

Article

# On State Occupancies, First Passage Times and Duration in Non-Homogeneous Semi-Markov Chains

Andreas C. Georgiou <sup>1,\*</sup>, Alexandra Papadopoulou <sup>2</sup>, Pavlos Kolias <sup>2</sup>, Haris Palikrousis <sup>2</sup>  
and Evanthia Farmakioti <sup>2</sup><sup>1</sup> Quantitative Methods and Decision Analytics Lab, Department of Business Administration, University of Macedonia, 54636 Thessaloniki, Greece<sup>2</sup> Department of Mathematics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece; apapado@math.auth.gr (A.P.); pakolias@math.auth.gr (P.K.); palihar7@gmail.com (H.P.); evanthiafarmakioti93@gmail.com (E.F.)

\* Correspondence: acg@uom.edu.gr

**Abstract:** Semi-Markov processes generalize the Markov chains framework by utilizing abstract sojourn time distributions. They are widely known for offering enhanced accuracy in modeling stochastic phenomena. The aim of this paper is to provide closed analytic forms for three types of probabilities which describe attributes of considerable research interest in semi-Markov modeling: (a) the number of transitions to a state through time (Occupancy), (b) the number of transitions or the amount of time required to observe the first passage to a state (First passage time) and (c) the number of transitions or the amount of time required after a state is entered before the first real transition is made to another state (Duration). The non-homogeneous in time recursive relations of the above probabilities are developed and a description of the corresponding geometric transforms is produced. By applying appropriate properties, the closed analytic forms of the above probabilities are provided. Finally, data from human DNA sequences are used to illustrate the theoretical results of the paper.

**Keywords:** semi-Markov modeling; occupancy; first passage time; duration; non-homogeneity; DNA sequences



**Citation:** Georgiou, A.C.; Papadopoulou, A.; Kolias, P.; Palikrousis, H.; Farmakioti, E. On State Occupancies, First Passage Times and Duration in Non-Homogeneous Semi-Markov Chains. *Mathematics* **2021**, *9*, 1745. <https://doi.org/10.3390/math9151745>

Academic Editor: Alexander Zeifman

Received: 20 May 2021

Accepted: 21 July 2021

Published: 24 July 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Human populations can be divided into categories (states and classes) taking into account some of their basic characteristics, such as place of residence, social class or rank in a hierarchy system. People usually move from a category to another category in a probabilistic manner and a person's history contains a sequence of sojourn times in the various categories and a set of transitions that have taken place. These are the basic parameters that construct a semi-Markov chain (SMC), according to which a mathematical model can be developed for the study of those systems [1,2]. These systems do not necessarily have to include humans, instead, they can describe any potential system characterized by and composed of historical observations, such as stay times in situations as well as transitions from one category to another. If, for the study of a population system, we reside on a Markov chain, we assume that the probability of transition from one category in another does not depend on the length of stay. Nonetheless, this time dependence is, in some cases, desirable to include in the process since it provides additional useful information. In this case, the transitions of such a system are not merely described by a typical Markov chain procedure and Semi-Markov models are introduced as the stochastic tools that provide a more rigorous framework accommodating a greater variety of applied probability models [3–5]. Various applications of semi-Markov processes include manpower planning, credit risk, word sequencing and DNA analysis [6–14].

In addition to semi-Markov processes, the non-homogeneous semi-Markov system (NHMS) was defined, introducing a class of broader stochastic models [15,16] that provide

a more general framework to describe the complex semantics of the system involved. Semi-Markov systems, which deploy a number of Markov chains evolving in parallel, are mostly applied in manpower planning, where the most important issues pertain to the evolution, control and asymptotic behavior [17–19]. In the last two decades, there has been an extended body of literature regarding the theory and results about NHMS [20–29]. The dynamic characteristics of the semi-Markov systems influence the number of times the chain occupies a state, of how long it takes to leave a state as well as the probability of first passage to a state. Therefore, in order to accompany the basic parameters of the semi-Markov chain and to enhance the modeling framework, additional attributes of critical interest are the occupancy, first passage time and duration probabilities, which are described as follows

1. *Occupancy probabilities.* These probabilities describe the distribution of the random variables that define the number of times the SMC has visited a specific state during an arbitrary time interval.
2. *First passage time probabilities.* These are the probabilities that describe the transition from a state to a different state for the first time. The properties of the first passage time probabilities have been investigated for Markov processes and some specific types of semi-Markov processes [30–35]. Details for the first passage time probabilities have been also presented for various stochastic processes [36].
3. *Duration probabilities.* These probabilities describe the distribution of random variables that define the time needed for the SMC to transfer to a different state.

DNA sequences are usually studied using probabilistic models, as nucleotide appearances are inter-correlated and attempts to use Markov models to model them have been reported [10,37]. One of the earliest studies applied a Markov model on the nucleotide alphabet  $\{A, C, G, T\}$  to estimate the transition probability matrix and the number of doublets and triplets [38]. Several statistics have been proposed to test the dependency order of the sequence, e.g., the Markov order, such as the phi-divergent statistics and conditional mutual information [39–41]. More advances in the subject include hidden-Markov models that are able to model different regions of DNA sequences [42]. Word occurrences are also of interest in DNA analysis [43]. Previous studies have examined the distribution, moments and properties of successive word occurrences [44,45]. Papadopoulou has provided some examples of semi-Markov models on modeling biological sequences [46]. Furthermore, algorithmic applications for estimating the first passage time probabilities in genomic sequences have been reported [47].

The aim of this study is to provide insight on the actual mechanism of the recursive relations of the probabilities mentioned above. Section 2 presents the basic parameters of a SMC, the interval transition probabilities and the entrance probabilities. Section 3 presents the main results of the paper, that is, the closed analytic solutions for the occupancy, duration and first passage time probabilities. The final section applies these theoretical results to human genome DNA strands. For the first illustration, the aim is to find the corresponding probabilities between nucleotide words and their symmetric complements by using the analytic form of the first passage time probabilities. Finally, for the second illustration, the frequency of the dinucleotide GC is examined for two distinct DNA sequences, using the occupancy probabilities.

## 2. Basic Framework

We can consider the semi-Markov chain  $\{X_t\}_{t \geq 1}$  with state space  $S = \{1, 2, \dots, N\}$  as a discrete stochastic process in which the successive states are defined by the transition probability matrix and the sojourn time in each state is described by a random variable conditioned on the current and the next state to be transitioned into. Thus, during the transition times, the process is equivalent to a Markov process. We call this Markovian process the *embedded* process. Let transition probabilities  $p_{ij}(t)$  be the probability of a SMC provided that it entered state  $i$  during its last transition at time  $t$  to transition to state  $j$  in the next transition. The transition probabilities should satisfy the same equations of a

Markovian process, that is,  $p_{ij} \geq 0, \forall i, j \in S$  and  $\sum_{j=1}^N p_{ij} = 1, \forall i \in S$ . When the process enters state  $i$  at time  $t$ , we assume that this state determines the next transition to state  $j$ , which occurs according to the transition probabilities. However, before making the transition from state  $i$  to state  $j$  and after the next state  $j$  is selected, the chain holds in state  $i$  for time  $\tau_{ij}$ . The sojourn time  $\tau_{ij}$  is a positive random variable with density function  $h_{ij}(\cdot)$ , which is called the function of sojourn time to transition from state  $i$  to state  $j$ . Thus,  $Prob[\tau_{ij} = m] = h_{ij}(m)$ , for  $m = 1, 2, \dots$ , and  $i, j \in S$ . We assume that the mean values of the distributions of sojourn times are finite and  $h_{ij}(0) = 0$ . In matrix notation, the basic parameters of the semi-Markov chain are the sequence of transition matrices  $\{P(t)\}_{t=0}^{\infty}$  and the sequence of sojourn time matrices  $\{H(m)\}_{m=1}^{\infty}$ . The probabilities of the *waiting times*  $w_i(t, m)$  are defined as follows:

$$w_i(t, m) = \sum_{j=1}^N p_{ij}(t)h_{ij}(m) = Prob[\tau_i = m|t],$$

where  $\tau_i$  is the holding time of the SMC in state  $i$ . The *core matrix* of the SMC connects the transition probabilities and the sojourn times and it is defined as follows:

$$C(t, m) = \{c_{ij}(t, m)\}_{ij \in S} = P(t) \circ H(m).$$

The operator  $\{\circ\}$  denotes the element-wise product of matrices (Hadamard product). Using the *core matrix*, we define  $q_{ij}(k|t, n)$ , which is the joint probability that the SMC will be in state  $j$  at time  $t + n$  and that it has made  $k$  transitions during the time interval  $(t, t + n]$ , given that at time  $t$  the process has entered state  $i$ . In order to calculate the probability  $q_{i,j}(k|t, n)$ , we distinguish two cases. First, we consider that during the time interval  $(t, t + n]$  the number of transitions is zero. Then, in order for the process at time  $t + n$  to be in state  $j$ , given that no transitions were made, it must be that the states  $i, j$  are the same. Secondly, assume that the SMC makes the first transition to state  $r$  at time  $t + m, 0 < m < n$ . Then, in the time interval  $(t, t + m]$ , we have one transition to state  $r$  and, in the remaining time interval  $(t + m, t + n]$ , we have the remaining  $k - 1$  transitions, with a final transition to state  $j$ . Thus, the resulting formula is as follows:

$$q_{ij}(k|t, n) = \delta_{ij}\delta(k) \overset{>}{w}_i(t, n) + \sum_{r=1}^N \sum_{m=0}^n c_{ir}(t, m)q_{rj}(k - 1|t + m, n - m).$$

where  $\overset{>}{w}_i(t, n) = \sum_{k=n+1}^{\infty} w_i(t, k)$  indicates the survival function of  $w_i(t, n)$  and  $\delta(k) = 1$  if  $k$  is zero, otherwise it is zero. If we are not interested in counting the number of transitions up to the final state  $j$ , we can deduce the following recursive relationship.

