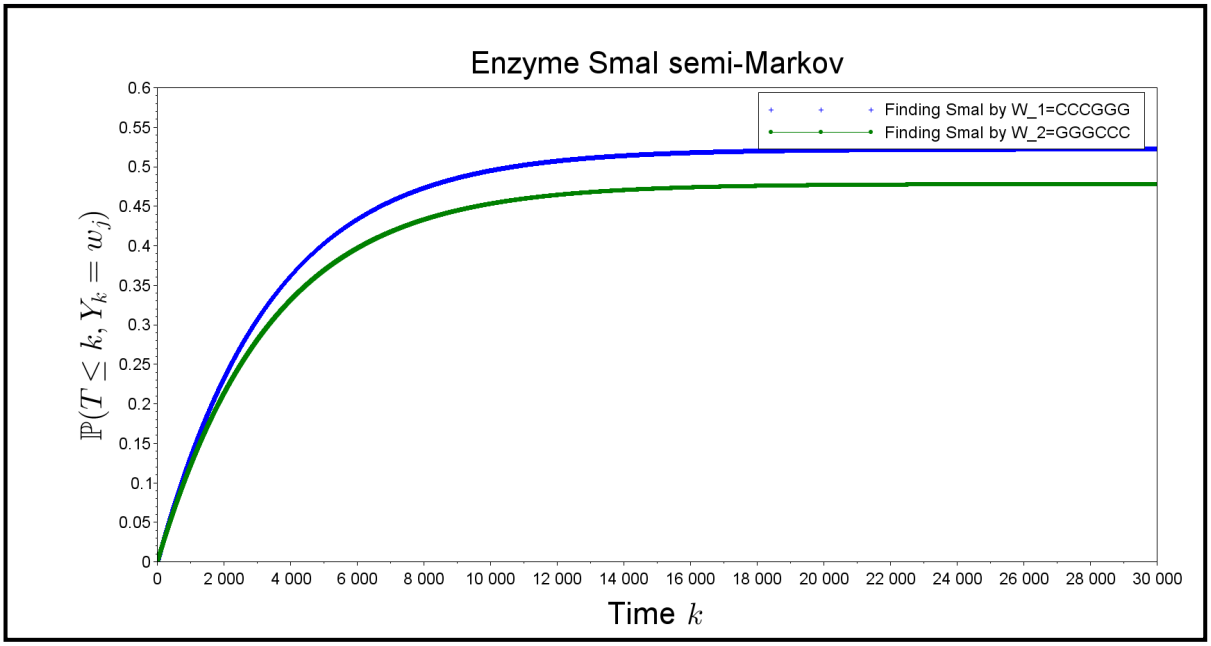
Properties of words and competing risk processes under semi-Markov hypothesis

Brenda Ivette Garcia Maya

Par **Brenda Ivette GARCIA MAYA**

Properties of words and competing risk processes under semi-Markov hypothesis

Thèse présentée
pour l'obtention du grade
de Docteur de l'UTC



Soutenue le 29 septembre 2020
Spécialité : Mathématiques Appliquées et Statistique : Laboratoire de Mathématiques Appliquées de Compiègne (Unité de recherche EA-2222)
D2569

SORBONNE UNIVERSITÉS, UNIVERSITÉ DE TECHNOLOGIE DE COMPIÈGNE

Laboratoire de Mathématiques Appliquées de Compiègne (LMAC)

École Doctorale "Sciences Pour L'Ingenieur"

Properties of words and competing risk processes under semi-Markov hypothesis

T H È S E

présentée par

Brenda Ivette GARCIA MAYA

pour l'obtention du grade de Docteur

Spécialité : Mathématiques Appliquées et Statistique

Sous la direction de :

Nikolaos Limnios,

Compiègne, France. 29 septembre 2020

SORBONNE UNIVERSITES, UNIVERSITE DE TECHNOLOGIE DE COMPIEGNE

Laboratoire de Mathématiques Appliquées de Compiègne (LMAC)

École Doctorale "Sciences Pour L'Ingenieur"

Propriétés des mots et problèmes de risques compétitifs dans l'hypothèse de semi-Markov

T H E S I S

presented by

Brenda Ivette GARCIA MAYA

to obtain the degree of Doctor

Spécialité : Mathématiques Appliquées et Statistique

Under de direction of:

Nikolaos Limnios,

Compiègne, France. 29 septembre 2020

Acknowledgements

I want to give a big thank you to the institutions that supported this work from the begining.....

Contents

Résumé

Notre thèse est dédiée, en grande partie, à certains problèmes de biologie (séquences biologiques et analyse de la durée de vie avec risks compétitifs) sous l'hypothèse semi-markovienne.

Au cours des années récentes, calculer les propriétés des mots dans les séquences stochastiques a été un sujet d'intérêt à l'intersection de mathématiques appliquées et de biologie. Dans la littérature, un grand nombre de méthodes ont abordé cette problématique sous l'hypothèse que la séquence des symboles soit modélisée par un processus de Markov. Cependant, l'hypothèse markovienne a quelques inconvénients. Dans un processus de Markov, le temps de séjour dans un état est modélisé par la loi exponentielle (géométrique) en temps continu (discret). Au contraire, dans un processus de semi-Markov le temps de séjour peut être modélisé par n'importe quelle loi de probabilité. Donc, pour calculer les propriétés des mots dans des séquences aléatoires d'une façon plus générale, dans cette thèse, on a considéré que la séquence biologique est modélisé par un processus semi-markovien. On a calculé la loi et le nombre moyen des fois que les éléments d'un ensemble spécifique apparaissent dans une séquence des lettres. En suit, nous avons obtenu la loi des grandes nombres et nous avons aussi présenté le théorème de la limite central pour la fréquence d'apparition des mots. Pour montrer l'applicabilité de notre modèle, on a cherché une enzyme spécifique dans une séquence d'ADN provenant d'un bactériophage.

Les problèmes de risques compétitives forment un autre sujet d'intérêt en durée de vie. En général, les problèmes de risques compétitives ont été abordés à partir d'un point de vu statistique. Dans cette thèse, on présente les problèmes de risques compétitives dans le cadre de semi-Markov. On considère des processus de semi-Markov en temps continu et discret avec un nombre fini d'états transitoires et absorbants. Chaque état absorbant représente un mode de défaillance (dans la fiabilité d'un système) ou la cause de mort d'un individu (dans le cadre d'analyse de survie). On exprime la probabilité qu'une défaillance apparaisse au temps donné en raison d'une cause spécifique. On donne la loi jointe de la durée de vie et de la cause de défaillance en utilisant la fonction de transition d'un processus semi-markovien en temps continu et en temps discret, respectivement. Quelques exemples sont donnés pour illustration.

Nous présentons également une méthode de résolution des équations de renouvellement markovien en temps continu, en se basant sur les algorithmes bien établis des équations correspondantes en temps discret. Le grand avantage tiré par cette approche est que la série infinie de la fonction de renouvellement, en temps continu, est remplacée, en temps discret, par une série finie. Des résultats pour l'estimation de l'erreur sont également établis. Pour illustrer cette approche nous proposons une

application numérique concernant les cyber-attaques où les fonctions de transitions conditionnelles sont de lois de Weibull.

Mots clés. Processus semi-markovien, premier temps d'arrivée, théorème de la limite central, loi forte des grands nombres, théorème ergodique, fonction de renouvellement markovien, méthode de discrétisation, risques compétitives, analyse de survie.

Organisation de la thèse.
La thèse est organisée comme suit: dans le chapitre 1 on présente une introduction aux processus de semi-Markov en temps continu et discret, on donne une bref description des principaux travaux où les propriétés des mots dans une séquence stochastique sont abordées et on présente la théorie classique des risques compétitives. Dans le chapitre 2, on calcule la loi du temps (position) de la première occurrence d'un mot à travers d'une séquence semi-markovienne, on présente également la variance de cette variable aléatoire. Dans le chapitre 3 on étend les résultats présentés au deuxième chapitre, et on donne le théorème de la limite centrale, la loi forte des grandes nombres et le théorème ergodique pour un sous-ensemble de mots tirés d'un alphabet fini. Dans le chapitre 4 on abord le sujet des risques compétitives à partir du point de vu des processus semi-markoviens. Dans le chapitre 5 on présente une méthode de discrétisation pour calculer les fonctions de renouvellement markovien au temps continu, et on donne quelques exemples numériques. Finalement dans le chapitre 6 on présente quelques générales conclusions et perspectives.

Summary

Our thesis is dedicated in big part, to solving certain problems in biology (biologic sequences and lifespan using the competing risk framework) under semi-Markovian hypothesis.

In recent years, computing the properties of words through random sequences has become a topic of interest in the intersection between mathematics and biology. In the literature, a vast number of methods have tackled this problem under the assumption that sequences of symbols are modeled by Markov processes. Nevertheless, the markovian hypothesis has some disadvantages. In Markov processes, the sojourn time is modeled by the exponential (geometric) distribution in continuous (discrete) time. By contrast, in semi-Markov processes the sojourn time in a state can be modeled by any probability law. Therefore, in order to propose a more general approach to compute the properties of words through a random sequence, in this PhD work we consider that biological sequences are modeled by semi-Markovian discrete processes. We also compute the average number of times that the elements from a specific set of words appear through a sequence of letters. To achieve our goal, we use the strong law of large numbers and we provide the central limit theorem. To prove the application of our proposed model, we find a particular enzyme in a bacteriophage DNA sequence.

Competing risk problems conform another interesting topic in the lifespan domain. In general, competing risk problems have been dealt with a statistic approach. In this thesis, we present competing risk models within a semi-Markov framework. We consider continuous and discrete time semi-Markov processes with a finite number of transient and absorbing states. Each absorbing state represents a failure mode (in reliability of a system) or a cause of death of an individual (in survival analysis). We express the probability that a failure occurs at a given time due to a unique cause. We give the joint distribution of the lifetime and the failure cause via the transition function of the semi-Markov process in continuous and discrete-time respectively. Some examples are given for illustration.

We also present a method for solving continuous time Markovian renewal equations, based on well-established algorithms in their discrete time corresponding counterparts. The great advantage drawn by this approach is that the infinite series of the renewal function, in continuous time, is replaced, in discrete time, by a finite series. Results for error estimation are also established. To illustrate this approach we propose a digital application concerning cyber-attacks where the functions of conditional transitions are of the Weibull type.

Key Words. semi-markovian process, first hitting time, central limit theorem, strong law of large numbers, ergodic theorem, continuous time markovian renewal function,

discretization method, competing risk, extended semi-Markov Ph-distributions, survival analysis.

Theses organization.
This thesis is organized as follows: In chapter 1 we give an introduction of continuous and discrete time semi-Markov processes, we present a brief description of the main works which tackle properties of words in stochastic sequences and we present the classical theory of competing risk processes. In chapter 2 we compute the first hitting time (position) of the first apparition of a word through a semi-Markov sequence, we also present the variance of this random variable. In chapter 3 we extend the results presented in the second chapter, and we also provide the central limit theorem, strong law of large numbers and the ergodic theorem for a set of words taken from a fine alphabet. In chapter 4 we present competing risk problems from the point of view of semi-Markov processes. In chapter 5 we present a discretization method to handle discrete time Markov renewal functions and we give some numerical examples. Finally in chapter 6 we present some general conclusions and perspectives.

Chapter 1

Introduction

In this chapter we present the main models used and developed in the present thesis, as well the corresponding bibliography.

1.1 Semi-Markov models in continuous time

The most suited mathematical models for describing the stochastic behaviors of a system along time are based on stochastic processes. The most popular stochastic process to model a system is the Markov process. In the Markov theory, the systems have different states, the probability to go from one state to another only depends on the present state. The time spent by the system in each state has an exponential distribution function in a continuous time process (geometric distribution in a discrete sequence). In real life, this hypothesis not always holds true. This is the reason why semi-Markov processes fit better than the Markov hypothesis. They offer the possibility of any distribution function to model the sojourn time between states.

