J. Appl. Prob. **53,** 315–326 (2016)
*doi:*10.1017/jpr.2016.2

A STATIONARY DISTRIBUTION ASSOCIATED TO A SET OF LAWS WHOSE INITIAL STATES ARE GROUPED INTO CLASSES. AN APPLICATION IN GENOMICS

SERVET MARTÍNEZ,* *Universidad de Chile*

Abstract

Let $\mathcal{I}$ be a finite set and $\mathcal{S}$ be a nonempty strict subset of $\mathcal{I}$ which is partitioned into classes, and let $C(s)$ be the class containing $s \in \mathcal{S}$. Let $(\boldsymbol{P}_s : s \in \mathcal{S})$ be a family of distributions on $\mathcal{I}^{\mathbb{N}}$, where each $\boldsymbol{P}_s$ applies to sequences starting with the symbol s. To this family, we associate a class of distributions $\mathbb{P}^{(\pi)}$ on $\mathcal{I}^{\mathbb{N}}$ which depends on a probability vector π. Our main results assume that, for each $s \in \mathcal{S}$, $\boldsymbol{P}_s$ regenerates with distribution $\boldsymbol{P}_{s'}$ when it encounters $s' \in \mathcal{S} \setminus C(s)$. From semiregenerative theory, we determine a simple condition on π for $\mathbb{P}^{(\pi)}$ to be time stationary. We give a similar result for the following more complex model. Once a symbol $s' \in \mathcal{S} \setminus C(s)$ has been encountered, there is a decision to be made: either a new region of type $C(s')$ governed by $\boldsymbol{P}_{s'}$ starts or the region continues to be a $C(s)$ region. This decision is modeled as a random event and its probability depends on s and s'. The aim in studying these kinds of models is to attain a deeper statistical understanding of bacterial DNA sequences. Here $\mathcal{I}$ is the set of codons and the classes $(C(s) : s \in \mathcal{S})$ identify codons that initiate similar genomic regions. In particular, there are two classes corresponding to the start and stop codons which delimit coding and noncoding regions in bacterial DNA sequences. In addition, the random decision to continue the current region or begin a new region of a different class reflects the well-known fact that not every appearance of a start codon marks the beginning of a new coding region.

Keywords: Markov chain; stationary distribution; regenerative process; Palm theory; genomics

2010 Mathematics Subject Classification: Primary 60J10; 60J20; 92D10; 92D20

1. Introduction

We propose a model in which sequences are segmented into different types of regions, each initiated by a particular class of start symbols. Consecutive regions are not allowed to be of the same type. The input to the model consists of the laws for the different types of regions and our aim is to establish a law for the global organization of such sequences. This problem is inspired by bacterial genomes where there are two types of regions: coding and noncoding. Start codons mark the sites where translation into a polypeptide sequence begins and stop codons define where the translation ends. So, stop codons define the starting points of noncoding regions. Hence, given the distribution of these two types of regions, we propose a law for the global organization of the genome.

Received 22 November 2013; revision received 18 August 2015.
* Postal address: Departamento Ingeniería Matemática and Centro Modelamiento Matemático, Universidad de Chile, UMI 2807 CNRS, Casilla 170-3, Correo 3, Santiago, Chile. Email address: smartine@dim.uchile.cl

The general setting is as follows. Let $\mathcal{I}$ be a finite set of symbols and $\mathcal{S} \subset \mathcal{I}$ be a strict subset of symbols which mark the beginning of specific regions in the infinite sequences in $\mathcal{I}^{\mathbb{N}}$. We assume $\mathcal{S}$ is partitioned into equivalence classes $(C(s): s \in \mathcal{S})$, each class defining a different type of region. We have as input a class of distributions $(\boldsymbol{P}_s : s \in \mathcal{S})$ on $\mathcal{I}^{\mathbb{N}}$. We say that the law $\boldsymbol{P}_s$ governs a region starting with s and this region is of type $C(s)$. We associate to this class a family of distributions $\mathbb{P}^{(\pi)}$ depending on a probability vector $\pi = (\pi_s : s \in \mathcal{S})$ and this family holds candidates for modeling the global distribution on the set of sequences. The selection of a distinguished distribution $\mathbb{P}^{(\pi^*)}$ will be made under the hypothesis of regeneration and by imposing stationarity.

The main results in the general setting are given in Sections 3 and 4. In Section 3 we assume that the laws $(\boldsymbol{P}_s : s \in \mathcal{S})$ have the following regenerative structure. If we start at $s \in \mathcal{S}$, the sequence of letters evolves with the distribution $\boldsymbol{P}_s$ until T^1 which is the time (or site) at which a state $s^1 \in \mathcal{S} \setminus C(s)$ is first encountered. We assume the law restarts at T^1 with law $\boldsymbol{P}_{s^1}$ until time T^2 when it first reaches $s^2 \in \mathcal{S} \setminus C(s^1)$, and so on. The time stationarity of the distribution $\mathbb{P}^{(\pi)}$ is studied through the chain of states $\{s^1, s^2, \ldots\}$ at times $\{T^1, T^2, \ldots\}$. By using results from regenerative processes and Palm theory, we are able to prove that $\mathbb{P}^{(\pi)}$ is time stationary if and only if $(\pi_s / \boldsymbol{E}_s(T^1) : s \in \mathcal{S})$ is invariant for this chain (Theorem 3.1). In Section 4 we consider a richer model. Here, a choice must be made at each site where a region of type C encounters a symbol $s' \notin C$: either it starts a new region governed by $\boldsymbol{P}_{s'}$ or it continues the current region of type C. We are able to treat this model by imposing a natural regenerative structure at times where a new region starts, and an analogous result to Theorem 3.1 can be stated.

For applications of our results to bacterial genomes, one takes the alphabet $\mathcal{I} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}^3$, which consists of 64 triplets of the bases {A, C, G, T}. Each such triplet is called a codon. The set of initial symbols is $\mathcal{S} = \{\mathrm{ATG}, \mathrm{GTG}, \mathrm{TTG}, \mathrm{TAA}, \mathrm{TAG}, \mathrm{TGA}\}$ which is partitioned into two classes. The triplets {ATG, GTG, TTG} constitute the class of start codons for coding regions while the other triplets {TAA, TAG, TGA} form the class of stop codons which mark the end points of coding regions and which are essentially the start points of noncoding regions.

Since the distinguished distribution $\mathbb{P}^{(\pi^*)}$ comes up by assuming regeneration and imposing stationarity, the need arises to argue about the validity of these properties on codon sequences. This is done by referring to some empirical statistical analyses of annotated genome sequences and by taking account of some of their theoretical consequences.

Regeneration. One might be tempted to think that the regenerative framework we have imposed in Section 4 is too strong. But in a recent joint paper [10] it was shown that the sequence of codons marking the beginnings of regions of annotated bacterial genomes is a homogeneous Markov chain, and this is consistent with one of the main consequences of this regenerative framework.

Stationarity. We show in Proposition 5.1 and Proposition 5.2 in Section 5 that Chargaff's second parity rule (CSPR) implies time stationarity of nucleotide and codon sequences and, when this rule is only assumed to be valid for n-tuples, then time stationarity holds for $(n/3-1)$-tuples of codon sequences.

