Chapter 9
Investigating Some Attributes of Periodicity in DNA Sequences via Semi-Markov Modelling

Pavlos Kolias and Alexandra Papadopoulou

Abstract Periodicity of DNA segments and sequences have been studied thoroughly during the past decades. One of the main problems is the identification of protein coding and non-coding regions inside genes, using mathematical techniques. Periodicity plays an important role in the structure of DNA, as specific regions have been shown to have periodic patterns. In this paper, we consider that a DNA sequence is described by a semi-Markov chain (SMC), with discrete state space consisting of the four nucleotides. Equations in closed analytic form are derived, in order to characterize strong or weak d-periodic and quasiperiodic behaviour of our model for both the homogeneous and non-homogeneous case. The model is applied to 3-base periodic sequences, which characterize the protein-coding regions of the gene. The related probabilities and the corresponding indexes are provided, which yield a description of the underlying periodic pattern. Last, the previous theoretical results are illustrated with data from synthetic and real DNA sequences.

Keywords Semi-Markov chains · Quasiperiodicity · Partial non-homogeneity · DNA sequence

MSC 2020 60K15

9.1 Introduction

Periodicity is a structural property of DNA sequences. It is expressed as either nucleotides or words of nucleotides, that have a tendency to appear with specific distances in-between. It is worth noting that, in DNA analysis, periodicity refers to a tendency of letters or words to reappear at certain distances, in contrast with

P. Kolias (✉) · A. Papadopoulou
Department of Mathematics, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece
e-mail: pakolias@math.auth.gr

A. Papadopoulou
e-mail: apapado@math.auth.gr

A. Malyarenko et al. (eds.), *Stochastic Processes, Statistical Methods, and Engineering Mathematics*, Springer Proceedings in Mathematics & Statistics 408,
https://doi.org/10.1007/978-3-031-17820-7_9

the formal mathematical definition of periodicity. Repetitive patterns in DNA often cause human diseases and algorithmic techniques have been applied to detect such patterns using statistically based criteria [5]. Also, there exist some structural types of periodicity in DNA, that are not linked to diseases. Mainly, there have been observed two such types of periodic behaviour in the DNA. The first one was discovered in 1980 [26] regarding the signal of nucleosomes contained in the nucleus. The authors observed that certain dinucleotides in the DNA of chromatin tend to appear at approximately every 10 to 11 bases. Subsequent studies suggested that the period of chromatin sequences converges to 10.4 bases [10]. Also, a more recent study, which investigated the genome of three organisms (A. thaliana, C. elegans and H. sapiens), suggested that the dinucleotide *AA* has almost perfect 10.5-base periodic behaviour in sequences of these organisms [22]. One explanation about this type of periodicity is that the distance of 10.5 bases is the "step" of the double strand, which suppress the long DNA sequence into the area of the nucleus [14]. Previous studies have used the Fourier transformation and spectral density analysis, as the main tool for exploring the periodic behaviour in DNA sequences [7–9, 24, 31]. The second type of periodicity has been observed in areas of the genome that are transcribed and later translated into proteins, called *coding regions* . Previous studies, using similar methods, have shown that in coding regions, there is a tendency of certain nucleotides to reappear every 3-bases [32]. Also, this type of periodicity has only been observed in coding regions, while for non-coding regions there was not found any similar periodic behaviour [12, 25, 27]. As each of the amino acids is encoded with a triplet of nucleotides (codon) and some specific amino acids are more abundant than others, the authors concluded that the periodic behaviour, in fact exists, due to the abundance of certain amino acids and the period of 3-bases is due to the triplet nature of the DNA [2]. As the whole genome sequence of each organism is frequently of several billions bases and the coding regions only constitute a small part of DNA, the information about the periodic behaviour of the coding regions of the DNA could be helpful for detecting these regions and distinguish between protein coding regions and non-coding regions [20]. Also, other well-known and highly accurate probabilistic algorithms use hidden-Markov models, in order to predict the different gene structures inside DNA [6]. Markov chains have been previously used in the analysis of letter and DNA sequences and some of the models could be found in the book of Waterman [29] and also in [1, 3, 13, 23, 30]. In this paper we consider that a DNA sequence is described by a semi-Markov chain (SMC) X_t, with discrete state space $S = \{A, C, G, T\}$, where t denotes the index position inside the sequence and $\boldsymbol{C}(m) = \{c_{i,j}(m)\}$ is the *core* matrix of the SMC. Previously, in [19] a similar modelling was examined to derive distributions of the word location and frequency of occurrences. The applied semi-Markov model was of a discrete finite state space S with elements specific words i.e. finite combinations of letters taken from the alphabet with known length and non-overlapping occurrences. An overview of probabilistic and statistical properties of words, as occurrences in biological sequences, is provided in [21]. Semi-Markov chains are a generalization of the Markov chains and allow the sojourn time between transitions to follow arbitrary distributions. An overview of the basic theory of homogeneous semi-Markov chains could be found in

the book of Howard [15]. Further theory and applications of semi-Markov modelling can be found in [4, 11, 16–18]

In Sect. 9.2.1, a recursive equation of the homogeneous semi-Markov model that could be used as an identification tool for regions that have strong or weak d-periodic behaviour is constructed. In Sect. 9.2.2 the previous theoretical results are generalized for the non-homogeneous case, considering the triplet nature of the DNA and assuming each coding position corresponds to a different transition matrix $\boldsymbol{P}(k)$. In Sect. 9.3, the case of quasiperiodicity of a state is included. The above tool is structured considering the fact that it is possible for the chain to lose its periodic behaviour for a number of cycles or the state to appear not exactly after a period of d positions, but in a radius of $d \pm \varepsilon$ positions. This could be due to the fact of genetic mutations that could shift the way the sequence is read. In Sect. 9.4, we present illustrations with data from both synthetic and real DNA sequences, regarding the 3-base periodicity. In the final section, conclusions are provided.