The Semi-Markov processes (SMPs) have an extensive history. They were simultaneously introduced by Levy [1954], Smith [1955], and Takács [1954]. Feller [1964] generalized the classic renewal theory to semi-Markovian processes. Limit theorems were proposed by Yackel [1966], Grigorescu and Oprişan [1976], Athreya and Ney [1978], Nummelin [1978] and Malinovskii [1987]. Other complementary theories were proposed by Çinlar [1969], Kaplan and Sil'vestrov [1980] and Shurenkov and Eleiko [1979]. These processes have been highly used by engineers in mechanics, informatics, communication, etc. SMPs offer the possibility of any distribution function to model the sojourn time between states. This is the main feature of SMPs in fact, a process which conserves the Markov hypothesis at jump points and where the sojourn time in a state can be modeled by any distribution function is a SMP. In the sequel we shall formally define continuous-time semi-Markov processes.

1.1.1 continuous-time semi-Markov framework

In this subsection, we shall introduce the basic definitions for continuous-time semi-Markov processes.

Consider a (finite) set, say E, and an E-valued jump stochastic process $Z = (Z_t)_{t\in\mathbb{R}_+}$. Let $0 =: S_0 \leq S_1 \leq ... \leq S_n \leq S_{n+1} \leq ...$ be the jump times of Z, and $J_0, J_1, J_2, ...$ the successive visited states of Z at jump points. Let $\mathbb{N} := \{0, 1, 2, ...\}$ be the set of nonnegative integers.

DEFINITION 1. (Limnios and Oprişan [2001]). The stochastic process $(J_n, S_n)_{n\in\mathbb{N}}$ is said to be a Markov renewal process (MRP), with state space E, if it satisfies almost sure (a.s.), the following equality

$$\mathbb{P}(J_{n+1} = j, S_{n+1} - S_n \leq t \mid J_0, ..., J_n; S_1, ..., S_n) = \mathbb{P}(J_{n+1} = j, S_{n+1} - S_n \leq t \mid J_n)$$

for all $j \in E$, all $t \in \mathbb{R}_+$ and all $n \in \mathbb{N}$.

We assume that the above probability is independent of n and S_n. In this case the MRP is called time homogeneous. The MRP $(J_n, S_n)_{n\in\mathbb{N}}$ is determined by its transition kernel

$$Q_{ij}(t) := \mathbb{P}(J_{n+1} = j, S_{n+1} - S_n \leq t \mid J_n = i) \tag{1.1}$$

and its initial distribution α, where $\alpha(i) := \mathbb{P}(J_0 = i)$, $i \in E$. It is worth noting here that $Q_{ii}(t) \equiv 0$, for all $i \in E$.
Let us define also the counting process $N(t)$, $t \geq 0$, of the number of jumps, i.e.,

$$N(t) = \sup\{n \geq 0 : S_n \leq t\}. \tag{1.2}$$

The SMP Z is connected to $(J_n, S_n)_{n\in\mathbb{N}}$ by

$$Z_t = J_{N(t)}, \text{ and } J_n = Z_{S_n}, \ n \geq 0.$$

Therefore (J_n) is called the embedded Markov chain (EMC) of process (Z_t).

The distribution function of the sojourn time in state $i \in E$ is given by

$$H_i(t) := \sum_{j\in E} Q_{ij}(t), \ t \geq 0.$$

It defines the distribution function of the sojourn time spent by (Z_t) in $i \in E$. Let $F_{ij}(t) := \mathbb{P}(S_{n+1} - S_n \leq t \mid J_n = i, J_{n+1} = j)$ be the conditional distribution function of the holding time in state i before visiting state j. Let

$$p_{ij} := \mathbb{P}(J_{n+1} = j \mid J_n = i), \ i, j \in E, \ n \in \mathbb{N},$$

be the transition probability of the EMC (J_n). The semi-Markov kernel is also a function of the transition probability matrix of process (J_n), and the conditional

distribution $F_{ij}(t)$ as we can observe in the following equation:

$$\begin{aligned} F_{ij}(t) &= \mathbb{P}(S_{n+1} - S_n \leq t \mid J_{n+1} = j, J_n = i) \\ &= \frac{Q_{ij}(t)}{p_{ij}}, \end{aligned}$$

then $Q_{ij}(t) = p_{ij}F_{ij}(t)$.
Another important function is the semi-Markov transition function at continuous time $\mathbf{P}(t) = (P_{ij}(t); i, j \in E, t \in \mathbb{R}_+)$ defined by

$$P_{ij}(t) := \mathbb{P}(Z_t = j \mid Z_0 = i), \;\; i, j \in E, t > 0, \tag{1.3}$$

which is the conditional marginal law of the process. We shall study this function in the next subsection.
The mean sojourn time of Z in state i is denoted by m_i. If the EMC (J_n) is ergodic, i.e., irreducible and positive recurrent, with stationary probability $\nu = (\nu_i, i \in E)$, and the mean sojourn time in every state is finite, i.e., for every $i \in E$,

$$m_i := \int_0^\infty (1 - H_i(t))dt < \infty, \;\; \text{and} \;\; m := \sum_{i \in E} \nu_i m_i > 0, \;\; m < \infty.$$

Therefore, it can be proved, see e.g., Limnios and Oprişan [2001], that

$$\lim_{t \to \infty} P_{ij}(t) = \frac{\nu_i m_i}{m} =: \pi_i$$

where π is the stationary distribution of process Z.
It is worth noticing that, in general, the stationary distribution π of the SMP Z is not equal to the stationary distribution ν of the embedded Markov chain (J_n).

1.1.2 Backward and forward recurrence times processes in continuous time

Other processes of interest for the SMPs are the backward and forward recurrence time processes. In the following we shall introduce these important processes.

DEFINITION 2. (Limnios and Oprişan [2001]). Given the MRP (J_n, S_n), for $t \in \mathbb{R}_+$ we define the following recurrence times processes

$$B_t = t - S_{N(t)} \;\; \text{and} \;\; V_t = S_{N(t)+1} - t,$$

where $N(t)$ is defined in Equation (1.2). The process $(B_t)_{t \in \mathbb{R}_+}$ is called the backward recurrence time process of process (Z_t) and $(V_t)_{t \in \mathbb{R}_+}$ is the forward recurrence time process of process (Z_t). Figure 1.1 presents a SM trajectory in which we observe the recurrence times processes. At time $S_0 = 0$ the process (Z_t) starts at state i, then the process (Z_t) makes its first jump at time S_1 to state j. At time S_n the process

(Z_t) arrives to state κ. At time t the process (Z_t) has spend u times in κ therefore the value for the backward process at t is $B_t = u$. The process (Z_t) will change its state at time S_{n+1} therefore, the value for the forward process at t is $V_t = v$. We can notice that the total sojourn time in a state at time t is the sum between the backward and the forward recurrence time in a state.

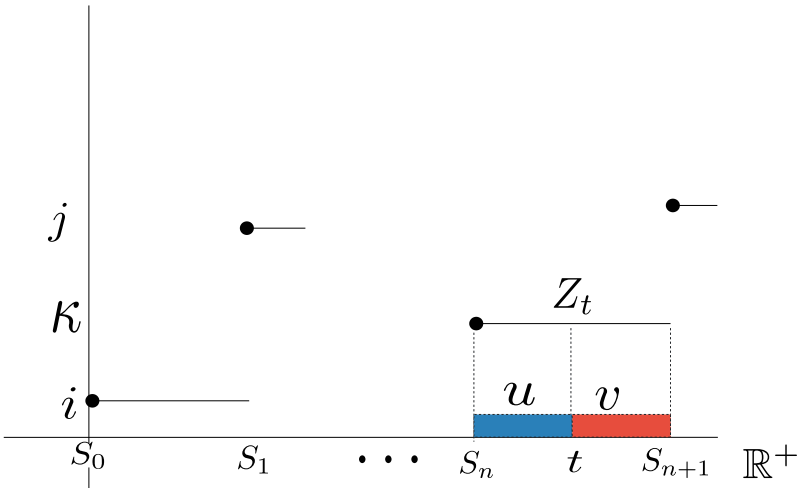


FIGURE 1.1: Sample path of a semi-Markov process where we can observe the backward and forward recurrence time at time t.

1.1.3 Nature of different states of a MRP

Let us now discuss the nature of different states of a MRP.

- A MRP is irreducible, if and only if, its EMC (J_n) is irreducible.

- A state i is recurrent (transient) in the MRP, if and only if, it is recurrent(transient) in the EMC.

- For an irreducible finite MRP, a state i is positive recurrent in the MRP, if, and only if, it is recurrent in the EMC, and if for all $j \in E$, $m_j < \infty$.

- If the EMC of a MRP is irreducible and recurrent, then all the states are positive recurrent, if and only if, $m := \nu m := \sum_{i \in E} \nu_i m_i < \infty$, and null-recurrent, if and only if, $m = \infty$ (where ν is the stationary probability of EMC (J_n)).

- A state i is said to be periodic with $a > 0$ if $G_{ii}(\cdot)$ (the distribution function of the random variable $S_2^i - S_1^i$) is discrete concentrated on $\{ka : k \in \mathbb{N}\}$, where process $(S_n^i)_{n \geq 0}$ represents the successive times of visit to state i. Such

a distribution is said to be periodic. In the opposite case it is called aperiodic. We can notice that the term period has a completely different meaning from the corresponding one of the classic Markov chain theory.

1.1.4 Continuous time Markov renewal theory

In the following, we shall define the Markov renewal equations (MREs) and give some elements of the Markov renewal theory.
Let us consider a real-valued measurable function $\varphi : E \times \mathbb{R}_+ \longrightarrow \mathbb{R}_+$, and define its Stieltjes' convolution by $Q(t)$ as follows

$$Q * \varphi(i,t) = \sum_{k \in E} \int_0^t Q_{ik}(ds)\phi(k, t-s), \tag{1.4}$$

see Limnios [2012b]. Now, for any $i, j \in E$ the n-fold Stieltjes' convolution of $Q_{ij}(t)$ by itself is

$$Q_{ij}^{(n)}(t) = \begin{cases} \delta_{ij} & \text{if } n = 0, \\ Q_{ij}(t) & \text{if } n = 1, \\ \sum_{k \in E} \int_0^t Q_{ik}(ds) Q_{kj}^{(n-1)}(t-s) & \text{if } n \geq 2. \end{cases}$$

We can observe that we have also the following fundamental equality

$$Q_{ij}^{(n)}(t) = \mathbb{P}_i(J_n = j, S_n \leq t),$$

see e.g., Limnios and Oprişan [2001]. Here $\mathbb{P}_i$ means the conditional probability on the event $\{Z_0 = i\}$.
Let us define the Markov renewal function $\psi_{ij}(t), i, j \in E, t \geq 0$, by

$$\begin{aligned} \psi_{i,j}(t) &:= \mathbb{E}_i \sum_{n=0}^{\infty} \mathbb{1}_{\{J_n = j, S_n \leq t\}} \\ &= \sum_{n=0}^{\infty} \mathbb{P}_i(J_n = j, S_n \leq t) \\ &= \sum_{n=0}^{\infty} Q_{ij}^{(n)}(t), \end{aligned} \tag{1.5}$$

we can observe that $\mathbb{E}_i$ means the conditional expected value on the event $\{Z_0 = i\}$.
Let us write the Markov renewal function, see Equation (1.5), in matrix form

$$\psi(t) = (I(t) - Q(t))^{(-1)} = \sum_{n=0}^{\infty} Q^{(n)}(t), \tag{1.6}$$

where the notation $A^{(-1)}$ means the inverse matrix function in the convolution sense. This can also be written as

$$\psi(t) = I(t) + Q * \psi(t), \tag{1.7}$$

where $I(t) = I$, is the identity matrix for $t \geq 0$ and $I(t) = 0$, for $t < 0$.
Equation (1.7) is a special case of a MRE. A general MRE is as follows

$$\Theta(t) = L(t) + Q * \Theta(t),$$

where $\Theta(t) = (\Theta_{ij}(t))_{i,j\in E}$, $L(t) = (L_{ij}(t))_{i,j\in E}$ are matrix-valued measurable functions, with $\Theta_{ij}(t) = L_{ij}(t) = 0$ for $t < 0$. The function $L(t)$ is a given matrix-valued function and $\Theta(t)$ is an unknown matrix-valued function.