So, the basis of arguing for stationarity is CSPR. CSPR is an empirical law which was first observed experimentally in *Bacillus subtilis* [19] and confirmed in sufficiently long sequences for small polymer chains in [17]; more recent studies assessing its validity can be found in [1], [9], and [15]. We refer the reader to [8, Chapter 4] for a detailed discussion of CSPR. There, the author [8, p. 77] states that:

> The number of occurrences of each n-tuple of nucleotides in a given strand approaches that of its complementary n-tuple in the same strand. This symmetry is true for all long

https://doi.org/10.1017/jpr.2016.2 Published online by Cambridge University Press

> sequences at small n (e.g. $n = 1, 2, 3, 4, 5$). It extends to sets of n-tuples of higher-order n with increase in length of the sequence.

On the contrary, a number of mechanisms causing violation of CSPR in short polymers are described in [4]. For further discussion on various mechanisms that could support the origins of CSPR, see [20].

We are aware that our stationarity view could be somewhat surprising because there exists ample literature devoted to DNA analysis which, from the very beginning, asserts the nonstationarity of DNA sequences. For instance, see [2], or [14, p. 121], where it is stated that, 'The fact that DNA and protein sequences are nonstationary is overlooked on a large body of works' or in [21] where it is claimed that 'Standard statistical tests have been used to verify that the genomic sequences are indeed non stationary'. Hence, one of the aims of this work is to supply arguments for revisiting the question of stationarity of DNA sequences also in response to the subtle observation made in [12, p. 678], 'the assumption of stochastic stationarity is problematic in view of the great degree of local and global heterogeneity in nucleotide sequences'. In fact stationarity (or some degree of stationarity) in the structure of DNA sequences is a complex issue which requires deeper study.

Perhaps one of the things that speaks most strongly against stationarity is the existence of two types of regions. For instance in [5, p. 200], it is stated:

> In this paper, we address, in the light of non-stationary time-series analysis, the questions of (i) the existence of long-range correlations in DNA sequences and (ii) whether they are present in both coding and non-coding segments or only in the latter.

The analysis of this question constitutes one of the main issues of this work, the existence of regions with different statistical behaviors does not have to contradict stationarity as is shown in Theorem 3.1.

Some of the most relevant works in the statistical analysis of DNA sequences have been devoted to describing the statistical differences between regions of different types. In [13] and [16], it was discovered that noncoding sequences have long-range correlations while short-range correlations prevail in coding sequences. A detailed statistical discussion about the stationarity or nonstationarity of coding and noncoding regions can be found in [6] and [7]. We point out that in the general model studied in Section 4, we do not impose any constraint on the initial laws $(\boldsymbol{P}_s : s \in \mathcal{S})$. They may have long- or short-range correlations, or neither, and they do not need to satisfy any kind of Markovianness or stationarity.

We are aware that the models we introduce and study do not have the necessary degree of complexity to realistically describe nucleotide or codon organization in DNA sequences of bacterial genomes, but they do provide some insight for the analysis of some of their main features. We wish to emphasize that, with respect to genomic analysis, our study is focused on the statistical description of DNA sequences of bacterial genomes and for this purpose we use annotated bacteria. We are not proposing automatic algorithms to identify the coding and noncoding regions, rather we are attempting to better understand how a single strand is statistically organized.

Thus, when one considers double-stranded DNA, a more sophisticated analysis is required, because even if the second strand of nucleotides is the reverse complement of the primary strand, the interaction between both strands is extremely difficult to state in terms of genome organization. In [11] we have analyzed a theoretical probabilistic toy model of a DNA duplex inspired by the GLIMMER (Gene Locator and Interpolated Markov ModelER) algorithm. In particular, it offered a statistical analysis of overlaps between potential coding regions on the two

https://doi.org/10.1017/jpr.2016.2 Published online by Cambridge University Press

strands and illuminated a bias towards runs of consecutive coding regions within each strand, rather than consecutive coding regions alternating between the two strands. Although the toy model succeeded in capturing some probabilistic features appearing in annotated bacterial genomes, it needs to be drastically updated with more sophisticated automatic models for selecting potential coding regions in single DNA strands such as the one proposed here, as well as with regard to the statistical correlation between gene candidates on each strand. A more in-depth discussion of this program is beyond the scope of this work.

There is a large bibliography on the statistics of codon and nucleotide sequences of bacterial DNA. Here, we have only cited papers that have a direct relationship to the present study. For a more complete view of this body of work, the reader is directed to the references contained in those that we have cited.

2. Main concepts

From now on $\mathcal{I}$ denotes a finite alphabet. Let us fix some notation and basic concepts. Every countable set L is endowed with the discrete σ-field $\mathbb{S}(L) = \{K\colon K \subseteq L\}$. We set $\mathbb{N} = \{0, 1, 2, \ldots\}$ and $\mathbb{N}^* = \{1, 2, \ldots\}$.

Define $X_n\colon \mathcal{I}^{\mathbb{N}} \to \mathcal{I}, x \to x_n$ to be the nth coordinate function, so $X_n(x) = x_n$ for $x \in \mathcal{I}^{\mathbb{N}}$. For each $n \in \mathbb{N}$, $\mathcal{B}_n^X = \sigma(X_0, \ldots, X_n)$ denotes the σ-field generated by the coordinates $X_0, \ldots, X_n$. The product set $\mathcal{I}^{\mathbb{N}}$ is endowed with the σ-field $\mathcal{B}_\infty^X = \sigma(X_n\colon n \in \mathbb{N})$ generated by all the coordinates.

For $q \in \mathbb{N}$, the q-shift is

$$\Theta_q\colon \mathcal{I}^{\mathbb{N}} \to \mathcal{I}^{\mathbb{N}}, \qquad (\Theta_q x)_n = x_{n+q} \quad \text{for all } n \in \mathbb{N}. \tag{2.1}$$

Below we use the usual convention $\inf \varnothing = \infty$.

Let $\mathcal{J}\colon \mathcal{I} \to \mathbb{S}(\mathcal{I}), i \to \mathcal{J}(i)$, be a map. Then, the function $\mathcal{I}^{\mathbb{N}} \to \mathbb{S}(\mathcal{I}), x \to \mathcal{J}(x_0)$ is $\sigma(X_0)$-measurable. So, $\mathcal{J}(x_0)$ is a random set. Let $T_{\mathcal{J}} = \inf\{n > 0\colon X_n \in \mathcal{J}(X_0)\}$ be the random time to hit $\mathcal{J}$ in the future, so $T_{\mathcal{J}}(x) = \inf\{n > 0 : x_n \in \mathcal{J}(x_0)\}$. It defines the sequence of successive returns to $\mathcal{J}$,

$$T_{\mathcal{J}}^1 = T_{\mathcal{J}} \quad \text{and} \quad T_{\mathcal{J}}^{n+1} = T_{\mathcal{J}}^n + T_{\mathcal{J}} \circ \Theta_{T_{\mathcal{J}}^n} \quad \text{for } n \in \mathbb{N}^*. \tag{2.2}$$

Here $T_{\mathcal{J}}^n = \infty$ implies $T_{\mathcal{J}}^{n'} = \infty$ for $n' \geq n$. We will set $T_{\mathcal{J}}^0 = 0$. Sometimes, the dependence on X_0 is important and we then write $T_{\mathcal{J}(X_0)}$ instead of $T_{\mathcal{J}}$. It is easy to see that for every random set $\mathcal{J} = \mathcal{J}(X_0)$, the return time $T_{\mathcal{J}}(X_0) = \inf\{n > 0 : X_n \in \mathcal{J}(X_0)\}$ is a stopping time; that is, it satisfies $\{T_{\mathcal{J}} \leq n\} \in \mathcal{B}_n^X$ for all $n \in \mathbb{N}$.