$$q_{ij}(t, n) = \delta_{ij} \overset{>}{w}_i(t, n) + \sum_{r=1}^N \sum_{m=0}^n c_{ir}(t, m)q_{rj}(t + m, n - m).$$

We also define the quantity  $e_{i,j}(k|t, n)$ , which is the probability that the SMC enters state  $j$  at time  $t + n$  and the total number of transitions in the time interval  $(t, t + n]$  is  $k$ , given that the SMC has entered state  $i$  at the initial position. Here, we can distinguish two cases. First, we assume that the number of transitions in the time interval  $(t, t + n]$  is zero. Then, to enter in state  $j$  at time  $t + n$ , the states  $i$  and  $j$  must be the same since state  $i$  was entered at the initial time. For the second case, suppose that the SMC at time  $t + m, 0 < m < n$  makes its first transition to state  $r$ . Then, at the time interval  $(t, t + m]$  we have a transition to state  $r$  and, at the time interval  $(t + m, t + n]$ , we have the remaining  $k - 1$  transitions, with the final transition to state  $j$ . These facts result in the following recursive relationship.

$$e_{ij}(k|t, n) = \delta_{ij}\delta(n)\delta(k) + \sum_{r=1}^N \sum_{m=0}^n c_{ir}(t, m)e_{rj}(k - 1|t + m, n - m).$$

If we are not interested in the number of transitions up to the final state  $j$ , we can reduce the recursive relationship to the quantity  $e_{ij}(t, n)$ , which are the probabilities that the SMC will enter state  $j$  at time  $n$ , provided that, at the initial position at time  $t$ , the SMC has entered state  $i$ . The equation for calculating the probabilities  $e_{ij}(t, n)$  is given by the following.

$$e_{ij}(t, n) = \delta_{ij}\delta(n) + \sum_{r=1}^N \sum_{m=0}^n c_{ir}(t, m)e_{rj}(t + m, n - m).$$

The interval transition probabilities and entrance probabilities are connected by the following relationship.

$$q_{ij}(k|t, n) = \sum_{m=0}^n e_{ij}(k|t, m)w_j(t + m, n - m).$$

### 3. Theoretical Results: Analytic Solutions of the Recursive Equations

#### 3.1. First Passage Time

The first passage times provide a measure of how long it takes to reach a given state from another. We can think of first passage times either in terms of transitions or of time or both. Thus, let  $f_{ij}(k|t, n)$  be the probability that  $k$  transitions and time  $n$  will be required for the first passage from state  $i$  to state  $j$  given that the SMC entered state  $i$  at time  $t$ . Applying a probabilistic argument, we can provide the following recursive formula.

$$f_{ij}(k|t, n) = \sum_{r \neq j}^N \sum_{m=0}^n c_{ir}(t, m)f_{rj}(k - 1|t + m, n - m) + \delta(k - 1)c_{ij}(t, n). \tag{1}$$

The first term of equation (1) corresponds to the case where  $k > 1$  and the SMC makes a transition to some state  $r$  different from  $j$  at time  $t + m$  and then makes a first passage from  $r$  to  $j$  in  $k - 1$  transitions during the interval  $(t + m, n - m)$ . The term is summed over all states and holding times that could describe the first transition. The second term corresponds to the case where  $k = 1$  and the process moves directly to state  $j$  at time  $t + n$ . If we are not interested in counting the transitions, then the recursive formula of the probabilities  $f_{ij}(t, n)$  is provided by the following.

$$f_{ij}(t, n) = \sum_{r \neq j}^N \sum_{m=0}^n c_{ir}(t, m)f_{rj}(t + m, n - m) + c_{ij}(t, n). \tag{2}$$

**Theorem 1.** For each non-homogeneous SMC with discrete state space  $S = 1, 2, \dots, N$ , a sequence of transition probability matrices  $\{\mathbf{P}(t)\}_{t=0}^\infty$  and a sequence of sojourn time matrices  $\{\mathbf{H}(m)\}_{m=1}^\infty$ , the probability matrices of first passage times  $\mathbf{F}(k|t, n) = \{f_{ij}(k|t, n)\}_{i,j \in S}$  are given by the following relationships:

1.  $\mathbf{F}(1|t, n) = \mathbf{C}(t, n)$ , for every  $n$ .
2.  $\mathbf{F}(k|t, n) = \mathbf{0}$ , if  $k > n$  or  $k = 0$ .
3.  $\mathbf{F}(k|t, n) = \sum_{m_1=1}^{n-k+1} * \sum_{m_2=1+m_1}^{n-k+2} * \dots * \sum_{m_{k-1}=1+m_{k-2}}^{n-1} \prod_{r=0}^{k-1} \{\mathbf{B}\} \mathbf{C}(t + m_{k-r-1}, m_{k-r} - m_{k-r-1})$ , for each  $1 < k \leq n$ ,

where  $\mathbf{B} = \mathbf{U} - \mathbf{I}$ ,  $\mathbf{U} = \{u_{ij} = 1\}_{i,j \in S}$ ,  $\mathbf{I}$  is the  $N \times N$  identity matrix and

$$\begin{aligned} & \prod_{r=0}^{k-1} \{\mathbf{B}\} \mathbf{C}(s + m_{k-r-1}, m_{k-r} - m_{k-r-1}) = \\ & = \mathbf{C}(s, m_1) \{ \mathbf{C}(s + m_1, m_2 - m_1) \{ \dots \{ \mathbf{C}(s + m_{k-1}, n - m_{k-1}) \circ \mathbf{B} \} \circ \mathbf{B} \} \dots \} \circ \mathbf{B}. \end{aligned}$$

**Proof.** Appendix A.1.  $\square$

### 3.2. Duration

Transitions of a SMC can be divided into two categories: virtual and real. The first category refers to transitions made from one state to the same state, while the second category refers to transitions from one state to a different state. Based on those two categories, one can define the duration as the number of transitions or the time required for the SMC to leave the initial state and to move to a different state, i.e., a real transition to take place for the first time and not a virtual one. Therefore, it is of interest to study the duration probability  $d_i(k|t, n)$  defined as the probability that the SMC moves for the first time to a different state that the initial one after  $n$  time units and  $k$  transitions during the interval  $(t, t + n]$ , given that the process entered state  $i$  at time  $t$ . We note here that out of the total  $k$  transitions in the above case,  $k - 1$  transitions are virtual and one transition is real. The duration probabilities for  $k \leq n$  are provided by the following.

$$d_i(k|t, n) = \sum_{m=0}^n c_{ii}(t, m)d_i(k - 1|t + m, n - m) + \delta(k - 1)(w_i(t, n) - c_{ii}(t, n)). \quad (3)$$

In the case that  $k > n$  or  $k = 0$ , then  $d_i(k|t, n) = 0$ . The rationale of this relationship can be deconstructed into two parts. In the first part, we can assume that the SMC has at least one virtual intermediate transition, while it starts from state  $i$  at time  $t$ , holds at the state  $i$  for  $m$  time units and finally transfers to state  $i$  again. At this point, the associated probability is  $d_i(k - 1|t + m, n - m)$ . In the second scenario, we assume that the SMC makes no transition up to time  $t + n$ . Therefore, the chain holds at state  $i$  for exactly  $n$  time units and then moves to a state  $j$  different than  $i$ . Thus, the duration defined in the present measures how long it takes to leave a given state.

**Theorem 2.** For each non-homogeneous SMC with discrete state space  $S = 1, 2, \dots, N$ , a sequence of transition probability matrices  $\{\mathbf{P}(t)\}_{t=0}^\infty$  and a sequence of sojourn time matrices  $\{\mathbf{H}(m)\}_{m=1}^\infty$ , the duration probability matrices  $\mathbf{D}(k|t, n) = \text{diag}\{d_i(k|t, n)\}_{i \in S}$  are provided by the following relationships:

1.  $\mathbf{D}(1|t, n) = [\mathbf{W}(t, n) - \mathbf{C}(t, n) \circ \mathbf{I}]$ , for every  $n$ .
2.  $\mathbf{D}(k|t, n) = \mathbf{0}$ , if  $k > n$  or  $k = 0$ .
3.  $\mathbf{D}(k|t, n) = \sum_{m_1=1}^{n-k+1} * \sum_{m_2=1+m_1}^{n-k+2} * \dots * \sum_{m_{k-1}=1+m_{k-2}}^{n-1} (\mathbf{C}(t, m_1) \circ \mathbf{I})(\mathbf{C}(t + m_1, m_2 - m_1) \circ \mathbf{I}) \dots (\mathbf{C}(t + m_{k-2}, m_{k-1} - m_{k-2}) \circ \mathbf{I})(\mathbf{W}(t + m_{k-1}, n - m_{k-1}) - \mathbf{C}(t + m_{k-1}, n - m_{k-1}) \circ \mathbf{I})$ , for each  $1 < k \leq n$ ,

where  $\mathbf{W}(t, n) = \text{diag}\{w_i(t, n)\}_{i \in S}$ .

