SOLVABLE MODELS OF NEIGHBOR-DEPENDENT SUBSTITUTION PROCESSES

JEAN BÉRARD, JEAN-BAPTISTE GOUÉRÉ, DIDIER PIAU

ABSTRACT. We prove that a wide class of Markov models of neighbor-dependent substitution processes on the integer line is solvable. This class contains some models of nucleotidic substitutions recently introduced and studied empirically by molecular biologists. We show that the polynucleotidic frequencies at equilibrium solve some finite-size linear systems. This provides, for the first time up to our knowledge, explicit and algebraic formulas for the stationary frequencies of non degenerate neighbor-dependent models of DNA substitutions. Furthermore, we show that the dynamics of these stochastic processes and their distribution at equilibrium exhibit some stringent, rather unexpected, independence properties. For example, nucleotidic sites at distance at least three evolve independently, and all the sites, when only encoded as purines and pyrimidines, evolve independently.

INTRODUCTION

In most models of nucleotidic substitution processes, one assumes that each site along the DNA sequence evolves independently of the others, according to some specified rates of substitution. To be more precise, we introduce at the onset some notations, common for the biologist and useful to describe conveniently these models as well as, later on, the more sophisticated ones which are the subject of this paper. For these definitions and for other nomenclatures used in this paper, see [5].

Definition 1. *The nucleotidic alphabet is* $\mathcal{A} := \{A, T, C, G\}$. *These letters stand for Adenine, Thymine, Cytosine and Guanine, respectively. Adenine and guanine are purines, often abbreviated by the letter R, and cytosine and thymine are pyrimidines, often abbreviated by the letter Y. Substitutions of the form* $R \to R$ *and* $Y \to Y$ *are called transitions, substitutions of the form* $R \to Y$ *and* $Y \to R$ *are called transversions. Finally, for every subsets X and Z of* $\mathcal{A}$, *XpZ is the collection of dinucleotides in* $\mathrm{X} \times \mathrm{Z}$.

For instance, YpR dinucleotides are formed by a purine followed by a pyrimidine, hence there are four of them, which are CpG, TpA, TpG, and CpA.

Experimental facts are that, in many cases, transitions are more frequent than transversions (typical ratios 3:1), and that substitutions to C and to G occur at different rates than substitutions to A and to T, see Duret and Galtier [6] and

1991 *Mathematics Subject Classification.* 60J25,92D20.

Key words and phrases. Markov processes, Poisson processes, Genetics, DNA sequences, CpG deficiency.

the references therein. Some models take these facts into account, for instance, in Tamura's model, one assumes that the matrix of substitution rates is

$$(1) \qquad \begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{pmatrix} \cdot & v_2 & v_1 & \kappa v_1 \\ v_2 & \cdot & \kappa v_1 & v_1 \\ v_2 & \kappa v_2 & \cdot & v_1 \\ \kappa v_2 & v_2 & v_1 & \cdot \end{pmatrix} \end{array},$$

where v_1, v_2 and κ are nonnegative real numbers. The matrix notation means, for instance, that each A is replaced by T at rate v_2, by C at rate v_1, and by G at rate κv_1.

In this model and in related ones with no interaction between the sites, the probability distribution of the nucleotide attached to any given site converges, for large times, to the stationary measure of the Markov chain described by the matrix of the rates and, at equilibrium, the sites are independent. This last consequence is unfortunate in a biological context, since the frequencies $F(x)$ of the nucleotides and the frequencies $F(xy)$ of the dinucleotides, observed in actual sequences, are often such that, for many nucleotides x and y,

$$F(xy) \neq F(x)\,F(y).$$

In fact, it is well known that the nucleotides in the immediate neighborhood of a site can affect drastically the substitution rates at this site. For instance, in the genomes of vertebrates, the increased rates of substitution of cytosine by thymine and of guanine by adenine in CpG dinucleotides are often quite noticeable (typical ratios 10:1 when compared to the other rates of substitution). The chemical reasons of this so-called CpG-methylation-deamination process are well known and one can guess that, at equilibrium, the number of CpG is decreased while the number of TpG and CpA is increased when one adds high rates of CpG substitutions to Tamura's model, for example.

The need to incorporate these effects into more realistic models of nucleotidic substitutions seems to be widely acknowledged. However, the exact consequences of the introduction of such neighbor-dependent substitution processes (in the case above, $\mathrm{CpG} \to \mathrm{CpA}$ and $\mathrm{CpG} \to \mathrm{TpG}$), while crucial for a quantitative assessment of these models, remain virtually unknown, at least up to our knowledge and on a theoretical ground. 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 to account for the $\mathrm{CpG} \to \mathrm{CpA}$ substitutions, and at sites i and $i+1$ to account for the $\mathrm{CpG} \to \mathrm{TpG}$ substitutions, whose joint distribution, in turn, depends a priori on the values of some trinucleotides, and so on.

Duret and Galtier [6] introduced and analyzed a class of models, which we call Tamura+CpG, that adds to Tamura's rates of substitution the availability of substitutions $\mathrm{CpG} \to \mathrm{CpA}$ and $\mathrm{CpG} \to \mathrm{TpG}$, both at the additional rate $r \geqslant 0$. (The authors use a parameter $\kappa_1 \geqslant \kappa$, related to our parameter r and to the ratio $\theta := v_1/(v_1+v_2)$ defined in [6] by the simple equation $r = (1-\theta)\,(\kappa_1 - \kappa)$.) To evade the curse, explained above, of recursive calls to the frequencies of longer and longer

words, Duret and Galtier use as approximate frequencies $F(xyz)$ of trinucleotides the values

$$F(xyz) \approx F(xy)\,F(yz)/F(y).$$

Interestingly, these approximations would be exact if the sequences at equilibrium followed a Markov model with respect to the space index i. This is not the case but, relying on computer simulations, Duret and Galtier study the G+C content at equilibrium for this model and other quantities of interest in a biological context, they compare the simulated values to the values predicted using their truncation procedure, and they show that their approximation captures some features of the behavior of the true model. In particular, they highlight that these models have inherent, and previously unexpected, consequences on the frequency of the TpA dinucleotide as well. We mention that Arndt and his coauthors consider similar models and their biological implications, see Arndt [1], Arndt, Burge and Hwa [2] and Arndt and Hwa [3] for instance.

In this paper we introduce a wide extension of the Tamura+CpG model of neighbor-dependent substitution processes, which we call RN+YpR models (RN stands for Rzhetsky and Nei, see below), and we show that these models are solvable. More precisely, we prove that the frequencies of polynucleotides at equilibrium solve explicit finite-size linear systems. Thus, the infinite regressions to the frequencies of longer and longer polynucleotides described above disappear, and one can compute analytically several quantities of interest related to these models. For instance, one can assess rigorously the effect of neighbor-dependent substitutions. As noted above, the very possibility of such a solution comes as a surprise and, at least up to our knowledge, our formulas are the only ones of their kind for neighbor-dependent models of evolution. Additionally, our analysis provides some stringent independence properties of these models at equilibrium. Finally, we mention two connected, more ambitious, questions, which we plan to study elsewhere: first, one can develop a perturbative analysis of non RN+YpR models which are close to a model in this class; second, for every RN+YpR model at equilibrium, one can estimate the evolution distance which separates an ancestor sequence from a sequence derived from it, this estimation being the first step towards a rigorous construction of phylogenies based on neighbor-dependent evolution models.

In section 1 we describe the models under study, in section 2 we state our main results, in section 3 we give an overview of the rest of the paper.

Acknowledgements. We thank the biologists who contributed to this work, in particular Laurent Duret, Nicolas Galtier, Manolo Gouy and Jean Lobry. When the need arose, they willingly provided facts and references and they shared with us their numerous insights about the subject of this study. However, they are not responsible for any remaining misconception in this paper, on biological matters or otherwise.

1. Description of the models

1.1. **RN models.**

Definition 2 (RN models)**.** *A matrix of substitution rates is RN if and only if there exists nonnegative rates v_x and w_x such that the matrix is*

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

This means, for instance, that each nucleotide A is replaced by T at rate v_T, by C at rate v_C, and by G at rate w_G. The rates v_x and w_x are indexed by the nucleotide x that the corresponding substitution creates.

The following characterization is direct.

Proposition 1. *A matrix of substitution rates is RN if and only if one can complete it by diagonal elements such that, on the one hand, the joint distributions of the elapsed times before a substitution occurs and of the nucleotide that this substitution yields coincide for the two purines, and, on the other hand, the joint distribution of the elapsed times before a substitution occurs and of the nucleotide that this substitution yields coincide for the two pyrimidines.*
In other words, for every x and y that are not both purines nor both pyrimidines, the rate of transversion of x to y depends only on y.

1.2. **Situation of the RN models.** Rzhetsky and Nei [12] introduced the class of matrices of substitution rates delineated by proposition 1 and called them eight-parameter models. We use the initials RN of these authors as a shorthand for this class. As mentioned by Rzhetsky and Nei (see table 1 on page 132 of [12]), many well known models belong to this class. For instance, Tamura-Nei's model (usually abbreviated as TN93) corresponds to the restriction of RN such that

$$w_A\, v_G = w_G\, v_A, \qquad w_T\, v_C = w_C\, v_T.$$

Hence special cases of TN93 are also special cases of RN. Furthermore, Felsenstein's model (F84) corresponds to the restriction of TN93 such that

$$w_A + v_T + v_C + w_G = v_A + w_T + w_C + v_G.$$

Hasegawa-Kishino-Yano's model (HKY85) corresponds to the restriction of RN such that w_x/v_x does not depend on x. In other words, both F84 and HKY85 are subcases of TN93, which is a subcase of RN.

Kimura's model with two parameters (K2P, also known as K80) corresponds to the restriction of RN such that v_x and w_x do not depend on x. Jukes-Cantor's model (JC69) corresponds to the restriction of RN such that $v_x = w_x$ and v_x does not depend on x. Finally, Tamura's model, as introduced by Duret and Galtier in [6], is intermediate between HKY85 and K2P since it corresponds to the restriction of RN such that $v_A = v_T$, $v_C = v_G$, and w_x/v_x does not depend on x.

We summarize all this as a proposition.

Proposition 2. *The model JC69 is a strict subcase of K80, which is a strict subcase of both HKY85 and F84, which are both strict subcases of TN93. All these models as well as Tamura's model in the sense above, are strict subcases of RN.*

The general time-reversible model (GTR) is not comparable with RN, in other words some matrices of rates of substitutions are GTR but not RN, and vice versa. The intersection of the GTR and RN classes is TN93. As a consequence, the complement of TN93 in RN contains only non-reversible models.

Finally, we mention a rather odd, at least to us, relationship between the RN condition and the so-called balanced-transversion condition introduced by Lake [7] in the context of evolutionary parsimony methods, see also Navidi and Beckett-Lemus [9] for an assessment of these methods when the balanced-transversion condition is violated. Lake considers discrete time processes ruled by matrices of substitution probabilities such that, for every nucleotide x, the probabilities of both transversions from x are equal. In RN models, the continuous time processes are ruled by matrices of substitution rates such that, for every nucleotide x, the probabilities of both transversions which produce x are equal.

1.3. **YpR substitutions.** We generalize the CpG mechanism of substitutions considered by Duret and Galtier [6], adding 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.

The rationale for these notations is that the rate r_x^y is indexed by the nucleotide x produced by the substitution, and by the nucleotide y completing the YpR dinucleotide xy that the substitution yields, when x is a pyrimidine, or completing the YpR dinucleotide yx that the substitution yields, when x is a purine. In other words, y is the nucleotide not affected by the substitution.

Definition 3. *RN+YpR models correspond to the superposition of rates of substitutions of an RN model and of rates of substitution r_x^y of dinucleotides YpR, as described above.*

To recover the model of Duret and Galtier, one should assume that

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

RN+YpR models use 16 mutation rates, namely, 4 parameters v_x for the transversions, 4 parameters w_x for the transitions, and 8 parameters r_x^y for the mutations involving YpR dinucleotides. To multiply these 16 parameters by the same scalar changes the time scale, but not the evolution itself nor the stationary distributions, hence the class RN+YpR is a simplex of dimension 15. We allow for possibly negative values of the rates r_x^y since the model makes sense as long as the following inequalities are satisfied for every x and y:

$$v_x \geqslant 0, \quad w_x \geqslant 0, \quad w_x + r_x^y \geqslant 0.$$

Finally, note that the rates r_x^y describe transitions, and never transversions.

1.4. **Sets of sites.** DNA sequences are represented by sequences of letters of the alphabet $\mathcal{A}$. Although real DNA sequences are obviously finite, their typical length is rather large, hence one often considers them as doubly infinite sequences of letters, that is, as elements of the set $\mathcal{A}^{\mathbb{Z}}$. Then sites on the DNA sequence are identified to integer numbers in $\mathbb{Z}$. Our mathematical results are most conveniently expressed in this setting but, as we explain below, most of them are in fact also valid for suitable finite sets of sites.

2. Description of the main results

Roughly speaking, our main findings are that, for every RN+YpR model, one can compute the exact value at equilibrium of the frequency of every polynucleotide since these frequencies solve explicit finite-size linear systems; that one can simulate exact samples of sequences of nucleotides at equilibrium; that some surprising independence properties between sites hold at equilibrium; and that, furthermore, the dynamics converges to the unique equilibrium, for every starting distribution.

To be more specific, we prove the following results. We consider an RN+YpR model.