Let $\mathcal{S}$ be a fixed nonempty strict subset of $\mathcal{I}$. Its elements are called initial symbols. We suppose that $\mathcal{S}$ is partitioned into equivalence classes and we denote by $C(s)$ the class containing $s \in \mathcal{S}$.

Let $(\boldsymbol{P}_s\colon s \in \mathcal{S})$ be a family of probability distribution on $\mathcal{I}^{\mathbb{N}}$. Under $\boldsymbol{P}_s$, the process $X = (X_n\colon n \in \mathbb{N})$ starts from s, so $\boldsymbol{P}_s(X_0 = s) = 1$. We denote by $\boldsymbol{E}_s$ the expectation defined by $\boldsymbol{P}_s$. Let $\pi = (\pi_s\colon s \in \mathcal{S})$ be a probability vector on $\mathcal{S}$, we denote by $\boldsymbol{P}_\pi = \sum_{s\in\mathcal{S}} \pi_s \boldsymbol{P}_s$ the distribution starting from π and $\boldsymbol{E}_\pi$ is the expectation defined by $\boldsymbol{P}_\pi$.

On the set $\{X_0 \in \mathcal{S}\}$, we define the random time

$$T = T_{\mathcal{S}\setminus C(X_0)} = \inf\{n > 0\colon X_n \in \mathcal{S} \setminus C(X_0)\}.$$

We assume the set $\mathcal{S} \setminus C(s)$ is attained in finite time $\boldsymbol{P}_s$-almost surely (a.s.),

$$\boldsymbol{P}_s(T < \infty) = 1 \quad \text{for all } s \in \mathcal{S}.$$

https://doi.org/10.1017/jpr.2016.2 Published online by Cambridge University Press

So, for $X_0 \in \mathcal{S}$, T is a stopping time which is finite $\boldsymbol{P}_s$-a.s. for all $s \in \mathcal{S}$. The sequence of successive returns is

$$T^1 = T \quad \text{and} \quad T^{n+1} = T^n + T \circ \Theta_{T^n} \quad \text{for } n \in \mathbb{N}^*.$$

By definition, we have $T^{n+1} < \infty$ which implies that $C(X_{T^{n+1}}) \neq C(X_{T^n})$. We will usually set $T^0 = 0$. (At the end of the next section T^0 will have another meaning, as explained there).

We will assume that

$$\boldsymbol{E}_s(T) < \infty \quad \text{for all } s \in \mathcal{S}. \tag{2.3}$$

We have $\boldsymbol{E}_s(T) = \sum_{n\in\mathbb{N}} \boldsymbol{P}_s(T > n)$. Then, every probability vector $\pi = (\pi_s : s \in \mathcal{S})$ on $\mathcal{S}$ defines the probability vector $(\pi_s \mathbb{E}_s(T)^{-1} \boldsymbol{P}_s(T > n) : s \in \mathcal{S}, n \in \mathbb{N})$ on $\mathcal{S} \times \mathbb{N}$. Hence, the following expression defines a distribution $\mathbb{P}^{(\pi)}$ on $\mathcal{I}^{\mathbb{N}}$ depending on π:

$$\mathbb{P}^{(\pi)}(B) = \sum_{s\in\mathcal{S}} \pi_s \mathbb{E}_s(T)^{-1} \left(\sum_{n\in\mathbb{N}} \boldsymbol{P}_s(T > n,\, B \circ \Theta_n^{-1}) \right) \quad \text{for all } B \in \mathcal{B}_\infty^X.$$

We denote by $\mathbb{E}^{(\pi)}$ its mean expected value.

Let us give a trajectorial description of $\mathbb{P}^{(\pi)}$. Let $X^{s,n} = (X_l : l \in \mathbb{N})$ be trajectories of the process X starting from $s \in \mathcal{S}$ with law $\boldsymbol{P}_s(\cdot \mid T > n)$. Then, the process $\mathbb{X}$ defined in $\mathcal{I}^{\mathbb{N}}$ by

$$\mathbb{X} = X^{s,n} \circ \Theta_n \quad \text{with probability } \pi_s \mathbb{E}_s(T)^{-1} \boldsymbol{P}_s(T > n),$$

has distribution $\mathbb{P}^{(\pi)}$. Then, T can be defined for $\mathbb{X}$ and it satisfies

$$\mathbb{P}^{(\pi)}(T < \infty) = \sum_{s\in\mathcal{S}} \pi_s \boldsymbol{E}_s(T)^{-1} \left(\sum_{n\in\mathbb{N}} \boldsymbol{P}_s(T > n, T < \infty) \right) = \sum_{s\in\mathcal{S}} \pi_s = 1.$$

Then, the sequence of times $(T^n : n \in \mathbb{N}^*)$ is defined in $\mathbb{X}$ and it is finite $\mathbb{P}^{(\pi)}$-a.s.

We seek the conditions such that some distribution $\mathbb{P}^{(\pi)}$ is time stationary; that is, it satisfies for all $m \in \mathbb{N}$, all $(i_0, \ldots, i_m) \in \mathcal{I}^{m+1}$, and all $t \in \mathbb{N}^*$, $\mathbb{P}^{(\pi)}(X_{k+t} = i_k, k = 0, \ldots, m) = \mathbb{P}^{(\pi)}(X_k = i_k, k = 0, \ldots, m)$. We note that this property is satisfied once it holds for $t = 1$. In the case $\mathbb{P}^{(\pi)}$ is time stationary we can extend it to the set of bi-infinite sequences $\mathcal{I}^{\mathbb{Z}}$ by putting

$$\mathbb{P}^{(\pi)}(X_{k+t} = i_k, k = 0, \ldots, m) = \mathbb{P}^{(\pi)}(X_k = i_k, k = 0, \ldots, m) \tag{2.4}$$

for all $t \in \mathbb{Z}$, $m \in \mathbb{N}$, and $(i_k : k = 0, \ldots, m) \in \mathcal{I}^{m+1}$.

3. Regeneration and conditions for time stationarity

In what follows we will assume that $(\boldsymbol{P}_s : s \in \mathcal{S})$ semiregenerates at times (T^n). This means that if we start from some $\boldsymbol{P}_\pi$, at time T^1 the process will restart with distribution $\boldsymbol{P}_{s_1}$, where $s_1 = X_{T^1}$, and in general at T^n the process will restart with distribution $\boldsymbol{P}_{s_n}$, where $s_n = X_{T^n}$. We note that this condition implies that in order to have trajectories $(X_n : n \in \mathbb{N})$ distributed with laws $(\boldsymbol{P}_s : s \in \mathcal{S})$ we only require a countable set of independent copies of the cycles $(X_0, \ldots, X_T)$ starting from each one of these laws.

https://doi.org/10.1017/jpr.2016.2 Published online by Cambridge University Press

Let us introduce the regeneration condition on $\boldsymbol{P}_\pi$ in a more formal way. Let $(i_0, i_1, \ldots, i_m) \in \mathcal{S} \times \mathcal{I}^m$. Define fixed times $(\tau^n : n \geq 0)$ with $\tau^0 = 0$ and $\tau^{n+1} = \inf\{k > \tau^n : k \leq m,\ i_k \in \mathcal{S} \setminus C(i_{\tau^n})\}$. Note that there is a finite $r \leq m$ such that $\tau^0, \ldots, \tau^r$ are finite and $\tau^{r+1} = \infty$. The regeneration property of $\boldsymbol{P}_\pi$ is as follows:

$$\boldsymbol{P}_\pi(X_l = i_l : l = 0, \ldots, m) = \sum_{s \in \mathcal{S}} \pi_s \mathbf{1}_{\{i_0 = s\}} \left(\prod_{k=0}^{r-1} \boldsymbol{P}_{i_{\tau^k}}(X_{\tau^k + t} = i_{\tau^k + t}, t = 1, \ldots, \tau^{k+1} - \tau^k) \right) \times \boldsymbol{P}_{i_{\tau^r}}(X_{\tau^r + t} = i_{\tau^r + t}, t = 1, \ldots, m - \tau^r). \tag{3.1}$$