PROPOSITION 1. (Limnios and Oprişan [2001]). The transition function $P(t) = (P_{ij}(t); i, j \in E, t \in \mathbb{R}_+)$ satisfies the following MRE

$$P(t) = I(t) - H(t) + Q * P(t), \tag{1.8}$$

where $H(t) = \text{diag}(H_i(t))$ is a diagonal matrix. Last Equation (1.8) has the unique solution

$$P(t) = \psi * (I(t) - H(t)).$$

In the next subsection, we shall give the equivalent definition for a discrete time SMP.

1.2 Discrete time semi-Markov processes

There are some processes which are observed in fixed periods of time. For instance, the profit of a company in a month, the number of car accidents per week, the number of infections of a certain disease per semester. These processes can be considered like stochastic processes at discrete-time. i.e., if the events are modeled by a stochastic process Z_k, they will be observed at discrete time points $k = 0, 1, 2, ...$ In the semi-Markov domain many authors have proposed these kind of models for different applications, see e.g., Janssen [2013], Ross [2013], Rachelson et al. [2008], Limnios and Oprişan [2001], etc. In this thesis we are specially interested in modeling DNA sequences, where every nucleotide in the DNA can be considered as a time unit. Before presenting this model in next subsection, we shall introduce the basic definitions for the semi-Markov chains.

1.2.1 Semi-Markov framework at discrete time

Let us formally define discrete time SMPes (as we have written before, we shall use the term chain for a discrete-time stochastic process). Let us consider a random

system $(Z_k)_{k\in\mathbb{N}}$ with finite state space $E = \{1, 2, ..., s\}$. Let us denote by $(S_n)_{n\in\mathbb{N}}$ the successive time points when a state changes in (Z_k), i.e., by definition let $S_0 := 0$, and

$$S_{n+1} := \inf\{k > S_n : Z_k \neq Z_{S_n}\}, \quad n \geq 0,$$

with the convention $\inf \emptyset = +\infty$. Process (S_n) is also called renewal points or jump points of process (Z_k). Let $(J_n)_{n\in\mathbb{N}}$ be the chain which records (Z_k) at points (S_n), i.e., $J_n = Z_{S_n}$. Let $(X_n)_{n\in\mathbb{N}}$ be the successive sojourn times in the visited states. By convention $X_0 := S_0 := 0$ and $X_{n+1} := S_{n+1} - S_n$, $n \in \mathbb{N}$. The relation between process (Z_k) and process (J_k) is given by

$$Z_k = J_{N(k)}, \quad \text{or equivalently, } J_n = Z_{S_n}, \quad n, k \in \mathbb{N},$$

where $N(k) := \max\{n \in \mathbb{N} : S_n \leq k\}$ is the counting process of the number of jumps in $[1, k] \subset \mathbb{N}$. If the following relation holds true a.s.

$$\begin{aligned} & \mathbb{P}(J_{n+1} = j, S_{n+1} - S_n = k \mid J_0 = \cdot, ..., J_n = i; S_0 = \cdot, ..., S_n = \cdot) \\ = \; & \mathbb{P}(J_{n+1} = j, S_{n+1} - S_n = k \mid J_n = i). \end{aligned} \tag{1.9}$$

The process (J_n, S_n) is called the Markov renewal chain (MRC) of process (Z_k). In other words, if process (Z_k) has entered to state $i \in E$ at its last jump $n \in \mathbb{N}$, the probability that the process passes $k \in \mathbb{N}$ units of time in $i \in E$, after passing to state $j \in E$, is independent of process up to the nth jump. In other words, we basically have the Markovian property with the difference that the memoryless property does not act on the calendar points k. The memoryless property acts at visited states $(J_0, J_1, ..., J_n, J_{n+1}, ...)$. This is what we called before as a more flexible Markovian hypothesis. Noticing that we use index $k \in \mathbb{N}$ for the calendar time, and index $n \in \mathbb{N}$ for the number of jumps of (Z_k). Therefore if Equation (1.9) holds true, then (Z_k) is called semi-Markov chain (SMC). Moreover, if the rigth-hand-side term of Equation (1.9) is independent of n, then (Z_k) and (J_n, S_n) are said to be (time homogeneous) and we define the discrete-time semi-Markov kernel $\mathbf{q} = (q_{ij}(k); i, j \in E, k \in \mathbb{N})$ by

$$q_{ij}(k) := \mathbb{P}(J_{n+1} = j, S_{n+1} - S_n = k \mid J_n = i), \quad n \geq 0, k \in \mathbb{N}. \tag{1.10}$$

The semi-Markov kernel satisfies the following three properties:

1. $0 \leq q_{ij}(k), \qquad i, j \in E, k \in \mathbb{N},$
2. $q_{ij}(0) = 0, \qquad i, j \in E,$
3. $\displaystyle\sum_{k=0}^{\infty} \sum_{j\in E} q_{ij}(k) = 1, \qquad i \in E.$

The semi Markov chain is defined by its semi-Markov kernel and its initial distribution $\alpha(i) := \mathbb{P}(Z_0 = i) = \mathbb{P}(J_0 = i), i \in E$.

We define the cumulative semi-Markov kernel by

$$Q_{ij}(k) := \mathbb{P}(J_{n+1} = j, X_{n+1} \leq k \mid J_n = i) = \sum_{l=0}^{k} q_{ij}(l), \;\; i, j \in E, k \in \mathbb{N}.$$

It is worth noticing that the semi-Markov kernel considered here is independent of n, which means that the SMC is homogeneous in time. If semi-Markov kernel is time homogeneous then (J_n) is an homogeneous Markov chain. We denote by $\mathbf{p} = (p_{ij})_{i,j \in E}$ its transition probability matrix, i.e.,

$$p_{ij} := \mathbb{P}(J_{n+1} = j \mid J_n = i), \;\; i, j \in E, \;\; n \in \mathbb{N}. \tag{1.11}$$

We do not allow transitions to the same state, i.e., $p_{ii} = 0$ for any $i \in E$. Note that p_{ij} can be expressed in terms of the semi-Markov kernel by $p_{ij} = \sum_{k=0}^{\infty} q_{ij}(k)$. Let us denote by

$$f_{ij}(k) = \mathbb{P}(X_{n+1} = k \mid J_n = i, J_{n+1} = j) \tag{1.12}$$

the conditional sojourn time distribution, conditioned by the next state to be visited. We want the chain spends at least one time unit in a state, that is, $f_{ij}(0) = 0$, for any states $i, j \in E$. Obviously, for any states $i, j \in E$ and non-negative integer k we have

$$q_{ij}(k) = p_{ij} f_{ij}(k).$$

The sojourn time distribution in state i is

$$h_i(k) := \mathbb{P}(X_{n+1} = k \mid J_n = i) = \sum_{j \in E} q_{ij}(k), \;\; k \in \mathbb{N}.$$

The cumulative distribution function of sojourn time in state $i \in E$ is

$$H_i(k) := \mathbb{P}(X_{n+1} \leq k \mid J_n = i) = \sum_{l=0}^{k} h_i(k), \;\; k \in \mathbb{N}. \tag{1.13}$$

The conditional cumulative distribution of the waiting time X_{n+1}, $n \in \mathbb{N}$, is

$$F_{ij}(k) := \mathbb{P}(X_{n+1} \leq k \mid J_n = i, J_{n+1} = j) = \begin{cases} \frac{Q_{ij}(k)}{p_{ij}}, & \text{if } p_{ij} \neq 0, \\ 1_{\infty}(k), & \text{if } p_{ij} = 0. \end{cases} \tag{1.14}$$

The main difference between Markov and semi-Markov discrete time processes is the distribution function $F_{ij}(k)$. In a Markov chain this function is geometric with parameter $1 - p_{ii}$, where p is the transition probability matrix; on the other hand, in the SMP the distribution function $F_{ij}(k)$ can be of any type.

Let us define the support of h_i and the maximum sojourn time in $i \in E$.

DEFINITION 3. (Garcia-Maya et al. [Submitted in 2020]). For any $i \in E$ the smallest subset C_i in $\mathbb{N}$ such that

$$\sum_{k \in C_i} h_i(k) = 1,$$

is the support of h_i. By consequence

$$r_i := \sup C_i. \tag{1.15}$$

is the maximum sojourn time in state i.

Other important quantity for investigating the evolution of SMCs is the probability that starting from state $i \in E$ at time zero, the SMC will do the nth jump at time k to state j, i.e.,

$$\mathbb{P}(J_n = j, S_n = k \mid J_0 = i), \quad i, j \in E; \;\; k, n \in \mathbb{N}. \tag{1.16}$$

After giving an expression for this last probability, we shall introduce the definition of the convolution between two functions. Let $\varphi(i,k),\ i,j \in E, k \in \mathbb{N}$, be a measurable function and define the convolution of φ by q as

$$(q * \varphi)_{ij}(k) := \sum_{r \in E} \sum_{l=0}^{k} q_{ir}(l)\phi_{rj}(k-l).$$

The n-fold convolution of q by itself is defined recursively by

$$q_{ij}^{(0)}(k) := \delta_{ij} \mathbb{1}_{\{k=0\}},$$

$$q_{ij}^{(1)}(k) := q_{ij}(k),$$

and

$$q_{ij}^{(n)}(k) := \sum_{r \in E} \sum_{l=0}^{k} q_{ir}(l) q_{rj}^{(n-1)}(k-l), \qquad n \geq 2.$$

The following proposition computes an expression for probability Equation (1.16)

PROPOSITION 2. (Barbu and Limnios [2008]). For all $i, j \in E$, for all $n, k \in \mathbb{N}$, we have

$$\mathbb{P}(J_n = j, S_n = k \mid J_0 = i) = q_{ij}^{(n)}(k). \tag{1.17}$$

Proof. We prove the result by induction. For $n = 0$, we have

$$\mathbb{P}(J_0 = j, S_0 = k \mid J_0 = i) = q_{ij}^{(0)}(k).$$

Obviously, for $k \neq 0$ or $i \neq j$, this probability is zero. On the other hand, if $i = j$ and $k = 0$, the probability is one, thus the result follows.
For $n = 1$, the result obviously holds true, by definition.
For $n \geq 2$

$$\begin{aligned}
& \mathbb{P}(J_n = j, S_n = k \mid J_0 = i) \\
= & \sum_{r \in E} \sum_{l=1}^{k-1} \mathbb{P}(J_n = j, S_n = k, J_1 = r, S_1 = l \mid J_0 = i) \\
= & \sum_{r \in E} \sum_{l=1}^{k-1} \mathbb{P}(J_n = j, S_n = k \mid J_1 = r, S_1 = l, J_0 = i) \mathbb{P}(J_1 = r, S_1 = l \mid J_0 = i) \\
= & \sum_{r \in E} \sum_{l=1}^{k-1} \mathbb{P}(J_{n-1} = j, S_{n-1} = k - l \mid J_0 = r) \mathbb{P}(J_1 = r, X_1 = l \mid J_0 = i) \\
= & \sum_{r \in E} \sum_{l=1}^{k-1} q_{rj}^{(n-1)}(k-l) q_{ir}(l) = q_{ij}^{(n)}(k). \ \blacksquare
\end{aligned}$$

As a direct application of the previous proposition, we have the following lemma.

LEMMA 1. (Barbu and Limnios [2008]). Let $\mathcal{M}_E$ be the set of real matrices on $E \times E$ and let $\mathcal{M}_E(\mathbb{N})$ be the set of real matrices on $E \times E$ which evolves in a discrete time $k \in \mathbb{N}$. Let us consider the Markov renewal chain $(J_n, S_n)_{n \in \mathbb{N}}$ and $\mathbf{q} \in \mathcal{M}_E(\mathbb{N})$ its associated semi-Markov kernel. Then, for all $n, k \in \mathbb{N}$ such that $n \geq k + 1$ we have $q^{(n)}(k) = 0$.