9.2 The Basic Framework

We assume that the DNA sequence is a realization of a semi-Markov chain X_t with state space the four nucleotides $S = \{A, C, G, T\}$. The semi-Markov chain is described by a sequence of Markov transition matrices $\{\boldsymbol{P}(t)\}_{t=0}^{\infty}$ and a sequence of conditional holding time matrices $\{\boldsymbol{H}(m)\}_{m=1}^{\infty}$, such as $\boldsymbol{P}(t) = \{p_{i,j}(t)\}, i, j \in S, t \in \mathcal{N}$, where

$$p_{i,j}(t) = Prob[\text{ the SMC will make its next transition to state } j \\ / \text{ the SMC entered state } i \text{ at time} t],$$

with $p_{i,j}(t) \geq 0,\ \forall i, j \in S,\ t \in \mathcal{N}$ and $\sum_{j \in S} p_{i,j}(t) = 1,\ \forall i,\ t \in \mathcal{N}$ and

$$\boldsymbol{H}(m) = \{h_{i,j}(m)\}, \qquad i, j \in S,\ m \in \mathcal{N},$$
$$h_{i,j}(m) = Prob[\text{The SMC will stay in state } i \text{ for } m \text{ time units} \\ \text{before moving to state } j].$$

We define the probabilities of the waiting time $w_i(t, m)$, which are the probabilities for the SMC to hold for m time units in state i, before making its next transition, while it entered state i at time t, to be $w_i(t, m) = \sum_{j \in S} p_{i,j}(t) h_{i,j}(m)$. Also the cumulative distribution for the waiting time is ${}^{>}w_i(t, n) = \sum_{m=n+1}^{\infty} w_i(t, m) = \sum_{m=n+1}^{\infty} \sum_{j \in S} p_{i,j}(t) h_{i,j}(m)$. The basic parameter of the SMC is the *core matrix* and

it is defined as $\boldsymbol{C}(t,m) = \{c_{i,j}(t,m)\}_{i,j\in S} = \boldsymbol{P}(t) \circ \boldsymbol{H}(m)$, where the operator $\{\circ\}$ denotes the element-wise product of matrices (Hadamard product). Also, we define the interval transition probabilities $q_{i,j}(t,n)$, which are the probabilities for the SMC to be in state j after n time units, while it entered state i in time t, to be

$$\begin{aligned} \boldsymbol{Q}(t,n) &= \{q_{i,j}(t,n)\}_{i,j\in S} \\ &= {}^{>}\boldsymbol{W}(t,n) + \sum_{m=0}^{n} [\boldsymbol{P}(t) \circ \boldsymbol{H}(m)]\, \boldsymbol{Q}(t+m, n-m), \end{aligned} \tag{9.1}$$

where ${}^{>}\boldsymbol{W}(t,n) = diag\{{}^{>}w_i(t,n)\}$. The elements of the matrix $\boldsymbol{Q}(t,n)$ are

$$q_{i,j}(t,n) = \delta_{i,j}\, {}^{>}w_i(t,n) + \sum_{r\in S}\sum_{m=1}^{n} c_{i,r}(t,m) q_{r,i}(t+m, n-m),\ i,j \in S,\ t,n \in \mathcal{N}.$$

9.2.1 The Homogeneous Case

In the following, we consider the DNA sequence to be a homogeneous semi-Markov chain, therefore we have $p_{i,j}(t) = p_{i,j}, \forall t \in \mathcal{N}$. Furthermore, we assume that DNA sequences do not contain virtual transitions, therefore subsequent appearances of the same state count as holding and $p_{i,i}(t) = 0,\ \forall i \in S,\ t \in \mathcal{N}$. For the purpose of the present, the parameter of time indicates the position, based on the nature of the DNA sequences, as their evolution depends on the index position of every letter in the sequence. In order to study the d-periodic behaviour of a DNA sequence, we would like to examine the probability of a letter reappearance after d positions. Also, for a sequence with strong d-periodic behaviour, it is expected that for every periodic state, the frequency of the state appearances, every kd positions, would be high. Therefore, an interesting question is whether the chain is in the same state, not only for the first cycle of length d, but also for a number of n successive cycles of the same length. Thus, we define the following probabilities.

Definition 1 Let $p_i(1,d)$ be the probability that the SMC will be in state i in position d, while in the initial position it was observed to be in state i, that is

$$p_i(1,d) = Prob[\text{the SMC will be in state } i \text{ in position } d / \text{ the initial state was observed to be } i].$$

Similarly, we define the probability that the SMC will be in state i every d positions for n cycles, while in the initial position it was observed to be in state i, as follows

$$p_i(n,d) = Prob[\text{the SMC will be in state i every d positions for n cycles / the initial state was observed to be } i\,].$$

It is important to note that for a given DNA sequence, we do not know if the initial position is due to a letter transition or reappearance of the same letter, therefore we have to include both cases in order to calculate the probability above. If we observed the process to be in state i in the initial position, it would be unlikely that upon the first observation the SMC had just entered this state. On the other hand, it would be more plausible to think that we started to observe the process in a position, where the entrance to a state has already been achieved. As a result, the process will stay in state i for the remaining positions and then make a transition to state j. The basic parameters of the SMC under random starting concern only the behaviour of the process until the first transition. Hence, let us denote by ${}_r p_{i,j}(\cdot)$ the transition probabilities under random starting and ${}_r h_{i,j}(\cdot)$ the distributions of the holding positions under random starting. A more detailed specification of the SMC under random starting could be found in the book of Howard [15].

Lemma 1 *Let $\boldsymbol{P}(1,d)$ and $\boldsymbol{P}(n,d)$ be the $(N \times 1)$ vectors, which consist of the probabilities $p_i(d)$ and $p_i(n,d)$, $i \in S$ respectively, following Definition 1. Then,*

$$\text{(a)} \quad \boldsymbol{P}(1,d) = \left[{}^{>}_{r}\boldsymbol{W}(d) + \sum_{x=1}^{d} \boldsymbol{I} \circ \left[{}_r\boldsymbol{C}(x)[\boldsymbol{Q}(d-x) \circ (\boldsymbol{U}-\boldsymbol{I})] \right] \right] \cdot \mathbf{1}. \tag{9.2}$$

$$\text{(b)} \quad \boldsymbol{P}(n,d) = \boldsymbol{P}(n-1,d) \circ \boldsymbol{P}(1,d), \tag{9.3}$$

where $\boldsymbol{I}$ is the identity matrix, ${}^{>}_{r}\boldsymbol{W}(d) = diag\{{}^{>}_{r}w_i(d)\}$ denotes the survival function of the waiting time distribution under random starting, ${}_r\boldsymbol{C}(x) = \{{}_r c_{i,j}(x)\}$ denotes the core matrix of the SMC under random starting, which consist of the elements ${}_r c_{i,j}(x) = {}_r p_{i,j} \cdot {}_r h_{i,j}(x)$, $\boldsymbol{U} = \{u_{i,j}\}$, where $u_{i,j} = 1$, for every $i, j \in S$ and $\mathbf{1} = [1, 1, \cdots, 1]^T$.