Under the law $\boldsymbol{P}_\pi$, the regeneration property (3.1) implies that the sequence $(X_{T^n} : n \in \mathbb{N}^*)$ taking values in $\mathcal{S}$ satisfies

$$\boldsymbol{P}_\pi(X_{T^{k+1}} = s_{k+1} \mid X_{T^k} = s_k, \ldots, X_{T^1} = s_1) = \boldsymbol{P}_{s_k}(X_T = s_{k+1}).$$

Hence, $(X_{T^n} : n \in \mathbb{N}^*)$ is a Markov chain and its transition matrix $Q = (q_{ss'} : s, s' \in \mathcal{S})$ is given by $q_{ss'} = \boldsymbol{P}_s(X_T = s')$ for $s, s' \in \mathcal{S}$. Since $X_{T^{k+1}} \in \mathcal{S} \setminus C(X_{T^k})$, it follows that $q_{ss'} > 0$ implies $s' \notin C(s)$.

Moreover, we can check that under $\boldsymbol{P}_\pi$ the sequences of cycles $((X_{T^k}, \ldots, X_{T^{k+1}-1}) : k \in \mathbb{N})$ are independent. Note that the distribution of the cycle $(X_{T^k}, \ldots, X_{T^{k+1}-1})$ under $\boldsymbol{P}_\pi$ is the same as the distribution of the cycle $(X_0, \ldots, X_{T-1})$ under $\boldsymbol{P}_{\pi\, Q^k}$ (here $\pi\, Q^k$ is the evolution of π under Q^k). It holds that $(X_n : n \in \mathbb{N}, T^n : n \in \mathbb{N}^*)$ is a semiregenerative process because it satisfies $(X_{T^k + n} : n \in \mathbb{N})$ given $T^1, \ldots, T^k, X_{T^1}, \ldots, X_{T^k} = s$ is distributed as $(X_n : n \in \mathbb{N})$ under $\boldsymbol{P}_s$. For definition and properties of semiregenerative processes, see [3, Chapter VII.5].

Recall that a positive vector ρ is invariant for the transition matrix Q if it satisfies the set of equalities

$$\rho_s = \sum_{s' \in \mathcal{S}} \rho_{s'} q_{s's} \quad \text{for all } s \in \mathcal{S}.$$

There always exist invariant positive vectors. We will assume that Q is irreducible, so up to a multiplicative constant the invariant positive vector is unique. Then, there exists a unique probability vector noted $\pi^* = (\pi^*_s : s \in \mathcal{S})$ such that

$$(\pi^*_s \boldsymbol{E}_s(T)^{-1} : s \in \mathcal{S}) \quad \text{is an invariant vector of } Q.$$

Theorem 3.1. *It holds that $\mathbb{P}^{(\pi)}$ is time stationary if and only if $\pi = \pi^*$.*

Proof. Let π^* be the unique probability vector such that $(\pi^*_s \boldsymbol{E}_s(T)^{-1} : s \in \mathcal{S})$ is invariant for the stochastic matrix Q and so $\gamma = (\gamma_s : s \in \mathcal{S})$, given by

$$\gamma_s = \frac{\pi^*_s \boldsymbol{E}_s(T)^{-1}}{\sum_{s' \in \mathcal{S}} \pi^*_{s'} \boldsymbol{E}_{s'}(T)^{-1}},$$

is the unique invariant distribution for Q. Hence, under $\boldsymbol{P}_\gamma$ the sequence $(X_{T^n} : n \in \mathbb{N})$ is stationary so every X_{T^n} is distributed as γ. It is also straightforward to check that under $\boldsymbol{P}_\gamma$ the increments $(T^{n+1} - T^n : n \in \mathbb{N})$ are independent and equally distributed as is the case for the cycles $((X_n : T^k \leq n < T^{k+1}) : k \in \mathbb{N})$. Also, under $\boldsymbol{P}_\gamma$, $(X_n : n \in \mathbb{N}, T^n : n \in \mathbb{N}^*)$ is a regenerative process as defined in [3, Chapter VI.1]. In terms of Palm theory (see [3, Chapter VII.6]), $\boldsymbol{P}_\gamma$ is an event stationary distribution. By using [3, Theorem 2.1, Chapter VI.2] and [3, Lemma 3.2, Chapter V], it can be checked that $\mathbb{P}^{(\pi^*)}$ is the time stationary distribution associated to $\boldsymbol{P}_\gamma$ in [3, Theorem 6.4, Chapter VII.6].

https://doi.org/10.1017/jpr.2016.2 Published online by Cambridge University Press

When the distribution of T is aperiodic; that is, the greatest common divisor $\{l > 0 \colon \boldsymbol{P}_\gamma(T = l) > 0\} = 1$, from [3, Proposition 5.2(ii), Chapter VII.5] it follows that $\mathbb{P}^{(\pi^*)}$ is also the limiting distribution of the semiregenerative processes $(X_n \colon n \in \mathbb{N},\ T^n \colon n \in \mathbb{N}^*)$. This implies that $\mathbb{P}^{(\pi)}$ is time stationary only when $\pi = \pi^*$. If the distribution of T is periodic it suffices to take the mean along the period as in [3, Corollary 1.5(ii), Chapter VI.1] to obtain the same result. □

Since $\mathbb{P}^{(\pi^*)}$ can be defined on $\mathcal{I}^{\mathbb{Z}}$ by (2.4), we can define a stationary Markov renewal process $(T^n \colon n \in \mathbb{Z})$ with $T^0 = \sup\{T^n \colon T^n \leq 0\}$ such that X_{T^n} is distributed as π. (In this definition and the following equation the variable T^0 is not identically 0.) Then, by using Palm theory, we have

$$\mathbb{P}^{(\pi^*)}(T^0 = -n) = \boldsymbol{E}_\gamma(T)^{-1}\boldsymbol{P}_{\pi^*}(T > n),$$

$$\mathbb{P}^{(\pi^*)}(X_0 = i_k, k = 0, \ldots, m \mid T^0 = -n) = \boldsymbol{P}_\gamma(X_n = i_{k+n}, k = 0, \ldots, m \mid T > n).$$

4. Random model

We will modify the model studied in Sections 2 and 3 so as to capture some of the phenomena which occur in sequences of codons within real bacterial genomes. In Section 3 we assumed that a region of a new type starts when a state belonging to a different class is hit. Nevertheless, it is known from genome annotation that when a noncoding region hits a start codon, only a small proportion of these start codons mark the beginning of a new coding region. Some signals must be present in the neighborhood of the start codon to trigger a true beginning. Nowadays, there is a lot of active research being conducted into predicting the locations of genuine coding regions, focused either on lists of motifs or on their locations near starting codons. A recent discussion on this topic can be found in [18].

So, when a site containing a state belonging to a different class is hit, a decision must be made: either a new region starts, or this state is treated as though it does not belong to $\mathcal{S}$ and the sequence continues to be governed by the law of the current region. We will model this decision by a random choice whose distribution can depend on the state that is hit and on the initial state of the region. Toward this end, we use a sequence of independent Bernoulli random variables.