Theorem 1 (Construction and general properties)**.** *There exists a unique Markov process on $\mathcal{A}^{\mathbb{Z}}$, denoted by $(X(s))_{s\geqslant 0}$, associated to the transition rates defined in section 1. Under a generic non-degeneracy condition (ND), stated in section 4, this process is ergodic, that is, it has a unique stationary distribution μ and, for any initial distribution, $X(s)$ converges to μ in distribution as s goes to infinity. Moreover, μ is invariant and ergodic with respect to the translations of $\mathbb{Z}$.*

From theorem 1, equilibrium properties of the model are well-defined, for instance the equilibrium frequency of polynucleotides.

Theorem 2 (Dynamics)**.** *There exists an i.i.d. sequence of marked Poisson point processes $(\xi_i)_i$ on the real line, indexed by the sites i, and a measurable function Φ with values in $\mathcal{A}$ such that, for every couple of times $s \leqslant t$ and every site i, the value $X_i(t)$ of site i at time t is*

$$X_i(t) = \Phi(X_{i-1}(s), X_i(s), X_{i+1}(s); \xi_{i-1} \cap [s,t], \xi_i \cap [s,t], \xi_{i+1} \cap [s,t]).$$

Theorem 3 (Statics)**.** *Assume that the distribution of $(X_i)_i$ is stationary. Then there exists an i.i.d. sequence of marked Poisson point processes $(\xi_i)_i$ on the real halfline, indexed by the sites i, and a measurable function Ψ with values in $\mathcal{A}$ such that, for every site i,*

$$X_i = \Psi(\xi_{i-1}, \xi_i, \xi_{i+1}).$$

Theorems 2 and 3 describe some structural properties, at the basis of our subsequent results. We begin with some direct consequences.

Definition 4. *For every subset J of $\mathbb{Z}$, let $*J*$ denote the set of integers i such that either $i-1$ or i or $i+1$ belongs to J.*

Proposition 3 (Dynamics)**.** *Consider sequences indexed by I. The restriction of the dynamics to a subset of sites J does not depend on I, as soon as I contains $*J*$.*

Proposition 3 shows for instance that the behavior of the sites in $\{1,\dots,n\}$ is the same, whether one considers that these sites are embedded in a sequence indexed by $\mathbb{Z}$, or in a sequence indexed by $\{0,1,\dots,n+1\}$. If the sites are embedded in $\{0,1,\dots,n+1\}$ or in a larger finite set, this means that the boundary conditions have no effect on the behavior of the sites in $\{1,\dots,n\}$, hence one can consider at will discrete intervals or discrete circles. For instance, one can decide, either that the only neighbor of 0 is 1 and the only neighbor of $n+1$ is n, or that 0 has neighbors 1 and $n+1$ and that $n+1$ has neighbors n and 0. This decision will modify the evolution at the sites 0 and $n+1$ but not at the sites in $\{1,\dots,n\}$. This remark concerns all the results that we state later on in this section.

We come back to the consequences of theorems 2 and 3.

Corollary 1 (Statics)**.** *Assume that the distribution of $(X_i)_i$ is stationary. Fix some sets of sites I_k such that the sets $*I_k*$ are disjoint. Then the families $(X_i)_{i\in I_k}$ are independent from each other.*

The condition in the corollary means that, for every $k\neq k'$, $|i-i'|\geqslant 3$ for every i in I_k and every i' in $I_{k'}$. For instance, the sequence $(X_{3i})_{i\in\mathbb{Z}}$ is i.i.d. at equilibrium.

Our next results deal with what is probably the main concern of biologists in relation to this model, namely, the actual computation of equilibrium frequencies.

Theorem 4. *The equilibrium frequencies of polynucleotides solve explicit finite-size linear systems and can be expressed as rational functions of the parameters of the model.*

Theorem 5 (Nucleotides and YpR dinucleotides)**.** *At equilibrium, the frequencies of the nucleotides are explicit affine functions of the frequencies of the YpR dinucleotides. Furthermore, the equilibrium frequencies of the YpR dinucleotides solve an explicit 4×4 linear system.*

We state theorem 5 more precisely as theorem 7 in section 8.2. Our last result is a consequence of the inner properties of our basic construction.

Theorem 6 (R/Y sequences)**.** *Encode the sequence of nucleotides as an R/Y sequence of purines and pyrimidines. If the sequence of nucleotides is at equilibrium, then one obtains an i.i.d. R/Y sequence described by the weights t_R and $t_Y=1-t_R$, with*

$$t_Y:=\frac{v_C+v_T}{v_A+v_T+v_C+v_G},\qquad t_R:=\frac{v_A+v_G}{v_A+v_T+v_C+v_G}.$$

In particular, the values of t_Y and t_R do not depend on the values of the YpR substitution rates. This fact reflects, once again, the specific property of RN+YpR substitution models that is at the basis of our analysis, namely, the fact that the global model is equivalent to the superposition of the double substitution processes described by the rates r_x^y on top of the simple substitution processes described by the rates v_x and w_x, see section 4. This remark shows that one cannot compute analytically, at least along these principles, the stationary measures of substitution models that are not in the RN+YpR class, and in fact, we suspect that there exists no closed form of these.

3. Description of the paper

Here is a moderately detailed description of the content and organization of the rest of the paper.

Part A is devoted to a rigorous construction of the processes described informally in section 1, based on Poisson processes. We stress at the onset that one could rely on the general principles in Liggett [8, chapter 1], based on infinitesimal generators, to build Markov processes on the space of finite or infinite nucleotidic sequences that correspond to the jump rates defined above. However, the specific construction given in Part A plays a key role in establishing several of our main results.

In section 4, we give some notations and definitions. In section 5, we explain the construction of the process when the number of sites is finite, and give the proof of its key structural properties. In section 6, we deduce from this the case when the number of sites is infinite. Finally, section 7 deals with the simplifications related to the encoding of the nucleotidic sequences as sequences of purines and pyrimidines.

Part B is devoted to the actual computation of some quantities of interest for the model at equilibrium. From structural properties of the process, described in part A, computing the equilibrium frequencies of polynucleotides amounts to solving finite-size linear systems. Moreover, one can express the nucleotidic frequencies as functions of the YpR dinucleotidic frequencies, and these, in turn, solve a 4×4 linear system. Section 8 explains this in the general case.

The remaining sections of part B are devoted to restricted versions of the model, which involve a reduced number of free parameters. Section 9 deals with models that are invariant by the classical symmetry of DNA strands. This case has biological significance since the invariance reflects the well known complementarity of the two strands of DNA molecules. Section 10 deals with the simplest non trivial version of the RN+YpR model, namely, the case when all the simple substitution rates coincide (Jukes-Cantor model) and when there are no double substitutions except from CpG to CpA and TpG, both at the same rate. In this setting, we provide the exact value of the 16 dinucleotidic frequencies (that is, not only the 4 YpR frequencies), we develop a perturbative analysis of the frequency of every polynucleotide at vanishing CpG substitution rates, and we describe the non degenerate limit of the model at high CpG substitution rates.

Part C deals with some coupling properties of these systems, that are used to study the speed of convergence to equilibrium and to simulate the stationary distribution via coupling-from-the-past techniques, as explained in section 11.

Part A Construction

4. Notations and definitions

For every topological space E and every real number s, $\mathcal{D}(s, E)$ denotes the space of càdlàg (right continuous with left limits) functions from $[s, +\infty)$ to E, equipped with the Skorohod topology and the corresponding Borel σ-algebra.

Let I denote the collection of nucleotidic sites, thus I may be either the integer line $\mathbb{Z}$ or a finite interval of integers. For technical reasons that will become apparent when we explain the construction of the dynamics, I must contain at least 3 sites.

We recall from definition 1 in the introduction that $\mathcal{A} := \{A, T, C, G\}$ denotes the nucleotidic alphabet, that A and G are purines, encoded by the letter R, and that C and T are pyrimidines, encoded by the letter Y. We now define some mappings on $\mathcal{A}$.

Definition 5. *The mapping π is defined on $\mathcal{A}$ by*

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

Let ϱ denote the application which fuses the two purines together, and η the application which fuses the two pyrimidines together, that is

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

and

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

For every nucleotide x in $\mathcal{A}$, let x^* denote the unique nucleotide such that $\{x, x^*\} = \{A, G\}$ or $\{x, x^*\} = \{C, T\}$. In other words, $x \mapsto x^*$ is the involution of $\mathcal{A}$ such that

$$A^* := G, \quad T^* := C, \quad C^* := T, \quad G^* := A.$$

For every subsets K and J of the integer line such that $J \subset K$, let Π_K^J denote the canonical projection from $\mathcal{A}^K$ to $\mathcal{A}^J$. When $J = \{a, \dots, b\}$ for two integers $a \leqslant b$, we often omit the mention of K and write $\Pi^{a,b}$ for Π_K^J. In other words, if K contains $\{a, \dots, b\}$ and if $\mathbf{x} := (x_k)_{k\in K}$, then $\Pi^{a,b}(\mathbf{x}) := (x_k)_{a\leqslant k\leqslant b}$. Let $\mathcal{S}^+$ denote the set of countably infinite locally finite subsets N of the real line $\mathbb{R}$, and let $\mathcal{S} := \mathcal{S}^+ \cup \{\emptyset\}$. We equip $\mathcal{S}$ with the usual σ–algebra in the context of point processes, namely, the smallest σ-algebra such that, for every Borel subset B of the real line, the function $N \mapsto \mathrm{card}(N \cap B)$ is measurable.

We consider collections ξ of elements of $\mathcal{S}$, defined as follows. Let $\mathcal{A}_0$ denote the disjoint union of five copies of $\mathcal{A}$, say $\mathcal{A}_0 := \mathcal{A}_U \cup \mathcal{A}_V \cup \mathcal{A}_W \cup \mathcal{A}_R \cup \mathcal{A}_Q$. Let Ω_0 denote the space $\Omega_0 := \mathcal{S}^{\mathcal{A}_0 \times I}$. We call $\mathcal{F}_0$ the product σ–algebra on Ω_0 inherited from that of $\mathcal{S}$. Let ξ be a generic element of Ω_0. Hence ξ may be written as $\xi =: (\xi_i)_{i\in I}$, where each ξ_i belongs to $\mathcal{S}^{\mathcal{A}_0}$. For every x in $\mathcal{A}$, let $\mathcal{U}_i^x(\xi)$ denote the x–coordinate of ξ_i in $\mathcal{A}_U$, hence $\mathcal{U}_i^x(\xi)$ belongs to $\mathcal{S}$. Likewise, $\mathcal{V}_i^x(\xi)$, $\mathcal{W}_i^x(\xi)$, $\mathcal{R}_i^x(\xi)$, and $\mathcal{Q}_i^x(\xi)$ respectively denote the x–coordinates of ξ_i in $\mathcal{A}_V$, $\mathcal{A}_W$, $\mathcal{A}_R$, and $\mathcal{A}_Q$, and belong to $\mathcal{S}$ as well. Thus, for every ξ in Ω_0 and every i in I, $\xi_i =: \big(\mathcal{U}_i^x(\xi), \mathcal{V}_i^x(\xi), \mathcal{W}_i^x(\xi), \mathcal{R}_i^x(\xi), \mathcal{Q}_i^x(\xi)\big)_{x\in\mathcal{A}}$. In words, for every site i, ξ_i denotes a collection of points on the real line and every point in this collection is colored by a nucleotide x and by a type of move (see below) encoded by one of the letters U, V, W, R and Q.

Recall that, for every real number r, the positive part r^+ and the negative part r^- of r are both nonnegative and defined by the relations $r = r^+ - r^-$ and $|r| = r^+ + r^-$.

Definition 6. *For every nucleotide x in $\mathcal{A}$, the combined rate of substitution to x is the nonnegative real number c_x defined as $c_x := w_x - \max\{(r_x^y)^-, (r_x^{y^*})^-\}$, where $\{y, y^*\} = \{C, T\}$ if $\pi(x) = R$ and $\{y, y^*\} = \{A, G\}$ if $\pi(x) = Y$.*

Finally, the probability measure $\mathbb{Q}$ on $(\Omega_0, \mathcal{F}_0)$ is such that, for every site i in I and every nucleotide x in $\mathcal{A}$, the following properties hold.

- $\mathcal{U}_i^x$ is a homogeneous Poisson process on the real line with rate $\min(v_x, c_x)$,
- $\mathcal{V}_i^x$ is a homogeneous Poisson process on the real line with rate $(v_x - c_x)^+$,
- $\mathcal{W}_i^x$ is a homogeneous Poisson process on the real line with rate $(c_x - v_x)^+$,
- $\mathcal{R}_i^x$ is a homogeneous Poisson process on the real line with rate $|r_x^y|$, where the rate r_x^y corresponds to YpR substitutions starting from CpG or TpA,
- $\mathcal{Q}_i^x$ is a homogeneous Poisson process on the real line with rate $|r_x^y|$, where the rate r_x^y corresponds to YpR substitutions starting from TpG or CpA.

Thus, $\mathbb{Q}$ is uniquely specified by the following additional condition.

- The Poisson processes $\mathcal{U}_i^x$, $\mathcal{V}_i^x$, $\mathcal{W}_i^x$, $\mathcal{R}_i^x$, and $\mathcal{Q}_i^x$, for every site i and every nucleotide x, are independent.

One sees that every $\mathcal{U}_i^x \cup \mathcal{V}_i^x$ is a homogeneous Poisson process with constant rate v_x and that every $\mathcal{U}_i^x \cup \mathcal{W}_i^x$ is a homogeneous Poisson process with constant rate c_x.

We now provide a brief and intuitive description of the construction of the dynamics of the process, using these Poisson processes. As usual, the points in the processes ξ_i are the ringing times of exponential clocks that rule the evolution of the sites. There exists five types of moves, labelled as U, V, W, R, and Q.