Proof Let $S_x = \underbrace{i\ i\ i\ i\ \cdots\ i}_{x-times}\ j\ u\ u \cdots u\ i$, be the sequence of states of length d, where $x = 1, 2, ..., d$, j denotes any state different than i and u denotes any state from the state space S. For a given sequence, let us now consider the following instances which are mutually exclusive and exhaustive events:

$$\begin{aligned}
S_1 &= i\ j\ u\ u \cdots u\ i \\
S_2 &= i\ i\ j\ u\ u \cdots u\ i \\
S_3 &= i\ i\ i\ j\ u\ u \cdots u\ i \\
&\vdots \\
S_{d-2} &= i\ i\ \cdots i\ j\ u\ i \\
S_{d-1} &= i\ i\ i\ \cdots i\ j\ i
\end{aligned}$$

$$S_d = i\,i\,i\,i\,i\,\cdots\,i$$

According to the previous, the semi-Markov chain, with initial observed state i, will be in state i after d positions, if either it holds for more than d steps in the initial state or makes a transition to a different state j at position x before the end of the cycle, but in any case to occupy state i in the final position. Thus, using probabilistic argument and summing over all possible states and holding times, we can conclude to the equation $p_i(1,d) = {}^{>}{}_r w_i(d) + \sum_{j\neq i}^{N}\sum_{x=1}^{d} {}_r c_{i,j}(x) q_{j,i}(d-x)$. Let the element of the ith row of a vector $\boldsymbol{P}(1,d)$ be the probability $p_i(1,d)$. The matrix notation in Eq. 9.2 can immediately be deduced by keeping only the non-diagonal elements, i.e. multiplying by the matrix $[\boldsymbol{U}-\boldsymbol{I}]$. Similarly, concerning Eq. 9.3, let us consider that the elements of the matrix $\boldsymbol{P}(n,d)$ to be the probabilities $p_i(n,d)$. Hence, in order for the SMC to be in the same state after n successive cycles of length d, we have $p_i(n,d) = \left[{}^{>}{}_r w_i(d) + \sum_{j\neq i}^{N}\sum_{x=1}^{d} {}_r c_{i,j}(x) q_{j,i}(d-x)\right]^n$. The matrix form is deduced immediately by the result above. □

Remark 1 For the interval transition probability matrix $\boldsymbol{Q}(n)$, instead of using the recursive formula 9.1, one can apply the closed analytic form, as proposed by Vassiliou and Papadopoulou [28]

$$\begin{aligned}\boldsymbol{Q}(n) = {}&{}^{>}\boldsymbol{W}(n) + \boldsymbol{C}(n) + \sum_{j=2}^{n}\{\boldsymbol{C}(j-1) + \sum_{k=1}^{j-2}\boldsymbol{S}_j(k,m_k)\}\\ &\times\{{}^{>}\boldsymbol{W}(n-j+1) + \boldsymbol{C}(n-j+1)\},\end{aligned} \tag{9.4}$$

where $\boldsymbol{S}_j(k,m_k) = \sum_{m_k=2}^{j-k}\sum_{m_{k-1}=1+m_k}^{j-k+1}\cdots\sum_{m_1=1+m_2}^{j-1}\prod_{r=-1}^{k-1}\boldsymbol{C}(m_{k-r-1}-m_{k-r})$ for $j \geqslant k+2$, while if $j \leqslant k+2$ we have $\boldsymbol{S}_j(k,m_k) = 0$.

9.2.2 The Case of Partial Non Homogeneity

The partial non-homogeneous semi-Markov chain (PNHSMC) is constructed based on the fact that every amino acid consists of three nucleotides (codon). Using this information, we can create three discrete coding positions $k = \{1,2,3\}$ and for the PNHSMC, we have three stochastic matrices $\boldsymbol{P}(k),\ k = 1,2,3$ for the embedded Markov chain. Similar to the homogeneous case, it would be of interest to find the probability for the PNHSMC to be in the same state after a length of d positions and also for n successive cycles of length d.

Definition 2 Let us define the quantity $p_i(k, 1, d)$ to be the probability that the PNHSMC will be in state i in position d, while in the initial position it was observed to be in state i, in coding position k, that is

$$p_i(k, 1, d) = Prob[\text{the SMC will be in state } i \text{ in position } d \text{ / the initial state was observed to be } i \text{ in coding position } k].$$

Furthermore, we define the quantity $p_i(k, n, d)$ to be the probability that the PNHSMC will be in state i every d positions for n cycles, while in the initial position it was observed to be in state i, in coding position k, that is

$$p_i(k, n, d) = Prob[\text{the SMC will be in state i every d positions for n cycles/ the initial state was observed to be } i \text{ in coding position } k].$$

Lemma 2 *Let $\boldsymbol{P}(k, 1, d)$ and $\boldsymbol{P}(k, n, d)$ be $(N \times 1)$ vectors, consisting of the probabilities $p_i(k, 1, d)$ and $p_i(k, n, d)$, $i \in S$ respectively, following Definition 2. Then*

(a) $$\boldsymbol{P}(k, 1, d) = \left[{}^{>}_{r}\boldsymbol{W}(k, d) + \sum_{x=1}^{d} \boldsymbol{I} \circ \left[{}_{r}\boldsymbol{C}(k, x)[\boldsymbol{Q}(k + x \bmod s, d - x) \circ (\boldsymbol{U} - \boldsymbol{I})] \right] \right] \cdot \mathbf{1} \tag{9.5}$$

(b) $$\boldsymbol{P}(k, n, d) = \boldsymbol{P}(k, n - 1, d) \circ \boldsymbol{P}(k, 1, d), \tag{9.6}$$

where ${}^{>}_{r}\boldsymbol{W}(k, d) = diag\{{}^{>}_{r}w_i(k, d)\}$ denotes the survival function of the waiting time distribution of the PNHSMC under random starting, ${}_{r}\boldsymbol{C}(k, x) = \{{}_{r}c_{i,j}(k, x)\}$ denotes the core matrix of the PNHSMC under random starting, which consist of the elements ${}_{r}c_{i,j}(k, x) = {}_{r}p_{i,j}(k) \cdot {}_{r}h_{i,j}(x)$ and $\boldsymbol{U} = \{u_{i,j}\}$, where $u_{i,j} = 1$.

Proof Let

$$S_x = \underbrace{i_k \; i_{k+1} \; \cdots \; i_{k+x-1 \bmod s}}_{x-times} \; j_{k+x \bmod s} \; u_{k+x+1 \bmod s} \cdots \; u_{k+d-1 \bmod s} \; i_{k+d \bmod s}$$

be the sequence of states of length d, where $x = 1, 2, ..., d$, j denotes any state different than i, u denotes any state from the state space S, k denotes the coding position and s denotes the total number of different coding positions. For a given sequence, let us define the following instances which are mutually exclusive and exhaustive events:

$$
\begin{aligned}
S_1 &= i_k\ j_{k+1}\ u_{k+3}\ u_{k+3}\ u_{k+4}\ \cdots\ u_{k+d-1 \bmod s}\ i_{k+d \bmod s}\\
S_2 &= i_k\ i_{k+1}\ j_{k+2}\ u_{k+3}\ u_{k+4} \cdots\ u_{k+d-1 \bmod s}\ i_{k+d \bmod s}\\
S_3 &= i_k\ i_{k+1}\ i_{k+2}\ j_{k+3}\ u_{k+4}\ \cdots\ u_{k+d-1 \bmod s}\ i_{k+d \bmod s}\\
&\vdots\\
S_{d-2} &= i_k\ i_{k+1}\ i_{k+2}\ \cdots\ j_{k+d-2 \bmod s}\ u_{k+d-1 \bmod s}\ i_{k+d \bmod s}\\
S_{d-1} &= i_k\ i_{k+1}\ i_{k+2}\ i_{k+3}\ \cdots\ j_{k+d-1 \bmod s}\ i_{k+d \bmod s}\\
S_d &= i_k\ i_{k+1}\ i_{k+2}\ i_{k+3}\ i_{k+4}\ i_{k+5}\ \cdots\ i_{k+d \bmod s},
\end{aligned}
$$