In this section we retain all the notions and assumptions made in Section 2.

From now on we assume for each pair of symbols $s_0 \in \mathcal{S}$ and $s \in \mathcal{S} \setminus C(s_0)$, there exists a well-defined probability $\varepsilon_{s_0}(s) \in [0, 1]$ that at s a new region starts when s is hit in a region initiated in s_0. We define $\varepsilon_{s_0}(s) = 0$ for $s_0 \in \mathcal{S}$ and $s \in C(s_0)$. We note that

$$F_{s_0} = \{s \in \mathcal{S} \colon \varepsilon_{s_0}(s) > 0\}.$$

We impose the irreducibility conditions: $F_{s_0} \neq \varnothing$ for all $s_0 \in \mathcal{S}$ and $\mathcal{S} = \bigcup_{s_0 \in \mathcal{S}} F_{s_0}$. Note that the case $\varepsilon_{s_0}(s) = 1$ for all pairs (s_0, s) such that $C(s_0) \neq C(s)$, means that when a region started at s_0 encounters a site containing s, then a new region governed by $\boldsymbol{P}_s$ will always start, which corresponds to the situation already examined in Section 3.

Let $\mathcal{E} = \{0, 1\}^{\mathcal{S}}$ and $e = (e^s \colon s \in \mathcal{S})$ be an element on $\mathcal{E}$. On $\mathcal{E}$ we define a family of Bernoulli product measures $(\boldsymbol{b}_{s_0} \colon s_0 \in \mathcal{S})$ given by

$$\boldsymbol{b}_{s_0}(e) = \prod_{s \in \mathcal{S} \colon e^s = 1} \varepsilon_{s_0}(s) \times \prod_{s \in \mathcal{S} \colon e^s = 0} (1 - \varepsilon_{s_0}(s)).$$

The measure $\boldsymbol{b}_{s_0}$ is supported by the set $\mathcal{E}_{s_0} = \{e \in \mathcal{E} \colon e^s = 1 \Rightarrow s \in F_{s_0}\}$. Let $(P^{\boldsymbol{b}_{s_0}} = \boldsymbol{b}_{s_0}^{\otimes \mathbb{N}^*} \colon s_0 \in \mathcal{S})$ be a family of product measures on $\mathcal{E}^{\mathbb{N}^*}$.

https://doi.org/10.1017/jpr.2016.2 Published online by Cambridge University Press

Let $\mathcal{K} = \mathcal{I} \times \mathcal{E}$ and (i, e) be an element in $\mathcal{K}$. The product space $\mathcal{K}^{\mathbb{N}} = (\mathcal{I} \times \mathcal{E})^{\mathbb{N}}$ is endowed with the product σ-field $\mathcal{B}_\infty^{\mathcal{X}} = \sigma(\mathcal{X}_n : n \in \mathbb{N})$. Let $u = (u_n : n \in \mathbb{N}) \in \mathcal{K}^{\mathbb{N}}$ and note that $u_n = (x_n, w_n)$ for $n \in \mathbb{N}$, so $w = (w_n : n \in \mathbb{N}) \in \mathcal{E}^{\mathbb{N}}$ with $w_n = (w_n^s : s \in \mathcal{S}) \in \mathcal{E}$.

Let $\mathcal{X}_n : \mathcal{K}^{\mathbb{N}} \to \mathcal{K}, u \in \mathcal{K}^{\mathbb{N}} \to \mathcal{X}_n(u) = u_n \in \mathcal{K}$ be the projection onto the nth component. We set $X_n : \mathcal{K}^{\mathbb{N}} \to \mathcal{I}, u \to x_n$ and $W_n : \mathcal{K}^{\mathbb{N}} \to \mathcal{E}, u \to w_n$. So, we can write $\mathcal{X}_n = (X_n, W_n)$. This is an abuse of notation because we will continue writing $X_n : \mathcal{I}^{\mathbb{N}} \to \mathcal{I}, x \in \mathcal{I}^{\mathbb{N}} \to x_n \in \mathcal{I}$ and also set $W_n : \mathcal{E}^{\mathbb{N}} \to \mathcal{E}, w \in \mathcal{E}^{\mathbb{N}} \to w_n \in \mathcal{E}$. We keep the same notation for the q-shift $\Theta_q : \mathcal{K}^{\mathbb{N}} \to \mathcal{K}^{\mathbb{N}}$, $(\Theta_q u)_n = u_{n+q}$, as the one introduced in (2.1) for $\mathcal{I}^{\mathbb{N}}$.

Let

$$\mathcal{V}^{s_0} = \{(s, e) \in F_{s_0} \times \mathcal{E}_{s_0} : e^s = 1\} \quad \text{and} \quad \mathcal{V} = \bigcup_{s_0 \in \mathcal{S}} \mathcal{V}^{s_0}.$$

It holds that $\mathcal{V}$ is a proper subset of $\mathcal{I} \times \mathcal{E}$ and will play the role of the set of starting states. For $(s, e) \in \mathcal{V}$, the class $\mathcal{C}(s, e)$ is defined to be

$$\mathcal{C}(s, e) = \{(s', e') \in \mathcal{V} : C(s) = C(s')\}.$$

On the set $\{\mathcal{X}_0 \in \mathcal{V}\}$, we define the random time

$$\mathcal{T} := T_{\mathcal{V} \setminus \mathcal{C}(\mathcal{X}_0)} = \inf\{n > 0 : \mathcal{X}_n \in \mathcal{V} \setminus \mathcal{C}(\mathcal{X}_0)\} = \inf\{n > 0 : \mathcal{X}_n \in \mathcal{V}, C(X_n) \neq C(X_0)\}$$

(it can take the value ∞). The time $\mathcal{T}$ is a stopping time for the natural sequence of σ-fields. As already stated for a random time in (2.2), we define the sequence of times

$$\mathcal{T}^1 = \mathcal{T} \quad \text{and} \quad \mathcal{T}^{n+1} = \mathcal{T}^n + \mathcal{T} \circ \Theta_{\mathcal{T}^n} \quad \text{for } n \in \mathbb{N}^*,$$

which are also stopping times. Note that $\mathcal{T}^{n+1}$ finite implies that $C(X_{\mathcal{T}^{n+1}}) \neq C(X_{\mathcal{T}^n})$. We set $\mathcal{T}^0 = 0$.

Let $(\boldsymbol{P}_s : s \in \mathcal{S})$ be a family of probability distribution on $\mathcal{I}^{\mathbb{N}}$ satisfying the conditions stated in Section 2: for all $s \in \mathcal{S}$, $\boldsymbol{P}_s(X_0 = s) = 1$, $\boldsymbol{P}_s(T < \infty) = 1$, and $\boldsymbol{E}_s(T) < \infty$. We emphasize that no regeneration property is assumed on this family.