Proof. It is clear that the jump time process $(S_n)_{n \in \mathbb{N}}$ verifies the relation $S_n \geq n$, $n \in \mathbb{N}$. Writting Equation (1.17) for n and $k \in \mathbb{N}$ such that $n \geq k + 1$, we obtain the desired result. $\blacksquare$

1.2.2 Discrete time backward and forward recurrence times processes

Now, let (B_k) be the backward recurrence times process for the SMC (Z_k) (also called the current life or age), defined by

$$B_k := k - \max\{S_m : S_m \leq k\} \text{ and } B_k := k \text{ if } S_1 > k. \tag{1.18}$$

In essence, the backward time at position k is the number of steps spent in state Z_k since the last jump. Notice that, if k coincides with a renewal point, the value of B_k is zero. After a renewal point, B_k grows one by one until another renewal point, and so on.
Let r_i be the maximum sojourn time in state i, see Equation (1.15), then the

maximum value for the backward time in state $i \in E$, is

$$l_i := r_i - 1. \tag{1.19}$$

Let (V_k) be the forward recurrence time of the SMC (Z_k) (also called the residual or excess lifetime), defined by

$$V_k := S_{N(k)+1} - k$$

In essence, the forward time at position k is the number of steps the sequence (Z_k) will spend in state Z_k until next jump.

1.2.3 Classification for states in SMC

Until this point we are prepared to define the basic characteristics of the associated Markov renewal chain (MRC) (J_n, S_n): communication between classes, transitivity, recurrence and periodicity. After introducing these points we shall define the first passage time in state $j \in E$.
For any $j \in E$, let

$$S_0^j := \inf\{k \in \mathbb{N}^* : J_{N(k)} = j\} \tag{1.20}$$

be a random variable which represents the first hitting time of state j.

We consider $v_{ij}(\cdot)$ its distribution function i.e.,

$$v_{ij}(k) := \mathbb{P}_i(S_0^j = k), k \geq 1. \tag{1.21}$$

We set $\mathbb{V}_{ij}(k) := \sum_{\ell=1}^{k} v_{ij}(\ell)$ for the corresponding cumulative distribution function and μ_{ij} for the mean first passage time from state i to state j for the SMC (Z_k), i.e., $\mu_{ij} := \mathbb{E}_i(S_0^j) = \sum_{k>0} v_{ij}(k)$. Observe that if $i = j$, $v_{ii}(\cdot)$ represents the return time to state i. Observe that the sojourn time in state $i \in E$ is also a random variable, we shall define its mean by

$$m_i := \mathbb{E}[S_1 \mid J_0 = i] = \sum_{k \geq 0} (1 - H_i(k)). \tag{1.22}$$

We shall give the classification for states in SMC (Z_k). For this purpose, note that $\mathbb{V}_{ij}(\infty)$ is the probability that the SMC will go from state i to j at some $k \in \mathbb{N}$. If $\mathbb{V}_{ij}(\infty) = 0$ that means that there is zero probability that the SMC will arrive at state j starting from i. Using this two remarks for $\mathbb{V}_{ij}(\cdot)$ we have the following definition.

DEFINITION 4. (Barbu and Limnios [2008]). Let $(Z_k)_{k\in\mathbb{N}}$ be a SMC with state space E and $(J_n, S_n)_{n\in\mathbb{N}}$ the associated MRC.

1. If $\mathbb{V}_{ij}(\infty)\mathbb{V}_{ij}(\infty) > 0$, we shall say that i and j communicate and it is denoted by the symbol $i \leftrightarrow j$. The communication is an equivalent relation on E. The

elements which communicates between them belongs to the same communication class. All classes are closed.

2. The SMC (MRC) is said to be irreducible if there is only one class.

3. A state i is said to be recurrent if $\mathbb{V}_{ij}(\infty) = 1$ and transient if $\mathbb{V}_{ii}(\infty) < 1$. A recurrent state i is positive recurrent if $\mu_{ij} < \infty$ and null recurrent if $\mu_{ii} = \infty$. If all the states are (positive/null) recurrent, the SMC (MRC) is said to be (positive/null) recurrent.

4. The SMC (MRC) is said to be ergodic if it is irreducible and positive recurrent.

5. Let $d > 1$ be a positive integer. A state $i \in E$ is said to be d-periodic (aperiodic) if the distribution $v_{ij}(\cdot)$ is d-periodic (aperiodic).

6. The SMC is d-periodic, $d > 1$, if all states are d-periodic. Otherwise, it is called aperiodic.

We can also define the limit distribution of a SMC

DEFINITION 5. (Barbu and Limnios [2008]). For a SMC $(Z_k)_{k\in\mathbb{N}}$, the limit distribution $\pi = (\pi_1, ..., \pi_{|E|})^T$ is defined, when it exits, by $\pi_j = \lim_{k\to\infty} P_{ij}(k)$, for every $i, j \in E$.

Note that in the case where the EMC (J_n) is ergodic (recurrent positive, irreducible and aperiodic), for any state $i \in E$ we have the following relation between the mean recurrence time of i in the Markov chain (J_n) denoted by μ^*_{ii} and the stationary distribution

$$\mu^*_{ii} = \frac{1}{\nu(i)},$$

see Barbu and Limnios [2008].

1.2.4 Discrete-time Markov renewal theory

In this section, we shall study Markov Renewal Equations (MRE). Our objective is to investigate the existence and the uniqueness of solution for this type of equations. We also find an explicit form of the transition function $\mathbf{P}$ of the SMC (Z_k), written in terms of the semi-Markov kernel $\mathbf{q}$. First we shall present some important results in the sense of the convolution functions.

DEFINITION 6. (Barbu and Limnios [2008]). Let $\mathbf{A} \in \mathcal{M}_E(\mathbb{N})$ be a matrix function. If there exist a matrix function $\mathbf{B} \in \mathcal{M}_E(\mathbb{N})$ such that

$$\mathbf{B} * \mathbf{A} = \mathbf{I} \tag{1.23}$$

then $\mathbf{B}$ is called the left inverse of $\mathbf{A}$, in the convolution sense, and it is denoted by $B = \mathbf{A}^{(-1)}$.

The left inverse of a matrix not always exists, next proposition gives the necessary conditions to guarantee its existence.

PROPOSITION 3. (Barbu and Limnios [2008]). The left inverse of a matrix $\mathbf{A} \in \mathcal{M}_E(\mathbb{N})$ exists and is unique iff $\det\mathbf{A}(0) \neq 0$. If we denoted by $\mathbf{B}$ the partial inverse of $\mathbf{A}$, i.e., $\mathbf{B} = \mathbf{A}^{(-1)} \in \mathcal{M}_E(\mathbb{N})$. The partial inverse is given by the recursive formula

$$\mathbf{B}(n) = \begin{cases} [\mathbf{A}(0)]^{-1} & \text{if } n = 0, \\ -(\sum_{l=0}^{n-1} \mathbf{B}(l)\mathbf{A}(n-l))[\mathbf{A}(0)]^{-1} & \text{if } n \geq 1. \end{cases} \tag{1.24}$$

Proof. To compute the left inverse of $\mathbf{A}$ we have to solve Equation (1.23), where $\mathbf{B} \in \mathcal{M}_E(\mathbb{N})$ is an unknown matrix function. Equation (1.23) is equivalent to

$$\sum_{l=0}^{n} \mathbf{B}(n-l)\mathbf{A}(l) = \mathbf{I}(n), n \in \mathbb{N},$$

where for $n = 0$ we have $\mathbf{B}(0)\mathbf{A}(0) = \mathbf{I}(0) = \mathbf{I}$, which holds iff $\mathbf{A}(0)$ is invertible. Therefore $\mathbf{B}(0) = [\mathbf{A}(0)]^{-1}$ and for $n > 1$ we have

$$\sum_{l=0}^{n} \mathbf{B}(n-l)\mathbf{A}(l) = \mathbf{I}(n) = \mathbf{0},$$

which yields $\mathbf{B}(n) = -\left(\sum_{l=0}^{n-1} \mathbf{B}(l)\mathbf{A}(n-l)\right)[\mathbf{A}(0)]^{-1}$. ■

Let us write the Markov renewal function (MRF)

$$\psi(k) := (\mathbf{I} - \mathbf{q})^{(-1)}(k), \tag{1.25}$$

where $\mathbf{I}$ is the identity matrix of dimension $|E| \times |E|$. The following result provides the mathematical expression of $\psi(k)$.

PROPOSITION 4. (Barbu and Limnios [2008]). The matrix-valued function $\psi = (\psi(k); k \in \mathbb{N})$ is given by

$$\psi(k) = \sum_{n=0}^{k} \mathbf{q}^{(n)}(k), \quad k \in \mathbb{N}. \tag{1.26}$$

Proof. Applying Proposition 3, we obtain that the left inverse of the matrix-valued function $(\mathbf{I} - \mathbf{q})$ exist and is unique.

For all $n, k \in \mathbb{N}$, we have

$$\begin{aligned}\left(\sum_{n=0}^{\infty}\mathbf{q}^{(n)}\right) * (I-\mathbf{q})(k) &= \left(\sum_{n=0}^{\infty}\mathbf{q}^{(n)}\right)(k) - \left(\sum_{n=0}^{\infty}\mathbf{q}^{(n)}\right) * \mathbf{q}(k) \\ &= \sum_{l=0}^{\infty}\mathbf{q}^{(l)}(k) - \sum_{l=1}^{\infty}\mathbf{q}^{(l)}(k) \\ &= \mathbf{q}^{(0)}(k) = \mathbf{I}(k). \end{aligned} \tag{1.27}$$

As the left inverse of $(\mathbf{I}-\mathbf{q})(k)$ is unique, see Proposition 3, we obtain that

$$\psi(k) = \sum_{n=0}^{\infty}\mathbf{q}^{(n)}(k).$$

Applying Lemma 1, we have $\mathbf{q}^{(n)}(k) = 0$, for $n > k$ hence we obtain that ψ is given by

$$\psi(k) = \sum_{n=0}^{k}\mathbf{q}^{(n)}(k). \ \blacksquare \tag{1.28}$$

We would like to obtain another expression for ψ. For any states $i, j \in E$ (not necessary distinct) and any positive integer $k \in \mathbb{N}$, from Proposition 2 and Lemma 1 we get

$$\psi_{i,j}(k) = \mathbb{P}\left(\bigcup_{n=0}^{k}\{J_n = j, S_n = k\} \mid J_0 = i\right) \leq 1.$$

In other words, $\psi_{i,j}(k)$ represents the probability that starting at time 0 in state $i \in E$, the SMC will do a jump to state j at time k.
Using $\psi_{i,j}(k)$ the following proposition computes the probability that SMP (Z_k) stays in state $j \in E$ at time $k \in \mathbb{N}$ with backward time u knowing that it started at initial time in state $i \in E$.

PROPOSITION 5. (Barbu and Limnios [2008]). For all $i, j \in E$ and $k \in \mathbb{N}$ we have

$$\mathbb{P}_i(Z_k = j, B_k = u) = \begin{cases} [1 - H_j(u)]\psi_{i,j}(k-u), & \text{for } u = 0, 1, ..., k \\ 0 & \text{elsewhere} \end{cases}$$

therefore

$$\mathbb{P}(Z_k = j, B_k = u) = \psi_{\cdot j}(k-u)[1 - H_j(u)].$$

Proof. For all $u = 0, 1, ..., k$ we have

$$\begin{aligned}
\mathbb{P}_i(Z_k = j, B_k = u) &= \sum_{n=0}^{k-u} \mathbb{P}(J_n = j, S_n = k-u, S_{n+1} > k \mid J_0 = i) \\
&= \sum_{n=0}^{k-u} \mathbb{P}(S_{n+1} > k \mid J_n = j, S_n = k-u, J_0 = i) \\
&\quad \cdot \mathbb{P}(J_n = j, S_n = k-u \mid J_0 = i) \\
&= \sum_{n=0}^{k-u} \mathbb{P}(S_{n+1} - S_n > u \mid J_n = j) \mathbb{P}(J_n = j, S_n = k-u \mid J_0 = i) \\
&= [1 - H_j(u)] \psi_{i,j}(k-u). \ \blacksquare
\end{aligned}$$

The chain $(Z_k, B_k)_{k \in \mathbb{N}}$ is a Markov chain with state $E \times \mathbb{N}$, see e.g., Barbu and Limnios [2008]. The following theorem provides its transition probability matrix.

