, the comparison with the independent Kimura model is done.

Fourthly, simple computations yield in the independent model with the same overall rate of substitutions

$$(C,C)(t) = \frac{1}{(4+\delta)^2}\left(1+(3+\delta)\mathrm{e}^{-(4+\delta)t}\right),$$

and

$$(A,A)(t) = \frac{2+\delta}{(8+2\delta)^2}\left(2+\delta+(4+\delta)\mathrm{e}^{-4t}+2\mathrm{e}^{-(4+\delta)t}\right).$$

thus the convergence of $(C,C)(t)$, respectivly $(A,A)(t)$ to equilibrium when $t \to +\infty$ in the model above is like $\mathrm{e}^{-(4+\delta)t}$, respectivly e^{-4t}. Since $4+\delta > 4$ for every $r > 0$, the comparison with this model is complete.

6. Evolutions of $(C,C)(t)$ and $(A,A)(t)$

We provide a linear differential system which rules the evolution of $(C,C)(t)$ and $(A,A)(t)$. Introduce the constant matrices

$$M = \begin{pmatrix} -4 & -r & 0 \\ 1 & -(8+2r) & 1 \\ 0 & -r & -4 \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \text{and} \quad L = (0,1,0),$$

and the time-dependent vectors

$$U_C(t) = \begin{pmatrix} (C,C)(t) \\ (C*,CG)(t) \\ (C*,*G)(t) \end{pmatrix} \quad \text{and} \quad V_C(t) = \begin{pmatrix} (A,G)(t) \\ (*A,CG)(t) \\ (*A,C*)(t) \end{pmatrix}.$$

The proof of theorem6.1 is in section 8.

Theorem 6.1. *The evolutions of $U_C(t)$ and $V_C(t)$ are ruled by the linear differential system*

$$U_C'(t) = MU_C(t) + (C)(0)B, \quad V_C'(t) = MV_C(t) + (A)(0)B.$$

The evolution of $(A,A)(t)$ can be deduced from the one of $V_C(t)$ through the differential equation

$$(A,A)'(t) = -4(A,A)(t) + rLV_C(t) + (A)(0).$$

The initial conditions are

$$U_C(0) = \begin{pmatrix} (C)(0) \\ (CG)(0) \\ (CG)(0) \end{pmatrix}, \qquad V_C(0) = \begin{pmatrix} 0 \\ 0 \\ (CA)(0) \end{pmatrix} \quad \textit{and} \quad (A,A)(0) = (A)(0).$$

Theorem 6.1 is valid out of equilibrium. We use it at stationarity hence, in particular, for the initial values

$$(C)(0) = (C) = \frac{4+r}{16+5r}, \qquad (A)(0) = (A) = \frac{4+3r/2}{16+5r},$$
$$(CG)(0) = (CG) = \frac{1}{16+5r}, \qquad (CA)(0) = (CA) = \frac{1+7r/16}{16+5r}.$$

Solving the system in theorem 6.1 yields expressions of $(C,C)(t)$ and $(A,A)(t)$. Consider the positive real numbers u, u_+ and u_- defined as

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

Corollary 6.2. *In the stationary regime,*

$$\begin{aligned}(C,C)(t) &= c_0 e^{-4t} + c_+ e^{-u_+ t} + c_- e^{-u_- t} + (C)^2,\\ (A,A)(t) &= a_0 e^{-4t} + a_+ e^{-u_+ t} + a_- e^{-u_- t} + (A)^2,\end{aligned}$$

with

$$\begin{aligned}c_0 &= \frac{3+r}{2(16+5r)}, \qquad a_0 = \frac{80+31r}{32(16+5r)},\\ c_\pm &= \frac{3+r}{4u(16+5r)^2}\left(u(16+3r) \pm (32+14r+3r^2)\right),\\ a_\pm &= \frac{512+384r+106r^2+13r^3 \pm u(256+18r+13r^2)}{64u(16+5r)^2}.\end{aligned}$$

As expected,

$$c_+ + c_- + c_0 = (C) - (C)^2 \quad \text{and} \quad a_+ + a_- + a_0 = (A) - (A)^2.$$

Furthermore, for every positive r,

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

For every positive r, the parameters $c_\pm$, c_0, $a_\pm$ and a_0 are positive. This proves the following proposition.

Proposition 6.3. *The functions $t \mapsto (C,C)(t)$ and $t \mapsto (A,A)(t)$ are decreasing diffeomorphisms.*

7. Evolution of $[C,C](t)$

We wish to compute a linear differential system which would rule the evolution of $t \mapsto [C,C](t)$. To this end, we use repeatedly the fact that for every sets W and W' of words and every time t, $[W,W'](t) = [W',W](t)$, since the evolution of the two sequences from their common ancestor is exchangeable.