For every probability vector π on $\mathcal{S}$ we define the following probability measure $\boldsymbol{P}_\pi^\dagger$ on $\mathcal{K}^{\mathbb{N}}$: for all $m \in \mathbb{N}$, $(i_0, \dots, i_m) \in \mathcal{I}^{m+1}$, and $(e_0, \dots, e_m) \in \mathcal{E}^{m+1}$,

$$\begin{aligned}\boldsymbol{P}_\pi^\dagger&(X_k = i_k, W_k = e_k, k = 0, \dots, m)\\ &= \sum_{s \in \mathcal{S}} \pi_s \, \mathbf{1}_{\{i_0 = s,\, e_0^s = 1\}} \, \boldsymbol{P}_s(X_k = i_k, k = 1, \dots, m) P^{\boldsymbol{b}_s}(W_k = e_k, k = 1, \dots, m).\end{aligned}$$

Recall that the set $\mathcal{S} \setminus C(s)$ is attained in finite time $\boldsymbol{P}_\pi$-a.s., so by applying the Borel–Cantelli lemma to the independent random variables $(W_n : n \in \mathbb{N}^*)$, we obtain

$$\boldsymbol{P}_s^\dagger(\mathcal{T} < \infty) = 1 \quad \text{for all } s \in \mathcal{S}.$$

Let $\boldsymbol{E}_s^\dagger$ be the expected value defined by $\boldsymbol{P}_s^\dagger$. The assumption (2.3) implies that

$$\boldsymbol{E}_s^\dagger(\mathcal{T}) < \infty \quad \text{for all } s \in \mathcal{S}. \tag{4.1}$$

Note that $\boldsymbol{P}_s^\dagger(X_{\mathcal{T}} \in F_{X_0}, W_{\mathcal{T}}^{X_{\mathcal{T}}} = 1) = 1$, so $\mathcal{T}$ models the time where a region of a new type starts, satisfying the conditions announced at the beginning of this section.

https://doi.org/10.1017/jpr.2016.2 Published online by Cambridge University Press

We will introduce a class of probability distributions $(\widehat{\boldsymbol{P}}_s^\dagger : s \in \mathcal{S})$ that semiregenerates at times $(\mathcal{T}^n)$ where the classes of a new type start. For this purpose let $((i_0, e_0), \dots, (i_m, e_m)) \in (\mathcal{S} \times \mathcal{E}) \times (\mathcal{I} \times \mathcal{E})^m$ be a finite sequence. Define a sequence of fixed times $(\tau^n : n \in \mathbb{N})$ by $\tau^0 = 0$, $\tau^{n+1} = \inf\{k > \tau^n : k \le m,\ (i_k, e_k) \in \mathcal{V} \setminus \mathcal{C}(i_{\tau^n}, e_{\tau^n})\}$. There exists $r \le m$ such that $\tau^0, \dots, \tau^r$ are finite and $\tau^{r+1} = \infty$. Then, inspired by (3.1), we define $\widehat{\boldsymbol{P}}_s^\dagger$ by

$$\begin{aligned}
&\widehat{\boldsymbol{P}}_s^\dagger(X_l = i_l, W_l = e_l : l = 0, \dots, m) \\
&\quad = \mathbf{1}_{\{i_0 = s,\, e_0^s = 1\}} \left(\prod_{k=0}^{r-1} \boldsymbol{P}_{i_{\mathcal{T}^k}}^\dagger (X_{\tau^k + t} = i_{\tau^k + t}, W_{\tau^k + t} = e_{\tau^k + t}, t = 1, \dots, \tau^{k+1} - \tau^k) \right) \\
&\qquad \times \boldsymbol{P}_{i_{\tau^r}}^\dagger (X_{\tau^r + t} = i_{\tau^r + t}, W_{\tau^r + t} = e_{\tau^r + t}, t = 1, \dots, m - \tau^r).
\end{aligned}$$

As usual we set $\widehat{\boldsymbol{P}}_\pi^\dagger = \sum_{s \in \mathcal{S}} \pi_s \widehat{\boldsymbol{P}}_s^\dagger$ and note that $\widehat{\boldsymbol{E}}_\pi^\dagger$ is the associated expectation. The times $(\mathcal{T}^n : n \in \mathbb{N}^*)$ are finite $\widehat{P}_\pi^\dagger$-a.s. We note that the sequence $(X_{\mathcal{T}^n} : n \in \mathbb{N}^*)$ is a Markov chain with transition matrix $Q^\dagger = (q_{ss'}^\dagger : s, s' \in \mathcal{S})$ given by

$$q_{ss'}^\dagger = \widehat{\boldsymbol{P}}_s^\dagger(X_{\mathcal{T}} = s') = \boldsymbol{P}_s^\dagger(X_{\mathcal{T}} = s') \quad \text{for all } s, s' \in \mathcal{S}.$$

By the definition of $\mathcal{T}$, we have $C(X_{\mathcal{T}^{k+1}}) \ne C(X_{\mathcal{T}^k})$, so $q_{ss'}^\dagger > 0$ implies that $C(s') \ne C(s)$. An invariant vector ρ for $Q^\dagger$ satisfies $\rho_s = \sum_{s' \in \mathcal{S}} \rho_{s'} q_{s's}^\dagger$ for all $s \in \mathcal{S}$. We have assumed that Q is irreducible and so $Q^\dagger$ is also irreducible. Then, up to a multiplicative constant the invariant vector is unique.

Assumption (4.1) implies that

$$\widehat{\boldsymbol{E}}_s^\dagger(\mathcal{T}) < \infty \quad \text{for all } s \in \mathcal{S}.$$

Then, every probability vector $\pi = (\pi_s : s \in \mathcal{S})$ defines the probability vector

$$(\widehat{\boldsymbol{E}}_s^\dagger(\mathcal{T})^{-1} \widehat{\boldsymbol{P}}_s^\dagger(\mathcal{T} > n) \pi_s : s \in \mathcal{S}, n \in \mathbb{N}) \quad \text{on } \mathcal{S} \times \mathbb{N}.$$

Hence, the following distribution is well defined on $\mathcal{K}^{\mathbb{N}}$:

$$\widehat{\mathbb{P}}^{\dagger(\pi)}(B) = \sum_{s \in \mathcal{S}} \pi_s \widehat{\boldsymbol{E}}_s^\dagger(\mathcal{T})^{-1} \left(\sum_{n \in \mathbb{N}} \widehat{\boldsymbol{P}}_s^\dagger(\mathcal{T} > n, B \circ \Theta_n^{-1}) \right) \quad \text{for all } B \in \mathcal{B}_\infty^{\mathcal{X}}, \tag{4.2}$$

where Θ_n is the n-shift on $\mathcal{K}^{\mathbb{N}}$.

Let us denote by $\pi^{\dagger *}$ the unique probability vector that satisfies

$$(\pi_s^{\dagger *} \widehat{\boldsymbol{E}}_s^\dagger(\mathcal{T})^{-1} : s \in \mathcal{S})$$

is invariant for $Q^\dagger$.

In a similar way as we did in Theorem 3.1, we can prove the following condition for time stationarity of $\mathbb{P}^{*\dagger}$.

Theorem 4.1. *It holds that $\mathbb{P}^{*\dagger(\pi)}$ is time stationary if and only if $\pi = \pi^{\dagger *}$.*

5. Stationarity and Chargaff's second parity rule

In the genomic setting, $L = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ is the set of nucleotides, $\mathcal{I} = L^3$ is the list of codons and the complement mapping $\varphi : \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\} \to \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ is given by $\varphi(\mathrm{A}) = T = \varphi^{-1}(\mathrm{A})$, $\varphi(\mathrm{C}) = \mathrm{G} = \varphi^{-1}(\mathrm{C})$. For a DNA sequence, CSPR means that the frequency of appearance of any k-tuple $(l_0, \dots, l_{k-1}) \in L^k$ is equal to the frequency of its reverse complement $(\varphi(l_{k-1}), \dots, \varphi(l_0))$.

https://doi.org/10.1017/jpr.2016.2 Published online by Cambridge University Press

Below, in the theoretical framework of CSPR, we show in the first part of the proof of Proposition 5.1 that CSPR implies that the probability distribution on a nucleotide sequence is time stationary. In the second part of the proof of Proposition 5.1, we prove that CSPR also implies that the probability distribution on a codon sequence is time stationary. This is the time stationarity property studied in Sections 3 and 4.