THEOREM 1. (Barbu and Limnios [2008]). For every $i, j \in E$ and $k \in \mathbb{N}$ such that $\mathbb{P}(Z_k = i, B_k = u) > 0$ the transition probability matrix of the Markov chain $(Z_k, B_k)_{k \in \mathbb{N}}$ is

$$\begin{aligned}
&\mathbb{P}(Z_{k+1} = j, B_{k+1} = u^{'} \mid Z_k = i, B_k = u) \\
&= \begin{cases} \frac{q_{ij}(u+1)}{1-H_i(u)}, & \text{if } u^{'} = 0 \text{ and } i \neq j \\ \frac{1-H_i(u+1)}{1-H_i(u)}, & \text{if } u^{'} = u+1 \text{ and } i = j \\ 0, & \text{elsewhere.} \end{cases}
\end{aligned}$$

Proof. see Chryssaphinou et al. [2008].
Another point of interest, strictly related to ψ, it is the Markov renewal function, defined as the expected number of visits to a certain state, up to a given time. More precisely, we have the following definition.

DEFINITION 7. (Barbu and Limnios [2008]). **Markov Renewal Function (MRF).** Let us define the Markov renewal function $\mathbf{\Psi} = (\Psi_{i,j}(k), i, j \in E, k \geq 0)$ by

$$\Psi_{i,j}(k) := \mathbb{E}_i[N_j^*(k)], \ \ i, j \in E, \ \ k \in \mathbb{N},$$

where $N_j^*(k)$ is the number of visits of (Z_k) to state $j \in E$ in the interval $[0, k]$. To be specif,

$$N_j^*(k) := \sum_{n=0}^{N(k)} \mathbb{1}_{\{J_n = j\}} = \sum_{n=0}^{k} \{J_n = j, S_n \leq k\}.$$

It is easy to see that the Markov renewal function can be expressed as follows:

$$\Psi(k) = \sum_{l=0}^{k} \psi(l). \tag{1.29}$$

Indeed, we have

$$\begin{aligned}
\Psi_{i,j}(k) &:= \mathbb{E}_i[N_j^*(k)] \\
&= \mathbb{E}_i\left[\sum_{n=0}^{k} \mathbb{1}_{\{J_n=j, S_n\leq k\}}\right] \\
&= \sum_{n=0}^{k} \mathbb{P}(J_n = j, S_n \leq k \mid J_0 = i) \\
&= \sum_{n=0}^{k}\sum_{l=0}^{k} \mathbb{P}(J_n = j; S_n = l \mid J_0 = i) \\
&= \sum_{l=0}^{k}\sum_{n=0}^{k} q_{ij}^{(n)}(l)
\end{aligned}$$

From Lemma 1 we know that $q_{ij}^{(n)}(l) = 0$ for $n > l$ and we get

$$\Psi_{i,j}(k) = \sum_{l=0}^{k}\sum_{n=0}^{l} q_{ij}^{(n)}(l) = \sum_{l=0}^{k} \psi(l).$$

Remark. One can check that a state is recurrent iff $\Psi(\infty) = \infty$ and transient iff $\Psi(\infty) < \infty$.

DEFINITION 8. (Barbu and Limnios [2008]). **Discrete-time Markov renewal equation.**
Let $\mathbf{L} = (L_{ij}(k); i, j \in E, k \in \mathbb{N}) \in \mathcal{M}_E(\mathbb{N})$ be an unknown matrix-valued function and $\mathbf{G} = (G_{ij}(k); i, j \in E, k \in \mathbb{N}) \in \mathcal{M}_E(\mathbb{N})$ be a known matrix-valued function. The equation

$$\mathbf{L}(k) = \mathbf{G}(k) + \mathbf{q} * \mathbf{L}(k), \quad k \in \mathbb{N}, \tag{1.30}$$

is called a discrete-time Markov renewal equation (DTMRE).

In the sequel, we shall see that ψ and Ψ are solutions of MRE. With the above definition, we shall compute an easy expression for $\mathbf{P}(k)$, see Equation (1.37). Observing Equation (1.27) it is clear that $(\mathbf{I} - \mathbf{q}) * \psi(k) = \mathbf{I}(k)$, so $\psi = (\psi(k); k \in \mathbb{N})$ is the solution of the MRE

$$\psi(k) = \mathbf{I}(k) + \mathbf{q} * \psi(k), \quad k \in \mathbb{N}. \tag{1.31}$$

Second, writing the previous equation for $k \in \mathbb{N}$, $0 < k < \nu$, $\nu \in \mathbb{N}$ fixed, and taking the sum, we obtain

$$\sum_{k=0}^{\nu} \psi(k) = \sum_{k=0}^{\nu} \mathbf{I}(k) + \sum_{k=0}^{\nu} \mathbf{q} * \psi(k). \tag{1.32}$$

This means that the matrix renewal function $\Psi = (\Psi(k); k \in \mathbb{N})$ is the solution of the MRE

$$\Psi(\nu) = I + \mathbf{q} * \Psi(\nu), \quad \nu \in \mathbb{N}. \tag{1.33}$$

The next theorem shows that DTMRE, see definition (8), has a unique solution.

THEOREM 2. (Barbu and Limnios [2008]). The DTMRE, see Equation (1.30), has a unique solution $\mathbf{L} = (L_{ij}(k); i, j \in E, k \in \mathbb{N}) \in \mathcal{M}_E(\mathbb{N})$, where

$$\mathbf{L}(k) = \psi * \mathbf{G}(k).$$

Proof. By Equation (1.30), for $k \in \mathbb{N}$

$$\begin{aligned} \mathbf{L}(k) &= \mathbf{G}(k) + \mathbf{q} * \mathbf{L} \\ (\mathbf{I} - \mathbf{q}) * \mathbf{L}(k) &= \mathbf{G}(k) \\ \mathbf{L}(k) &= (\mathbf{I} - \mathbf{q})^{(-1)} * \mathbf{G}(k). \end{aligned}$$

Therefore by definition of ψ, see Equation (1.25), we have

$$\mathbf{L}(k) = \psi * \mathbf{G}(k). \tag{1.34}$$

Then, we show that $\psi * \mathbf{G}(k)$ is the unique solution of the renewal equation. Let $\mathbf{L}'$ be another solution of Equation (1.30). We obtain

$$(\mathbf{L} - \mathbf{L}')(k) = \mathbf{q}^{(n)} * (\mathbf{L} - \mathbf{L}')(k), \quad k \in \mathbb{N}, \tag{1.35}$$

with n an arbitrary positive integer. Taking $n > k$ in Equation (1.30) and recalling that $\mathbf{q}^{(n)}(k) = 0$ for $n > k$, see Lemma 1, we get $\mathbf{L}(k) = \mathbf{L}'(k), k \in \mathbb{N}$. ■

DEFINITION 9. (Barbu and Limnios [2008]). The transition function of the SMC (Z_k) is the matrix-valued function $\mathbf{P} = (P_{i,j}(k); i, j \in E, k \in \mathbb{N}) \in \mathcal{M}_E(\mathbb{N})$ defined by

$$P_{ij}(k) := \mathbb{P}(Z_k = j \mid Z_0 = i), \quad i, j \in E, k \in \mathbb{N}.$$

The following result consists in a recursive formula for computing the transition function $\mathbf{P}$ of the SMC (Z_k) which is an example of a MRE.

PROPOSITION 6. (Barbu and Limnios [2008]). For all $i, j \in E$ and for all $k \in \mathbb{N}$, we have

$$P_{ij}(k) = [1 - H_i(k)]\delta_{ij} + \sum_{r \in E} \sum_{l=0}^{k} q_{ir}(l) P_{rj}(k - l), \tag{1.36}$$

where for all $k \in \mathbb{N}$, let us define $\mathbf{H}(k) := \text{diag}(H_i(k); i \in E)$, $\mathbf{H} := (\mathbf{H}(k); k \in \mathbb{N})$, where $H_i(\cdot)$ is the sojourn time cumulative distribution function in state $i \in E$, see definition. In matrix-valued function, Equation (1.36) becomes,

$$\mathbf{P}(k) = (\mathbf{I} - \mathbf{H})(k) + \mathbf{q} * \mathbf{P}(k), k \in \mathbb{N}. \tag{1.37}$$

Proof. For all $i, j \in E$ and for all $k \in \mathbb{N}$, we have

$$\begin{aligned}
P_{ij}(k) &= \mathbb{P}(Z_k = j \mid Z_0 = i) \\
&= \mathbb{P}(Z_k = j, S_1 > k \mid Z_0 = i) + \mathbb{P}(Z_k = j, S_1 \leq k \mid Z_0 = i) \\
&= (1 - H_i(k))\delta_{ij} + \sum_{r \in E} \sum_{l=0}^{k} \mathbb{P}(Z_k = j, Z_{S_1} = r, S_1 = l \mid Z_0 = i) \\
&= (1 - H_i(k))\delta_{ij} + \sum_{r \in E} \sum_{l=0}^{k} \mathbb{P}(Z_k = j \mid Z_{S_1} = r, S_1 = l, Z_0 = i)\mathbb{P}(J_1 = r, S_1 = l \mid J_0 = i) \\
&= (1 - H_i(k))\delta_{ij} + \sum_{r \in E} \sum_{l=0}^{k} \mathbb{P}(Z_{k-l} = j \mid Z_0 = r)\mathbb{P}(J_1 = r, X_1 = l \mid J_0 = i) \\
&= (1 - H_i(k))\delta_{ij} + \sum_{r \in E} \sum_{l=0}^{k} P_{rj}(k - l)q_{ir}(l).
\end{aligned}$$

Observe the last equation is a DTMRE, see Definition 8, where $\mathbf{L}(k) = \mathbf{P}(k)$ and $\mathbf{G}(k) = (\mathbf{I} - \mathbf{H})$, therefore by Theorem 2 we have

$$\mathbf{P}(k) = \psi * (\mathbf{I} - \mathbf{H})(k) = (\delta\mathbf{I} - \mathbf{q})^{(-1)} * (\mathbf{I} - \mathbf{H}(k)), \quad k \in \mathbb{N}.$$

1.2.5 Construction of the Estimators

Let us consider an estimator for the semi-Markov kernel $q_{ij}(k)$, see Equation (1.10); the conditional sojourn time distribution $f_{ij}(k)$, see Equation (1.12); and the transition probability p_{ij}, see Equation (1.11). Let $M \in \mathbb{N}^*$ be a fixed arbitrary time and $N(M)$ the discrete-time counting process of the number of jumps in $[1, M]$. For any states $i, j \in E$ and positive integer $k \in \mathbb{N}$, $k \leq M$, we define the following empirical estimators

$$\hat{p}_{ij}(M) := \frac{N_{ij}(M)}{N_i(M)}, \tag{1.38}$$

$$\hat{f}_{ij}(k, M) := \frac{N_{ij}(k, M)}{N_{ij}(M)}, \tag{1.39}$$

$$\hat{q}_{ij}(k, M) := \frac{N_{ij}(k, M)}{N_i(M)}, \tag{1.40}$$

where $N_{ij}(k, M)$ is the number of transitions of the EMC from i to j, up to time M, with sojourn time in state i equal to k, $1 \leq k \leq M$, i.e.,

$$N_{ij}(k, M) := \sum_{n=1}^{N(M)} \mathbb{1}_{\{J_{n-1}=i, J_n=j, X_n=k\}} = \sum_{n=1}^{M} \mathbb{1}_{\{J_{n-1}=i, J_n=j, X_n=k, S_n \leq M\}}.$$

Therefore

$$N_{ij}(M) := \sum_{k=1}^{M} N_{ij}(k, M)$$

and

$$N_i(M) := \sum_{j \in E} N_{ij}(M).$$

Note that the proposed estimators are natural estimators. For instance, the probability p_{ij} that the system goes from state i to state j is estimated by the number of transitions from state i to j, up to time M divided by the total number of transitions from state i to any state j up to time M. Estimators (1.38), (1.39) and (1.40) verify nice asymptotic properties as consistency and asymptotic normality, see Barbu and Limnios [2008].