The PNHSMC, with initial observed state i in coding position k, will be in state i after d positions, either if it holds for more than d positions in the initial state or moves to a different state j at position $x + k \bmod s$ before the end of the cycle, but in any case to occupy state i in the final position. Thus, using probabilistic argument and summing over all possible states and holding positions, we obtain

$$
p_i(k, 1, d) = {}^{>}{}_r w_i(k, d) + \sum_{j \neq i}^{N} \sum_{x=1}^{d} {}_r c_{i,j}(k, x) q_{j,i}((k + x) \bmod s, d - x).
$$

Let the element of the ith row of a vector $\boldsymbol{P}(k, 1, d)$ to be the probability $p_i(k, 1, d)$. The matrix notation in Eq. (9.5) can be deduced immediately by multiplying with the matrix $[\boldsymbol{U} - \boldsymbol{I}]$. Similarly, concerning equation (9.6), let us consider the elements of the matrix $\boldsymbol{P}(k, n, d)$ to be the probabilities $p_i(k, n, d)$. In order for the PNHSMC to be in the same state after n successive cycles of length d, we have

$$
p_i(k, n, d) = \left[{}^{>}{}_r w_i(k, d) + \sum_{j \neq i}^{N} \sum_{x=1}^{d} {}_r c_{i,j}(k, x) q_{j,i}(k + x \bmod s, d - x)\right]^n.
$$

The matrix form in (9.6) is deduced immediately by applying the Hadamard product over n matrices of the form $\boldsymbol{P}(k, 1, d)$. □

Remark 2 For the interval transition probability matrix $\boldsymbol{Q}(t, n)$, instead of using the recursive formula, we can apply the closed analytic form, which is [28]

$$
\begin{aligned}
\boldsymbol{Q}(k, n) = {}^{>}\boldsymbol{W}(k, n) + \boldsymbol{C}(k, n) + \sum_{j=2}^{n} \{\boldsymbol{C}(k, j - 1) + \sum_{x=1}^{j-2} \boldsymbol{S}_j(x, k, m_x)\}\\
\times \{{}^{>}\boldsymbol{W}(k + j - 1, n - j + 1) + \boldsymbol{C}(k + j - 1, n - j + 1)\},
\end{aligned}
\tag{9.7}
$$

where $\boldsymbol{S}_j(x, k, m_x) = \sum_{m_x=2}^{j-x} \sum_{m_{x-1}=1+m_x}^{j-x+1} \cdots \sum_{m_1=1+m_2}^{j-1} \prod_{r=-1}^{x-1} \boldsymbol{C}(k + m_{x-r} - 1, m_{x-r-1} - m_{x-r})$ for $j \geqslant x + 2$, while if $j \leqslant x + 2$ we have $\boldsymbol{S}_j(x, k, m_x) = 0$.

9.3 Quasiperiodicity

The previous results, for both the homogeneous and non-homogeneous case, correspond to the probability of a state i to reappear again after d positions and n successive cycles. However, for the model to be more coherent, we also have to include the event that the periodicity is not strict and the state i does not appear exactly after d positions, but in the interval $(d-\varepsilon, d+\varepsilon)$. Also, we are interested in the quasiperiodic behaviour of the SMC, not only for a cycle of length d, but also for a number of n successive cycles. For simplicity we assume that $\varepsilon = 1$, although the results for $\varepsilon > 1$ are straightforward. For this purpose, let us define the entrance probabilities under random starting ${}_re_{i,j}(n)$, which are the probabilities that the SMC will enter state j at position n, given that, in the initial position, the SMC was observed to be in state i [15]. The equation for calculating the probabilities is ${}_re_{i,j}(n) = \delta_{i,j}\delta(n) + \sum_{r=1}^{N}\sum_{m=0}^{n} {}_rc_{i,r}e_{r,j}(n-m)$. Furthermore, let us define the first passage time probabilities $f_{i,j}(n)$, which are the probabilities that the SMC will transition to state j for the first time after n positions, given that it had entered state i in the initial position [15]. The recursive formula of the probabilities $f_{i,j}(n)$ is given by $f_{i,j}(n) = \sum_{r\neq j}^{N}\sum_{m=0}^{n} p_{i,r}h_{i,r}(m)f_{r,j}(n-m) + p_{i,j}h_{i,j}(n)$.

Definition 3 Let us define the quantity ${}_\varepsilon p_i(1,d)$, assuming $\varepsilon = 1$, to be the probability that the SMC will be in state i at least once in the position interval $d \pm \varepsilon$, while in the initial position, the SMC was observed to be in state i. Also, let us define the probability ${}_\varepsilon p_i(n,d)$ to be the probability that the SMC will be in the state i in the interval $(d-1, d+1)$ for n successive cycles, that is

(a) ${}_\varepsilon p_i(1,d) = Prob$[the SMC will be in state i either in position $d-1$, or d, or $d+1$ / the initial state was observed to be i] (9.8)

(b) ${}_\varepsilon p_i(n,d) = Prob$[the SMC to be in state i either in position $d-1$, or d, or $d+1$ for n cycles/ the initial state was observed to be i] (9.9)

Theorem 1 *Let* ${}_\varepsilon \boldsymbol{P}(1,d)$ *and* ${}_\varepsilon \boldsymbol{P}(n,d)$ *be* $(N \times 1)$ *vectors, consisiting of the probabilities* ${}_\varepsilon p_i(1,d)$ *and* ${}_\varepsilon p_i(n,d)$, $i \in S$ *respectively, following Definition 3. Then*

(a) ${}_\varepsilon \boldsymbol{P}(d) = \boldsymbol{P}(d-1) +$
$$\Big[\sum_{m=1}^{d-1} \boldsymbol{I} \circ \Big[{}_r \boldsymbol{E}(m)\big[[\boldsymbol{F}(d-m) + \boldsymbol{F}(d+1-m)] \circ (\boldsymbol{U}-\boldsymbol{I})\big]\Big] \cdot \boldsymbol{1}\Big] \quad (9.10)$$

(b) ${}_\varepsilon \boldsymbol{P}(n,d) = {}_\varepsilon \boldsymbol{P}(n-1,d)) \circ$
$$\Big[\boldsymbol{P}(d-1) + \Big[\sum_{m=1}^{d-1} \boldsymbol{I} \circ \Big[{}_r \boldsymbol{E}(m)\big[[\boldsymbol{F}(d-m) + \boldsymbol{F}(d+1-m)] \circ (\boldsymbol{U}-\boldsymbol{I})\big]\Big] \cdot \boldsymbol{1}\Big]\Big], \quad (9.11)$$

where ${}_r\boldsymbol{E}(\cdot) = \{{}_re_{i,j}(\cdot)\}$ *is the matrix which consists of the entrance probabilities under random starting and* $\boldsymbol{F}(\cdot) = \{f_{i,j}(\cdot)\}$ *is the matrix with the first passage time probabilities.*

Proof Let us define the events A_0, A_1, A_2 as

$A_0 =$ [the SMC is in state i in position $d-1$/the initial state was observed to be$i$].