The validity of CSPR has been checked for k-tuples of nucleotides with small k but it extends to sets of k-tuples of higher-order k with an increase in length of the sequences, as mentioned in [8]. With this fact in mind, in Proposition 5.2, we state stationarity for the class of $(k-1)$-tuples in the nucleotide sequence, which implies stationarity for $(k/3-1)$-tuples of codons.

Let us supply CSPR in a general theoretical framework. Let L be an alphabet and $Y_n\colon L^{\mathbb{N}} \to L$ be the nth coordinate function: $Y_n(y) = y_n$ for $y \in L^{\mathbb{N}}$. Let $\varphi\colon L \to L$ be an involution, this means that φ is one-to-one and $\varphi^{-1} = \varphi$. Since φ is a bijection, we have $L = \{\varphi(h)\colon h \in L\}$.

Let $P_{L^{\mathbb{N}}}$ be a probability measure on $L^{\mathbb{N}}$. We say that $P_{L^{\mathbb{N}}}$ satisfies the CSPR with respect to φ if for all $m \in \mathbb{N}$, all $(l_0, \ldots, l_m) \in L^{m+1}$, and all $t \in \mathbb{N}$,

$$P_{L^{\mathbb{N}}}(Y_{k+t} = l_k, k = 0, \ldots, m) = P_{L^{\mathbb{N}}}(Y_{k+t} = \varphi(l_{m-k}), k = 0, \ldots, m). \tag{5.1}$$

We claim that (5.1) is satisfied if it holds for $t = 0$. That is, if for all $m \in \mathbb{N}$ and all $(l_0, \ldots, l_m) \in L^{m+1}$,

$$P_{L^{\mathbb{N}}}(Y_k = l_k, k = 0, \ldots, m) = P_{L^{\mathbb{N}}}(Y_k = \varphi(l_{m-k}), k = 0, \ldots, m). \tag{5.2}$$

In fact, from (5.2), we obtain

$$\begin{aligned} &P_{L^{\mathbb{N}}}(Y_k = h_k, k = 0, \ldots, t-1; Y_{t+k} = l_k, k = 0, \ldots, m; Y_{t+m+k} = c_k, k = 0, \ldots, t-1) \\ &\quad = P_{L^{\mathbb{N}}}(Y_k = \varphi(c_{t-1-k}), k = 0, \ldots, t-1; Y_{k+t} = \varphi(l_{m-k}), k = 0, \ldots, m; \\ &\qquad Y_{k+t+k} = \varphi(h_{t-1-k}), k = 0, \ldots, t-1). \end{aligned}$$

Hence, by summing on $(h_0, \ldots, h_{t-1}) \in L^t$ and $(c_0, \ldots, c_{t-1}) \in L^t$, we obtain (5.1).

Let $d \in \mathbb{N}^*$ be fixed. In our results we shall also consider the following setting of d-mers, where $\mathcal{I} := L^d$ is a new alphabet. We take the following transformation: $\zeta\colon L^{\mathbb{N}} \to \mathcal{I}^{\mathbb{N}}$, $y \to x = \zeta y$ with $x_n = (\zeta y)_n = (y_{dn}, \ldots, y_{d(n+1)-1})$. So $P_{L^{\mathbb{N}}} \circ \zeta^{-1}$ is the induced law by $P_{L^{\mathbb{N}}}$ on $\mathcal{I}^{\mathbb{N}}$.

Proposition 5.1. *If $P_{L^{\mathbb{N}}}$ verifies the CSPR then $P_{L^{\mathbb{N}}}$ and $P_{L^{\mathbb{N}}} \circ \zeta^{-1}$ are time stationary.*

Proof. Assume that $P_{L^{\mathbb{N}}}$ satisfies the CSPR. For all $m \in \mathbb{N}^*$, we have

$$\begin{aligned} &P_{L^{\mathbb{N}}}(Y_{k+1} = l_k, k = 0, \ldots, m) \\ &\quad = \sum_{h \in L} P_{L^{\mathbb{N}}}(Y_0 = h, Y_{k+1} = l_k, k = 0, \ldots, m) \\ &\quad = \sum_{h \in L} P_{L^{\mathbb{N}}}(Y_{m+1} = \varphi(h), Y_{m-k} = \varphi(l_k), k = 0, \ldots, m) \\ &\quad = P_{L^{\mathbb{N}}}(Y_{m+1} \in L, Y_{m-k} = \varphi(l_k), k = 0, \ldots, m) \\ &\quad = P_{L^{\mathbb{N}}}(Y_{m-k} = \varphi(l_k), k = 0, \ldots, m) \\ &\quad = P_{L^{\mathbb{N}}}(Y_k = l_k, k = 0, \ldots, m). \end{aligned}$$

Then, $P_{L^{\mathbb{N}}}$ is time stationary.

https://doi.org/10.1017/jpr.2016.2 Published online by Cambridge University Press

Let us now prove that $P_{L^{\mathbb{N}}} \circ \zeta^{-1}$ is time stationary. Let $X_n \colon \mathcal{I}^{\mathbb{N}} \to \mathcal{I}$ be the nth coordinate function. We must prove that, for all $m \in \mathbb{N}$ and $((l_{dk}, \dots, l_{d(k+1)-1}) \colon k = 0, \dots, m) \in \mathcal{I}^{m+1}$, we have

$$\begin{aligned} P_{L^{\mathbb{N}}} \circ \zeta^{-1}(X_k = (l_{dk}, \dots, l_{d(k+1)-1}), k = 0, \dots, m) \\ = \sum_{(c_0,\dots,c_{d-1}) \in L^d} P_{L^{\mathbb{N}}} \circ \zeta^{-1}(X_0 = (c_0, \dots, c_{d-1}), \\ X_{k+1} = (l_{dk}, \dots, l_{d(k+1)-1}), k = 0, \dots, m). \end{aligned}$$

This relation is equivalent to,

$$\begin{aligned} P_{L^{\mathbb{N}}}(Y_t = l_t, t = 0, \dots, d(m+1) - 1) \\ = \sum_{(c_0,\dots,c_{d-1}) \in L^d} P_{L^{\mathbb{N}}}(Y_0 = c_0, \dots, Y_{d-1} = c_{d-1}; Y_{t+d} = l_t, t = 0, \dots, d(m+1) - 1), \end{aligned}$$

which is equivalent to the equality

$$P_{L^{\mathbb{N}}}(Y_t = l_t, t = 0, \dots, d(m+1) - 1) = P_{L^{\mathbb{N}}}(Y_{t+d} = l_t, t = 0, \dots, d(m+1) - 1).$$

This last relation follows straightforwardly from the time stationarity of $P_{L^{\mathbb{N}}}$, completing the proof. □

Let us state that a weaker condition of CSPR implies a weaker stationary property. Assume that the CSPR is verified only for tuples of length smaller or equal to r_0. This means that, for all $m < r_0$, all $(l_0, \dots, l_m) \in L^{m+1}$, and all $u \in \mathbb{N}$,

$$P_{L^{\mathbb{N}}}(Y_{k+u} = l_k, k = 0, \dots, m) = P_{L^{\mathbb{N}}}(Y_{k+u} = \varphi(l_{m-k}), k = 0, \dots, m).$$

Let us prove that in this case the stationarity only holds for the cylinders of length strictly smaller than r_0.