1.3 Properties of words through a sequence

Looking for a specific word (pattern or motif) through a sequence of symbols is useful in many areas. For instance, in digital transactions lots of patterns have to be compared and identified to make possible a particular operation, see e.g., Jungck and Helms [2013], de Haan and Rotmans [2011], Jansen et al. [2007]; in web navigation identification of patterns is useful for determining particular behaviors in web customers see e.g., Papadopoulou [2013], Bobbitt et al. [1969], Paraskevopoulos et al. [2013]; in communications, identifying a particular pattern is useful to mediate the spread of information see e.g., Jaffe et al. [1999], Bailey et al. [2018]; in databases it could be the keyword for certain research see e.g., Pederson [2016]; in biology a word can determine a particular instruction for a given biology function, or it can be the responsible for a particular illness see e.g., Willett [1995], Farré et al. [2003], Kitano [2002], etc. In one hand, searching a particular word over a chain formed by a big quantity of symbols is a very cumbersome task and in most cases it can not be achieved by simple inspection. In the other hand, it is of interest to know general properties of words in a sequence, i.e., frequency, mean number of words, mean number of symbols between two successive words, etc. There are different approaches to study the searching problem, for instance one of them is using probabilistic models. In probabilistic models the efficiency and precision depend chiefly on the kind of dependency between symbols. For example independently and identically distributed Bernoulli trials or Markovian sequences. Whatever be the distribution of symbols in a stochastic sequence, there are different techniques to study the searching problem, for instance some algorithms propose an automatic treatment. To mention one of them, in the article Crochemore and Stefanov [2003], the authors computed the probability of the first hitting time of a word in a binary alphabet using a finite deterministic automata. In the following we shall describe this model.

Crochemore and Stefanov [2003] used a binary alphabet $\mathcal{A} = \{a, b\}$. Any infinite sequence formed by the elements of $\mathcal{A}$ is denoted by $\mathsf{X} = \mathcal{A}^{\mathbb{N}}$. In this article, the authors considered the symbols in the sequence to be independent and identically distributed (i.i.d.) Bernoulli trials. They searched the word $w = abaabab$. To achieve their goal, the authors used the prefixes of the word which consist in writing the word symbol by symbol from its first letter until the word is completed, i.e., the prefixes set is $\{a, ab, aba, abaa, abaab, abaaba, abaabab\}$. The idea of computing the first hitting time of the word in the sequence of symbols X is to embed a Markov prefix sequence through the sequence of symbols. To arrive until this point they used a deterministic finite automata, which is defined as follows

DEFINITION 10. (Hopcroft et al. [2001]). **Deterministic finite automata**
We call a deterministic finite automata (DFA) a five tuple $(\mathcal{A}, \mathcal{Q}, s, \mathcal{F}, \delta)$ where

1. $\mathcal{A}$ is a finite set of symbols (in particular cases $\mathcal{A}$ is a finite alphabet),

2. $\mathcal{Q}$ is a finite set of states,

3. $s \in \mathcal{Q}$ is an initial state,

4. $\mathcal{F} \subset \mathcal{Q}$ is a non empty set of final or accepted states,

5. $\delta : \mathcal{Q} \times \mathcal{A} \to \mathcal{Q}$ is a transition function that takes like an input a state and a symbol and returns a state.

For every sequence of symbols $\mathsf{X} = a_1 a_2 \cdots a_d \in \mathcal{A}^d$, $d \geq 2$ and $q \in \mathcal{Q}$, we recursively define $\delta(q, a_1 a_2 \cdots a_d) = \delta(\delta(q, a_1 \cdots a_{d-1}), a_d)$. A word $w = w_1 \cdots w_h$ is accepted (or recognized) by a DFA if $\delta(s, w) \in \mathcal{F}$. The set of all words accepted by a DFA is called its language.
Crochemore and Stefanov [2003] worked with a DFA: $(\mathcal{A}, \mathcal{Q}, s, \mathcal{F}, \delta)$, where they make the following considerations

Considerations for a DFA 1.

1. $\mathcal{A}$ is the alphabet. For this particular work $\mathcal{A} = \{a, b\}$,

2. $\mathcal{Q}$ is the prefix set. For this particular work $\mathcal{Q} = \{\varepsilon = 1, a = 2, ab = 3, aba = 4, abaa = 5, abaab = 6, abaaba = 7, abaabab = 8\}$. The prefixes are enumerated to make the distinction between a simple letter and a prefix, where the symbol ε is used in case of non one of the symbols $a, ab, aba, abaa, abaab, abaaba, abaabab$ appear in the sequence X,

3. $s = \varepsilon$ is the initial state,

4. $\mathcal{F} = \{w\}$ is the final state,

5. the function $\delta : \mathcal{Q} \times \mathcal{A} \to \mathcal{Q}$ is defined as the longest prefix that can be formed with the concatenation between a symbol and a prefix.

Figure 1.2 shows a graphical representation of DFA previously described.

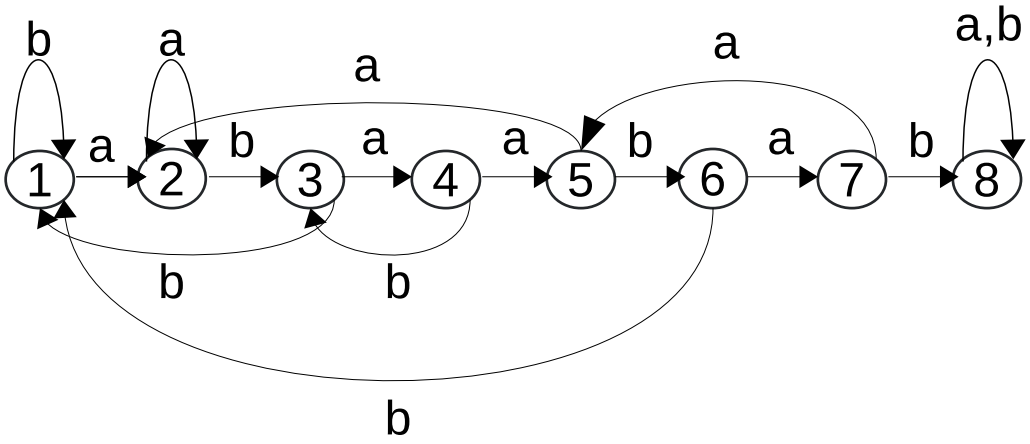


FIGURE 1.2: Graphical representation of the DFA with $\mathcal{A} = \{a, b\}$ for computing the first hitting time of $w = abaabab$.

In Crochemore and Stefanov [2003] it is shown that even if the sequences of letters X is formed by i.i.d. Bernoulli trails, the sequence of prefixes is a Markov sequence where the probability to pass from prefix q to prefix q' is the probability to pass from the last letter of q to the last letter of q', if there is $a \in \mathcal{A}$ such that $\delta(q, a) = q'$. In other words, if $\mathsf{X} = \mathsf{X}_1\mathsf{X}_2 \cdots \mathsf{X}_i \cdots$ is an i.i.d. sequence formed by the elements of $\mathcal{A}$, then the sequence $Y = Y_0Y_1 \cdots Y_i$ defined by

$$Y_0 = s \quad \text{and} \quad Y_i = \delta(Y_{i-1}, \mathsf{X}_i), \quad i \geq 1$$

is a Markov chain with transition matrix

$$\mathsf{P}(p, p') = \begin{cases} \mathbb{P}(\mathsf{X}_1 = a) & \text{if } \delta(p, a) = p', \\ 0 & \text{if } p' \notin \delta(p, \mathcal{A}). \end{cases}$$

Therefore the transition matrix has the expression

$$\mathsf{P} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\ \begin{pmatrix} q & p & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & q & p & 0 & 0 & 0 & 0 & 0 \\ q & 0 & 0 & p & 0 & 0 & 0 & 0 \\ 0 & 0 & q & 0 & p & 0 & 0 & 0 \\ 0 & q & 0 & 0 & 0 & p & 0 & 0 \\ q & 0 & 0 & 0 & 0 & 0 & p & 0 \\ 0 & 0 & 0 & 0 & q & 0 & 0 & p \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

where the letter p denotes the probability to have a success. Of course we can notice that $p+q=1$. It is not difficult to see that the time to absorption at state eight can be viewed as the waiting time until reaching the word *abaabab* in an independent Bernoulli trial. The authors define the time to abortion in state eight as follows

$$\tau := \inf\{n \in \mathbb{N} : \mathsf{X}_n = 8\}. \tag{1.41}$$

Therefore they obtained the following table for the cumulative distribution function (cdf) $F(x)$ of the first hitting time of the word *abaabab*

x	F(x)	x	F(x)
6	0.00000000	7	0.00781250
8	0.01171875	9	0.01757812
15	0.04943848	16	0.05516052
24	0.09879988	25	0.10415334
44	0.19994631	45	0.20469458
66	0.29815051	67	0.30231602
92	0.39878769	93	0.40235591
122	0.49711254	123	0.50009720
160	0.59891921	161	0.60129964
208	0.69860187	209	0.70039068
276	0.79893126	277	0.80012461
393	0.89979901	394	0.90039371
509	0.94976751	510	0.95006564
780	0.98999111	781	0.99005051

In the model described above we can observe that Crochemore and Stefanovc computed in an easy way the first hitting time of a word in a sequences of letters. Nevertheless they considered a binary alphabet and they used the hypothesis that the sequence of letters is modeled by i.i.d. Bernoulli trails, but in most of real applications this hypothesis does not hold true. For this reason Chryssaphinou et al. [2008], proposed a mathematical model where the sequence of symbols is modeled by a semi-Markov sequence. In this work, for $v, h \in \mathbb{N}$, the authors consider a finite set

of words $\mathcal{W} = \{w^1, w^2, ..., w^v\}$ of equal length h. They focus on the waiting time for the first word occurrence from set $\mathcal{W}$ through a semi-Markov sequence (Z_k). The corresponding probability distribution as well as the mean waiting time and the variance were obtained. In the following we shall describe some details of this last work.