$A_1 =$ [the SMC is in state i in position d and in state $r \neq i$ in position $d-1$/ the initial state was observed to be i].

$A_2 =$ [the SMC is in state i in position $d+1$ and in state $r \neq i$ in positions $d-1$ and d/the initial state was observed to be i].

Schematically, we can visualize the events defined above, as the following sequences

$$A_0 = i\ u\ u\ u\ \cdots\ u\ i$$
$$A_1 = i\ u\ u\ u\ \cdots\ u\ r\ i$$
$$A_2 = \underbrace{i\ u\ u\ u\ \cdots\ u\ r}_{d-1}\ r\ i\ ,$$

where u denotes any state from state space S and r denotes a state different from i. It is obvious that the events are mutually exclusive, therefore $Prob[A_0 \cup A_1 \cup A_2] = Prob[A_0] + Prob[A_1] + Prob[A_2]$. The probability for the event A_0 is defined as

$$Prob[A_0] = p_i(1, d-1) = Prob[\text{the SMC will be in state } i \text{ in position} d-1/ \text{ the initial state was observed to be } i].$$

For the event A_1 to happen, it is required for the SMC to be in a state $r \neq i$ in position $d-1$ and transition to state i in position d. Therefore, the SMC could have entered state $r \neq i$ at a position $m \leq d-1$ and then transitioned to state i for the first time after the remaining $d-m$ positions. Using probabilistic argument and summing over all the different positions and states, we can deduce the following equation $Prob[A_1] = \sum_{r \neq i}^{N} \sum_{m=0}^{d-1} {}_re_{i,r}(m) f_{r,i}(d-m)$. Similarly we can deduce the probability of the event A_2 to happen, $Prob[A_2] = \sum_{r \neq i}^{N} \sum_{m=0}^{d-1} {}_re_{i,r}(m) f_{r,i}(d+1-m)$. For the sum of the probabilities of the three events we can derive the following expression

$$
\begin{aligned}
&{}_{\varepsilon}p_i(d) = Prob[A_0] + Prob[A_1] + Prob[A_2] \\
&= p_i(d-1) + \sum_{r\neq i}^{N}\sum_{m=0}^{d-1} {}_{r}e_{i,r}(m) f_{r,i}(d-m) + \sum_{r\neq i}^{N}\sum_{m=0}^{d-1} {}_{r}e_{i,r}(m) f_{r,i}(d+1-m) = \\
&= p_i(d-1) + \sum_{r\neq i}^{N}\sum_{m=0}^{d-1} {}_{r}e_{i,r}(m)[f_{r,i}(d-m) + f_{r,i}(d+1-m)].
\end{aligned}
$$

Equation (9.2) can be written in matrix form as

$$
\begin{aligned}
{}_{\varepsilon}\boldsymbol{P}(d) =& \boldsymbol{P}(d-1) + \\
&\Big[\sum_{m=1}^{d-1} \boldsymbol{I} \circ \Big[\boldsymbol{E}(m)\big[[\boldsymbol{F}(d-m) + \boldsymbol{F}(d+1-m)] \circ (\boldsymbol{U}-\boldsymbol{I})\big]\Big] \cdot \mathbf{1}\Big].
\end{aligned}
$$

Last and by applying Lemmas 1 and 2, we can derive the corresponding equations for the probabilities ${}_{\varepsilon}p_i(n,d)$, which are described in matrix notation, as follows

$$
\begin{aligned}
&{}_{\varepsilon}\boldsymbol{P}(n,d) =_{\varepsilon} \boldsymbol{P}(n-1,d)) \circ \\
&\Big[\boldsymbol{P}(d-1) + \sum_{m=1}^{d-1} \boldsymbol{I} \circ \Big[\boldsymbol{E}(m)\big[[\boldsymbol{F}(d-m) + \boldsymbol{F}(d+1-m)] \circ (\boldsymbol{U}-\boldsymbol{I})\big]\Big]\mathbf{1}\Big].
\end{aligned}
$$

□

9.4 Illustrations of Real and Synthetic Data

For the illustrations of the homogeneous semi-Markov model, synthetic DNA sequences as well as real genomic and mRNA sequences were used. The coding sequence used was human dystrophin mRNA and the non-coding sequence, which was used for comparison, was the human b-nerve growth factor gene (BNGF). These sequences have already been examined using the spectral density analysis by Tsonis [27]. We assumed that each of the sequences could be described by a homogeneous semi-Markov chain $\{X_t\}_{t=0}^{\infty}$, with state space $S = \{A, C, G, T\}$ and the index t denotes the position of each nucleotide inside the sequence. The basic parameters $\boldsymbol{P}_{i,j}(s)$ and $\boldsymbol{H}_{i,j}(m)$ of the SMC were estimated using the empirical estimators $\widehat{p}_{i,j}(k) = \dfrac{N(i(k) \to j)}{\sum_{x\in S} N(i(k) \to x)}$ and $\widehat{h}_{i,j}(m) = \dfrac{N(i \to j, m)}{\sum_{x\in S} N(i \to x, m)}$ where $N(i(k) \to j)$ denotes the number of transitions from state i to state j, starting from coding position k and $N(i \to j, m)$ denotes the number of transitions from state i to state j, while the SMC remained in state i for m positions. In order to estimate the initial condition, which are the probabilities of the matrix $\boldsymbol{P}(1,d)$, the first 10