Proposition 5.2. *Let $r_0 \geq 2$. Assume that $P_{L^{\mathbb{N}}}$ verifies the CSPR for cylinders defined by tuples of length smaller or equal to r_0. Then, for all $m < r_0 - 1$ and all $(l_0, \dots, l_m) \in L^{m+1}$, we have*

$$P_{L^{\mathbb{N}}}(Y_{k+u} = l_k, k = 0, \dots, m) = P_{L^{\mathbb{N}}}(Y_k = l_k, k = 0, \dots, m) \quad \text{for all } u \in \mathbb{N}^*.$$

To state the result in the d-mers setting assume that $\lfloor r_0/d \rfloor \geq 2$. Then, for $m < \lfloor r_0/d \rfloor - 1$ and all $((l_{dk}, \dots, l_{d(k+1)-1}), k = 0, \dots, m) \in \mathcal{I}^{m+1}$, we have

$$\begin{aligned} P_{L^{\mathbb{N}}} \circ \zeta^{-1}(X_{k+u} = (l_{dk}, \dots, l_{d(k+1)-1}), k = 0, \dots, m) \\ = P_{L^{\mathbb{N}}} \circ \zeta^{-1}(X_k = (l_{dk}, \dots, l_{d(k+1)-1}), k = 0, \dots, m) \quad \text{for all } u \in \mathbb{N}^*. \end{aligned}$$

Proof. Let us prove the first relation by induction on $u \in \mathbb{N}^*$. For $u = 1$ the proof is the same as the first part of the proof of Proposition 5.1 when we showed that $P_{L^{\mathbb{N}}}$ is stationary. Assume it has been shown up to u, let us prove it for $u + 1$. Since $m + 2 \leq r_0$, we obtain

$$\begin{aligned} P_{L^{\mathbb{N}}}(Y_{u+1+k} = l_k, k = 0, \dots, m) \\ = \sum_{h \in L} P_{L^{\mathbb{N}}}(Y_u = h, Y_{u+1+k} = l_k, k = 0, \dots, m) \\ = \sum_{h \in L} P_{L^{\mathbb{N}}}(Y_{u+m-k} = \varphi(l_k), k = 0, \dots, m; Y_{u+1+m} = \varphi(h)) \\ = P_{L^{\mathbb{N}}}(Y_{u+m-k} = \varphi(l_k), k = 0, \dots, m) \\ = P_{L^{\mathbb{N}}}(Y_{u+k} = l_k, k = 0, \dots, m). \end{aligned}$$

https://doi.org/10.1017/jpr.2016.2 Published online by Cambridge University Press

Then, from an inductive argument, we obtain $P_{L^{\mathbb{N}}}(Y_{u+1+k} = l_k, k = 0, \ldots, m) = P_{L^{\mathbb{N}}}(Y_k = l_k, k = 0, \ldots, m)$. Hence, the first equation is shown. The second relation follows straightforwardly from the first one, and this is done as in the second part of the proof of Proposition 5.1 when we showed that $P_{L^{\mathbb{N}}} \circ \zeta^{-1}$ is time stationary. □

Acknowledgements

The author thanks the Center for Mathematical Modeling (CMM) Basal CONICYT Program PFB 03 and INRIA-CHILE program CIRIC for supporting this work. He is indebted to Andrew Hart for fruitful discussions and several improvements made to the presentation of the manuscript. The author is deeply grateful to Prof. Søren Asmussen for first calling his attention to the relation between the regenerative construction in (3.1) at the beginning of Section 3, and semiregenerative processes. This allowed the original proof of Theorem 3.1 to be substantially reduced by using Palm theory and helped to improve the presentation of the paper.

References

[1] Albrecht-Buehler, G. (2006). Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. *Proc. Nat. Acad. Sci. USA* **103,** 17828–17833.
[2] Allegrini, P., Buiatti, M., Grigolini, P. and West, B. J. (1998). Fractional Brownian motion as a nonstationary process: an alternative paradigm for DNA sequences. *Phys. Rev. E* **57,** 4558–4567.
[3] Asmussen, S. (2003). *Applied Probability and Queues*, 2nd edn. Springer, New York.
[4] Bell, S. J. and Forsdyke, D. R. (1999). Deviations from Chargaff's second parity rule correlate with direction of transcription. *J. Theoret. Biol.* **197,** 63–76.
[5] Bouaynaya, N. and Schonfeld, D. (2007). Non-stationary analysis of DNA sequences. In *Proc. IEEE Statistical Signal Processing Workshop*, IEEE, New York, pp. 200–204.
[6] Bouaynaya, N. and Schonfeld, D. (2008). Emergence of new structure from non-stationary analysis of genomic sequences. In *Proc. IEEE International Workshop on Genomic Signal Processing and Statistics*, IEEE, New York, pp. 1–4.
[7] Bouaynaya, N. and Schonfeld, D. (2008). Nonstationary analysis of coding and noncoding regions in nucleotide sequences. *IEEE J. Selected Topics Signal Process.* **2,** 357–364.
[8] Forsdyke, D. R. (2011). *Evolutionary Bioinformatics*, 2nd edn. Springer, New York.
[9] Hart, A. and Martínez, S. (2011). Statistical testing of Chargaff's second parity rule in bacterial genome sequences. *Stoch. Models* **27**, 272–317.
[10] Hart, A. and Martínez, S. (2014). Markovianness and conditional independence in annotated bacterial DNA. *Statist. Appl. Genetics Molec. Biol.* **13,** 693–716.
[11] Hart, A. G., Martínez, S. and Videla, L. (2006). A simple maximization model inspired by algorithms for the organization of genetic candidates in bacterial DNA. *Adv. Appl. Prob.* **38,** 1071–1097.
[12] Karlin, S. and Brendel, V. (1993). Patchiness and correlations in DNA sequences. *Science* **259,** 677–680.
[13] Li, W. and Kaneko, K. (1992). Long-range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence. *Europhys. Lett.* **17,** 655–660.
[14] Milenkovic, O. (2008). Data storage and processing in cells: an information theoretic approach. In *Advances in Information Recording* (DIMACS Ser. Discrete Math. Theoret. Comput. Sci. **73**), American Mathematical Society, Providence, RI, pp. 105–146.
[15] Mitchell, D. and Bridge, R. (2006). A test of chargaff's second rule. *Biochem. Biophys. Res. Commun.* **340,** 90–94.
[16] Peng, C.-K. *et al.* (1992). Long-range correlations in nucleotide sequences. *Nature* **356,** 168–170.
[17] Prabhu, V. V. (1993). Symmetry observations in long nucleotide sequences. *Nucleic Acids Res.* **21,** 2797–2800.
[18] Richardson, E. J. and Watson, M. (2013). The automatic annotation of bacterial genomes. *Briefings Bioinformatics* **14,** 1–12.
[19] Rudner, R., Karkas, J. D. and Chargaff, E. (1968). Separation of B. subtilis DNA into complementary strands. 3. Direct analysis. *Proc. Nat. Acad. Sci. USA* **60,** 921–922.
[20] Zhang, S.-H. and Huang, Y.-Z. (2010). Limited contribution of stem-loop potential to symmetry of single-stranded genomic DNA. *Bioinformatics* **26,** 478–485.
[21] Zielinski, J. S., Bouaynaya, N., Schonfeld, D. and O'Neill, W. (2008). Time-dependent ARMA modeling of genomic sequences. *BMC Bioinformatics* **9,** S14.