**Proof.** Appendix A.2.  $\square$

### 3.3. Occupancy

We define  $v_{ij}(t, n)$  to be the number of times the SMC makes transitions to a state  $j$  in time interval of length equal to  $n$ , provided that in the initial time  $t$  the SMC had entered state  $i$ . If the initial state is the same as  $j$ , that is when  $i = j$ , then the initial state is not counted in  $v_{ij}(t, n)$ . We call the quantity  $v_{ij}(t, n)$  as the *occupancy measure* of state  $j$  at time  $t + n$ , provided that the SMC entered state  $i$  at time  $t$ . Clearly, the quantity  $v_{ij}(t, n)$  is a discrete random variable. We define as  $\omega_{ij}(\cdot|t, n)$  the probability mass distribution of  $v_{ij}(t, n)$ , which is  $\omega_{ij}(x|t, n) = \text{Prob}[v_{ij}(t, n) = x]$ . The recursive relationship of the occupancy probabilities is given by the following:

$$\begin{aligned} \omega_{ij}(x|t, n) &= \sum_{\substack{r=1 \\ r \neq j}}^N \sum_{m=0}^n c_{ir}(t, m)\omega_{rj}(x|t + m, n - m) + \\ &+ \sum_{m=0}^n c_{ij}(t, m)\omega_{jj}(x - 1|t + m, n - m) + \delta(x) > w_i(t, n), \end{aligned} \quad (4)$$

where  $i, j \in S, n = 0, 1, \dots$ , and  $x = 0, 1, \dots$

**Assumption 1.** In what follows, we assume that the embedded Markov chain is homogeneous, i.e.,  $\{P(t)\}_{t=0}^\infty = P$ , for each  $t$ .

Considering the above assumption, one can use the double geometric transform of the occupancy probabilities as follows.

$$\omega_{ij}^{gg}(y|z) = \sum_{x=0}^\infty \sum_{n=0}^\infty \omega_{ij}(x|n)z^n y^x.$$

Moreover, from the Equation (4), we can write the double geometric transform of the occupancy probabilities as follows.

$$\omega_{ij}^{gg}(y|z) = \sum_{r=1}^N c_{ir}^g(z)\omega_{rj}^{gg}(y|z) - (1-y)c_{ij}^g(z)\omega_{jj}^{gg}(y|z) + >w_i^g(z).$$

In matrix notation, we can use the previous results to obtain the following [3]:

$$\Omega^{gg}(y|z) = \frac{1}{1-z}\mathbf{U} - \frac{1-y}{1-z}[\mathbf{I} - \mathbf{C}^g(z)]^{-1}\mathbf{C}^g(z)(y\mathbf{I} + (1-y)[\mathbf{I} - \mathbf{C}^g(z)]^{-1} \circ \mathbf{I})^{-1},$$

where  $\mathbf{U}$  is the unit matrix,  $\Omega^{gg}(y|z) = \{\omega_{ij}^{gg}(y|z)\}_{i,j \in S}$  is the double geometric transform of  $\Omega(x|n) = \{\omega_{ij}(x|n)\}_{i,j \in S}$  and  $\mathbf{C}^g(z) = \{c_{ij}^g(z)\}_{i,j \in S}$ .

The occupancy probabilities are connected with the corresponding homogeneous first passage time probabilities through the following relationship.

$$\omega_{ij}(x|n) = \delta(x) >f_{ij}(n) + \sum_{m=0}^n f_{ij}(m)\omega_{jj}(x-1|n-m).$$

Using the double geometric transform, we can present the occupancy probabilities in matrix form according to the geometric transforms of the first passage time probabilities:

$$\Omega^{gg}(y|z) = >\mathbf{F}^g(z) + y\mathbf{F}^g(z)[>\mathbf{F}^g(z) \circ \mathbf{I}][\mathbf{I} - y(\mathbf{F}^g(z) \circ \mathbf{I})]^{-1},$$

which could be further simplified by using  $>f_{ij}^g(z) = \frac{1-f_{ij}^g(z)}{1-z}$  (Appendix B.1) resulting in matrix notation in (Appendix B.2).

$$\Omega^{gg}(y|z) = \frac{1}{1-z}\mathbf{U} - \frac{1-y}{1-z}\mathbf{F}^g(z)[\mathbf{I} - y\mathbf{F}^g(z) \circ \mathbf{I}]^{-1}.$$

We now provide Theorem 3 and Lemma 1 that will be used to prove the main Theorem 4 of the occupancy probabilities with respect to the core matrix.

**Theorem 3.** For a SMC with core matrix  $\mathbf{C}(\cdot)$ , we have the following:

$$\begin{aligned} \Omega^g(z|n) &= (z-1) \sum_{j=1}^{n-1} [\mathbf{C}(j) + [\sum_{i=2}^j (\mathbf{C}(i-1) + \sum_{k=1}^{i-2} \mathbf{S}_i(k, m_k)) \mathbf{C}(j+1-i)]] [\Omega^g(z|n-j) \circ \mathbf{I}] \\ &+ z [\mathbf{C}(n) + \sum_{j=2}^n (\mathbf{C}(j-1) + \sum_{k=1}^{j-2} \mathbf{S}_j(k, m_k)) \mathbf{C}(n+1-j)] + \\ &+ [\sum_{j=2}^n (\mathbf{C}(j-1) + \sum_{k=1}^{j-2} \mathbf{S}_j(k, m_k))] >\mathbf{W}(n+1-j) + >\mathbf{W}(n), \end{aligned}$$

where  $\mathbf{S}_i(k, m_k) = \sum_{m_k=2}^{i-k} \sum_{m_{k-1}=1+m_k}^{i-k+1} \dots \sum_{m_1=1+m_2}^{i-1} \prod_{r=1}^{k-1} \mathbf{C}(m_{k-r-1} - m_{k-r}), \forall i, j \in S$  and  $n = 0, 1, 2, \dots$ . Please note that the  $(j, r)$  element of  $\mathbf{S}_i(k, m_k)$  is the probability of moving from state  $j$  to state  $r$  after  $i-1$  time units and  $k$  intermediate transitions during the interval  $(t, t+i-1]$  for every  $t$  due to the time-homogeneity assumption.

**Proof.** Appendix A.3. □

**Lemma 1.** The product  $\Omega^g(z|n) \circ \mathbf{I}$  is equal to the following:

$$\begin{aligned} \Omega^g(z|n) \circ \mathbf{I} &= -(z-1) \sum_{j=1}^{n-1} \left[ \left[ \sum_{i=1}^j \mathbf{a}_{1i}^{-1} \mathbf{C}(j+1-i) \right] \circ \mathbf{I} \right] [\Omega^g(z|n-j) \circ \mathbf{I}] \\ &\quad - z \sum_{j=1}^n \left[ \mathbf{a}_{1j}^{-1} \mathbf{C}(n+1-j) \right] \circ \mathbf{I} + \sum_{j=1}^n \left[ -\mathbf{a}_{1j}^{-1} \mathbf{W}(n+1-j) \right] \circ \mathbf{I}, \end{aligned}$$

$$\forall i, j \in S \text{ and } n = 0, 1, 2, \dots, \text{ where} \\ -\mathbf{a}_{1i}^{-1} = \mathbf{C}(i-1) + \sum_{k=1}^{i-2} \mathbf{S}_i(k, m_k).$$

**Proof.** Appendix A.4. □

We now provide Theorem 4, which describes the analytic solutions of the occupancy probabilities. In order to facilitate the presentation and proof of Theorem 4, we begin with some aggregate notation. Let the following be the case:

$$\begin{aligned} \mathbf{A}_j &= \mathbf{C}(j) + \sum_{i=2}^j \left( \mathbf{C}(i-1) + \sum_{k=1}^{i-2} \mathbf{S}_i(k, m_k) \right) \mathbf{C}(j+1-i), \\ \mathbf{B}_{n,j} &= \left[ \sum_{w=2}^{n-j} \left[ \left( \mathbf{C}(w-1) + \sum_{k=1}^{w-2} \mathbf{S}_w(k, m_k) \right) \mathbf{W}(n-j+1-w) \right] \circ \mathbf{I} + \mathbf{W}(n-j) \right] \circ \mathbf{I}, \\ \mathbf{M}_u &= - \left[ \mathbf{C}(u-1) + \sum_{i=2}^{u-1} \left( \mathbf{C}(i-1) + \sum_{k=1}^{i-2} \mathbf{S}_i(k, m_k) \right) \mathbf{C}(u-i) \right] \circ \mathbf{I} + \sum_{k=1}^{u-2} (-1)^{k+1} \mathbf{R}_u(k, m_k), \\ \mathbf{M}'_u &= \left[ \mathbf{C}(u-1) + \sum_{i=2}^{u-1} \left( \mathbf{C}(i-1) + \sum_{k=1}^{i-2} \mathbf{S}_i(k, m_k) \right) \mathbf{C}(u-i) \right] \circ \mathbf{I}, \\ \mathbf{M}''_u &= \left[ \mathbf{C}(u-1) + \sum_{i=2}^{u-1} \left( \mathbf{C}(i-1) + \sum_{k=1}^{i-2} \mathbf{S}_i(k, m_k) \right) \mathbf{C}(u-i) \right] \circ \mathbf{I} + \left[ \sum_{k=1}^{u-2} (k+1)(-1)^k \mathbf{R}_u(k, m_k) \right], \\ \mathbf{M}'''_u &= \left[ \mathbf{C}(u-1) + \sum_{i=2}^{u-1} \left( \mathbf{C}(i-1) + \sum_{k=1}^{i-2} \mathbf{S}_i(k, m_k) \right) \mathbf{C}(u-i) \right] \circ \mathbf{I} - \left[ \sum_{k=1}^{u-2} (k+2)(-1)^k \mathbf{R}_u(k, m_k) \right], \\ \mathbf{E}_n &= \sum_{j=2}^n \left( \mathbf{C}(j-1) + \sum_{k=1}^{j-2} \mathbf{S}_j(k, m_k) \right) \mathbf{W}(n+1-j) + \mathbf{W}(n), \\ \mathbf{F}_{x,u} &= x(x-1) \sum_{k=x-3}^{u-2} \left[ \prod_{r=-1}^{x-4} (k-r) \right] (-1)^{(k-x+3)} \mathbf{R}_u(k, m_k) - x \left[ \sum_{k=x-2}^{u-2} \left[ \prod_{r=-1}^{x-3} (k-r) \right] (-1)^{(k-x+2)} \mathbf{R}_u(k, m_k) \right], \\ \mathbf{G}_{u,n,j} &= \mathbf{C}(n-j+1-u) \circ \mathbf{I} + \sum_{w=2}^{n-j+1-u} \left[ \left( \mathbf{C}(w-1) + \sum_{k=1}^{w-2} \mathbf{S}_w(k, m_k) \right) \mathbf{C}(n-j+2-u-w) \right] \circ \mathbf{I}, \\ \mathbf{H}_{x,u} &= x \sum_{k=x-2}^{u-2} \left[ \prod_{r=-1}^{x-3} (k-r) \right] (-1)^{k-(x-2)} \mathbf{R}_u(k, m_k) - \sum_{k=x-1}^{u-2} \left[ \prod_{r=-1}^{x-2} (k-r) \right] (-1)^{k-(x-1)} \mathbf{R}_u(k, m_k), \\ \mathbf{Q}_{u,n,j} &= \sum_{w=2}^{n-j+1-u} \left[ \left( \mathbf{C}(w-1) + \sum_{k=1}^{w-2} \mathbf{S}_w(k, m_k) \right) \mathbf{W}(n-j+2-u-w) \right] \circ \mathbf{I} + \left[ \mathbf{W}(n-j+1-u) \right] \circ \mathbf{I}, \end{aligned}$$

where

$$\mathbf{R}_u(k, m_k) = \sum_{m_k=2}^{u-k} \sum_{m_{k-1}=1+m_k}^{u-k+1} \dots \sum_{m_1=1+m_2}^{u-1} \prod_{r=-1}^{k-1} \left[ \sum_{i=1}^{m_{k-r-1}-m_{k-r}} \left( -\mathbf{a}_{1i}^{-1} \right) \mathbf{C}(m_{k-r-1} - m_{k-r} + 1 - i) \right] \circ \mathbf{I},$$

$$\mathbf{S}_i(k, m_k) = \sum_{m_k=2}^{i-k} \sum_{m_{k-1}=1+m_k}^{i-k+1} \dots \sum_{m_1=1+m_2}^{i-1} \prod_{r=-1}^{k-1} \mathbf{C}(m_{k-r-1} - m_{k-r}),$$

and

$$-\mathbf{a}_{1i}^{-1} = \mathbf{C}(i - 1) + \sum_{k=1}^{i-2} \mathbf{S}_i(k, m_k).$$

**Theorem 4.** For a SMC with core matrix  $\mathbf{C}(\cdot)$ , by adopting the above notations, we have that the following:

$$\begin{aligned} \Omega(0|n) &= - \sum_{j=1}^{n-1} \mathbf{A}_j \left[ \mathbf{B}_{n,j} + \sum_{u=2}^{n-j} \mathbf{M}_u \mathbf{Q}_{u,n,j} \right] + \mathbf{E}_n, \\ \Omega(1|n) &= \sum_{j=1}^{n-1} \mathbf{A}_j \left[ \mathbf{B}_{n,j} - \mathbf{G}_{1,n,j} - \sum_{u=2}^{n-j} [\mathbf{M}_u + \mathbf{G}_{u,n,j}] - 2 \sum_{u=2}^{n-j} \mathbf{M}_u''' \mathbf{Q}_{u,n,j} \right], \\ \Omega(2|n) &= \sum_{j=1}^{n-1} \mathbf{A}_j \left[ 2\mathbf{G}_{1,n,j} + \sum_{u=2}^{n-j} \sum_{k=1}^{u-2} [(-2k - 4)(-1)^k \mathbf{R}_u(k, m_k) \mathbf{G}_{u,n,j}] - 4 \sum_{u=2}^{n-j} \mathbf{M}_u' \mathbf{G}_{u,n,j} \right. \\ &\quad \left. + 2 \sum_{u=2}^{n-j} \mathbf{M}_u' \mathbf{Q}_{u,n,j} - \sum_{u=2}^{n-j} \sum_{k=1}^{u-2} (k + 1)(k + 2)(-1)^{k-1} \mathbf{R}_u(k, m_k) \mathbf{Q}_{u,n,j} \right], \\ \Omega(3|n) &= \sum_{j=1}^{n-1} \mathbf{A}_j \left[ 6 \sum_{u=2}^{n-j} \mathbf{M}_u'' \mathbf{G}_{u,n,j} - 3 \sum_{u=2}^{n-j} \sum_{k=1}^{u-2} k(k + 1)(-1)^{k+1} \mathbf{R}_u(k, m_k) \mathbf{G}_{u,n,j} \right. \\ &\quad \left. - \sum_{u=2}^{n-j} (k - 1)k(k + 1)(-1)^{k-2} \mathbf{R}_u(k, m_k) \mathbf{Q}_{u,n,j} + 3 \sum_{u=2}^{n-j} \sum_{k=1}^{u-2} k(k + 1)(-1)^{k-1} \mathbf{R}_u(k, m_k) \mathbf{Q}_{u,n,j} \right], \end{aligned}$$

and

$$\Omega(x|n) = \sum_{j=1}^{n-1} \left[ \mathbf{A}_j \sum_{u=2}^{n-j} [\mathbf{F}_{x,u} + \mathbf{G}_{u,n,j} + \mathbf{H}_{x,u} \mathbf{Q}_{u,n,j}] \right], \quad \forall x \geq 4.$$

**Proof.** Appendix A.5. □

#### 4. Illustration

In this section we will accompany the theoretical results of the paper with two applications related to DNA sequences. It is known that a DNA strand consists of a sequence of adenine (A), guanine (G), cytosine (C) and thymine (T), which are the four nucleotides. We assume that a DNA sequence could be described by a homogeneous discrete SMC  $\{X_t\}_{t=0}^\infty$  with state space  $S = \{w_1, w_2, \dots, w_N\}$ , where  $w_i, i = 1, 2, \dots, N$  is a specific word that is a combination of the letters of the DNA alphabet  $S = \{A, C, G, T\}$  with length  $l$  and  $t$  denoting the position of the word inside the sequence.

##### 4.1. Inverted Repeats

The main focus of the following approach is the appearance of specific words formed from the alphabet A, C, G, T and their symmetric complements (inverted repeats). Inverted repeats are commonly found in eukaryotic genomes [48]. The presence of inverted repeats could form DNA cruciforms that have been shown to play an important role in the regulation of natural processes involving DNA. The cruciform structures are important for various biological processes, including replication, regulation of gene expression and

nucleosome structure. They have also been implicated in the development of diseases including cancer, Werner's syndrome and others [49].

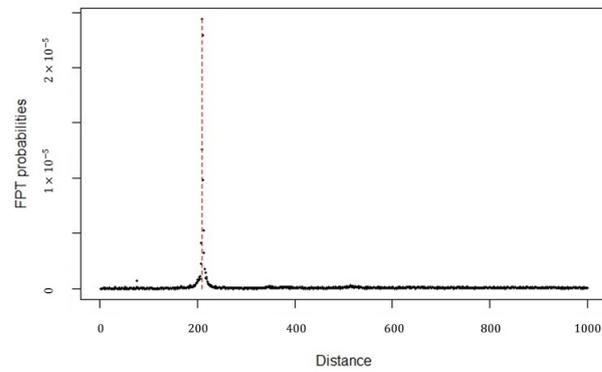
For each DNA word  $w$ , there exists a reversed complement of the word  $w'$ . For example, the word  $w = ACG$  has the word  $w' = CGT$  as an inverted repeat. The main question that we will attempt to address by applying the analytic relationships derived earlier is the following: Given that the SMC entered at the initial position in the word  $w$ , we want to estimate the probability of the reversed complement word  $w'$  appearing for the first time after a certain range of letters  $n$ . We define the distance,  $d$ , between two words as the number of letters between the first letter of the initial word that has appeared and the first letter of the following word that subsequently appears. For the sake of simplicity, we consider only the scenario where  $d > l$ . The DNA sequence that was used for this illustration is the first chromosome of the human genome consisting of 248,956,422 base-pairs that are publicly available from the website of the National Center for Biotechnology Information (NCBI) [50].