cycles of length 3 have been used and the basic parameters $\boldsymbol{P}$ and $\boldsymbol{H}(m)$ have been estimated. After that and for each cycle n, the core matrix $\boldsymbol{C}(m)$ has been estimated, using the letters of the sequence up until the position $30 + n \cdot d$. This specific process has been implemented, correcting the estimations, as in the current application the length of each period is small ($d = 3$), resulting in an non adequate sample size for each cycle. However, if we were interested in examining the periodic behaviour for larger periods, this correction procedure would not be necessary. Finally, the probability for the chain to be in the same state for every $n \cdot d$ positions has been calculated using the recursive equation $\boldsymbol{P}(n, d)$. Let us define the ratio by $\boldsymbol{R}(n) = \left[[\boldsymbol{P}(n-1, d)\mathbf{1}] \circ \boldsymbol{I}\right]^{-1} \cdot \boldsymbol{P}(n, d)$, where $\mathbf{1} = [1, 1, ..., 1]$. The quantity $\boldsymbol{R}(n)$ is a $(N \times 1)$ vector and the ith element of matrix $\boldsymbol{R}(n)$ is the ratio of the probability $p_i(n, d)$ over $p_i(n-1, d)$ for every n and illustrates the variations between the probabilities $p_i(n, d)$ and $p_i(n-1, d)$, in order to investigate the periodicity over a number of cycles. It is obvious that the probabilities $p_i(k, n, d)$ will converge to zero, as they are a product of n probabilities. The most important things in the periodic investigation, are the initial probability $\boldsymbol{P}(1, d)$, which contains the probabilities for the chain to be in the same state after d positions and also the ratio $\boldsymbol{R}(n)$, which measures the relationship between the probabilities of the current cycle and the previous one using the correction procedure. For higher values of $\boldsymbol{R}(n)$, the probabilities $p_i(n, d)$ decrease with a slower rate, while for lower values of $\boldsymbol{R}(n)$,the probabilities $p_i(n, d)$ converge to zero faster.

9.4.1 DNA Sequences of Synthetic Data

Example 1 *(Comparison between random and periodic DNA sequences)* Let L be a DNA sequence of length $N = 1000$ of the form: $L = \{U, U, U, ..., U\}$, where the letter U corresponds to any nucleotide, from a uniform distribution. Thus,

$$Prob[U = A] = Prob[U = C] = Prob[U = G] = Prob[U = T] = 1/4.$$

This kind of sequence would not exhibit any periodic behaviour, however the estimated probability matrix $\boldsymbol{P}(n, d)$, for $d = 3$, will be estimated for comparison. The estimation of the embedded Markov matrix is $\boldsymbol{P} = \begin{pmatrix} 0 & 0.2 & 0.8 & 0 \\ 0.375 & 0 & 0.5 & 0.125 \\ 0.125 & 0.5 & 0 & 0.375 \\ 0.25 & 0.75 & 0 & 0 \end{pmatrix}$ and the core matrix $\boldsymbol{C}(m)$ is $\boldsymbol{C}(1) = \begin{pmatrix} 0 & 0 & 0.8 & 0 \\ 0.375 & 0 & 0.5 & 0.125 \\ 0.125 & 0.375 & 0 & 0.375 \\ 0.25 & 0.5 & 0 & 0 \end{pmatrix}$, $\boldsymbol{C}(2) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0.125 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$

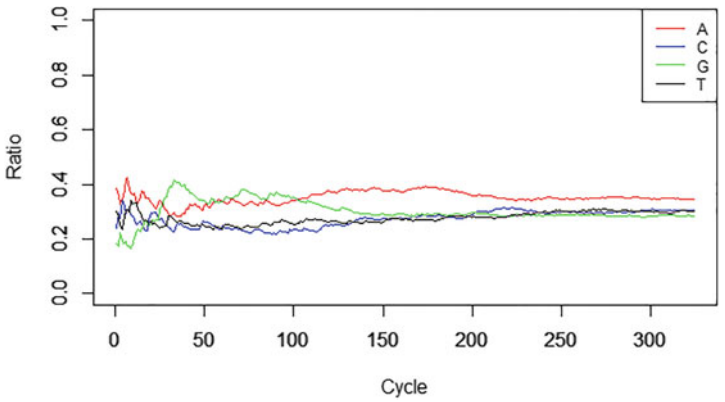

Fig. 9.1 $R(n)$ for the synthetic DNA sequence of a uniform distribution

while the only non zero element of $\boldsymbol{C}(3)$ is $c_{4,2}(3) = 0.25$. The initial condition is $\boldsymbol{P}(1,3) = \begin{pmatrix} 0.32 \\ 0.34 \\ 0.42 \\ 0.27 \end{pmatrix}$.

Figure 9.1 visualizes the ratio $\boldsymbol{R}(n)$ for the whole sequence. We observe that, as expected, there exist no clear tendency for any state to achieve a stronger periodic behaviour, compared to the other states. Now, let L be a DNA sequence of length $N = 1000$ of the form: $L = \{A, U, U, A, U, U, ...\}$, where the letter A corresponds to adenine and the letter U corresponds to any nucleotide from a uniform distribution, therefore

$$Prob[U = A] = Prob[U = C] = Prob[U = G] = Prob[U = T] = 1/4.$$

We will investigate the periodic behaviour, of period $d = 3$. One can notice that the letter A can possibly have a non-zero waiting time probability $w_A(m)$ for every m. On the other hand, for the other three letters C, G, T, the waiting time probabilities are zero if m exceeds two, as between every three letters, the letter A always appears at least once. The estimated embedded Markov transition matrix is $\boldsymbol{P} = \begin{pmatrix} 0 & 0.30 & 0.30 & 0.40 \\ 0.73 & 0 & 0.15 & 0.12 \\ 0.69 & 0.17 & 0 & 0.14 \\ 0.70 & 0.14 & 0.16 & 0 \end{pmatrix}$ and the core matrix is $\boldsymbol{C}(1) = \begin{pmatrix} 0 & 0.19 & 0.16 & 0.27 \\ 0.60 & 0 & 0.15 & 0.13 \\ 0.56 & 0.17 & 0 & 0.15 \\ 0.50 & 0.14 & 0.16 & 0 \end{pmatrix}$, $\boldsymbol{C}(2) = \begin{pmatrix} 0 & 0.08 & 0.11 & 0.09 \\ 0.13 & 0 & 0 & 0 \\ 0.13 & 0 & 0 & 0 \\ 0.20 & 0 & 0 & 0 \end{pmatrix}$ while the other matrices $\boldsymbol{C}(m)$ for $m > 2$, have non-zero elements only in the first row, that is for the letter A. The initial condition

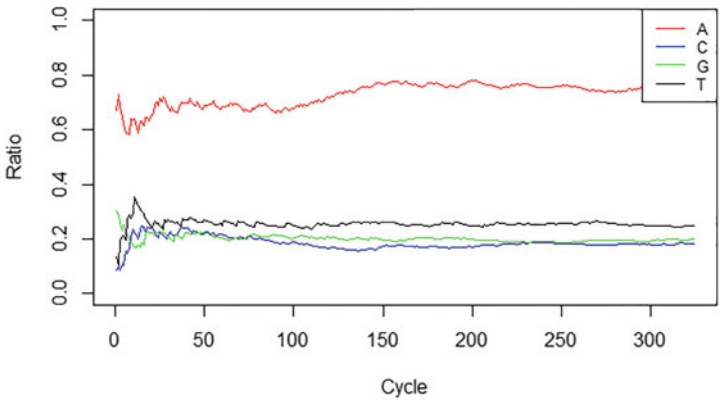

Fig. 9.2 $R(n)$ for the synthetic DNA sequence with 3-base periodicity of adenine

is $\boldsymbol{P}(1,3) = \begin{pmatrix} 0.83 \\ 0.18 \\ 0.20 \\ 0.25 \end{pmatrix}$. The probability for the chain to be in state A, every $d = 3$ positions, while starting from state A, is greater than the other three states, as we expected. This is also confirmed by the ratio, as presented in Fig. 9.2, that shows that state A exhibits higher values compared to the other states.