Additionally, we assume that the common ancestral sequence is at stationarity. A consequence is the symmetry of C and G in the Jukes-Cantor model with CpG influence. Thus, for every positive t,

$$[G,G](t) = [C,C](t) \quad \text{and} \quad [*G,CG](t) = [C*,CG](t).$$

Introduce the constant matrices

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

and the time-dependent vector

$$\widetilde{U}_C(t) = \begin{pmatrix} [C,C](t) \\ [C*,CG](t) \\ [CG,CG](t) \\ [C*,*G](t) \end{pmatrix}.$$

Theorem 7.1. *The evolution of $\widetilde{U}_C(t)$ is ruled by the linear differential system*

$$\widetilde{U}_C'(t) = \widetilde{M}\widetilde{U}_C(t) + \widetilde{B}.$$

Proof. As for theorem 6.1, one compares easily $[C,C](t)$ to $[C,C](t+s)$ up to the order s, for every positive t and vanishingly small positive s. □

Solving this system yields expressions of $[C,C](t)$, $[C*,CG](t)$, $[CG,CG](t)$ and $[C*,*G](t)$.

Corollary 7.2. *In the stationary regime,*

$$\begin{aligned} [C,C](t) &= c_0 e^{-8t} + c_- e^{-2u_+ t} + c_+ e^{-2u_- t} + (C)^2, \\ [C*,CG](t) &= c'_- e^{-2u_+ t} + c'_+ e^{-2u_- t} + (C)(CG), \\ [CG,CG](t) &= c''_- e^{-2u_+ t} + c''_+ e^{-2u_- t} + (CG)^2, \\ [C*,*G](t) &= -c_0 e^{-8t} + c_- e^{-2u_+ t} + c_+ e^{-2u_- t} + (C)(G), \end{aligned}$$

where the constants c_0 and $c_\pm$ are defined in section 6, and

$$\begin{aligned} c'_\pm &= c_\pm (2 + r \mp u), \\ c''_\pm &= (2c_\pm / r^2)\left(4 + 3r + r^2 \mp u(2+r)\right). \end{aligned}$$

Proposition 7.3. *The functions $t \mapsto [C,C](t)$, $t \mapsto [C*,CG](t)$ and $t \mapsto [CG,CG](t)$ are decreasing diffeomorphisms. The function $t \mapsto [C*,*G](t)$ is an increasing diffeormorphism.*

The proof of proposition 7.3 is in section 9.

8. Proof of theorem 6.1

To compute $(C,C)'(t)$, for example, one must compare $(C,C)(t)$ to $(C,C)(t+s)$ up to the order s, for every positive t and vanishingly small positive s. The probability that at least two substitutions occur at the same site between times t and $t+s$ is $o(s)$, hence these events do not appear in the limit we consider.

Let $(W'|W)(s)$ denote the probability that sites occupied by a word of W at time 0 are occupied by a word of W' at time s. For every letter x in $\mathscr{A}$, recall that $\overline{x} = \mathscr{A} \setminus \{x\}$.

For every x in $\overline{C}$, $(C|x)(s) = s + o(s)$, hence $(C|\overline{C})(s) = s + o(s)$. Two other cases arise, namely

$$(C*|CG)(s) = 1 - (3+r)s + o(s), \quad (C*|C\overline{G})(s) = 1 - 3s + o(s).$$

We are now ready to evaluate $(C,C)(t+s)$. Decomposing along the values at time t, one gets

$$(C,C)(t+s) = (C*,CG,C*)(t,s) + (C*,C\bar{G},C*)(t,s) + (C,\bar{C},C)(t,s).$$

The first contribution is

$$(C*,CG,C*)(t,s) = (C*,CG)(t)(C*|CG)(s) = (C*,CG)(t)(1-(3+r)s) + o(s).$$

Likewise, the second contribution is

$$(C*,C\bar{G},C*)(t,s) = (C*,C\bar{G})(t)(C*|C\bar{G})(s) = (C*,C\bar{G})(t)(1-3s) + o(s).$$

Finally,

$$(C,\bar{C},C)(t,s) = (C,\bar{C})(t)(C|\bar{C})(s) = (C,\bar{C})(t)s + o(s).$$

The sum of the contributions of order 1 is $(C,C)(t)$. The sum of the contributions of order s yields the derivative, hence

$$(C,C)'(t) = -(3+r)(C*,CG)(t) - 3(C*,C\bar{G})(t) + (C,\bar{C})(t).$$

Using the relations

$$(C,C)(t) = (C)(0) - (C,\bar{C})(t) = (C*,CG)(t) + (C*,C\bar{G})(t),$$

one gets that $(C,C)'(t)$ is the sum of $(C)(0)$ and of the first coordinate of $MU_C(t)$, as stated in theorem 6.1.