For the first illustration, three words of length  $l = 7$  were chosen that have been previously shown to exhibit different distances between them and their inverted complements [51]. The words were  $w_1 = GGCTCAC$ ,  $w_2 = ATATATG$  and  $w_3 = CCACAAT$ . For each word, the state space of the SMC consisted of the word and its reversed complement, e.g.,  $S = \{w_i, w'_i\}$ . First, the basic parameters of the SMC were estimated, namely the transition probability matrix and the sequence of sojourn times. The sojourn time was defined as the distance, i.e., the number of nucleotides that occur between each word and its inverted repeat. The transition matrix and the empirical distribution of the sojourn times were estimated using the empirical estimators. The sequence of the core matrices was calculated as the Hadamard product of the transition matrix with the sequence of the sojourn time matrices. For each word  $w \in S$ , the first passage time probability was calculated between the word  $w$  and its reversed complement  $w'$  according to the proposed analytic relationship (Theorem 1). For a maximum distance, ( $n = 1000$ ), the highest first passage time probabilities of the three words and their inverted repeats, along with the corresponding distances are illustrated in Figure 1. Concretely, the first passage time probabilities were calculated for the human Chromosome 1, aiming to estimate the most probable distances between words and their symmetrical complements. More specifically, as presented in Figure 1, we have noted that, for the first passage time probabilities, we have  $\operatorname{argmax}(f_{w_1 w'_1}) = 210$ ,  $\operatorname{argmax}(f_{w_2 w'_2}) = 10$  and  $\operatorname{argmax}(f_{w_3 w'_3}) = 132$  approximating the numerical results of previous studies with corresponding values for the arguments 210, 15 and 133 for the three words, respectively [51]. This highlights the fact that specific DNA words exhibit different behaviors and the distance between them and their inverted repeats demonstrates variability.

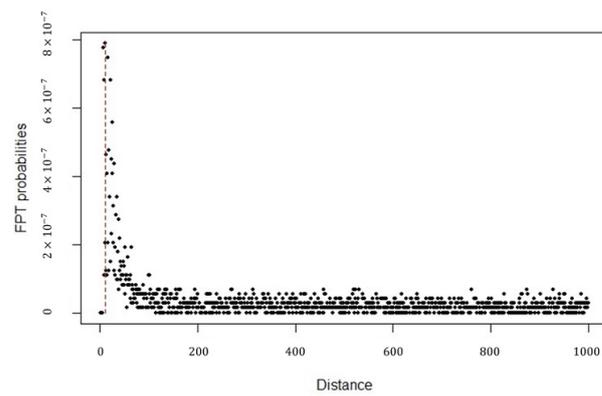
#### 4.2. CpG Islands

Usually, in vertebrate DNA sequences, the dinucleotide CG occurs less frequently than expected [52]. For the second illustration, we considered CpG islands, which are genomic regions that contain an elevated number of the dinucleotide CG. The human genome contains approximately 30 thousand CpG islands. The APRT gene is an example of a CpG region and it was used for this analysis [53]. This gene provides instructions for making an enzyme called adenine phosphoribosyltransferase (APRT). APRT contains approximately 2500 nucleotides and it had been shown to include an elevated amount of the dinucleotide GC [54]. We modeled the sequence of this DNA region as a homogeneous SMC with state space containing all the two-letter words from the DNA alphabet. The transition probability matrix and the sojourn times were estimated using the empirical estimators. The occupancy distribution  $\omega_{GCC}(x|n)$  for a fixed length of  $n = 100$  was calculated using the analytic relationship from Theorem 4 in order to estimate the occupancy distribution of specific words up to a specified sequence length. For comparison, we also applied the model to an intron sequence of human's phosphodiesterase gene (PDEA) [55]. The two sequences are publicly available from the NCBI. The occupancy probabilities are presented in Figure 2 up

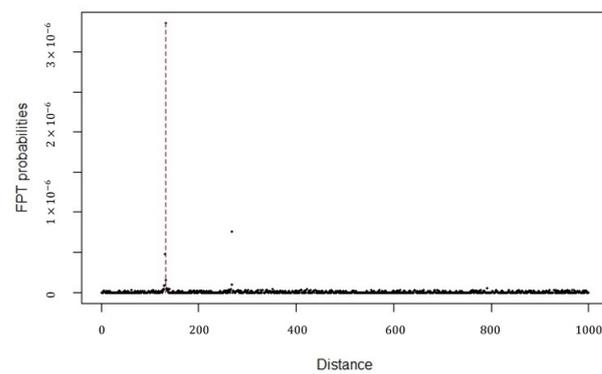
to length  $n = 50$ . It is confirmed that the number of occupancies of the dinucleotide GC will be greater in the CpG island compared to the intron sequence. As expected, the occupancy probabilities applied on the two sequences indicated that the occurrences of GCs were more frequent in the CpG sequence.



(a)  $w_1 = GGCTCAC$

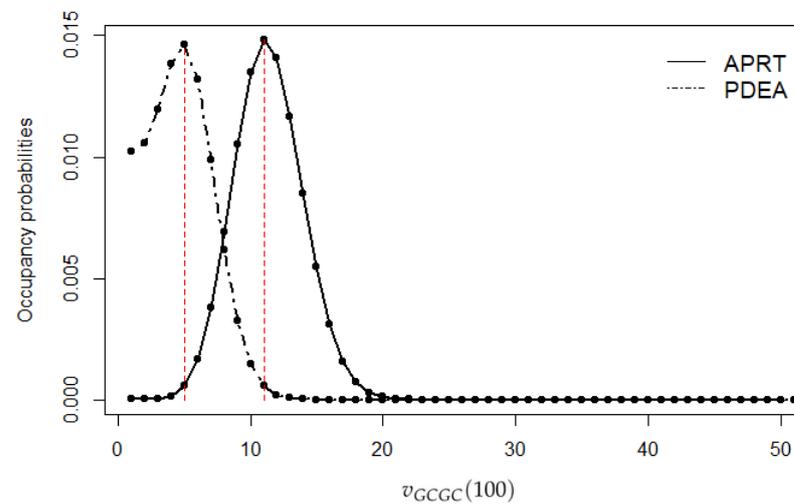


(b)  $w_2 = ATATATG$



(c)  $w_3 = CCACAAT$

**Figure 1.** First passage time (FPT) probabilities for distance  $n \leq 1000$ .



**Figure 2.** Occupancy probabilities of APRT and PDEA genes.

### 5. Concluding Remarks

In this article, three classes of important probabilities of a semi-Markov process, namely the first passage time, the occupancy and the duration probabilities were defined and their closed analytic forms were proved by using the basic parameters of the process. The study of the first passage time probability provides information regarding the distribution of the time elapsed to reach a state from another for the first time, either in terms of transitions or time. The second category of duration probabilities provides information about the distribution of the number of virtual transitions taking place before an actual transition to a different state occurs. Finally, the third class of probabilities provides insight information regarding the distribution of the number of times the SMC makes transitions to some state in a time interval of a given length. We provided analytic forms on the actual behavior of the recursive relations of the aforementioned probabilities and included these results into specific propositions and theorems.

The analytical results were accompanied with two illustrations on human genome DNA strands which are often studied using probabilistic modeling and, specifically, Markovian models. Although, in the relevant literature, there exist several algorithmic approaches analyzing the occupancy and appearance of words in DNA sequences, the results of the illustration section strongly suggest that the proposed modeling framework could also be used for the investigation of the structure of genome sequences.

Of course nothing comes without limitations and motivation for further research. For example, additional research effort could aim towards high-order dependencies since DNA sequences often show long-range correlations. This could result in a more coherent modeling approach. Furthermore, additional parameters could be included in the model, for example the length of sequence or specific mutations, resulting in more realistic representations regarding the different structures of complex genome of humans and other organisms. Finally, the proposed model could be applied in completely different contexts, such as natural language processing, linguistics, text similarity and anomaly detection, i.e., areas of machine learning that appear to be amongst the most popular areas in the last decade in data science and stochastic modeling.

**Author Contributions:** Conceptualization, A.C.G., A.P. and P.K.; Data curation, P.K.; Formal analysis, P.K.; Investigation, A.P. and P.K.; Methodology, A.C.G., A.P., H.P. and E.F.; Software, P.K.; Supervision, A.C.G. and A.P.; Validation, A.C.G.; Visualization, P.K.; Writing—original draft, A.P., P.K., H.P. and E.F.; Writing—review & editing, A.C.G., A.P. and P.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Publicly available datasets were analyzed in this study. This data can be found here: <https://www.ncbi.nlm.nih.gov/nuccore/CM000663>, <https://www.ncbi.nlm.nih.gov/gtr/genes/353/>, <https://www.ncbi.nlm.nih.gov/nuccore/1059792111>.