- Type U. When an exponential clock attached to $\mathcal{U}_i^x$ rings, the nucleotide at site i moves unconditionally to the value x.
- Type V. When an exponential clock attached to $\mathcal{V}_i^x$ rings, the nucleotide at site i moves to the value x provided that this move corresponds to a transversion.
- Type W. When an exponential clock attached to $\mathcal{W}_i^x$ rings, the nucleotide at site i moves to the value x provided that this move corresponds to a transition.
- Type R. When an exponential clock attached to $\mathcal{R}_i^x$ rings, the nucleotide at site i moves to the value x in the following cases: if this move corresponds to a YpR substitution from CpG or TpA when the associated rate r_x^y is positive, and if this move corresponds to a transition but not to a substitution from CpG or TpA when r_x^y is negative.
- Type Q. When an exponential clock attached to $\mathcal{Q}_i^x$ rings, the nucleotide at site i moves to the value x in the following cases: if this move corresponds to a YpR substitution from TpG or CpA when the associated rate r_x^y is positive, and if this move corresponds to a transition but not to a substitution from TpG or CpA when r_x^y is negative.

The rates of the Poisson processes are chosen in order to couple as strongly as possible the transitions and the transversions that yield the same nucleotide and to take properly into account the inhibitory effect of YpR mutations when some rates r_x^y are negative.

We introduce a subset Ω_1 of Ω_0, defined by the following conditions.

- The sets $\mathcal{P}_i^x$ are disjoint, for every nucleotide x, every site i, and for every symbol $\mathcal{P}$ in the set $\{\mathcal{U}, \mathcal{V}, \mathcal{W}, \mathcal{R}, \mathcal{Q}\}$.
- For every site i, there exists a symbol $\mathcal{P}$ in the set $\{\mathcal{U}, \mathcal{V}, \mathcal{W}, \mathcal{R}, \mathcal{Q}\}$ and a nucleotide x in $\mathcal{A}$ such that the set $\mathcal{P}_i^x$ is infinite.

We also introduce a non-degeneracy condition.

- (ND) For every nucleotide x in $\mathcal{A}$, v_x and c_x are positive.

Under condition (ND), $\mathbb{Q}(\Omega_0 \setminus \Omega_1) = 0$. To avoid the handling of tedious exceptions, we assume from now on that condition (ND) holds and we work exclusively on Ω_1, equipped with the Borel σ–field and the probability measure that are induced by those of $(\Omega_0, \mathcal{F}_0, \mathbb{Q})$. We denote this new probability space by $(\Omega_1, \mathcal{F}_1, \mathbb{Q})$.

5. Construction on finite intervals

Before considering the case of the full integer line, we define the dynamics on finite discrete segments, with periodic boundary conditions. The choice of boundary conditions is somewhat arbitrary, and one could use instead free boundary conditions or fixed boundary conditions. However, the dynamics with periodic boundary conditions is invariant by the translations of the discrete circle, and this fact will turn out to be useful to our algorithmic purposes later on, since it reduces the dimension of the linear system which yields the invariant distribution, see section 8.

5.1. **Definitions and notations.** In this whole section 5, we fix two integers a and b such that $a + 2 \leqslant b$ and we assume that $I := \{a, \dots, b\}$. The definitions below depend on the choice of I but, to alleviate the notations, we do not always mention explicitly the dependence.

For every site i in I, we introduce $*i$ as the neighbor of i to the left of i and $i*$ as the neighbor of i to the right of i. More precisely,

$$*i := i - 1 \text{ if } i \neq a, \quad *a := b, \quad i* := i + 1 \text{ if } i \neq b, \quad b* := a.$$

Since $a + 2 \leqslant b$, for every site i, the sites i, $*i$ and $i*$ are three different sites.

Fix a sequence $\mathbf{x} = (x_i)_{i \in I}$ in $\mathcal{A}^I$ and a site i in I. Our next definitions are related to the moves of types Q and R, defined in section 4. Say that:

- The sequence $\mathbf{x}$ accepts the substitutions of type R to G at site i if $x_{*i}x_i$ is TpA.
- The sequence $\mathbf{x}$ accepts the substitutions of type R to C at site i if $x_i x_{i*}$ is TpA.
- The sequence $\mathbf{x}$ accepts the substitutions of type R to A at site i if $x_{*i}x_i$ is CpG.
- The sequence $\mathbf{x}$ accepts the substitutions of type R to T at site i if $x_i x_{i*}$ is CpG.

Likewise:

- The sequence $\mathbf{x}$ accepts the substitutions of type Q to G at site i if $x_{*i}x_i$ is CpA.

- The sequence $\mathbf{x}$ accepts the substitutions of type Q to C at site i if $x_i x_{i*}$ is TpG.
- The sequence $\mathbf{x}$ accepts the substitutions of type Q to A at site i if $x_{*i} x_i$ is TpG.
- The sequence $\mathbf{x}$ accepts the substitutions of type Q to T at site i if $x_i x_{i*}$ is CpA.

Although this terminology makes little concrete sense when some rates r_x^y are negative, we use it even then.

5.2. **Construction.** The goal of this section is to define, for every real number s, a measurable map $\varphi_s^I \ : \ \mathcal{A}^I \times \Omega_1 \to \mathcal{D}(s, \mathcal{A}^I)$. Intuitively, each function $\varphi_s^I(\mathbf{x}, \xi)$ describes the dynamics in $\mathcal{A}^I$ that starts from the configuration $\mathbf{x}$ at time s and uses the moves prescribed by the realization ξ of the Poisson processes. To be specific about the notations, we write $\varphi_s^I(\mathbf{x}, \xi)(t) = \big(\varphi_s^I(\mathbf{x}, \xi, i, t)\big)_{i \in I}$. In other words, $\varphi_s^I(\mathbf{x}, \xi, i, t)$ stands for the i–coordinate of the value of the function $\varphi_s^I(\mathbf{x}, \xi)$ at time $t \geqslant s$, hence $\varphi_s^I(\mathbf{x}, \xi, i, t)$ belongs to $\mathcal{A}$.

From now on, we fix ξ in Ω_1 and $\mathbf{x}$ in $\mathcal{A}^I$, and, to alleviate the notations, we omit to mention the dependence with respect to a, b, s or ξ of various quantities. Introduce

$$\mathcal{T} := \bigcup_{i \in I} \mathcal{T}(i), \quad \mathcal{T}(i) := [s, +\infty) \cap \bigcup_{z \in \mathcal{A}} \mathcal{U}_i^z \cup \mathcal{V}_i^z \cup \mathcal{W}_i^z \cup \mathcal{R}_i^z \cup \mathcal{Q}_i^z.$$

Let $t_{-1} := s$ and $(t_0, t_1, \cdots)$ denote the ordered list of points in $\mathcal{T}$, that is, $s < t_0 < t_1 < \cdots$ and $T =: \{t_0, t_1, t_2, \ldots\}$. For every $n \geqslant 0$, c_n denotes the site where the nth move occurs that affects a site in I after the time s, and let M_n denote the description of this move. That is, $c_n = i$ and $M_n = (z, U)$ if t_n belongs to $\mathcal{U}_i^z$. Likewise, $c_n = i$ and $M_n = (z, V)$, $M_n = (z, W)$, $M_n = (z, R)$, and $M_n = (z, Q)$ respectively, if t_n belongs to $\mathcal{V}_i^z$, $\mathcal{W}_i^z$, $\mathcal{R}_i^z$, and $\mathcal{Q}_i^z$ respectively. We often consider that M_n belongs to $\mathcal{A}_0$, for instance the couple $M_n = (z, U)$ may be identified with the point z in $\mathcal{A}_U$. Alternatively, the second component of M_n is considered as a flag, it takes its values in the set $\{U, V, W, R, Q\}$, and it is often denoted by F. In any case, the variables c_n and M_n are uniquely defined for every ξ in Ω_1.

We now define a map $\gamma_I \ : \ \mathcal{A}^I \times \mathcal{A}_0 \times I \to \mathcal{A}$. Fix a sequence $\mathbf{x} := (x_i)_{i \in I}$ in $\mathcal{A}^I$, a move description $M = (z, F)$ in $\mathcal{A}_0$, and a site i in I. The definition of the i–coordinate of $\gamma_I(\mathbf{x}, z, F, i)$ differs from the definition of the other coordinates. More precisely, one sets $\gamma_I(\mathbf{x}, z, F, i)_i := z$, if one of the following conditions is met.

- The flag F is U.
- The flag F is V and x_i is a purine and z is a pyrimidine, or vice versa.
- The flag F is W and x_i and z are both purines or both pyrimidines,
- The flag F is R and the type R rate r_z^y is positive and $\mathbf{x}$ accepts the substitutions of type R to z at site i.
- The flag F is R and the type R rate r_z^y is negative and $\mathbf{x}$ does not accept the substitutions of type R to z at site i and x_i and z are both purines or both pyrimidines.

- The flag F is Q and the type Q rate r_z^y is positive and $\mathbf{x}$ accepts the substitutions of type Q to z at site i.
- The flag F is Q and the type Q rate r_z^y is negative and $\mathbf{x}$ does not accept the substitutions of type Q to z at site i and x_i and z are both purines or both pyrimidines.

In every other case, that is, if $j = i$ and none of the conditions above is met, or if $j \neq i$, one sets $\gamma_I(\mathbf{x}, z, F, c)_j := x_j$. This defines the map γ_I. We now construct the map φ_s^I, using an induction over increasing values of the time $t \geqslant s$. The initial condition is that, for every $s \leqslant t < t_0$, $\varphi_s^I(\mathbf{x}, \xi)(t) := \mathbf{x}$. Assume now that $n \geqslant 0$ and that $\varphi_s^I(\mathbf{x}, \xi)(t)$ is well-defined for every time t such that $s \leqslant t < t_n$. Then, for every time t such that $t_n \leqslant t < t_{n+1}$, one sets

$$\varphi_s^I(\mathbf{x}, \xi)(t) := \gamma_I \left(\varphi_s^I(\mathbf{x}, \xi)(t_{n-1}), M_n, c_n \right).$$

This defines a configuration $\varphi_s^I(\mathbf{x}, \xi)(t)$ for every time $t \geqslant s$. Since I is finite, measurability issues are obvious here.

5.3. **First properties.** Recall that in this section 5, $I = \{a, \dots, b\}$ is finite. We record some immediate properties of the family of maps φ_s^I, making use of still another notation.

Definition 7. *For every time set T, let $\xi T := (\xi_i T)_{i \in I}$, where*

$$\xi_i T := \left(\mathcal{U}_i^z(\xi) \cap T, \mathcal{V}_i^z(\xi) \cap T, \mathcal{W}_i^z(\xi) \cap T, \mathcal{R}_i^z(\xi) \cap T, \mathcal{Q}_i^z(\xi) \cap T \right)_{z \in \mathcal{A}}.$$

Proposition 4. **(1)** *For every $t' \geqslant t > s$,*

$$\varphi_s^I(\mathbf{x}, \xi)(t') = \varphi_t^I \left(\varphi_s^I(\mathbf{x}, \xi)(t), \xi \right)(t').$$

(2) *For every $t' \geqslant 0$ and every real number t,*

$$\varphi_{s+t}^I(\mathbf{x}, \xi + t)(s + t + t') = \varphi_s^I(\mathbf{x}, \xi)(s + t'),$$

where $\xi + t$ denotes the result of the addition of t to each component of ξ.
(3) *Finally, $\varphi_s^I(\mathbf{x}, \xi)(t)$ depends on ξ only through $\xi[s, t]$.*

The function $\varphi_0^I(\mathbf{x}, \cdot)$, defined by $s \mapsto \varphi_0^I(\mathbf{x}, \cdot)(s)$, is a random variable on $(\Omega_1, \mathcal{F}_1, \mathbb{Q})$ with values in $\mathcal{D}(0, \mathcal{A}^I)$. Let $\mathbb{P}_{\mathbf{x}}^I$ denote its distribution. Then, from proposition 4 and from the translation invariance of Poisson processes, the family of measures $\left\{ \mathbb{P}_{\mathbf{x}}^I, \mathbf{x} \in \mathcal{A}^I \right\}$ defines a Markov process in the sense of Liggett [8, chapter 1]. Moreover, it corresponds to the specification of the process on $\mathcal{A}^I$ given in our section 1 in terms of transition rates, with periodic boundary conditions. From now on, we use the notation $X_{\mathbf{x}}^I(s) := \varphi_0^I(\mathbf{x}, \cdot)(s)$.

Proposition 5. *For every finite interval I and every initial configuration $\mathbf{x}$, the Markov process $(X_{\mathbf{x}}^I(s))_{s \geqslant 0}$ is ergodic. In other words, there exists a unique stationary distribution μ_I on $\mathcal{A}^I$ and, for every $\mathbf{x}$ in $\mathcal{A}^I$, $X_{\mathbf{x}}^I(s)$ converges in distribution to μ_I when $s \to +\infty$.*

Proof of proposition 5. Immediate since $(X_{\mathbf{x}}^I(s))_{s \geqslant 0}$ lives on a finite state space and is irreducible from the non-degeneracy assumption (ND) at the end of section 4. □

5.4. **Dependencies.** We now make a fundamental observation.