Example 2 (*Detection of periodic regions inside a sequence*) Let L be a DNA sequence of length $N = 5000$ of the form: $L = \{U, U, U, ..., U\}$, where the letter U corresponds to any random nucleotide from a uniform distribution. In the position intervals 1500–2000 and 3000–3500, which correspond to the 3-base cycles 500–666 and 1000–1166 respectively, the letter U has been substituted with the letter A, starting from the first position and at every 3 positions thereafter. Figure 9.3 shows the values of the ratio $R(n)$ for the letter A, where the green regions are the 3-base cycles of the sequence $R(n)$ where the sequence is increasing, while the red regions are the 3-base cycles where the sequence $R(n)$ decreases. It is observed, that the regions, in which we have synthetically added periodic behaviour for the letter A, have an increasing ratio $R(n)$ for A, indicating that in these regions the periodic behaviour of A is stronger.

9.4.2 DNA Sequences of Real Data

The information about the periodic behaviour of the coding regions of the genome could possibly be used, in order to distinguish these regions, over a DNA sequence with great length. For the coding sequences of real DNA, the human dystrophin mRNA has been used, while for the non coding region, the human b-nerve growth

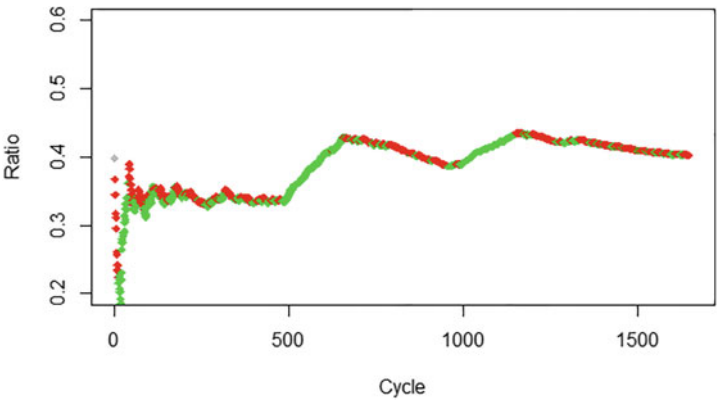

Fig. 9.3 $R(n)$ of the letter A of the synthetic sequence with periodicity in the cycles 500-666 and 1000-1166

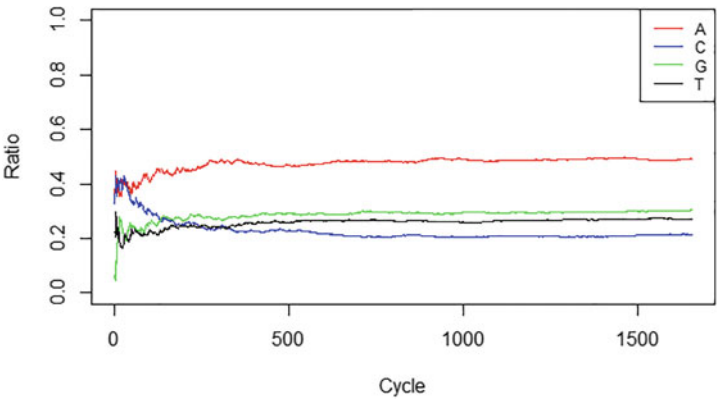

Fig. 9.4 $R(n)$ for the human dystrophin mRNA sequence

factor has been used. These sequences have a length greater than 5000 bases and they have already been studied for periodic behaviour [27]. One can notice through Fig. 9.4, that for the human dystrophin mRNA sequence, the nucleotide A has a higher chance to appear every 3 positions, while all the other nucleotides have almost the same behaviour. The ratio for the nucleotide A is higher compared to the other three states for the human dystrophin mRNA sequence, indicating the stronger periodic behaviour for adenine. However, Fig. 9.5 indicates that for the human b-nerve growth factor gene, which contains in more than 90% intronic sequences, the results are similar with the random sequence, that was created in the first example.

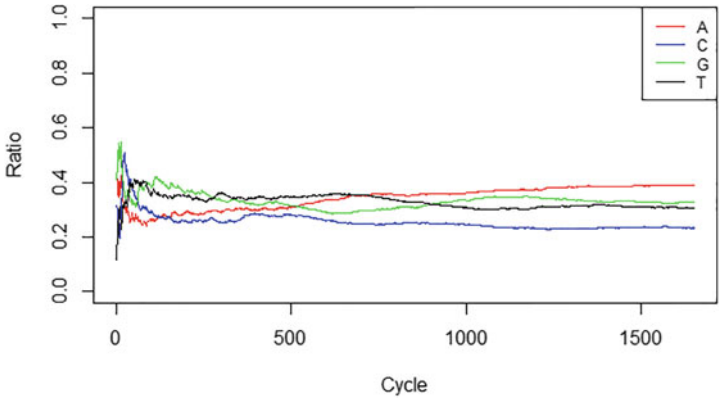

Fig. 9.5 $R(n)$ for the human b-nerve growth factor sequence

9.5 Conclusion

In the present paper, a method is developed, in order to investigate some attributes related to the periodicity of DNA sequences. The applied model is a semi-Markov chain of discrete and finite state space and discrete time, where the elements of the state space are the four nucleotides, i.e. $S = \{A, C, G, T\}$ and time denotes the index position in the sequence. The purpose of the model was to describe the periodic behaviour of a given DNA sequence, something that could possibly discriminate between coding and non-coding regions. It is known that the coding regions of the genome have different structure from the non-coding regions, as they exhibit a characteristic tendency of repetition of some nucleotides every 3 bases. Considering the previous fact and by modelling a DNA sequence with a semi-Markov chain, a recursive equation that could be used as an identification tool for regions that have strong or weak d-periodic behaviour is constructed. The corresponding probabilities are calculated in relation to the basic parameters of the model in closed analytic form. The theoretical results are also generalized for the non-homogeneous case, considering the triplet nature of the DNA and assuming each coding position corresponds to a different transition matrix $\boldsymbol{P}(k)$. In addition, the case of quasiperiodicity of a state is examined. The above theory is developed considering the fact that small perturbations in the cycle of the period may appear, such as a shift of the position of a letter due to genetic mutations and lead the chain to lose its periodic behaviour for a number of cycles. Therefore, the state will appear not exactly after a period of d positions, but in a radius of $d \pm \epsilon$ positions. The numerical results of the implementation of the model on actual data confirmed the previous studies, as it was apparent that periodic behaviour is a characteristic of the coding segments, unlike non-coding segments that did not show similar behaviour. For the estimation of the parameters, a correction procedure was applied, due to the short duration of the period ($d = 3$) for the specific application. The approach could potentially be used as an initial method