Chryssaphinou et al. [2008] consider a semi-Markov sequence $(Z_k)_{k\in\mathbb{N}}$ with EMC $(J_n, S_n)_{n\in\mathbb{N}}$ and semi-Markov kernel $\mathbf{q} = (q_{ij}(k); i, j \in \mathcal{A}, k \in \mathbb{N})$ where the alphabet $\mathcal{A} := \{a_1, a_2, ..., a_l\}$, $l \geq 2$, is the state space of SMC (Z_k). They take into account that if (Z_k) is a semi-Markov sequence therefore the chain (Z_k, B_k) is a Markov chain (MC) with transition probability matrix $\tilde{p}$, i.e.,

$$\tilde{p}((i,u),(j,v)) := \mathbb{P}(Z_{k+1} = j, B_{k+1} = v \mid Z_k = i, B_k = u)$$

where

$$\mathbb{P}(Z_{k+1} = j, B_{k+1} = v \mid Z_k = i, B_k = u) = \begin{cases} \frac{q_{ij}(u+1)}{1-H_i(u)} & \text{if } i \neq j \text{ and } v = 0, \\ \frac{1-H_i(u+1)}{1-H_i(u)} & \text{if } i = j \text{ and } v = u+1, \\ 0 & \text{elsewhere.} \end{cases}$$

see Proposition 5. In this article the authors proposed an h-dimensional Markov process $(\overline{Z}_k, \overline{B}_k)_{k\in\mathbb{N}}$, where $\overline{Z}_k := (Z_k, .., Z_{k+h-1})$ and $\overline{B}_k := (B_k, ..., B_{k+h-1})$ to compute the first hitting time of words taken form a subset $\mathcal{W} \subset \mathcal{A}^h$ where $\mathcal{A}^h$ represents the set of all words of size h formed by the letters in the alphabet $\mathcal{A}$. For every $\overline{z}_j \in \mathcal{A}^h$, let K_j be a subset of $\mathcal{A}^h \times \mathbb{N}^h$ where

$$K_j := \{(\overline{z}_j, \overline{b}) \in \mathcal{A}^h \times \mathbb{N} : \tilde{p}((a_1^{z_j}, b_1), (a_2^{z_j}, b_2)) \cdots \tilde{p}((a_{h-1}^{z_j}, b_{h-1}), (a_h^{z_j}, b_h)) > 0\}.$$

In other words, the set K_j represents the word $\bar{z}_j$ and all possible backwards time for each letter in the word. Clearly for all $i, j = 1, ..., l$ we have, $K_i \cap_{i\neq j} K_j = \emptyset$. Therefore the set which contains all words of size h and the corresponding backward time for each word is denoted by $K := \bigcup_{j=1}^{l} K_j$, this is the state space of MC $(\overline{Z}_k, \overline{B}_k)$. It can be noticed that $K \subset \mathcal{A}^h \times \mathbb{N}^h$.
The initial distribution of MC $(\overline{Z}_k, \overline{B}_k)_{k\in\mathbb{N}}$ is a function of Markov discrete process (Z_k, B_k). This initial distribution is denoted by α and has the following expression

$$\alpha(\overline{z}, \overline{b}) = [\mathbb{P}((Z_0, B_0) = (a_1^z, u_1))] \cdot \tilde{p}((a_1^z, b_1), (a_2^z, b_2)) \cdots \tilde{p}((a_{h-1}^z, b_{h-1}), (a_h^z, b_h)).$$

where $\overline{z} = a_1^z a_2^z \cdots a_h^z$ and $\overline{b} = b_1 b_2 \cdots b_h$. The transition probability of MC $(\overline{Z}_k, \overline{B}_k)$ is given by

$$\widetilde{P}((\overline{z}_i, \overline{b}), (\overline{z}_j, \overline{b}^{'})) = \tilde{p}((a_h^{z_i}, b_h), (a_h^{z_j}, b_h^{'})) \cdot \mathbb{1}_{\{a_s^{z_j} = a_{s+1}^{z_i}; s=1,...,h-1\}} \cdot \mathbb{1}_{\{b_s^{'} = b_{s+1}; s=1,...,h-1\}}.$$

It will be said that $w = w_1 w_2 \cdots w_h$ occurs at time k in the sequence (Z_k) if and only if $Z_{k-h+1} = w_1, ..., Z_k = w_h$, i.e., $\overline{Z}_{k-h+1} = w$. Let $T_{\mathcal{W}} := \min\{k \geq 0 : (\overline{Z}_{k-h+1}, \overline{B}_{k-h+1)} = (w^j, \cdot), w^j \in \mathcal{W}\}$ be the random variable which determines the

first hitting time of an element $w^j \in \mathcal{W}$ in the sequence $(\overline{Z}_k, \overline{B}_k)$. To give the law, the expected value and the variance of the random variable $T_{\mathcal{W}}$ the authors proposed a partition of the state space K according with the elements in $\mathcal{W}$. Let $K_{\mathcal{W}} := \bigcup_{w^j \in \mathcal{W}} K_j$ be the elements in K which contains an element from $\mathcal{W}$.

Now, the probability function, the mean and the variance of $T_{\mathcal{W}}$ can be computed .

PROPOSITION 7. (Chryssaphinou et al. [2008]). The hitting time to $\mathcal{W}$ is given by

$$\mathbb{P}(T_{\mathcal{W}} = k) = \begin{cases} 0 & \text{for } k < h-1, \\ \alpha_{K_{\mathcal{W}}} \cdot \mathbf{1}_{|K_{\mathcal{W}}|} & k = h-1, \\ \alpha_{K^c_{\mathcal{W}}} \cdot (\tilde{P}_{K_{\mathcal{W}^c K^c_{\mathcal{W}}}})^{k-h} \tilde{P}_{K^c_{\mathcal{W}} K_{\mathcal{W}}} \cdot \mathbf{1}_{|K^c_{\mathcal{W}}|} & \text{for } k \geq h, \end{cases}$$

its mean waiting time is

$$\mathbb{E}(T_{\mathcal{W}}) = h - 1 + \alpha_{K^c_{\mathcal{W}}} \cdot (I - \tilde{P}_{K^c_{\mathcal{W}} K^c_{\mathcal{W}}})^{-1} \mathbf{1}_{|K^c_{\mathcal{W}}|},$$

and its variance

$$\mathrm{Var}(T_{\mathcal{W}}) = \widetilde{P}_{K^c_{\mathcal{W}}} \cdot [2(I - \widetilde{P}_{K^c_{\mathcal{W}} K^c_{\mathcal{W}}})^{-1} - I](I - \widetilde{P}_{K^c_{\mathcal{W}} K^c_{\mathcal{W}}})^{-1} \mathbf{1}_{|K^c_{\mathcal{W}}|} - [\widetilde{P}_{K^c_{\mathcal{W}}} (I - \widetilde{P}_{K^c_{\mathcal{W}} K^c_{\mathcal{W}}})^{-1} \mathbf{1}_{|K^c_{\mathcal{W}}|}]^2,$$

where the set $K^c_{\mathcal{W}} = K \setminus K_{\mathcal{W}}$ are the elements in K which do not contain an element from $\mathcal{W}$, the vector $\mathbf{1}_{|K^c_{\mathcal{W}}|}$ is a column vector of ones with dimension the cardinal of $K^c_{\mathcal{W}}$, the initial distribution $\alpha = (\alpha_{K^c_{\mathcal{W}}}, \alpha_{K_{\mathcal{W}}})$ is the initial distribution of process $(\overline{Z}_k, \overline{B}_k)$, where $\alpha_{K^c_{\mathcal{W}}}$ represents the initial distribution of states $i \in K^c_{\mathcal{W}}$ and $\alpha_{K_{\mathcal{W}}}$ represents the initial distribution of states $i \in K_{\mathcal{W}}$, matrix $\tilde{P}_{K^c_{\mathcal{W}} K^c_{\mathcal{W}}}$ is the restriction of matrix $\tilde{P}$ on $K^c_{\mathcal{W}} \times K^c_{\mathcal{W}}$. Since $K^c_{\mathcal{W}}$ is a proper subset of the state space K of an irreducible and aperiodic Markov chain $(\overline{Z}_k, \overline{B}_k)$ the matrix $(I - \widetilde{P}_{K^c_{\mathcal{W}} K^c_{\mathcal{W}}})^{-1}$ exists.

The inconvenience with the model proposed by Chryssaphinou et al. is its implementation. If the length of the sequence of symbols is huge, it is almost impossible to compute the transition matrix $\tilde{P}$. In this thesis we shall improve this part. But identifying a word through a sequence of symbols is not the only point in which we are interested. It is well know that words are not always uniformly distributed through a random sequence, some words are more frequent than others. Therefore we are also interested in identifying how many times a word appears through a random sequence, i.e., we are interested in the frequency of a word. Statistical distribution of word counts in a Markovian sequence of letters, see Schbath [2000], and optimal Markov chain embedding through deterministic finite automata, see Nuel [2008], are used to compute the frequency. In the following we shall describe these two last mentioned works.

Schbath [2000] considered the number of overlapping occurrence of an h-letter word $w = w_1 w_2 \cdots w_h$ on the alphabet $\mathcal{A}$, through a Markov sequence $(\mathbb{X}_k)_{k \in \mathbb{N}}$ with state space $\mathcal{A}$, transition probability matrix

$$\mathbb{P}(a_i, a_j) := \mathbb{P}(\mathbb{X}_{k+1} = a_j \mid \mathbb{X}_k = a_i) \;\; a_i, a_j \in \mathcal{A}.$$

and stationary distribution: $\check{\pi}(a_i)$, $a_i \in \mathcal{A}$. For instance, in the sequence

$$ACGAATAATAAATAAGGCAATAA,$$

there are four occurrences of $AATAA$ (starting at positions 4,7,11 and 19).
Before computing the probability that the word w appears $c \in \mathbb{N}$ times in the MC $(\mathbb{X}_k)$ it is necessary to introduce the period of a word which is defined as follows

DEFINITION 11. (Schbath [2000]). A period of a word $w = w_1 w_2 \cdots w_h$ is denoted by $\varphi(w)$ and it is an integer $p \in \{1, ..., h-1\}$ such that $w_i = w_{i+p}$, for all $i \in \{1, ..., h-1\}$; a period then corresponds to a possible lag between two overlapping occurrences of w.

Schbat denoted by $N(w)$ the number of occurrences of w through the MC $(\mathbb{X}_k)$. The random variable which determines if an occurrence of w starts at position k in the sequence $(\mathbb{X}_k)$ is defined as follows

$$\mathbb{Y}_k := \mathbb{1}_{\{\text{an occurence of } w \text{ starts at position } k \text{ in the sequence}\}},$$

the count $N(w)$ is given by

$$N(w) = \sum_{i=1}^{n-h+1} \mathbb{Y}_i.$$

Therefore the probability that the word appears at certain position through the sequence and its stationary distribution is

$$\check{\pi}(w) = \check{\pi}(w_1) \prod_{j=i}^{h-1} \mathbb{P}(w_j, w_{j+1}).$$

The mean and the variance of the count $N(w)$ are given by the following expressions

$$\mathbb{E}[N(w)] = (n - h + 1)\check{\pi}(w),$$

and

$$\begin{aligned} \text{Var}[N(w)] \quad = \quad & (n-h+1)\pi(w) + 2 \sum_{\substack{p \in \varphi(w) \\ p \leq h-2}} (n-h-p+1)\check{\pi}(w_1 \cdots w_p w_1 \cdots w_h) \\ & + \check{\pi}^2(w) \left(-(n-h+1)^2 + \frac{2}{\check{\pi}(w)} \sum_{d=1}^{n-2h+1} (n-2h+2-d)\mathbb{P}^d(w_h, w_1) \right) \end{aligned}$$

where $\varphi(w)$ is the period set of w, see Definition 11.

Another way to compute the number of times that a word w appears through a MC is presented by Nuel [2008]. He considers a finite alphabet $\mathcal{A} := \{a_1, a_2, ..., a_l\}$, $l \geq 2$ and a word w formed by the elements of $\mathcal{A}$. As Crochemore and Stefanov [2003], he

embedded a Markov prefix sequence through the sequence of symbols using a DFA: $(\mathcal{A}, \mathcal{Q}, s, \mathcal{F}, \delta)$, which accepts or recognizes the word w, see definition 10. Nuel made the same considerations as Crochemore and Stefanov for the DFA, see *consideration for the DFA* 1.
In Nuel's work the sequence of letters $\mathbb{X}$ is modeled by an m-order Markov sequence. He introduced the definition of ambiguity as follows.

DEFINITION 12. (Nuel [2008]). A DFA $(\mathcal{A}, \mathcal{Q}, \mathcal{F}, s, \delta)$ in which there exist $q \in \mathcal{Q}$ and $a, b \in \mathcal{A}^m$ such that $a \neq b$ and $\delta(q, a) = \delta(q, b)$ is called *m-ambiguos*. A DFA which is not m-ambiguous is called m-unambiguos.

He also defined the partial m-inverse of a prefix $p \in \mathcal{Q}$.