Proposition 6. *For every initial sequence $\mathbf{x} := (x_i)_{i\in I}$, every site i in I and every time $t \geqslant s$, the nucleotide $\varphi_s^I(\mathbf{x}, \xi, i, t)$, which depends a priori on all the information contained in $\mathbf{x}$ and ξ, is in fact measurable with respect to the following restricted initial conditions and restricted sources of moves:*

$$x_{*i}, \quad x_i, \quad x_{i*}, \quad \xi_{*i}[s,t], \quad \xi_i[s,t], \quad \xi_{i*}[s,t].$$

More precisely, there exists measurable maps $\Theta_{s,t}$, independent of a and b, and such that, for every integers a and b such that $a+2 \leqslant b$, and for every $a \leqslant i \leqslant b$ and $t \geqslant s$,

$$\varphi_s^{a,b}(\mathbf{x}, \xi, i, t) = \Theta_{s,t}(x_{*i}, x_i, x_{i*}, \xi_{*i}[s,t], \xi_i[s,t], \xi_{i*}[s,t]).$$

A straightforward consequence of proposition 6 is proposition 7 below.

Proposition 7. *Let a, b, c, d, c' and d' denote integers such that $c+1 \leqslant a \leqslant b \leqslant d-1$, $\quad c'+1 \leqslant a \leqslant b \leqslant d'-1$. Fix some initial conditions $\mathbf{x}$ in $\mathcal{A}^{\{c,\dots,d\}}$ and $\mathbf{x}'$ in $\mathcal{A}^{\{c',\dots,d'\}}$ that coincide on the interval $\{a-1, a, \dots, b, b+1\}$, that is, such that $\Pi^{a-1,b+1}(\mathbf{x}) = \Pi^{a-1,b+1}(\mathbf{x}')$. Then, for every $t \geqslant s$,*

$$\Pi^{a,b}\left(\varphi_s^{c,d}(\mathbf{x}, \xi)(t)\right) = \Pi^{a,b}\left(\varphi_s^{c',d'}(\mathbf{x}', \xi)(t)\right).$$

5.5. **Proofs.** The proof of proposition 6 relies on the key lemma 1 below. Recall the definition of ϱ and η in definition 5.

Lemma 1. *For every site i in I and every time $t \geqslant s$, the functions*

$$\varrho\left[\varphi_s^I(\mathbf{x}, \xi, *i, t)\right], \quad \varphi_s^I(\mathbf{x}, \xi, i, t), \quad \eta\left[\varphi_s^I(\mathbf{x}, \xi, i*, t)\right],$$

which depend a priori on all the information in $\mathbf{x}$ and ξ, are in fact measurable with respect to the following restricted initial conditions and restricted source of moves:

$$\varrho(x_{*i}), \quad x_i, \quad \eta(x_{i*}), \quad \xi_{*i}[s,t], \quad \xi_i[s,t], \quad \xi_{i*}[s,t].$$

More precisely, one may define a measurable map $\Psi_{s,t}$, independent of a and b, and such that, for every integers a and b such that $a+2 \leqslant b$, every $a \leqslant i \leqslant b$, and every $t \geqslant s$, the triple

$$\left(\varrho\left[\varphi_s^{a,b}(\mathbf{x}, \xi, *i, t)\right], \varphi_s^{a,b}(\mathbf{x}, \xi, i, t), \eta\left[\varphi_s^{a,b}(\mathbf{x}, \xi, i*, t)\right]\right)$$

coincides with

$$\Psi_{s,t}\left(\varrho(x_{*i}), x_i, \eta(x_{i*}), \xi_{*i}[s,t], \xi_i[s,t], \xi_{i*}[s,t]\right).$$

Proof of lemma 1. Fix a source of moves ξ in Ω_1, an initial configuration $\mathbf{x}$ in $\mathcal{A}^I$, a site i in I, and let $\mathcal{T}^* = \mathcal{T}(*i) \cup \mathcal{T}(i) \cup \mathcal{T}(i*)$, where the sets $\mathcal{T}(\cdot)$ are defined in section 5.2. Let $(t_0^*, t_1^*, \cdots)$ denote the ordered list of points in $\mathcal{T}^*$, that is, $\mathcal{T}^* = \{t_0^*, t_1^*, t_2^*, \dots\}$ where $s < t_0^* < t_1^* < \cdots$, and set $t_{-1}^* := s$. Then, for all $n \geqslant 0$, we describe the move that occurs at time t_n^* through the description $M_n^* = (z_n^*, F_n^*)$ of this move and the site c_n^* where the move occurs, as defined in section 5.2. Note that the moves that affect the sites $*i$, i and $i*$, can only occur at one of the times t_n^* for $n \geqslant 0$.

Thus, to prove lemma 1, it is enough to prove that, for all $n \geqslant -1$,

$$\varrho\left[\varphi_s^I(\mathbf{x},\xi,*i,t_{n+1}^*)\right], \quad \varphi_s^I(\mathbf{x},\xi,i,t_{n+1}^*), \quad \eta\left[\varphi_s^I(\mathbf{x},\xi,i*,t_{n+1}^*)\right],$$

depend only on M_{n+1}^*, on c_{n+1}^*, and on

$$\varrho\left[\varphi_s^I(\mathbf{x},\xi,*i,t_n^*)\right], \quad \varphi_s^I(\mathbf{x},\xi,i,t_n^*), \quad \eta\left[\varphi_s^I(\mathbf{x},\xi,i*,t_n^*)\right].$$

To this aim, we first assume that $c_{n+1}^* = i$ and we examine the value of the flag F_{n+1}^*.

- If F_{n+1}^* is U, V or W, by the very construction of the process, $\varphi_s^I(\mathbf{x},\xi,i,t_{n+1}^*)$ is determined by $\varphi_s^I(\mathbf{x},\xi,i,t_n^*)$, f_n^* and z_n^*.
- If F_{n+1}^* is R or Q, $\varphi_s^I(\mathbf{x},\xi,i,t_{n+1}^*)$ is determined by $\varphi_s^I(\mathbf{x},\xi,i,t_n^*)$, by z_n^*, and by the knowledge of whether or not the sequence accepts the R or Q substitutions to z_n^* at site i. In turn, this only depends on

$$\varrho\left[\varphi_s^I(\mathbf{x},\xi,*i,t_n^*)\right], \quad \varphi_s^I(\mathbf{x},\xi,i,t_n^*), \quad \eta\left[\varphi_s^I(\mathbf{x},\xi,i*,t_n^*)\right].$$

This settles the case when $c_{n+1}^* = i$. Now we assume that $c_{n+1}^* = *i$ and we examine the value of the flag F_{n+1}^*.

- If F_{n+1}^* is U, V or W, $\varphi_s^I(\mathbf{x},\xi,*i,t_{n+1}^*)$ depends on z_n^* and on whether the nucleotide $\varphi_s^I(\mathbf{x},\xi,*i,t_n^*)$ is a purine or a pyrimidine. This information is provided by $\varrho\left[\varphi_s^I(\mathbf{x},\xi,*i,t_n^*)\right]$.
- If F_{n+1}^* is R or Q and z_n^* is C or T, whether the sequence accepts the substitutions of types R or Q at site $*i$ depends only on $\varrho\left[\varphi_s^I(\mathbf{x},\xi,*i,t_n^*)\right]$ and $\varphi_s^I(\mathbf{x},\xi,i,t_n^*)$.
- If F_{n+1}^* is R or Q and z_n^* is A or G, we observe that the corresponding substitution of types R or Q at site $*i$ could only turn an A to a G or vice versa. This does not affect the value of $\varrho\left[\varphi_s^I(\mathbf{x},\xi,*i,t_{n+1}^*)\right]$, which must be equal to $\varrho\left[\varphi_s^I(\mathbf{x},\xi,*i,t_n^*)\right]$.

This settles the case when $c_{n+1}^* = *i$. Symmetric arguments hold when $c_{n+1}^* = i*$, hence this concludes the proof. □

Proof of proposition 7. Let $\Pi^{a-1,b+1}(\mathbf{x}) = \Pi^{a-1,b+1}(\mathbf{x}') =: (x_i)_{a-1\leqslant i\leqslant b+1}$. The result follows from proposition 6 since, for every i in $\{a,\dots,b\}$, $\varphi_s^{c,d}(\mathbf{x},\xi,i,t)$ and $\varphi_s^{c',d'}(\mathbf{x}',\xi,i,t)$ are both equal to

$$\Theta_{s,t}(x_{i-1},x_i,x_{i+1},\xi_{i-1}[s,t],\xi_i[s,t],\xi_{i+1}[s,t]).$$

□

6. Construction on the integer line

The construction of the process on the integer line $\mathbb{Z}$ relies crucially on a consequence of proposition 7 above, namely the fact that, for every initial sequence $\mathbf{x}$, every site i in $\mathbb{Z}$, every source of moves ξ, and every couple of times $t \geqslant s$, the value of $\varphi_s^{a,b}(\Pi^{a,b}(\mathbf{x}),\xi,i,t)$ does not depend on a and b as soon as $a+1 \leqslant i \leqslant b-1$. Hence the projective limit of the system $(\varphi_s^{a,b})_{a,b}$ when $a \to -\infty$ and $b \to \infty$ exists

trivially and defines a mesurable map $\Phi_s : \mathcal{A}^{\mathbb{Z}} \times \Omega_1 \to \mathcal{D}(s, \mathcal{A}^{\mathbb{Z}})$. Some previous observations about $\varphi_s^{a,b}$ translate immediately to Φ_s.

Proposition 8. **(1)** *For every $t' \geqslant t \geqslant s$, $\Phi_s(\mathbf{x}, \xi)(t') = \Phi_t\left(\Phi_s(\mathbf{x}, \xi)(t), \xi\right)(t')$.*
(2) *For every $t' \geqslant 0$ and t, $\Phi_{s+t}(\mathbf{x}, \xi + t)(s + t + t') = \Phi_s(\mathbf{x}, \xi)(s + t')$.*
(3) *Finally, $\Phi_s(\mathbf{x}, \xi)(t)$ depends on ξ only through $\xi[s, t] = (\xi_i[s, t])_{i\in\mathbb{Z}}$.*

For every $\mathbf{x}$ in $\mathcal{A}^{\mathbb{Z}}$, let $\mathbb{P}_{\mathbf{x}}$ denote the distribution of $s \mapsto \Phi_0(\mathbf{x}, \cdot)(s)$, viewed as a random variable on $(\Omega_1, \mathcal{F}_1, \mathbb{Q})$ with values in $\mathcal{D}(0, \mathcal{A}^{\mathbb{Z}})$. Then, as can be checked from propositions 7 and 8 using the translation invariance of Poisson processes, the family $\left\{ \mathbb{P}_{\mathbf{x}},\ \mathbf{x} \in \mathcal{A}^{\mathbb{Z}} \right\}$ defines a Feller Markov process in the sense of Liggett [8, chapter 1]. From now on, we use the notation $X^{\mathbf{x}}(s) := \Phi_0(\mathbf{x}, \cdot)(s)$, and we sometimes omit the initial condition $\mathbf{x}$ of the Markov process $(X(s))_{s\geqslant 0}$.

It is straightforward to check that the construction in [8], based on infinitesimal generators, yields the same process. Let $\mathcal{G}$ denote its infinitesimal generator, based on the construction in [8], then $\mathcal{G}$ is well-defined and explicitly known at least for the Lipschitz functions on $\mathcal{A}^{\mathbb{Z}}$ (see section I.3 in [8]).

Let $Y := (Y_i)_{i\in\mathbb{Z}\setminus I}$ denote any element of $\mathcal{A}^{\mathbb{Z}\setminus I}$ and $(X_{\mathbf{x}}^{Y,I}(s))_{s\geqslant 0}$ the Markov process on $\mathcal{A}^{\mathbb{Z}}$ defined by

$$X_{\mathbf{x}}^{Y,I}(s)_i := \begin{cases} X_{\mathbf{x}}^{I}(s)_i & \text{if } i \in I, \\ Y_i & \text{otherwise.} \end{cases}$$

By definition, $X_{\mathbf{x}}^{Y,I}(s)$ converges to $X_s^{\mathbf{x}}$ as $I \to \mathbb{Z}$. On the other hand, since the process $(X_{\mathbf{x}}^{Y,I}(s))_{s\geqslant 0}$ involves moves only on the finite set of sites I, its infinitesimal generator $\mathcal{G}_I$ can be readily computed. Moreover, for every Y, $\mathcal{G}_I$ converges to $\mathcal{G}$ as $I \to \mathbb{Z}$, at least on the set of real valued Lipschitz functions defined on $\mathcal{A}^{\mathbb{Z}}$. This is enough to identify $(X(s))_{s\geqslant 0}$ with the process yielded by the construction of [8], according to Corollary 3.14 in [8].

A consequence of proposition 6 above is the following result.

Proposition 9. *For every sequence $\mathbf{x} := (x_i)_{i\in\mathbb{Z}}$, every integer i and every couple of times $t \geqslant s$, $\Phi_s(\mathbf{x}, \xi, i, t)$ depends on $\mathbf{x}$ and ξ only through*

$$x_{i-1}, \quad x_i, \quad x_{i+1}, \quad \xi_{i-1}[s, t], \quad \xi_i[s, t], \quad \xi_{i+1}[s, t].$$

Indeed, using the function $\Theta_{s,t}$ defined in proposition 6,

$$\Phi_s(\mathbf{x}, \xi, i, t) = \Theta_{s,t}(x_{i-1}, x_i, x_{i+1}, \xi_{i-1}[s, t], \xi_i[s, t], \xi_{i+1}[s, t]).$$

A simple consequence of proposition 9 is the following proposition.