**Acknowledgments:** The authors greatly acknowledge the comments and suggestions of the three anonymous referees, which improved the content and the presentation of the current paper.

**Conflicts of Interest:** The authors declare no conflicts of interest.

### Appendix A. Proofs

#### Appendix A.1. Proof of Theorem 1

The results for (1) and (2) are obvious. For the third part, we used the matrix notation of the first passage time probabilities:

$$F(k|t, n) = \sum_{m=1}^{n-k+1} C(t, m) \{ F(k-1|t+m, n-m) \circ B \} + \delta(k-1)C(t, m),$$

with  $F(k|t, n) = 0$  if  $k > n$  or  $k = 0$ . For  $k = 1$  and  $m = m_i$  we have shown the results for the case where  $k > 1$  can be proved by induction. Thus, we assume that this result holds for  $k - 1$  and we will show that it also holds for each  $k \leq n$ . Here we note that the recursive relationship of the first passage time probabilities could be reformulated as follows.

$$f_{ij}(k|t, n) = \sum_{m_1=1}^{n-k+1} \sum_{x_1 \neq j} c_{ix_1}(t, m_1) \left\{ \sum_{m_2=1+m_1}^{n-k+2} \sum_{x_2 \neq j} c_{x_1x_2}(t+m_1, m_2-m_1) \right\} \\ \left\{ \dots \left\{ \sum_{m_{k-1}=1+m_{k-2}}^{n-1} \sum_{x_{k-1} \neq j} c_{x_{k-2}x_{k-1}}(t+m_{k-2}, m_{k-1}-m_{k-2}) c_{x_{k-1}j}(t+m_{k-1}, n-m_{k-1}) \right\} \dots \right\} \\ + \delta(k-1)c_{ij}(t, n).$$

Using matrix notation, we can express the previous relationship as the following.

$$F(k|t, n) = \sum_{m_1=1}^{n-k+1} C(t, m_1) \left\{ \sum_{m_2=1+m_1}^{n-k+2} C(t+m_1, m_2-m_1) \right\} \\ \left\{ \dots \left\{ \sum_{m_{k-1}=1+m_{k-2}}^{n-1} C(t+m_{k-2}, m_{k-1}-m_{k-2}) \{ C(t+m_{k-1}, n-m_{k-1}) \circ B \} \right\} \dots \circ B \right\} \\ \text{for } 0 < k \leq n.$$

The initial conditions are  $F(k|t, n) = 0$  for  $k > n$  or  $k = 0$  and  $F(1|t, n) = C(t, n)$ . By using the following notation:

$$\sum_{m_1=1}^{n-k+1} \left\{ \sum_{m_2=1+m_1}^{n-k+2} \left\{ \dots \left\{ \sum_{m_{k-1}=1+m_{k-2}}^{n-1} \right. \right. \right. = \sum_{m_1=1}^{n-k+1} * \sum_{m_2=1+m_1}^{n-k+2} * \dots * \sum_{m_{k-1}=1+m_{k-2}}^{n-1} ,$$

we obtain the following.

$$F(k|t, n) = \sum_{m=1}^{n-k+1} C(t, m) \left\{ \left\{ \sum_{m_1=1}^{n-m-k+2} * \sum_{m_2=1+m_1}^{n-m-k+3} * \dots * \sum_{m_{k-2}=1+m_{k-3}}^{n-m-1} C(t, m_1) \right\} \right\} \\ \{ C(t+m_1, m_2-m_1) \} \dots \{ C(t+m+m_{k-2}, n-m-m_{k-2}) \circ B \} \circ B \dots \circ B.$$

By the appropriate substitution of the time indices and by the definition of the following operation  $\prod_{r=1}^2 \{B\} A_r = A_1 * B A_2 = A_2(A_1 \circ B)$  for the matrices  $A_1, A_2, B$ , we obtain the desired result.  $\square$

Appendix A.2. Proof of Theorem 2

The results for (1) and (2) are obvious. For the third part, we used induction. By using matrix notation on the recursive relationship, it holds that, for  $k = 2$ , we have the following.

$$D(2|t, n) = \sum_{m_1=1}^{n-1} (C(t, m_1) \circ I)(W(t + m_1, n - m_1) - C(t + m_1, n - m_1) \circ I)$$

Now assume that the relationship hold for  $k - 1$ , which is the following.

$$D(k - 1|t + m, n - m) = \sum_{m_1=1}^{n-m-k+2} * \sum_{m_2=1+m_1}^{n-m-k+3} * \dots * \sum_{m_{k-2}=1+m_{k-3}}^{n-m-1} (C(t + m, m_1) \circ I) \\ (C(t + m + m_1, m_2 - m_1) \circ I) \dots (C(t + m + m_{k-3}, m_{k-2} - m_{k-3}) \circ I) \\ (W(t + m + m_{k-2}, n - m - m_{k-2}) - C(t + m + m_{k-2}, n - m - m_{k-2}) \circ I).$$

Therefore, the following obtains.

$$D(k|t, n) = \sum_{m=1}^{n-k+1} * \sum_{m_1=1}^{n-m-k+2} * \dots * \sum_{m_{k-2}=1+m_{k-3}}^{n-m-1} (C(t + m, m_1) \circ I) \\ (C(t + m + m_1, m_2 - m_1) \circ I) \dots (C(t + m + m_{k-3}, m_{k-2} - m_{k-3}) \circ I) \\ (W(t + m + m_{k-2}, n - m - m_{k-2}) - C(t + m + m_{k-2}, n - m - m_{k-2}) \circ I).$$

By appropriately substituting the time indices with  $m'_0 = 0, m'_1 = m, m'_2 = m + m_1, \dots, m'_i = m + m_{i-1}, \dots, m'_{k-1} = m + m_{k-2}, i = 1, 2, \dots, k - 1$ , where  $1 + m_{i-1} \leq m_i \leq n - m - k + i + 1$ , we obtain the following:

$$D(k|t, n) = \sum_{m'_1=1}^{n-k+1} * \sum_{m'_2=1+m'_1}^{n-k+2} * \dots * \sum_{m'_{k-1}=1+m'_{k-2}}^{n-1} (C(m'_1) \circ I)(C(m'_2 - m'_1) \circ I)(C(m'_3 - m'_2) \circ I) \\ \dots (C(m'_{k-1} - m'_{k-2}) \circ I)(W(n - m'_{k-1}) - C(n - m'_{k-1}) \circ I),$$

which results in the stated relationship.  $\square$

Appendix A.3. Proof of Theorem 3

Assuming homogeneity in time, Equation (4) is provided by the following:

$$\omega_{ij}(x|n) = \sum_{r=1}^N \sum_{\substack{m=0 \\ r \neq j}}^n c_{ir}(m) \omega_{rj}(x|n - m) + \\ + \sum_{m=0}^n c_{ij}(m) \omega_{jj}(x - 1|n - m) + \delta(x) > w_i(n), \tag{A1}$$

where  $i, j \in S, n = 0, 1, \dots$  and  $x = 0, 1, \dots$  Equation (A1) can be written as follows.

$$\omega_{ij}(x|n) = \sum_{r=1}^N \sum_{m=0}^n c_{ir}(m) [\omega_{rj}(x|n - m)(1 - \delta_{rj}) + \omega_{rj}(x - 1|n - m)\delta_{rj}] + \delta(x) > w_i(n). \tag{A2}$$

Equation (A2) in matrix notation is the following.

$$\Omega(x|n) = \sum_{m=1}^n C(m) [\Omega(x|n - m) \circ (U - I) + \Omega(x - 1|n - m) \circ I] + \delta(x) > W(n).$$

By applying the geometric transform to the above, we obtain the following:

$$\Omega^g(z|n) = \sum_{m=1}^n C(m)\Omega^g(z|n-m) + (z-1) \sum_{m=1}^n C(m)[\Omega^g(z|n-m) \circ \mathbf{I}] \circ \mathbf{I} + \mathbf{W}(n),$$

with initial condition  $\Omega^g(z|0) = \mathbf{I}$ . Following the methodology of Vassiliou and Papadopoulou (1992), we derive the result of the Theorem 3. [15]

Appendix A.4. Proof of Lemma 1

By using the Hadamard product on Theorem 3, we have the following.

$$\begin{aligned} \Omega^g(z|n) \circ \mathbf{I} &= - (z-1) \sum_{j=1}^{n-1} \left[ \left[ \sum_{i=1}^j \mathbf{a}_{1i}^{-1} C(j+1-i) \right] [\Omega^g(z|n-j) \circ \mathbf{I}] \right] \circ \mathbf{I} \\ &\quad - z \sum_{j=1}^n \left[ \mathbf{a}_{1j}^{-1} C(n+1-j) \right] \circ \mathbf{I} - \sum_{j=1}^n \left[ \mathbf{a}_{1j}^{-1} \mathbf{W}(n+1-j) \right] \circ \mathbf{I}. \end{aligned}$$