for investigating periodicity for any DNA sequence and also it could be used to separate two different DNA segments, in terms of their periodic behaviour. Although the examples produced satisfactory results, they should be perceived with caution, due to the complexity of the structure of DNA and its various peculiarities. For example, additional parameters could be included in the model, such as the sequence length, the frequencies of each nucleotide, the open reading frames (ORFS), the target organism, specific mutations and others.

References

1. Almagor, H.: A Markov analysis of DNA sequences. J. Theor. Biol. **104**(4), 633–645 (1983)
2. Almirantis, Y.: A standard deviation based quantification differentiates coding from non-coding DNA sequences and gives insight to their evolutionary history. J. Theor. Biol. **196**(3), 297–308 (1999)
3. Avery, P.J., Henderson, D.A.: Fitting Markov chain models to discrete state series such as DNA sequences. J. R. Stat. Soc.: Ser. C (Appl. Stat.) **48**(1), 53–61 (1999)
4. Bartholomew, D., Forbes, A., McClean, S.: Statistical Techniques for Manpower Planning. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Wiley (1991)
5. Benson, G.: Tandem repeats finder: a program to analyze DNA sequences. Nucl. Acids Res. **27**(2), 573–580 (1999)
6. Burge, C., Karlin, S.: Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. **268**(1), 78–94 (1997)
7. Chechetkin, V.R., Yu. Turygin, A.: Search of hidden periodicities in DNA sequences. J. Theor. Biol. **175**(4), 477–94 (1995)
8. Chechetkin, V.R., Turygin, A.Y.: On the spectral criteria of disorder in nonperiodic sequences: application to inflation models, symbolic dynamics and DNA sequences. J. Phys. A: Math. Gen. **27**(14), 4875–4898 (1994)
9. Cheever, E.A., Overton, G.C., Searls, D.B.: Fast Fourier transform-based correlation of DNA sequences using complex plane encoding. Comput. Appl. Biosci.: CABIOS **7**(2), 143–54 (1991)
10. Cohanim, A.B., Trifonov, E.N., Kashi, Y.: Specific selection pressure at the third codon positions: contribution to 10-to 11-base periodicity in prokaryotic genomes. J. Mol. Evol. **63**(3), 393–400 (2006)
11. D'Amico, G., Petroni, F., Prattico, F.: First and second order semi-Markov chains for wind speed modeling. Phys. A: Stat. Mech. Its Appl. **392**(5), 1194–1201 (2013)
12. Eskesen, S.T., Eskesen, F.N., Kinghorn, B., Ruvinsky, A.: Periodicity of DNA in exons. BMC Mol. Biol. **5**(1), 12 (2004)
13. Garden, P.W.: Markov analysis of viral DNA/RNA sequences. J. Theor. Biol. **82**(4), 679–684 (1980)
14. Herzel, H., Weiss, O., Trifonov, E.N.: 10–11 bp periodicities in complete genomes reflect protein structure and DNA folding. Bioinformatics (Oxford, England) **15**(3), 187–193 (1999)
15. Howard, R.A.: Dynamic probabilistic systems: Markov models, vol. 2. Courier Corporation (1971)
16. Janssen, J.: Semi-Markov Models: Theory and Applications. Springer (1999)
17. Janssen, J., Manca, R.: Applied semi-Markov processes. Springer Science & Business Media (2006)
18. Papadopoulou, A.: Counting transitions–entrance probabilities in non-homogeneous semi-Markov systems. Appl. Stoch. Models Data Anal. **13**(3–4), 199–206 (1997)

19. Papadopoulou, A.A.: Some results on modeling biological sequences and web navigation with a semi Markov chain. Commun. Stat.-Theory Methods **42**(16), 2853–2871 (2013)
20. Provata, A., Almirantis, Y.: Scaling properties of coding and non-coding DNA sequences. Phys. A: Stat. Mech. Its Appl. **247**(1–4), 482–496 (1997)
21. Reinert, G., Schbath, S., Waterman, M.S.: Probabilistic and statistical properties of words: an overview. J. Comput. Biol. **7**(1–2), 1–46 (2000)
22. Salih, B., Tripathi, V., Trifonov, E.N.: Visible periodicity of strong nucleosome DNA sequences. J. Biomol. Struct. Dyn. **33**(1), 1–9 (2015)
23. Schbath, S., Prum, B., De Turckheim, E.: Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. J. Comput. Biol. **2**(3), 417–437 (1995)
24. Tavare, S., Giddings, B.W.: Some statistical aspects of the primary structure of nucleotide sequences. In: Waterman, M.S. (ed.) Mathematical Methods for DNA Sequences (1989)
25. Trifonov, E.N.: 3-, 10.5-, 200-and 400-base periodicities in genome sequences. Phys. A: Stat. Mech. Its Appl. **249**(1–4), 511–516 (1998)
26. Trifonov, E.N., Sussman, J.L.: The pitch of chromatin DNA is reflected in its nucleotide sequence. Proc. Natl. Acad. Sci. **77**(7), 3816–3820 (1980)
27. Tsonis, A.A., Elsner, J.B., Tsonis, P.A.: Periodicity in DNA coding sequences: implications in gene evolution. J. Theor. Biol. **151**(3), 323–331 (1991)
28. Vassiliou, P.C.G., Papadopoulou, A.: Non-homogeneous semi-Markov systems and maintainability of the state sizes. J. Appl. Probab. **29**(3), 519–534 (1992)
29. Waterman, M.: Introduction to Computational Biology: Maps, Sequences, and Genomes: Interdisciplinary Statistics. Chapman & Hall/CRC, New York (1995)
30. Wu, T.J., Hsieh, Y.C., Li, L.A.: Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition. Biometrics **57**(2), 441–448 (2001)
31. Yin, C., Wang, J.: Periodic power spectrum with applications in detection of latent periodicities in DNA sequences. J. Math. Biol. **73**(5), 1053–1079 (2016)
32. Yin, C., Yau, S.S.T.: Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. J. Theor. Biol. **247**(4), 687–694 (2007)