Proposition 10. *Fix $\mathbf{x}$ and some subsets J of the integer line at distance at least 3 from each other, that is, such that for every such pair $J \neq J'$ of subsets and every sites i in J and i' in J', $|i - i'| \geqslant 3$. Then, the collections $\left[(X_i^{\mathbf{x}}(s))_{s\geqslant 0}\right]_{i\in J}$ are independent.*

Proposition 11. *There exists a unique stationary distribution of $(X(s))_{s\geqslant 0}$ on $\mathcal{A}^{\mathbb{Z}}$.*

Definition 8. *Let μ denote the stationary distribution of $(X(s))_{s\geqslant 0}$. Let $\mu_{a,b}$ denote the measure μ_I whose existence is ensured by proposition 5, when $I = \{a, \dots, b\}$.*

A consequence of the uniqueness of μ is its invariance by the transformations that preserve the dynamics, for instance the translations of $\mathbb{Z}$. Likewise, $\mu_{a,b}$ is invariant with respect to the translations of the discrete circle $\{a,\dots,b\}$ viewed as $\mathbb{Z}/(b-a+1)\mathbb{Z}$. Moreover, if $\{a,\dots,b\}$ is the image of $\{c,\dots,d\}$ by a translation of $\mathbb{Z}$, $\mu_{a,b}$ coincide with the image of $\mu_{c,d}$ by this translation. Other properties are in proposition 12.

Proposition 12. **(1)** *Assume that the integers a, b, c and d are such that $c+1 \leqslant a \leqslant b \leqslant d-1$. Then, $\Pi^{a,b}(\mu_{c,d}) = \Pi^{a,b}(\mu)$.*
(2) *The Markov process $(X(s))_{s\geqslant 0}$ is ergodic. That is, for every initial condition $\mathbf{x}$ in $\mathcal{A}^{\mathbb{Z}}$, $X^{\mathbf{x}}(s)$ converges in distribution to μ as $s \to \infty$.*
(3) *Finally, the independence properties stated above hold with respect to μ as well.*

The considerations above justify the following definition.

Definition 9 (Stationary frequencies)**.** *Define the stationary frequency $F(x_1\cdots x_k)$ of every polynucleotide $x_1\cdots x_k$ as*

$$F(x_1\cdots x_k) := \mu_{0,k+1}(\mathcal{A}\times\{(x_1,\dots,x_k)\}\times\mathcal{A}).$$

Hence, $F(x_1\cdots x_k)$ is also

$$F(x_1\cdots x_k) = \mu\left(\mathcal{A}^{\{\dots,-1,0\}}\times\left\{(x_1,\dots,x_k)\right\}\times\mathcal{A}^{\{k+1,k+2,\dots\}}\right).$$

Due to the independence properties of μ stated in proposition 12, μ is clearly ergodic with respect to the translations in $\mathbb{Z}$, so we may as well define stationary frequencies as the following almost sure limits, when $a\to-\infty$ and $b\to+\infty$:

$$F(x_1\cdots x_k) = \lim\frac{1}{b-a}\sum_{i=a+1}^{b}\mathbf{1}(X_{i+1}\cdots X_{i+k} = x_1\cdots x_k),$$

where the distribution of $(X_i)_{i\in\mathbb{Z}}$ is μ and $\mathbf{1}(B)$ denotes the indicator function of B.

Proof of propositions 11 and 12. Fix a and b such that $a\leqslant b$. From proposition 7, for every $\mathbf{x}$ in $\mathcal{A}^{\mathbb{Z}}$ and every $s\geqslant 0$,

$$\Pi^{a,b}\left(X^{\mathbf{x}}(s)\right) = \Pi^{a,b}\left(X_{\mathbf{x}}^{a-1,b+1}(s)\right).$$

Now, according to proposition 5, $X_{\mathbf{x}}^{a-1,b+1}(s)$ converges in distribution towards $\mu_{a-1,b+1}$ as $s\to\infty$. As a consequence, $\Pi^{a,b}\left(X^{\mathbf{x}}(s)\right)$ converges to $\Pi^{a,b}(\mu_{a-1,b+1})$. Proposition 7 again shows that $\left(\Pi^{a,b}(\mu_{a-1,b+1})\right)_{a\leqslant b}$ is a coherent family of probability distributions. Hence there exists a unique probability distribution μ on $\mathcal{A}^{\mathbb{Z}}$ such that $\Pi^{a,b}(\mu) = \Pi^{a,b}(\mu_{a-1,b+1})$. The convergence of $\Pi^{a,b}\left(X^{\mathbf{x}}(s)\right)$ in distribution towards $\Pi^{a,b}(\mu)$ for every $a\leqslant b$ implies that, for every $\mathbf{x}$ in $\mathcal{A}^{\mathbb{Z}}$, $X^{\mathbf{x}}(s)$ converges in distribution to μ. The existence and the uniqueness of an invariant distribution, equal to μ, follow easily. The other properties are obvious. □

7. R/Y encodings

The situation of the R/Y process is even simpler.

Definition 10. *Introduce the R/Y configurations at time s as the $\{R,Y\}$-valued functions*

$$Z^I_{\mathbf{x}}(s) := \pi(\varphi^I_s(\mathbf{x},\xi,i,s)), \quad Z_{\mathbf{x}}(s) := \pi(\varphi_s(\mathbf{x},\xi,i,s)).$$

Proposition 13 (finite intervals)**.** *For every sequence $\mathbf{x} := (x_i)_{i\in I}$, every site i in I and every time $t \geqslant s$, the value of $Z^I_{\mathbf{x}}(t)_i$, which depends a priori on all the information in $\mathbf{x}$ and ξ, is in fact measurable with respect to $\pi(x_i)$ and $\xi_i[s,t]$.*

Proof of proposition 13. The substitutions associated to clocks of types W, R and Q do not change the value of $Z^I_{\mathbf{x}}(s)_i$. Hence the moves of the function $s \mapsto Z^I_{\mathbf{x}}(s)_i$ are determined by the clocks $\mathcal{U}_i$ and $\mathcal{V}_i$. Erasing the $\mathcal{R}_i$ and $\mathcal{Q}_i$ clocks is like setting every r^y_z to 0, in which case the sites evolve independently. This proves the proposition. □

We recall from theorem 6 in section 2 the notation

$$t_Y := \frac{v_C + v_T}{v_A + v_T + v_C + v_G}, \quad t_R := \frac{v_A + v_G}{v_A + v_T + v_C + v_G}.$$

Corollary 2 below is a straightforward consequence of Proposition 13 and is equivalent to theorem 6 in section 2.

Corollary 2. *There exists a unique stationary distribution of $(Z(s))_{s\geqslant 0}$ on the state space $\{R,Y\}^{\mathbb{Z}}$. This measure is the product mesure $\nu^{\otimes\mathbb{Z}}$, where*

$$\nu(Y) := t_Y, \quad \nu(R) := t_R.$$

Theorem 6 yields relations, which hold irrespective of the values of the mutation rates r^y_x, namely,

$$\begin{aligned} F(CG)+F(CA)+F(TG)+F(TA) &= t_Y\, t_R, \\ F(CC)+F(CT)+F(TC)+F(TT) &= t_Y^2, \\ F(AC)+F(AT)+F(GC)+F(GT) &= t_R\, t_Y, \\ F(CC)+F(CT)+F(TC)+F(TT) &= t_R^2. \end{aligned}$$

These relations are simple enough to write. However, they yield awkward formulas for the individual frequencies of nucleotides and dinucleotides, in full generality.

Part B Computations

8. General case

8.1. **Polynucleotidic frequencies.** From the construction given in part A, knowing the stationary distribution of the Markov process $(X^I(s))_{s\geqslant 0}$ with $I = \{a-1,\dots,b+1\}$ is enough to compute $\Pi^{a,b}(\mu)$. Since $(X^I(s))_{s\geqslant 0}$ lives on the state space $\mathcal{A}^I$, computing its stationary distribution amounts to solving a linear system of size $\#\mathcal{A}^I \times \#\mathcal{A}^I$. Computing the equilibrium frequency of polynucleotides of length N thus requires solving a $4^{N+2} \times 4^{N+2}$ linear system.

Theorem 4 in section 2 follows from these considerations and from Cramér's formula. However, even moderate lengths of polynucleotides ($N = 4$, say) lead to fairly

large linear systems, so finding ways of lowering the dimension of the system to be solved is a critical issue, if solvability of the model is to be considered something more than a mere theoretical possibility.

For single nucleotides and for the restricted class of YpR dinucleotides, an autonomous linear subsystem can be isolated, and this is discussed in section 8.2 below.

For general polynucleotides, symmetries can be used to reduce the computational burden. Using the invariance of $(X^I(s))_{s\geqslant 0}$ with respect to translations on the discrete circle, the linear system yielding the stationary distribution of $(X^I(s))_{s\geqslant 0}$ can be reduced to a linear system of size

$$m(N+2) \times m(N+2),$$

where $m(k)$ denotes the number of distinct orbits in $\mathcal{A}^k$ under translations. It is a well-known counting result, see [11] for example, that

$$m(k) = \frac{1}{k} \sum_{d|k} \phi(d)\, 4^{k/d},$$

where ϕ stands for Euler's function, hence $\phi(d)$ denotes the number of primes to d smaller than d. Hence, when $k \to \infty$,

$$m(k) \sim 4^k/k.$$

This remark also achieves significant improvement for small values of N, see the table below. The last columns give the value of $1/(N+2)$ and the exact ratio $m(N+2)/4^{N+2}$ of the sizes of the two linear systems.

N	4^{N+2}	$m(N+2)$	$1/(N+2)$	Ratio
2	256	70	25%	27.3%
3	1024	208	20%	20.3%
4	4096	700	16.7%	17.1%

To lower the size of the linear system, the reader could think of the following alternative reduction. From the results of part A, to find the stationary distribution of $\Pi^{a,b}\left(X^I(s)\right)$ with $I = \{a-1, \ldots, b+1\}$, it is sufficient to find the invariant distribution of the Markov chain

$$\left(\varrho(X^I(s)_{a-1}), \Pi^{a,b}\left(X^I(s)\right), \eta(X^I(s)_{b+1})\right).$$

This chain lives on a state space of size $4^N \times 3^2$. Hence, this remark reduces the size of the system by a factor $9/16 \approx 56\%$. On the other hand, the translation invariance is lost. All in all, using the translation invariance described above yields more effective reductions.

Two approaches to computing equilibrium frequencies of polynucleotides may be considered. One can solve the linear system numerically with fixed values of the parameters (with finite or infinite precision arithmetic), within reasonable time for $N \leqslant 4$. One can also solve this symbolically, a task that we performed only for $N = 2$, using a restricted version of the model possessing a single free parameter, and additional symmetries, see section 10.

8.2. **Nucleotidic and YpR dinucleotidic frequencies.** Recall that $F(x_1 \cdots x_k)$, introduced formally in definition 9 in section 6, denotes the stationary frequency of the polynucleotide $x_1 \cdots x_k$. In this section, we show how to compute $F(x)$ for every nucleotide x, and $F(xy)$ for every YpR dinucleotide xy.

Definition 11. *Introduce*

$$F(Y) := F(C) + F(T), \quad F(R) := F(G) + F(A).$$

Similar conventions are valid for polynucleotides, for instance

$$F(YR) := \sum_{\pi(x)=Y} \sum_{\pi(y)=R} F(xy).$$

Likewise, $F(Yy) := F(Cy) + F(Ty), \quad F(xR) := F(xA) + F(xG)$.

We introduce some notations, related to the rates of the simple substitutions.

Definition 12. *For every nucleotide x, let s_x denote the sum of the rates of mutations from x, hence $s_A := s_R =: s_G$ and $s_C := s_Y =: s_T$, with*

$$s_R := w_A + v_T + v_C + w_G, \quad s_Y := v_A + w_T + w_C + v_G.$$

Likewise, let

$$u_x := v_x - w_x, \quad v^* := \sum_x v_x, \quad w^* := \sum_x w_x.$$

Finally, let $t_A := t_R =: t_G$ and $t_C := t_Y =: t_T$, with $t_R + t_Y = 1$ and

$$t_R := (v_A + v_G)/v^*, \quad t_Y := (v_T + v_C)/v^*.$$

We turn to some notations related to the effect of the r_x^y substitutions on nucleotides.

Definition 13. *For every nucleotide x and every YpR dinucleotide yz, introduce $p_{yz}(x)$ as the rate at which x appears (or disappears if $p_{yz}(x)$ is negative) because of the dinucleotidic substitutions associated to yz. Hence,*

$$\begin{array}{ll} p_{CG}(T) := r_T^G =: -p_{CG}(C), & p_{CG}(A) := r_A^C =: -p_{CG}(G), \\ p_{TA}(C) := r_C^A =: -p_{TA}(T), & p_{TA}(G) := r_G^T =: -p_{TA}(A), \\ p_{CA}(T) := r_T^A =: -p_{CA}(C), & p_{CA}(G) := r_G^C =: -p_{CA}(A), \\ p_{TG}(A) := r_A^T =: -p_{TG}(G), & p_{TG}(C) := r_C^G =: -p_{TG}(T). \end{array}$$

Finally, for every x and every yz which is not a YpR dinucleotide, let

$$p_{yz}(x) := 0.$$

Equilibrium for the nucleotides yields the following relations.

Proposition 14. *For every nucleotide x,*

$$s_x\, F(x) = v_x - u_x\, t_x + \sum_{yz} p_{yz}(x)\, F(yz).$$

Furthermore, $F(R) = t_R$ and $F(Y) = t_Y$.

Note that the values of $F(R)$ and $F(Y)$ are independent of the YpR substitution rates r_x^y. The underlying reason for this a priori surprising fact is in corollary 2. The proof of proposition 14 is in section 8.3.