The same method applies to $(C*,CG)$ and to $(C*,*G)$. For instance,

$$\begin{aligned}(C*,CG)(t+s) &= (C*,CG,CG)(t,s) + (C*,C\bar{G},CG)(t,s) + \\ &\quad + (C*,\bar{C}G,CG)(t,s) + (C*,\bar{C}\bar{G},CG)(t,s).\end{aligned}$$

The last term is $o(s)$ since one asks that at least two substitutions occur between times t and $t+s$. The three other terms can be factored as

$$(C*,xy,CG)(t,s) = (C*,xy)(t)(CG|xy)(s).$$

Plugging into this the expansions

$$(CG|CG)(s) = 1-(6+2r)s + o(s),$$

and

$$(CG|C\bar{G})(s) = (CG|\bar{C}G)(s) = s + o(s),$$

and regrouping the first order terms yields

$$(C*,CG)'(t) = -(6+2r)(C*,CG)(t) + (C*,C\bar{G})(t) + (C*,\bar{C}G)(t).$$

Using the relations

$$(C*,C\bar{G})(t) = (C,C)(t) - (C*,CG)(t),$$

and

$$(C*,\bar{C}G)(t) = (C*,*G)(t) - (C*,CG)(t),$$

yields $(C*,CG)'(t)$ as the second coordinate of $MU_C(t)$, as stated in theorem 6.1.

As regards the evolution of $(C*,*G)$, using once again the relation

$$(C*,*G) = (C*,CG) + (C*,\bar{C}G),$$

one is left with the evolution of $(C*,\bar{C}G)$. Decomposing as before, one gets

$$\begin{aligned}(C*,\bar{C}G)(t+s) &= (C*,\bar{C}G,\bar{C}G)(t,s)+(C*,CG,\bar{C}G)(t,s)+\\ &\quad +(C*,\bar{C}\bar{G},\bar{C}G)(t,s)+(C*,C\bar{G},\bar{C}G)(t,s).\end{aligned}$$

The last term is $o(s)$ since one asks that at least two substitutions occur between times t and $t+s$. The three other terms can be factored like before. Using the expansions

$$(\bar{C}G|\bar{C}G)(s)=1-4s+o(s),\quad (\bar{C}G|CG)(s)=(3+r)s+o(s),$$

and

$$(\bar{C}G|\bar{C}\bar{G})(s)=s+o(s),$$

and regrouping the first order terms yields

$$(C*,\bar{C}G)'(t)=-4(C*,\bar{C}G)(t)+(3+r)(C*,CG)(t)+(C*,\bar{C}\bar{G})(t).$$

Coming back to $(C*,*G)$ yields $(C*,*G)'(t)$ as the sum of $(C)(0)$ and of the third coordinate of $MU_C(t)$, as stated in theorem 6.1.

9. Proof of proposition 7.3

The function $t\mapsto[C,C](t)$ is decreasing. Since the parameters c_0 and $c_\pm$ are positive, this is direct.

The function $t\mapsto[C*,CG](t)$ is decreasing. Simple computations yield

$$(16+5r)u\mathrm{e}^{2u_+t}\,[C*,CG]'(t)=-2u(r+3)-(u+2-r)(r+3)\big(\mathrm{e}^{4ut}-1\big).$$

Since $u+-r2>0$, the right hand side is a sum of negative terms, hence $[C*,CG]'(t)$ is negative for every nonnegative t.

The function $t\mapsto[CG,CG](t)$ is decreasing. As for $[C*,CG]$, one computes

$$(16+5r)u\mathrm{e}^{2u_+t}[CG,CG]'(t)=-4u(3+r)-2(3+r)(u-r-1)(\mathrm{e}^{4ut}-1),$$

and the fact that $u-r-1>0$ concludes the proof.