By using the following property:

$$(\mathbf{A}(\mathbf{B} \circ \mathbf{I})) \circ \mathbf{I} = (\mathbf{A} \circ \mathbf{I})(\mathbf{B} \circ \mathbf{I}).$$

we obtain the following:

$$\left[ \left[ \sum_{i=1}^j \mathbf{a}_{1i}^{-1} C(j+1-i) \right] [\Omega^g(z|n-j) \circ \mathbf{I}] \right] \circ \mathbf{I} = \left[ \left[ \sum_{i=1}^j \mathbf{a}_{1i}^{-1} C(j+1-i) \right] \circ \mathbf{I} \right] [\Omega^g(z|n-j) \circ \mathbf{I}].$$

which completes the proof. □

Appendix A.5. Proof of Theorem 4

An early version of the proof of Theorem 4 can be found in [56]. We analytically present here all necessary steps of the proof. Using the equations provided by the results of Theorem 3 and by substituting  $\Omega^g(z|n) \circ \mathbf{I}$  with the result found in Lemma 1, we can obtain the analytic relation for the geometric transforms of  $\Omega^g(z|n)$ , which is as follows:

$$\begin{aligned} \Omega^g(z|n) &= (z-1) \sum_{j=1}^{n-1} \mathbf{A}_j \left[ \begin{aligned} &z\mathbf{G}_{1,n,j} + z \sum_{u=2}^{n-j} \left[ (z-1)\mathbf{M}'_u + \sum_{k=1}^{u-2} (z-1)^{k+1} \mathbf{R}_u(k, m_k) \right] \mathbf{G}_{u,n,j} \\ &+ \mathbf{Q}_{1,n,j} + \sum_{u=2}^{n-j} \left[ (z-1)\mathbf{M}'_u + \sum_{k=1}^{u-2} (z-1)^{k+1} \mathbf{R}_u(k, m_k) \right] \mathbf{Q}_{u,n,j} \end{aligned} \right] + \quad (A3) \\ &\quad + z\mathbf{A}_n + \mathbf{E}_n, \end{aligned}$$

where

$$\begin{aligned} \mathbf{A}_j &= C(j) + \sum_{i=2}^j \left( C(i-1) + \sum_{k=1}^{i-2} \mathbf{S}_i(k, m_k) \right) C(j+1-i), \\ \mathbf{M}'_u &= \left[ C(u-1) + \sum_{i=2}^{u-1} \left( C(i-1) + \sum_{k=1}^{i-2} \mathbf{S}_i(k, m_k) \right) C(u-i) \right] \circ \mathbf{I}, \\ \mathbf{E}_n &= \sum_{j=2}^n \left( C(j-1) + \sum_{k=1}^{j-2} \mathbf{S}_j(k, m_k) \right) \mathbf{W}(n+1-j) \circ \mathbf{I} + \mathbf{W}(n), \\ \mathbf{G}_{u,n,j} &= C(n-j+1-u) \circ \mathbf{I} + \sum_{w=2}^{n-j+1-u} \left[ \left( C(w-1) + \sum_{k=1}^{w-2} \mathbf{S}_w(k, m_k) \right) C(n-j+2-u-w) \right] \circ \mathbf{I}, \\ \mathbf{Q}_{u,n,j} &= \sum_{w=2}^{n-j+1-u} \left[ \left( C(w-1) + \sum_{k=1}^{w-2} \mathbf{S}_w(k, m_k) \right) \mathbf{W}(n-j+2-u-w) \right] \circ \mathbf{I} + [\mathbf{W}(n-j+1-u) \circ \mathbf{I}]. \end{aligned}$$

Then, by applying properties of the inverse geometric transforms by using the equation  $\Omega(x|n) = \frac{1}{x!} \frac{d^n}{dz^n} \Omega^g(z|n) \Big|_{z=0}$  and by repeatedly taking the derivatives of  $\Omega^g(z|n)$  with respect to  $z$ , we obtain the result of the Theorem 5 for  $x \geq 1$ .

Finally, for the special case where  $x = 0$ , by substituting  $z = 0$  in expression (A3), we obtain the following:

$$\Omega(0|n) = - \sum_{j=1}^{n-1} \mathbf{A}_j \left[ \mathbf{B}_{n,j} + \sum_{u=2}^{n-j} \mathbf{M}_u \mathbf{Q}_{u,n,j} \right] + \mathbf{E}_n,$$

where the following results.

$$\mathbf{B}_{n,j} = \left[ \sum_{w=2}^{n-j} \left[ \left( \mathbf{C}(w-1) + \sum_{k=1}^{w-2} \mathbf{S}_w(k, m_k) \right) \mathbf{W}(n-j+1-w) \right] \circ \mathbf{I} + \mathbf{W}(n-j) \right] \circ \mathbf{I},$$

$$\mathbf{M}_u = - \left[ \mathbf{C}(u-1) + \sum_{i=2}^{u-1} \left( \mathbf{C}(i-1) + \sum_{k=1}^{i-2} \mathbf{S}_i(k, m_k) \right) \mathbf{C}(u-i) \right] \circ \mathbf{I} + \sum_{k=1}^{u-2} (-1)^{k+1} \mathbf{R}_u(k, m_k).$$

### Appendix B

#### Appendix B.1

$$\begin{aligned} >f_{ij}(n) = 1 - f_{ij}(n) \Rightarrow >f_{ij}^g(z) = \sum_{n=0}^{\infty} >f_{ij}(n)z^n = \sum_{n=0}^{\infty} (1 - f_{ij}(n))z^n = \\ &= \sum_{n=0}^{\infty} z^n - \sum_{n=0}^{\infty} f_{ij}(n)z^n = \frac{1}{1-z} - \sum_{n=0}^{\infty} \left( \sum_{m=0}^n f_{ij}(m) \right) z^n = \\ &= \frac{1}{1-z} - \sum_{m=0}^{\infty} \sum_{n=m}^{\infty} f_{ij}(m)z^m z^{n-m} = \frac{1}{1-z} - \frac{f_{ij}^g(z)}{1-z} = \frac{1-f_{ij}^g(z)}{1-z}. \end{aligned}$$

#### Appendix B.2

$$\begin{aligned} \omega_{ij}^{gg}(y|z) &= >f_{ij}^g(z) + yf_{ij}^g(z) \frac{>f_{ij}^g(z)}{(1-yf_{ij}^g(z))} \\ &= \frac{1-f_{ij}^g(z)}{1-z} + \frac{yf_{ij}^g(z)}{(1-yf_{ij}^g(z))} \frac{1-f_{ij}^g(z)}{1-z} \\ &= \frac{(1-f_{ij}^g(z))(1-yf_{ij}^g(z)) + yf_{ij}^g(z)(1-f_{ij}^g(z))}{(1-z)(1-yf_{ij}^g(z))} \\ &= \frac{1-yf_{ij}^g(z) - f_{ij}^g(z) + yf_{ij}^g(z)f_{ij}^g(z) + yf_{ij}^g(z) - yf_{ij}^g(z)f_{ij}^g(z)}{(1-z)(1-yf_{ij}^g(z))} \\ &= \frac{(1-yf_{ij}^g(z)) - f_{ij}^g(z) + yf_{ij}^g(z)}{(1-z)(1-yf_{ij}^g(z))} \\ &= \frac{(1-yf_{ij}^g(z)) - (1-y)f_{ij}^g(z)}{(1-z)(1-yf_{ij}^g(z))}. \end{aligned}$$

### References

1. Pyke, R. Markov renewal processes with finitely many states. *Ann. Math. Stat.* **1961**, *32*, 1243–1259. [\[CrossRef\]](#)
2. Cinlar, E. *Introduction to Stochastic Processes*; Courier Corporation: Chelmsford, MA, USA, 2013.
3. Howard, R.A. *Dynamic Probabilistic Systems: Semi-Markov and Decision Processes*; Dover Publications: Mineola, NY, USA, 2007; Volume 2.
4. McClean, S.I. A semi-Markov model for a multigrade population with Poisson recruitment. *J. Appl. Probab.* **1980**, *17*, 846–852. [\[CrossRef\]](#)
5. McClean, S.I. Semi-Markov models for manpower planning. In *Semi-Markov Models*; Springer: Berlin/Heidelberg, Germany, 1986; pp. 283–300.