From proposition 14, the values of $F(xy)$ for every YpR dinucleotide xy determine $F(z)$ for every nucleotide z. To compute $F(xy)$ for these 4 dinucleotides xy, we need some notations related to the effects of the r_x^y mutations on dinucleotides.

Definition 14. *For every couple of YpR dinucleotides xy and zt, introduce $p_{zt}(xy)$ as the rate at which xy appears (or disappears if $p_{zt}(xy)$ is negative), due to the existence of the dinucleotide zt. Hence, assuming that $\{x, x^*\} = \{C, T\}$ and that $\{y, y^*\} = \{A, G\}$,*

$$p_{xy}(xy) := -r_{y^*}^x - r_{x^*}^y, \quad p_{xy}(xy^*) := r_{y^*}^x, \quad p_{xy}(x^*y) := r_{x^*}^y, \quad p_{xy}(x^*y^*) := 0.$$

For instance,

$$p_{CG}(CG) := -r_T^G - r_A^C, \quad p_{CA}(CG) := r_G^C, \quad p_{TG}(CG) := r_C^G, \quad p_{TA}(CG) := 0.$$

Finally, let

$$q_{xy} := -p_{xy}(xy) = r_{y^*}^x + r_{x^*}^y.$$

Equilibrium for the YpR dinucleotides yields the following relations. Recall the notations in definition 11.

Proposition 15. *For every YpR dinucleotide xy,*

$$(v^* + w^*)\, F(xy) + u_x\, F(Yy) + u_y\, F(xR) = v_x\, F(y) + v_y\, F(x) + \sum_{zt} p_{zt}(xy)\, F(zt).$$

The proof of proposition 15 is in section 8.4.

From propositions 14 and 15, the 8 unknown frequencies we are looking for, namely the 4 frequencies of the nucleotides and the 4 frequencies of the YpR dinucleotides, solve a system of 8 linear equations. One can show that the determinant of this system is not zero, hence the 8 frequencies are entirely determined by this system.

A simpler way to proceed is to write the frequency of each nucleotide as an affine function of the 4 frequencies of YpR dinucleotides, then to plug these expressions in the 4 last equations. We state this as theorem 7 below, which precises theorem 5 in section 2.

Definition 15. *Let $\mathbb{F}$ denote the 4×1 vector of the frequencies of the YpR dinucleotides, that is,*

$$\mathbb{F} := \begin{pmatrix} F(CG) \\ F(CA) \\ F(TG) \\ F(TA) \end{pmatrix}.$$

For every YpR dinucleotide, let

$$v_x^* := v_x / s_R, \qquad v_y^* := v_y / s_Y.$$

The matrix $\mathbb{U}$ *is*

$$\mathbb{U} := \begin{pmatrix} u_C + u_G & u_G & u_C & 0 \\ u_A & u_C + u_A & 0 & u_C \\ u_T & 0 & u_T + u_G & u_G \\ 0 & u_T & u_A & u_T + u_A \end{pmatrix}.$$

The coefficients of the matrix $\mathbb{W}$ *are*

$$\mathbb{W}_{xy,zt} := -p_{zt}(xy) - v_x^* \, p_{zt}(y) - v_y^* \, p_{zt}(x).$$

Finally, the coefficients of the matrix $\mathbb{V}$ *are*

$$\mathbb{V}_{xy} := v_x \, \frac{v_y - u_y \, t_y}{s_y} + v_y \, \frac{v_x - u_x \, t_x}{s_x}.$$

Note that every $\mathbb{V}_{xy}$ *is positive.*

Theorem 7 (YpR frequencies)**.** *The YpR frequencies solve a linear system*

$$((v^* + w^*) \, \mathrm{Id} + \mathbb{U} + \mathbb{W}) \cdot \mathbb{F} = \mathbb{V},$$

where the 4×4 *matrices* $\mathbb{U}$ *and* $\mathbb{W}$ *and the* 4×1 *vector* $\mathbb{V}$ *are defined above and depend on the substitution rates* v_x*,* w_x *and* r_x^y*.*

We now write the coefficients of $\mathbb{W}$ more explicitly. Each column $\mathbb{W}_{\cdot,xy}$ of $\mathbb{W}$ depends on the YpR rates of substitution through $r_{y^*}^x$ and $r_{x^*}^y$ only, and through affine functions. More precisely,

$$\mathbb{W}_{xy,xy} = (1 + v_x^*) \, r_{y^*}^x + (1 + v_y^*) \, r_{x^*}^y,$$

Furthermore,

$$\mathbb{W}_{xy^*,xy} = -(1 + v_x^*) \, r_{y^*}^x + v_{y^*}^* \, r_{x^*}^y, \quad \mathbb{W}_{x^*y,xy} = -(1 + v_y^*) \, r_{x^*}^y + v_{x^*}^* \, r_{y^*}^x,$$

and

$$\mathbb{W}_{x^*y^*,xy} = -v_{x^*}^* \, r_{y^*}^x - v_{y^*}^* \, r_{x^*}^y.$$

8.3. **Proof of proposition 14.** Assume that the distribution of $(X_i)_{i \in \mathbb{Z}}$ is the stationary measure μ introduced in definition 8 in section 6. By the definition of the dynamics, equilibrium for nucleotide x at site i reads

$$\begin{aligned} s_x \, \mathbb{P}(X_i = x) \quad = \quad & \sum_{\pi(z) = \pi(x)} w_x \, \mathbb{P}(X_i = z) + \sum_{\pi(z) \neq \pi(x)} v_x \, \mathbb{P}(X_i = z) \\ & + \sum_{\pi(y) = \pi(x) \neq \pi(z)} p_{yz}(x) \, \mathbb{P}(X_i X_{i+1} = yz) \\ & + \sum_{\pi(y) \neq \pi(x) = \pi(z)} p_{yz}(x) \, \mathbb{P}(X_{i-1} X_i = yz). \end{aligned}$$

Note that one of the last two sums in the expression above is always zero, which one depending on whether x is a purine or a pyrimidine.

Using the translation invariance of the stationary distribution, and extracting constant factors from sums, this reduces to

$$s_x\, F(x)\, F(yz) = \sum_{yz} p_{yz}(x) + w_x \sum_{\pi(z)=\pi(x)} F(z) + v_x \sum_{\pi(z)\neq\pi(x)} F(z). \tag{2}$$

Using the fact that $p_{yz}(x) = -p_{yz}(x^*)$, and summing, on the one hand the above equilibrium equations for $x = A$ and $x = G$, and on the other hand for $x = C$ and $x = T$, we obtain a linear system of two equations involving $F(R)$ and $F(Y) = 1 - F(R)$. For instance,

$$s_R\, F(R) = (w_A + w_G)\, F(R) + (v_A + v_G)\, F(Y).$$

Solving this for $F(R)$ and $F(Y)$ yields the first assertion of the proposition, plugging these values into equation (2) yields the second assertion.

8.4. **Proof of proposition 15.** As in the proof of proposition 14 above, we assume that the distribution of $(X_i)_{i\in\mathbb{Z}}$ is the stationary measure μ. Equilibrium for a YpR dinucleotide xy located at the pair of sites $(i, i+1)$ reads as the equality of the exit rate and the entrance rate. Both are due to single substitutions and to double substitutions. The exit rate of single substitutions has size

$$(s_x + s_y)\, \mathbb{P}(X_i X_{i+1} = xy).$$

The entrance rate due to the single transitions has size

$$\sum_{\pi(z)=\pi(x)} w_x\, \mathbb{P}(X_i X_{i+1} = zy) + \sum_{\pi(t)=\pi(y)} w_y\, \mathbb{P}(X_i X_{i+1} = xt).$$

The entrance rate due to the single transversions has size

$$\sum_{\pi(z)\neq\pi(x)} v_x\, \mathbb{P}(X_i X_{i+1} = zy) + \sum_{\pi(t)\neq\pi(y)} v_y\, \mathbb{P}(X_i X_{i+1} = xt).$$

Finally, the rate of double substitutions, counted as an entrance rate, has size

$$\sum_{\pi(z)=\pi(x)} p_{zy}(x)\, \mathbb{P}(X_i X_{i+1} = zy) + \sum_{\pi(t)=\pi(y)} p_{xt}(y)\, \mathbb{P}(X_i X_{i+1} = xt).$$

An important point to notice is that no YpR mutation affecting the pairs of sites $(i-1, i)$ or $(i+1, i+2)$ can occur when $X_i X_{i+1} = xy$, nor can lead to $X_i X_{i+1} = xy$, since x is a pyrimidine and y is a purine. This fact rules out probabilities of trinucleotides from appearing in the above equation, which we could not avoid if xy was not an YpR dinucleotide.

Using the facts that

$$\sum_{\pi(z)\neq\pi(x)} \mathbb{P}(X_i X_{i+1} = zy) = \mathbb{P}(X_{i+1} = y) - \sum_{\pi(z)=\pi(x)} \mathbb{P}(X_i X_{i+1} = zy),$$

and that

$$\sum_{\pi(t)\neq\pi(y)} \mathbb{P}(X_i X_{i+1} = xt) = \mathbb{P}(X_i = x) - \sum_{\pi(t)=\pi(y)} \mathbb{P}(X_i X_{i+1} = xt),$$

and the translation invariance of the stationary distribution, we obtain the identity stated in proposition 15.

9. Symmetric rates

9.1. **General strand-symmetric models.** We consider property (c) below.

> **Property (c)** The substitution rates respect the strand complementarity of nucleotides.

Recall that strand complementarity $x \mapsto x_*$ is the involution of $\mathcal{A}$ such that

$$A_* := T, \quad T_* := A, \quad C_* := G, \quad G_* := C.$$

This means, first, that the rate of substitution from x to y and from x_* to y_* coincide, for every nucleotides x and y.

This also means that the rates of YpR substitutions from CpG to CpA and to TpG coincide, and that the rates of YpR substitutions from TpA to CpA and to TpG coincide, that is,

$$r_A^C = r_T^G =: r_W, \quad r_C^A = r_G^T =: r_S.$$

This means finally that the rates of YpR substitutions from CpA and from TpG to CpG coincide, and that the rates of YpR substitutions from CpA and from TpG to TpA coincide.

As regards the single substitutions, the most general RN model such that property (c) holds is described by matrices

$$\begin{pmatrix} \cdot & v_W & v_S & w_S \\ v_W & \cdot & w_S & v_S \\ v_W & w_W & \cdot & v_S \\ w_W & v_W & v_S & \cdot \end{pmatrix},$$

where v_S, v_W, w_S and w_W are nonnegative rates. For instance, every nucleotide A mutates to T at rate v_W, to C at rate v_S, and to G at rate w_S. The indices W and S refer to the classification of nucleotides according to the strength of their link in double stranded DNA, the link between C and G being strong (S) and the link between A and T being weak (W).

One recovers Tamura's matrix when $w_S\, v_W = v_S\, w_W$. On the other hand, assuming that the complementarity (c) holds, the RN condition corresponds to the additional requirements that the two substitution rates from a purine to C coincide, and that the two substitution rates from a purine to T coincide.

9.2. **RNc+CpG models.** In the rest of this section, we consider RN models such that the complementarity (c) holds, and such that the only active YpR nucleotide is CpG (i.e. such that r_W is the only non-zero YpR rate).

Specializing the results of Section 8.2 to the present situation, we prove some results about the ratio of the observed and expected frequencies of CpG, which seems to be a parameter universally used by molecular biologists.

Definition 16. *Introduce*

$$\mathrm{CpGo/e} := \frac{F(CG)}{F(C)\,F(G)}.$$

Definition 17. *Introduce the parameters*

$$\sigma_S := v_S + w_S, \quad \sigma_W := v_W + w_W, \quad v_0 := v_S + v_W, \quad w_0 := w_S + w_W,$$

and

$$\sigma := \sigma_S + \sigma_W = v_0 + w_0.$$

Proposition 16. *For RNc+CpG models,* $\mathrm{CpGo/e} \leqslant 1$ *for every* r_W*,* CpGo/e *is a non increasing function of* r_W*,* $\mathrm{CpGo/e} \to 1$ *when* $r_W \to 0$*, and* $\mathrm{CpGo/e} \to 0$ *when* $r_W \to \infty$*. Furthermore, when* $r_W \to 0$*,* $\mathrm{CpGo/e} = 1 - r_W\, K_{CG} + o(r_W)$*, where* K_{CG} *is positive and defined as*

$$K_{CG} := \frac{(\sigma + 3v_0)\,\sigma_W + \sigma\, w_W}{\sigma^2(\sigma + 2v_0)}.$$

Another quantity of interest is the ratio of the observed to expected frequencies of TpA.

Definition 18. *Introduce*

$$\mathrm{TpAo/e} := \frac{F(TA)}{F(T)\,F(A)}.$$

The situation for the ratio TpAo/e is less clear than for the ratio CpGo/e. For instance, one can show that, when $r_W \to 0$,

$$\mathrm{TpAo/e} = 1 - r_W\, K_{TA} + o(r_W), \qquad K_{TA} := \frac{L_{TA}}{\sigma^2\, w_0^2\,(\sigma + 2v_0)},$$

where

$$\begin{aligned} L_{TA} \;:=\; & \sigma_W \sigma_S^2(\sigma + 2v_0) + \sigma_W^2(\sigma(\sigma + 2v_0) + \sigma w_W + v\sigma_W) \\ & -\sigma^2(\sigma w_W + v\sigma_W) - 2\sigma^2\sigma_W v_W. \end{aligned}$$

The sign of L_{TA} is difficult to decipher. In fact, assume that there exists c such that $w_S = c\,v_S, \quad w_W = c\,v_W$. Then the expression of L_{TA} reduces to $L_{TA} = \sigma\,\sigma_W^2\,((3 + c)\,v_W - (1 + c)\,v_S)$. This shows the following result.