The function $t\mapsto[C*,*G](t)$ is increasing. We begin with the differential equation

$$[C*,*G]'(t)=-2r[C*,CG](t)-8[C*,*G](t)+2(C).$$

This yields

$$[C*,*G]''(t)=-2r[C*,CG]'(t)-8[C*,*G]'(t).$$

Hence,

$$\mathrm{e}^{8t}[C*,*G]'(t)=[C*,*G]'(0)-2r\int_0^t\mathrm{e}^{8s}[C*,CG]'(s)\mathrm{d}s.$$

One now computes $[C*,*G]'(0)$. Using our first equation, one gets

$$\begin{aligned}[C*,*G]'(0)&=-2r[C*,CG](0)-8[C*,*G](0)+2(C)\\ &=-2r(CG)-8(CG)+2(C)=2(C)-2(r+4)(CG)=0.\end{aligned}$$

Hence,

$$\mathrm{e}^{8t}[C*,*G]'(t)=-2r\int_0^t\mathrm{e}^{8s}[C*,CG]'(s)\mathrm{d}s.$$

Since the function $t \mapsto [C*, CG](t)$ is decreasing, the last integral above is negative and this yield the result.

REFERENCES

[ABH03] Peter F. Arndt, Christopher B. Burge, and Terence Hwa, *DNA sequence evolution with neighbor-dependent mutation*, Journal of Computational Biology **10** (2003), 313–322.

[AH05] Peter F. Arndt and Terence Hwa, *Identification and measurement of neighbor dependent nucleotide substitution processes*, Bioinformatics **21** (2005), 2322–2328.

[BGP08] J. Bérard, J.-B. Gouéré, and D. Piau, *Solvable models of neighbor-dependent nucleotide substitution processes*, Mathematical Biosciences **211** (2008), 56–88.

[DG00] 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** (2000), 1620–1625.

[Fel81] J. Felsenstein, *Evolutionary trees from DNA sequences : A maximum likelihood approach*, J. Mol. Evol. **17** (1981), 368–376.

[GGG96] N. Galtier, M. Gouy, and C. Gautier, *Seaview and phylo_win, two graphic tools for sequence alignment and molecular phylogeny*, Comput. Applic. Biosci. **12** (1996), 543–548.

[HH80] P. Hall and C. C. Heyde, *Martingale limit theory and its applications*, Academic Press, New York, 1980.

[HKY85] M. Hasegawa, H. Kishino, and T. Yano, *Dating of the human-ape splitting by a molecular clock of mitochondrial DNA*, J. Mol. Evol. **22** (1985), 160–174.

[JC69] T.H. Jukes and C.R. Cantor, *Mammalian protein metabolism*, ch. Evolution of Protein Molecules, pp. 21–132, Academic Press, New York, 1969.

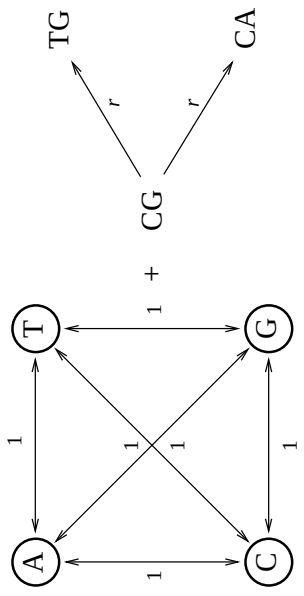
[JTT92] D. Jones, W. Taylor, and J. Thornton, *The rapid generation of mutation data matrices from protein sequences*, Comput. Appl. Biosci. **8** (1992), 275–282.

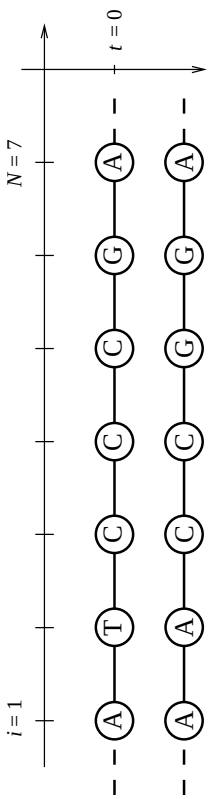
[Kim80] M. Kimura, *A Simple Method for Estimating Evolutionary Rates of Base Substitutions Through Comparative Studies of Nucleotide Sequences*, J. Mol. Evol. **10** (1980), 111–120.

[vdV98] A. W. van der Vaart, *Asymptotic Statistics*, Cambridge University Press, 1998.

[Yan06] Z. Yang, *Computational Molecular Evolution*, Oxford Series in Ecology and Evolution, 2006.

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


TG
CA
r
r
CG
+
T
G
A
C
1
1
1
1
1
1


i = 1
N = 7
t = 0
A
T
C
C
C
G
G
A
A
C
C
G
G
A