6. D'Amico, G.; Di Biase, G.; Janssen, J.; Manca, R. *Semi-Markov Migration Models for Credit Risk*; Wiley Online Library: Hoboken, NJ, USA, 2017.
7. Vassiliou, P.-C.G. Non-Homogeneous Semi-Markov and Markov Renewal Processes and Change of Measure in Credit Risk. *Mathematics* **2021**, *9*, 55. [[CrossRef](#)]
8. Janssen, J.; Manca, R. *Applied Semi-Markov Processes*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2006.
9. Janssen, J. *Semi-Markov Models: Theory and Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
10. Schbath, S.; Prum, B.; de Turckheim, E. Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *J. Comput. Biol.* **1995**, *2*, 417–437. [[CrossRef](#)]
11. De Dominicis, R.; Manca, R. Some new results on the transient behaviour of semi-Markov reward processes. *Methods Oper. Res.* **1986**, *53*, 387–397.
12. Vasileiou, A.; Vassiliou, P.-C.G. An inhomogeneous semi-Markov model for the term structure of credit risk spreads. *Adv. Appl. Probab.* **2006**, *38*, 171–198. [[CrossRef](#)]
13. Vassiliou, P.-C.G.; Vasileiou, A. Asymptotic behaviour of the survival probabilities in an inhomogeneous semi-Markov model for the migration process in credit risk. *Linear Algebra Appl.* **2013**, *438*, 2880–2903. [[CrossRef](#)]
14. Vassiliou, P.-C.G. Semi-Markov migration process in a stochastic market in credit risk. *Linear Algebra Appl.* **2014**, *450*, 13–43. [[CrossRef](#)]
15. Vassiliou, P.-C.G.; Papadopoulou, A. Non-homogeneous semi-Markov systems and maintainability of the state sizes. *J. Appl. Probab.* **1992**, *29*, 519–534. [[CrossRef](#)]
16. Vassiliou, P.-C.G. Asymptotic behavior of Markov systems. *J. Appl. Probab.* **1982**, *19*, 851–857. [[CrossRef](#)]
17. Dimitriou, V.; Georgiou, A.C. Introduction, analysis and asymptotic behavior of a multi-level manpower planning model in a continuous time setting under potential department contraction. *Commun. Stat. Theory Methods* **2021**, *50*, 1173–1199. [[CrossRef](#)]
18. Papadopoulou, A.; Vassiliou, P.-C.G. Asymptotic behavior of nonhomogeneous semi-Markov systems. *Linear Algebra Appl.* **1994**, *210*, 153–198. [[CrossRef](#)]
19. Papadopoulou, A.; Vassiliou, P.-C.G. On the variances and covariances of the duration state sizes of semi-Markov systems. *Commun. Stat. Theory Methods* **2014**, *43*, 1470–1483. [[CrossRef](#)]
20. Vassiliou, P.-C.G. Markov systems in a general state space. *Commun. Stat. Theory Methods* **2014**, *43*, 1322–1339. [[CrossRef](#)]
21. Dimitriou, V.A.; Georgiou, A.C.; Tsantas, N. The multivariate non-homogeneous Markov manpower system in a departmental mobility framework. *Eur. J. Oper. Res.* **2013**, *228*, 112–121. [[CrossRef](#)]
22. Symeonaki, M. Theory of fuzzy non homogeneous Markov systems with fuzzy states. *Qual. Quant.* **2015**, *49*, 2369–2385. [[CrossRef](#)]
23. Tsaklidis, G.; Vassiliou, P.-C.G. Asymptotic periodicity of the variances and covariances of the state sizes in non-homogeneous Markov systems. *J. Appl. Probab.* **1988**, *25*, 21–33. [[CrossRef](#)]
24. Vassiliou, P.-C.G. The evolution of the theory of non-homogeneous Markov systems. *Appl. Stoch. Model. Data Anal.* **1997**, *13*, 159–176. [[CrossRef](#)]
25. Vassiliou, P.-C.G.; Georgiou, A.C. Asymptotically attainable structures in nonhomogeneous Markov systems. *Oper. Res.* **1990**, *38*, 537–545. [[CrossRef](#)]
26. Ugwuowo, F.I.; McClean, S.I. Modelling heterogeneity in a manpower system: A review. *Appl. Stoch. Model. Bus. Ind.* **2000**, *16*, 99–110. [[CrossRef](#)]
27. Symeonaki, M.; Stamatopoulou, G. Describing labour market dynamics through Non Homogeneous Markov System theory. In *Demography of Population Health, Aging and Health Expenditures*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 359–373.
28. Ossai, E.; Uche, P. Maintainability of departmentalized manpower structures in Markov chain model. *Pac. J. Sci. Technol.* **2009**, *2*, 295–302.
29. Guerry, M.A.; De Feyter, T. Optimal recruitment strategies in a multi-level manpower planning model. *J. Oper. Res. Soc.* **2012**, *63*, 931–940. [[CrossRef](#)]
30. Hunter, J.J. Stationary distributions and mean first passage times of perturbed Markov chains. *Linear Algebra Appl.* **2005**, *410*, 217–243. [[CrossRef](#)]
31. Hunter, J.J. Simple procedures for finding mean first passage times in Markov chains. *Asia-Pac. J. Oper. Res.* **2007**, *24*, 813–829. [[CrossRef](#)]
32. Hunter, J.J. The computation of the mean first passage times for Markov chains. *Linear Algebra Appl.* **2018**, *549*, 100–122. [[CrossRef](#)]
33. Yao, D.D. First-passage-time moments of Markov processes. *J. Appl. Probab.* **1985**, *22*, 939–945. [[CrossRef](#)]
34. Zhang, X.; Hou, Z. The first-passage times of phase semi-Markov processes. *Stat. Probab. Lett.* **2012**, *82*, 40–48. [[CrossRef](#)]
35. Pitman, J.; Tang, W. Tree formulas, mean first passage times and Kemeny's constant of a Markov chain. *Bernoulli* **2018**, *24*, 1942–1972. [[CrossRef](#)]
36. Redner, S. *A Guide to First-Passage Processes*; Cambridge University Press: Cambridge, UK, 2001.
37. Waterman, M.S. *Introduction to Computational Biology: Maps, Sequences and Genomes*; CRC Press: Boca Raton, FL, USA, 1995.
38. Almagor, H. A Markov analysis of DNA sequences. *J. Theor. Biol.* **1983**, *104*, 633–645. [[CrossRef](#)]
39. Menéndez, M.; Pardo, L.; Pardo, M.; Zografos, K. Testing the order of Markov dependence in DNA sequences. *Methodol. Comput. Appl. Probab.* **2011**, *13*, 59–74. [[CrossRef](#)]
40. Skewes, A.D.; Welch, R.D. A Markovian analysis of bacterial genome sequence constraints. *PeerJ* **2013**, *1*, e127. [[CrossRef](#)]

41. Papapetrou, M.; Kugiumtzis, D. Markov chain order estimation with conditional mutual information. *Phys. A: Stat. Mech. Appl.* **2013**, *392*, 1593–1601. [[CrossRef](#)]
42. Boys, R.J.; Henderson, D.A.; Wilkinson, D.J. Detecting homogeneous segments in DNA sequences by using hidden Markov models. *J. R. Stat. Soc. Ser. C* **2000**, *49*, 269–285. [[CrossRef](#)]
43. Reinert, G.; Schbath, S.; Waterman, M.S. Probabilistic and statistical properties of words: An overview. *J. Comput. Biol.* **2000**, *7*, 1–46. [[CrossRef](#)]
44. Robin, S.; Daudin, J.J. Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Probab.* **1999**, *36*, 179–193. [[CrossRef](#)]
45. Schbath, S. An overview on the distribution of word counts in Markov chains. *J. Comput. Biol.* **2000**, *7*, 193–201. [[CrossRef](#)] [[PubMed](#)]
46. Papadopoulou, A. Some Results on Modeling Biological Sequences and Web Navigation with a Semi Markov Chain. *Commun. Stat. Theory Methods* **2013**, *42*, 2853–2871. [[CrossRef](#)]
47. Ricciardi, L.; Crescenzo, A.; Giorno, V.; Nobile, A. An outline of theoretical and algorithmic approaches to first passage time problems with applications to biological modeling. *Math. Jpn.* **1999**, *50*, 247–322.
48. Lavi, B.; Levy Karin, E.; Pupko, T.; Hazkani-Covo, E. The prevalence and evolutionary conservation of inverted repeats in proteobacteria. *Genome Biol. Evol.* **2018**, *10*, 918–927. [[CrossRef](#)] [[PubMed](#)]
49. Brázda, V.; Laister, R.C.; Jagelská, E.B.; Arrowsmith, C. Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Mol. Biol.* **2011**, *12*, 1–16. [[CrossRef](#)]
50. Homo Sapiens Chromosome 1, GRCh38.p13 Primary Assembly. Available online: <https://www.ncbi.nlm.nih.gov/nuccore/CM000663> (accessed on 17 December 2020).
51. Tavares, A.H.; Pinho, A.J.; Silva, R.M.; Rodrigues, J.M.; Bastos, C.A.; Ferreira, P.J.; Afreixo, V. DNA word analysis based on the distribution of the distances between symmetric words. *Sci. Rep.* **2017**, *7*, 1–11. [[CrossRef](#)] [[PubMed](#)]
52. Gardiner-Garden, M.; Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **1987**, *196*, 261–282. [[CrossRef](#)]
53. APRT adenine phosphoribosyltransferase. Available online: <https://www.ncbi.nlm.nih.gov/gtr/genes/353/> (accessed on 17 December 2020).
54. Broderick, T.P.; Schaff, D.A.; Bertino, A.M.; Dush, M.K.; Tischfield, J.A.; Stambrook, P.J. Comparative anatomy of the human APRT gene and enzyme: Nucleotide sequence divergence and conservation of a nonrandom CpG dinucleotide arrangement. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 3349–3353. [[CrossRef](#)] [[PubMed](#)]
55. Homo sapiens Human Phosphodiesterase (PDEA) Gene. Available online: <https://www.ncbi.nlm.nih.gov/nuccore/1059792111> (accessed on 17 December 2020).
56. Farmakioti, E. Probabilities of State Occupancies in Semi-Markov Chains. Master's Thesis, Aristotle University of Thessaloniki, Thessaloniki, Greece, 2018.