Proposition 17 (Values of TpAo/e)**.** *For RNc+CpG models, both cases* $\mathrm{TpAo/e} \leqslant 1$ *and* $\mathrm{TpAo/e} \geqslant 1$ *are possible. However, when* $w_S = w_W$ *and* $v_S = v_W$*,* $\mathrm{TpAo/e} \leqslant 1$*.*

10. The simplest model

In this section, we provide the values of the 16 frequencies of dinucleotides at equilibrium. To avoid awkward formulas, we consider the simplest non degenerate RN+YpR model, that is, we assume in this section that, for every nucleotide x,

$$v_x = w_x = 1,$$

and that all YpR substitutions but those starting from CpG are excluded, that is,

$$r_G^C = r_T^A = r_C^G = r_A^T = r_G^T = r_C^A = 0,$$

and finally that the rates of CpG to CpA and CpG to TpG are equal, that is,

$$r_A^C = r_T^G =: r.$$

10.1. **On symbolic resolutions.** Theoretically, one has to solve an appropriate linear system related to the dynamics on the discrete circle with $N+2=4$ vertices, that is, of size $4^{N+2} = 256$. The translation invariance yields a system of size $m(N+2) = 70$, see section 8.1. Using the invariance with respect to both translations of the discrete circle and nucleotidic complementarity, the size of the linear system to be solved is further reduced to 42. (This is because 14 of the 70 classes that the invariance by translations induces, are invariant by the nucleotidic complementarity as well, the 56 other classes being grouped into pairs. We omit the details of this enumeration.)

We solved this 42×42 system symbolically, using Maple$^{\mathrm{TM}}$. We computed the full invariant distribution of $(X^I(s))_{s\geqslant 0}$ with $I := \{1,2,3,4\}$ but we only give the equilibrium frequencies of dinucleotides because these are the quantities of greatest interest. Checking the results by human computations seemed prohibitively time-consuming and tedious but, to confirm the validity of the formulas, we performed some tests. In particular, we compared the formulas with the following:

- The exact formulas obtained by human computations for YpR dinucleotides.
- The results obtained by numerically solving the system with MATLAB® for various settings of the parameters.
- The results of extensive Monte-Carlo simulations, usually with 10^8 runs for each setting of the parameters.

All these tests confirmed the values given below.

10.2. **Frequencies.** We first recall the values of the frequencies of nucleotides, deduced from previous sections.

Definition 19. *Introduce* $K_0(xy) = 4\,U(xy) + 2\,R(x) + Y(x) + R(y) + 2\,Y(y)$. *where*

$$U = \mathbf{1}_{TG} + \mathbf{1}_{CA} - 2\,\mathbf{1}_{CG}, \quad R = \mathbf{1}_A - \mathbf{1}_G, \quad Y = \mathbf{1}_T - \mathbf{1}_C.$$

Introduce

$$a(r) := \frac{3}{96+19r}, \quad b(r) := \frac{4}{32+10r}.$$

Finally, for every polynucleotide $x_1 \cdots x_k$, *define a function* $K(x_1 \cdots x_k)$ *by the relation*

$$F(x_1 \cdots x_k) =: \frac{1}{4^k}\left(1 + r\,\frac{K(x_1 \cdots x_k)}{32+10r}\right).$$

Theorem 8. (1) *For every nucleotide* x, $K(x)$ *does not depend on* r, *and*

$$K(x) = 2\,(R(x) + Y(x)).$$

Hence $K(A) = K(T) = 2$ *and* $K(C) = K(G) = -2$.
(2) *For every YpR dinucleotide* xy, $K(xy)$ *does not depend on* r, *and* $K(xy) =$

$K_0(xy)$. Hence $K(CG) = -10, \quad K(CA) = K(TG) = 4, \quad K(TA) = 2$.
(3) *For every dinucleotide, $K(xy) \to K_0(xy)$ when $r \to 0$.*

Part (3) reads as

$$K_0(GG) = K_0(CC) = -3, \quad K_0(TT) = K_0(AA) = 3,$$

$$K_0(AG) = K_0(CT) = 1, \quad K_0(TC) = K_0(GA) = -1,$$

and

$$K_0(AC) = K_0(GT) = 0, \quad K_0(AT) = 4, \quad K_0(GC) = -4.$$

Here is a consequence of theorem 8.

Proposition 18. *The nucleotides C and G are always less frequent than A and T. More precisely, for every positive r,*

$$20\% \leqslant P(C) = P(G) < 25\% < P(A) = P(T) \leqslant 30\%.$$

Furthermore, the dinucleotides CG and TA are repulsive and the dinucleotides CA and TG are attractive, in the sense that

$$F(CG) \leqslant F(C)\,F(G), \quad F(TA) \leqslant F(T)\,F(A),$$

and

$$F(CA) \geqslant F(C)\,F(A), \quad F(TG) \geqslant F(T)\,F(G).$$

Definition 20. *For every dinucleotide xy, introduce $K_1(xy)$ as*

$$K(xy) =: K_0(xy) + r\,K_1(xy).$$

Theorem 9. *For every dinucleotide xy, $K_1(xy)$ assumes one of the five values 0, $\pm a(r)$, and $\pm b(r)$. More precisely,*

$$K_1(xy) = a(r)\,(R(x)\,R(y) + Y(x)\,Y(y)) + b(r)\,R(x)\,Y(y).$$

As regards, for instance, the dinucleotides Ax, this means that

$$K_1(AA) = a(r),\ K_1(AC) = -b(r),\ K_1(AG) = -a(r),\ K_1(AT) = b(r).$$

Going back to frequencies, this reads as

$$\begin{aligned}
F(AA) &= \frac{1}{16}\left(1 + \frac{r}{32+10r}\left(3 + \frac{3r}{96+19r}\right)\right),\\
F(AC) &= \frac{1}{16}\left(1 + \frac{r}{32+10r}\left(0 - \frac{4r}{32+10r}\right)\right),\\
F(AG) &= \frac{1}{16}\left(1 + \frac{r}{32+10r}\left(1 - \frac{3r}{96+19r}\right)\right),\\
F(AT) &= \frac{1}{16}\left(1 + \frac{r}{32+10r}\left(4 + \frac{4r}{32+10r}\right)\right).
\end{aligned}$$

Similar formulas are available for the 12 other dinucleotides.

Part C Coupling

11. Coupling from the past

A consequence of the construction of the previous sections is that we can simulate the restriction of the dynamics on $\mathbb{Z}$ to any finite interval of sites of length n, without truncation errors due to neglecting the influence of remote sites. One adds a site to the left and a site to the right, one performs simulations for the system on these $n+2$ sites, and the projection on the n original sites yields the desired simulation.

In this section, we show how the coupling-from-the-past (CFTP) methodology of Propp and Wilson [10] can be applied in our context. Our motivation is two-fold. First, estimates about the coupling times automatically yield estimates on the speed of convergence of the dynamics to the stationary distribution. In our context, this applies to the speed of convergence of the distribution of $\Pi^{a,b}(X(s))$ to $\Pi^{a,b}(\mu)$. Second, the CFTP technique allows to sample exactly from $\Pi^{a,b}(\mu)$. Despite the results of the previous sections, which show that the obtention of exact expressions of $\Pi^{a,b}(\mu)$ amounts to the inversion of a linear system, this task becomes computationally infeasible as soon as the number $b-a+1$ of sites is large, say greater than 6. Hence, Monte-Carlo simulations are still useful, if only to confirm the results obtained by inverting the linear system!

In the whole section, we fix $I = \{a, \dots, b\}$.

11.1. **Coupling events.** We first define the notions of coupling events and locked sites.

Definition 21 (Coupling events)**.** *We say that a coupling event occurs at site i and times (s_1, s_2, s_3) if the following assertions hold.*

- *$s_1 > s_2$ and $s_3 > s_2$,*
- *$-s_1$ belongs to $\mathcal{U}^A_{*i} \cup \mathcal{U}^G_{*i} \cup \mathcal{V}^A_{*i} \cup \mathcal{V}^G_{*i}$,*
- *$-s_3$ belongs to $\mathcal{U}^C_{i*} \cup \mathcal{U}^T_{i*} \cup \mathcal{V}^C_{i*} \cup \mathcal{V}^T_{i*}$,*
- *$-s_2$ belongs to $\bigcup_{z\in\mathcal{A}} \mathcal{U}^z_i$,*
- *$(-s_1, -s_2) \cap \bigcup_{z\in\mathcal{A}} \mathcal{U}^z_{*i} \cup \mathcal{V}^z_{*i}$ is empty,*
- *$(-s_3, -s_2) \cap \bigcup_{z\in\mathcal{A}} \mathcal{U}^z_{i*} \cup \mathcal{V}^z_{i*}$ is empty.*

Definition 22 (Locked sites)**.** *We say that the site i is locked at times (u, v) if, for every times s and t such that $s \leqslant -u$ and $t \geqslant -v$, the set*

$$\Phi_s(\mathcal{A}^I, \xi, i, t)$$

contains but one single element. In other words, $\Phi_s(\mathbf{x}, \xi, i, t) = \Phi_s(\mathbf{x}', \xi, i, t)$ for every initial configurations $\mathbf{x}$ and $\mathbf{x}'$ in $\mathcal{A}^I$.

Recall that Φ_s is introduced in section 6. Definitions 21 and 22 both involve I, through the definitions of the sites $*i$ and $i*$. Our next proposition shows that the notions of coupling event and locked site are related. The proofs of the results of the sections 11.1 to 11.3 are postponed to the sections 11.4 to 11.7.

Proposition 19. *If a coupling event occurs at site i and times (s_1, s_2, s_3), then i is locked at times (s_4, s_2), with $s_4 := \max(s_1, s_3)$.*

In words, for every initial condition imposed before time $-\max(s_1, s_3)$, the ith coordinate is the same after time $-s_2$. A consequence is that all the trajectories of the process have coalesced as far as site i is concerned.

11.2. **Locking times.** We wish to estimate, or at least control, the time before a given collection of sites becomes locked. We start with one site.

Definition 23 (Locking times)**.** *Let T_i be defined as the supremum of the times s_4 such that there exists three times s_1, s_2 and s_3, with the following properties: $s_2 \geqslant 0$, $s_4 = \max(s_1, s_3)$, and a coupling event occurs at site i and at times (s_1, s_2, s_3).*

We now define additional numerical parameters.

Definition 24 (Combined rates)**.** *Introduce*

$$\kappa_R := \sum_{z=A,G} \min(c_z, v_z), \quad \kappa_Y := \sum_{z=T,C} \min(c_z, v_z), \quad \kappa := \kappa_R + \kappa_Y,$$

and

$$\nu_R := \sum_{z=A,G} (v_z - c_z)^+, \quad \nu_Y := \sum_{z=T,C} (v_z - c_z)^+, \quad \nu := \nu_R + \nu_Y.$$

Note that $\kappa_R + \nu_R = v_A + v_G$ and $\kappa_Y + \nu_Y = v_C + v_T$. Finally, let

$$\alpha := t_Y\, t_R = \frac{(v_A + v_G) \times (v_C + v_T)}{(v_A + v_G + v_C + v_T)^2}.$$

These parameters allow us to control the distribution of the locking times, as follows.

Proposition 20. *Each locking time T_i is stochastically dominated by the random variable $H_1 + \cdots + H_Z$, where $(H_k)_{k\geqslant 1}$ is a sequence of i.i.d. Gamma $(3, \kappa)$ random variables, and $Z \geqslant 1$ is a geometric random variable with parameter α, independent from $(H_k)_{k\geqslant 1}$.*
In particular, T_i is almost surely finite and $\mathbb{E}(T_i) \leqslant 3/(\kappa\alpha)$.

We now proceed to bound the tail of T_i. For each $k \geqslant 1$, the distribution of $H_1 + \ldots + H_k$ is Gamma $(3k, \kappa)$.

Lemma 2. *Let $n_\alpha = -4\log(1-\alpha)/\alpha$, hence n_α is finite and $n_\alpha \geqslant 4$. For every nonnegative integer N, $\mathbb{P}(H_1 + \cdots + H_Z \geqslant N n_\alpha/\kappa) \leqslant 2\,(1-\alpha)^N$.*

Our interest lies in the locking times of whole intervals, defined below.

Definition 25 (Locking times of intervals)**.** *The locking time $T_{a,b}$ of the sites in the interior of the interval $I = \{a, \ldots, b\}$ is*

$$T_{a,b} = \max\{T_i\,;\, a+1 \leqslant i \leqslant b-1\}.$$

Proposition 21. *Introduce the integer $k_{a,b} = \left\lceil \frac{b-a-1}{3} \right\rceil$. For every integer N,*

$$\mathbb{Q}(T_{a,b} \geqslant N n_\alpha/\kappa) \leqslant 3\left[1 - \left(1 - 2\,(1-\alpha)^N\right)^{k_{a,b}}\right].$$

Remark 1. *The bound above has the nice property that it does not involve the rates r_z^y of the YpR mutations except, through the c_z, when they are negative.*

More readable forms of proposition 21 might be proposition 22 and corollary 3 below.

Proposition 22 (Control of the locking times)**.** *For every t,*

$$\mathbb{Q}((\alpha\kappa)\,T_{a,b} \geqslant 4\log(6\,k_{a,b}) - \log(1-\alpha) + t) \leqslant \exp(-t/4).$$

Definition 26. *Let $T_{(n)}$ denote the locking time of n consecutive sites.*

For instance $T_{(n)}$ is distributed as $T_{0,n+1}$. Note that $3k_{0,n+1} \leqslant n+2$ and that $\alpha \leqslant \frac{1}{4}$, hence $4\log 2 - \log(1-\alpha) \leqslant 6\log 2$. This yields the following corollary.

Corollary 3. *For every t,*

$$\mathbb{Q}((\alpha\kappa)\,T_{(n)} \geqslant t + \log(n+2) + 6\log 2) \leqslant \exp(-t/4).$$

11.3. **Consequences.** The now traditional Propp-Wilson method induces that, whatever the initial condition $\mathbf{x}$ in $\mathcal{A}^I$ with $I = \{a, \dots, b\}$, for every $J \subset I$,

$$\left[\varphi^{a,b}_{-T}(\mathbf{x}, \xi, j, 0)\right]_{j\in J}, \quad \text{where } T = \max_{j\in J} T_j,$$

is distributed according to the projection of the stationary distribution μ_I on $\mathcal{A}^J$. In particular, the distribution of $\Pi^{a+1,b-1}\left(\varphi^{a,b}_{-T_{a,b}}(\mathbf{x}, \xi)(0)\right)$ is $\Pi^{a+1,b-1}(\mu_{a,b})$, that is, by proposition 7, $\Pi^{a+1,b-1}(\mu)$. We state this as a proposition.

Proposition 23. *For every $a \leqslant b$, $\Pi^{a,b}(\mu)$ is the distribution of*

$$\Pi^{a,b}\left(\varphi^{a-1,b+1}_{-T_{a-1,b+1}}(\mathbf{x}, \xi)(0)\right).$$

Hence proposition 22 yields the result below.

Proposition 24. *For every t, the distance in total variation between the distribution of $\Pi^{a,b}(X^{\mathbf{x}}(s))$ at time $s = (t + t_{a-1,b+1} - \log(1-\alpha))/(\alpha\kappa)$, and the limiting distribution $\Pi^{a,b}(\mu)$, is at most* $\exp(-t/4)$.

11.4. **Proof of proposition 19.** We shall in fact prove the following assertion: assume that site i is locked at times (s_1, s_2, s_3), then, for every $s \leqslant -s_4$ and $t \geqslant -s_2$, the three sets below are singletons:

$$\varrho(\Phi_s(\mathcal{A}^I, \xi, *i, t)), \quad \Phi_s(\mathcal{A}^I, \xi, i, t), \quad \eta(\Phi_s(\mathcal{A}^I, \xi, i*, t)).$$

We first check the claim when $t = -s_2$. By the definition of s_1, at time $-s_1$, either a move of type U occurs at site $*i$, yielding an A or a G unconditionally, or a move of type V occurs, namely a transversion to a purine, yielding a purine if site i was not already occupied by a purine. As a consequence, the set $\varrho\left(\Phi_s(\mathcal{A}^I, \cdot, *i, -s_1)\right)$ is a singleton. Once again by the definitions, we ruled out the possibility that any move of type U or V occurred at site $*i$ between the times $-s_1$ and $-s_2$. Furthermore, moves of type W, R and Q, when applied to a purine, can only yield a (possibly different) purine. This implies that the set $\varrho(\Phi_s(\mathcal{A}^I, \xi, *i, -s_2))$ is a singleton as well. The same argument applies symmetrically to the set $\eta(\Phi_s(\mathcal{A}^I, \xi, i*, -s_2))$. As

regards $\Phi_s(\mathcal{A}^I, \xi, i, -s_2)$, this is a singleton since a move of type U occurs at site i at time $-s_2$. Lemma 1 above shows that, for every x in $\mathcal{A}^I$, the values of

$$\varrho(\Phi_s(\mathbf{x}, \xi, *i, t)), \quad \Phi_s(\mathbf{x}, \xi, i, t), \quad \eta(\Phi_s(\mathbf{x}, \xi, i*, t)),$$

for every $t \geqslant -s_2$, are completely determined by ξ and by the values of

$$\varrho(\Phi_s(\mathbf{x}, \xi, *i, -s_2)), \quad \Phi_s(\mathbf{x}, \xi, i, -s_2), \quad \eta(\Phi_s(\mathbf{x}, \xi, i*, -s_2)).$$

Since these values are the same for every $\mathbf{x}$ in $\mathcal{A}^I$, so is the case for $\varrho(\Phi_s(\cdot, \xi, *i, t))$, $\Phi_s(\cdot, \xi, i, t)$ and $\eta(\Phi_s(\cdot, \xi, i*, t))$ for every $t \geqslant -s_2$. This concludes the proof.

11.5. **Proof of proposition 20.** Recall the convention that $\sup \emptyset = -\infty$. Define $M_0 := 0$, and, inductively for $k \geqslant 1$, $-L_k := \sup\,(-\infty, -M_{k-1}) \cap \bigcup_{z \in \mathcal{A}} \mathcal{U}_i^z$; $-U_k := \sup\,(-\infty, -L_k) \cap \bigcup_{z \in \mathcal{A}} \mathcal{U}_{*i}^z \cup \mathcal{V}_{*i}^z$; $-V_k := \sup\,(-\infty, -L_k) \cap \bigcup_{z \in \mathcal{A}} \mathcal{U}_{i*}^z \cup \mathcal{V}_{i*}^z$; $-M_k := -\max(U_k, V_k)$.

Define K as the smallest integer $k \geqslant 1$ such that $-U_k$ belongs to $\mathcal{U}_{*i}^A \cup \mathcal{U}_{*i}^G \cup \mathcal{V}_{*i}^A \cup \mathcal{V}_{*i}^G$, and $-V_k$ belongs to $\mathcal{U}_{i*}^C \cup \mathcal{U}_{i*}^T \cup \mathcal{V}_{i*}^C \cup \mathcal{V}_{i*}^T$. Then, provided that K is finite, the integer K is such that a coupling event occurs at site i and times (U_K, L_K, V_K), hence $T_i \leqslant M_K$ as soon as K is finite. Furthermore, standard properties of Poisson processes and the independence of the Poisson processes that are associated to different sites show that the sequence

$$\big(L_k - M_{k-1}, U_k - L_k, V_k - L_k\big)_{k \geqslant 1}$$

is i.i.d., and that, for every given $k \geqslant 1$, $L_k - M_{k-1}$, $U_k - L_k$ and $V_k - L_k$ are mutually independent and exponentially distributed with parameters κ, $\kappa + \nu$, and $\kappa + \nu$ respectively. Finally, $K \geqslant 1$ is independent from $(L_k - M_{k-1}, U_k - L_k, V_k - L_k)_{k \geqslant 1}$, and geometrically distributed with parameter α and expectation $1/\alpha$. Writing

$$M_k - M_{k-1} = L_k - M_{k-1} + \max(U_k - L_k, V_k - L_k),$$

and recalling that $T_i \leqslant M_K$, one gets $\mathbb{E}(T_i) \leqslant \frac{1}{\alpha}\left(\frac{1}{\kappa} + \frac{3}{2(\kappa+\nu)}\right)$. Simpler upper bounds obtain as follows. Since $\kappa + \nu \geqslant \kappa$, the distribution of the random variable $\max(U_k - L_k, V_k - L_k)$ is (crudely) dominated by the distribution of the sum of two independent κ exponential random variables, hence the distribution of T_i is dominated by the distribution of the sum of three independent κ exponential random variables. One sees that $\mathbb{E}(T_i) \leqslant \frac{5}{2\alpha\kappa}$.

11.6. **Proof of lemma 2.** By the homogeneity of the Gamma distributions, we can, and we will, assume that $\kappa = 1$. The nonnegativity of the random variables $(H_k)_k$ implies that, for every integer $k \geqslant 0$ and every real number $t \geqslant 0$,

$$\mathbb{P}(H_1 + \cdots + H_Z \geqslant t) \leqslant \mathbb{P}(Z \geqslant k+1) + \mathbb{P}(H_1 + \cdots + H_k \geqslant t).$$

The first term on the right hand side is $(1-\alpha)^k$. By Cramér's bound and the value of the Laplace transform of the standard exponential distribution, evaluated at $0 \leqslant u < 1$, the second term is at most $(1-u)^{-3k} \exp(-ut)$. Assume that $t = n_\alpha N$ for an integer N, and choose $u = \alpha$ and $k = N$. Then the proof is complete, since for these values, $(1-\alpha)^k = (1-u)^{-3k} \exp(-ut) = (1-\alpha)^N$.

11.7. **Proof of proposition 21.** Write $I = I_0 \cup I_1 \cup I_2$, where I_j collects the sites in I that are equal to j modulo 3. For $j = 0$, 1 and 2, let $T_{(j)} = \max\{T_i\,;\, i \in I_j\}$.

By the independence properties of the collection $(T_i)_i$, for each j, the random variables $(T_i)_{i\in I_j}$ are i.i.d. Furthermore, for each j, $T_{(j)}$ involves at most $k_{a,b}$ sites in I. Hence, for every nonnegative t,

$$\mathbb{Q}(T_{(j)} \geqslant t) \leqslant 1 - (1 - \mathbb{Q}(T_i \geqslant t))^{k_{a,b}}.$$

Since $T_{a,b}$ is the maximum of the three random variables $T_{(j)}$,

$$\mathbb{Q}(T_{a,b} \geqslant t) \leqslant \mathbb{Q}(T_{(0)} \geqslant t) + \mathbb{Q}(T_{(1)} \geqslant t) + \mathbb{Q}(T_{(2)} \geqslant t).$$

One concludes, using the upper bound of $\mathbb{Q}(T_i \geqslant t)$ in lemma 2 above.

11.8. **Practical issues.** For the sake of brevity, we do not provide here the details of how the CFTP method of simulation is implemented in practice (see [4] for more details), but we mention that the properties of the coupling times defined above allow for efficient algorithmic schemes of coalescence detection.

References

[1] Peter F. Arndt (2004). *Identification and measurement of neighbor dependent nucleotide substitution processes*. Lecture Notes in Informatics P53, 227–234.
Available at `evogen.molgen.mpg.de/publications`.

[2] Peter F. Arndt, Christopher B. Burge, and Terence Hwa (2003). DNA sequence evolution with neighbor-dependent mutation. *Journal of Computational Biology* 10, 313–22.
Available at `arXiv:physics/0112029`.

[3] Peter F. Arndt and Terence Hwa (2005). Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics* 21, 2322–2328.
Available at `arXiv:q-bio.GN/0501018`.

[4] Jean Bérard, Jean-Baptiste Gouéré, and Didier Piau (2005). Solvable models of neighbor-dependent substitution processes. *ArXiv e-print.*
Available at `arXiv:math.PR/0510034`.

[5] Athel Cornish-Bowden (1985). Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Research* 13 (9), 3021–3030.
Available at `www.chem.qmul.ac.uk/iubmb/misc/naseq.html`.
Also published as: Nomenclature Committee of the International Union of Biochemistry (1986). Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences: Recommendations 1984. *Proceedings of the National Academy of Sciences of the USA* 83 (1), 4–8.
Available at `www.pnas.org/cgi/reprint/83/1/4`.

[6] Laurent Duret and Nicolas Galtier (2000). 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.
Available at `mbe.oxfordjournals.org/cgi/content/full/17/11/1620`.

[7] James A. Lake (1987). A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. *Molecular Biology and Evolution* 4, 167–191.
Available at `mbe.oxfordjournals.org/cgi/content/abstract/4/2/167`.

[8] Thomas M. Liggett (2005). *Interacting particle systems*. Reprint of the 1985 original, Springer, Berlin.

[9] William C. Navidi and Lynn Beckett-Lemus (1992). The effect of unequal transversion rates on the accuracy of evolutionary parsimony. *Molecular Biology and Evolution* 9, 1163–1175.
Available at `mbe.oxfordjournals.org/cgi/content/abstract/9/6/1163`.

[10] James G. Propp and David B. Wilson (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. Proceedings of the Seventh International Conference on Random Structures and Algorithms (Atlanta, GA, 1995). *Random Structures Algorithms* 9, no. 1–2, 223–252.
Available at `dbwilson.com/exact`.

[11] Joseph J. Rotman (1995). *An introduction to the theory of groups*. Fourth edition. Graduate Texts in Mathematics, 148. Springer-Verlag, New York.

[12] Andrey Rzhetsky and Masatoshi Nei (1995). Tests of applicability of several substitution models for DNA sequence data. *Molecular Biology and Evolution* 12, 131–151.
Available at `mbe.oxfordjournals.org/cgi/content/abstract/12/1/131`.

[13] Simon Whelan, Pietro Liò, and Nick Goldman (2001). Molecular phylogenetics: state-of-the art methods for looking into the past. *Trends in Genetics* 17 (5), 262–272.
Available at `doi:10.1016/S0168-9525(01)02272-7`.

(Jean Bérard) Institut Camille Jordan - UMR 5208, Université Claude Bernard Lyon 1, 43 boulevard du 11 novembre 1918, 69622 Villeurbanne, France; Université de Lyon, 69003 Lyon, France.
e-mail: `jean.berard@univ-lyon1.fr`

(Jean-Baptiste Gouéré) Laboratoire MAPMO - UMR 6628, Université d'Orléans, B.P. 6759, 45067 Orléans Cedex 2, France.
e-mail: `Jean-Baptiste.Gouere@univ-orleans.fr`

(Didier Piau) Institut Fourier - UMR 5582, Université Joseph Fourier Grenoble 1, 100 rue des Maths, BP 74, 38402 Saint Martin d'Hères, France.
e-mail: `Didier.Piau@ujf-grenoble.fr`