DEFINITION 13. (Nuel [2008]). For any DFA $(\mathcal{A}, \mathcal{Q}, s, \mathcal{F}, \delta)$, we define for any $q \in \mathcal{Q}$, and $m \geq 1$, its partial m-inverse as follows

$$\delta^{-m}(q) := \{a \in \mathcal{A}^m \text{ there exists } p \in \mathcal{Q}, \delta(p, a) = q\}$$

Hence, such a DFA is m-unambiguous if for all $\delta^{-m}(q)$ are singletons. Next theorem shows how to embed a pattern through an m-order MC.

THEOREM 3. (Nuel [2008]). If $\mathbb{X} = \mathbb{X}_1 \cdots \mathbb{X}_n$ is an m-order Markov sequence, $m \geq 1$, on $\mathcal{A}$ with transition probability P, if w is a pattern, and if $(\mathcal{A}, \mathcal{Q}, s, \mathcal{F}, \delta)$ is an m-unambiguos DFA which recognizes a w, then the sequence $\mathbb{Y}^* = \mathbb{Y}^*_m \cdots \mathbb{Y}^*_n$, $n > m$, defined by

$$\mathbb{Y}^*_0 = s \text{ and } \mathbb{Y}^*_i = \delta(\mathbb{Y}^*_{i-1}, \mathbb{X}_i) \text{ for all } 1 \leq i \leq n$$

is an 1-order Markov chain of prefixes with transition matrix

$$\tilde{P}^*(p, q) = \begin{cases} \mathbb{P}(\mathbb{X}_{m+1} = b \mid \mathbb{X}_1 \cdots \mathbb{X}_m = \delta^{-m}(p)), & \text{if } \delta(p, b) = q, \\ 0, & \text{if } q \notin \delta(p, \mathcal{A}). \end{cases} \tag{1.42}$$

and such that occurrences of w in $\mathbb{X}$ correspond to occurrences of a subset of letters in $\mathbb{Y}^*$.

Nicodeme et al. [2002] showed that it is possible to build an m-unambiguous DFA starting from a DFA m-ambiguousby duplicating states until the ambiguities are removed.
Let w be a word and let $(\mathcal{A}, \mathcal{Q}, s, \mathcal{F}, \delta)$ be a DFA which recognizes w. The transition probability matrix of the chain Y is $\tilde{P}^* = \mathscr{P} + \mathscr{Q}$, where $\mathscr{Q}$ contains all transitions towards final state w and $\mathscr{P}$ contains all transitions toward regular states. To compute the probability that a word w appears $c \in \mathbb{N}$ times through the m-order Markov sequence: $\mathbb{X}$, Nuel defined a Finite Markov Chain Embedding (FMCE): $\mathcal{Z}$, as follows

DEFINITION 14. (Nuel [2008]). For any $c \in \mathbb{N}$, we define the FMCE $\mathcal{Z}$ by

$$\mathcal{Z}_j := \begin{cases} (\mathbb{Y}^*_j, N_j) & \text{if } N_i < c, \\ f & \text{if } N_j \geq c, \end{cases}$$

where N_j is the number of pattern occurrences of w in $\mathbb{X}_1 \cdots \mathbb{X}_j$.

PROPOSITION 8. (Nuel [2008]). Ordering the $cL+1$ states of $\mathcal{Z}$ as $\{(1,0),...,(L,0),(1,1),...,(L,1),...,(1,c-1),...,(L,c-1),f=(w,c)\}$, the corresponding transition matrix is given by

$$\mathcal{P}((i_1,i_2)(j_1,j_2)) = \begin{cases} \mathscr{P}(i_1,j_1) & \text{if } i_1 \neq w, j_1 \neq w, i_2 = j_2, \\ \mathscr{Q}(i_1,j_1) & \text{if } i_1 \neq w, j_1 = w, j_2 = i_2+1 \\ 1 & \text{if } (i_1,i_2) = (j_1,j_2) = f \\ 0 & \text{otherwise.} \end{cases}$$

The next proposition computes the probability that the word w is repeated c times until position n through the Markov sequence $\mathbb{X}$.

PROPOSITION 9. (Nuel [2008]). Let us consider a partition of the state space of FMCE, $\mathcal{Z}$. Such that

$$U := \{(1,0),...,(L,0),(1,1),...,(L,1),...(1,c-1),...(L,c-1)\}$$

and

$$D := \{(w,c)\}.$$

Therefore, for $i \in U$

$$\mathbb{P}(N < c \mid X_0 = i) = \mathbb{P}(X_0 = i)\mathcal{P}_{u\times u}\mathbf{1}_{|U|}$$

where $\mathcal{P}_{u\times u}$ is the restriction of $\mathcal{P}$ in states $U \times U$ and $\mathbf{1}_{|U|}$ is a vector of ones of size $|U|$.

Even if Markov chains describe a sequence of symbols better than Bernoulli trials, the main drawback in Markov hypothesis is that it can not take into account general distributions in the sojourn time in a state. The sojourn time in a state must be governed by the geometric distribution. In contrast discrete-time semi-Markov processes generalize the Markov hypothesis, they allow the distribution function in a state to be of any type. For this reason in this thesis we shall focus on counting the number of times that a biological sequence is repeated through a DNA sequence by any one of its configurations. We provide the strong law of large numbers for a word sequence. To achieve our goal, we consider two cases: DNA is modeled by an ergodic Markov sequence, and DNA is modeled by a SMC. For both hypothesis we also present the Central Limit Theorem. Even more we are also interested in computing the number of times that the elements from a specific set of words appear through the DNA sequence. But these are not the only problems that we tackle in this theses. Likewise we are also interested in competing risk problems. Competing risk analysis refers to a special type of survival analysis that aims to calculate the probability of an event in the presence of competing events. In next subsection we shall introduce competing risk analysis.

1.4 Competing risk (CR)

In standard survival data, subjects are supposed to experience only one type of event over follow-up, that means in standard survival analysis we are interested in only one cause of failure. But, in real life, individuals (or machines) can experiment more than one cause of death (or more than one cause of failure). For instance, patients could die from a heart attack or breast cancer, or even a traffic accident. There are more than one pathway that can cause the death of a person, but finally the death occurs by one specific cause. Figure 1.3 exemplifies these three causes of death. In engineering, competing risks refer to the cause of breakdown for a machine. For instance if we consider a computer, it can stop working for different reasons, for instance, hardware problems, computational virus or software problems. Figure 1.4 represents these three causes of failure. When we are interested in determining the cause of death (or the cause of failure) in presence of many possible causes, we refer to these events as "competing events". Competing risk analysis refers to a special type of survival analysis that aims to correctly calculate the probability of an event in presence of competing events. In competing risks there are two random variables of interest T is the time to failure, and C the cause of failure.

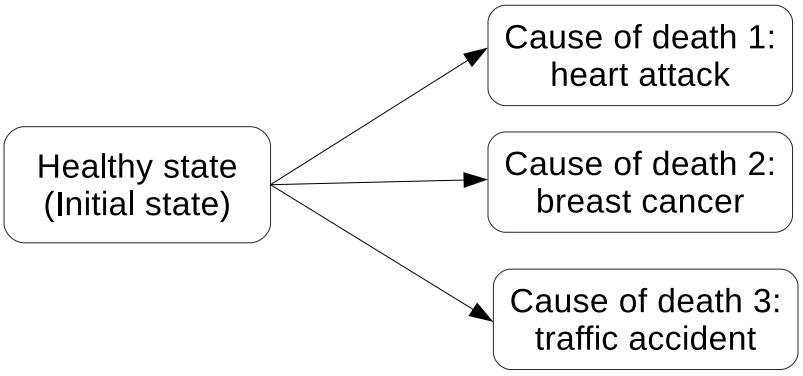


FIGURE 1.3: Example of three competing risk of death

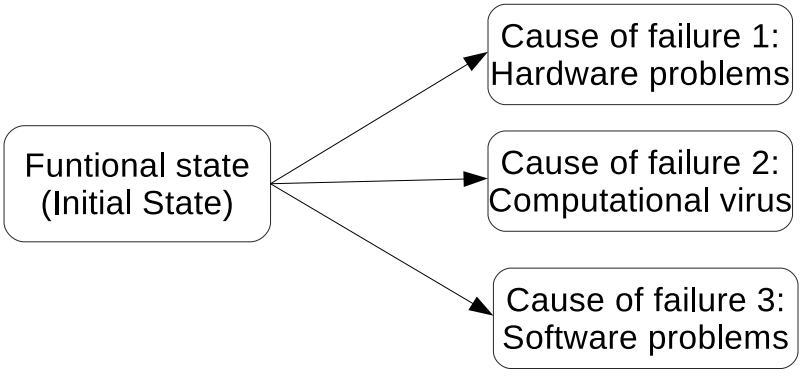


FIGURE 1.4: Example of three competing risk of failure

There are mainly three different ways to specify the distribution function of the time to failure T: the survivor function, the probability function, and the hazard function. The survivor function stands for the probability that the event occurs after a fixed time t, that is,

$$\overline{\mathsf{F}}(t) := \mathbb{P}(\mathsf{T} > t), \;\; 0 \leq t < \infty.$$

We can notice that if F denotes the cumulative distribution function (cdf) of the random variable T, therefore we also have $\overline{\mathsf{F}}(t) = 1 - \mathsf{F}(t)$ for $0 \leq t < \infty$. When T is a continuous variable, the probability function is defined as

$$\mathsf{f}(t) = \frac{d(1 - \overline{\mathsf{F}}(t))}{dt} = \frac{d\mathsf{F}(t)}{dt}, \;\; 0 \leq t < \infty.$$

Obviously, it holds that $\overline{\mathsf{F}}(t) = \int_t^\infty \mathsf{f}(u)du$. Finally, the hazard rate function stands for the rate of that event occurs instantaneously after the time t when it is known that it does not happen before t; that is,

$$\begin{aligned} \lambda(t) &:= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < \mathsf{T} \leq t + \Delta t \mid \mathsf{T} > t)}{\Delta t} \\ &= \frac{\mathsf{f}(t)}{\overline{\mathsf{F}}(t)} = \frac{-d}{dt} \log(\overline{\mathsf{F}}(t)), \;\; 0 \leq t < \infty. \end{aligned}$$

Integrating with respect to t and taking into account that $\overline{\mathsf{F}}(0) = 1$, it holds the equality

$$\overline{\mathsf{F}}(t) = \exp\left\{ - \int_0^t \lambda(u)du \right\} = \exp\{-\Lambda(t)\}, \;\; 0 \leq t < \infty,$$

where $\Lambda(t) = \int_0^t \lambda(u)du$ is known as the cumulative hazard function.

Traditional methods for competing risk estimate the function of interest. The Kaplan Meier (KM) method is one of the most popular, see, e.g., Pintilie [2011], Austin et al. [2016], Lacny et al. [2018], etc. KM method allows one to estimate the survival function. The typical formula for the KM estimator is

$$\hat{\overline{\mathsf{F}}}(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i},$$

where $t_1 < t_2 < t_3 < ...$ are the ordered time points at which an event was observed, n_i represents the number of patients at risk at time t_i and d_i is the number of events at time t_i. This formula can be transformed through algebraic manipulation to express the probability of event as:

$$\hat{\overline{\mathsf{F}}}(t) = 1 - \hat{\overline{\mathsf{F}}}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i} \hat{\overline{\mathsf{F}}}(t_{i-1}). \tag{1.43}$$

In the presence of CR there are at least 2 types of events: event of interest, identified with the subscript e, and the competing risk event, identified with the subscript

c. Prentice et al. [1978] introduced the formula for the probability of an event of interest in the presence of CR

$$\hat{P}_e(t) = \sum_{t_i \leq t} \frac{d_{ei}}{n_i} \hat{\bar{\mathsf{F}}}(t_{i-1}), \tag{1.44}$$

where d_{ei} is the number of events of interest. It is of interest to point out the relation between Equations (1.43) and (1.44). Since d_i is the number of all events at t_i, it can be conceived as the sum of the number of events of interest d_{ei} and the number of CR events d_{ci} at time t_i. As such, the probability of any type of event can be decomposed as follows:

$$\begin{aligned} \text{Probability of all events} &= \hat{P}_e(t) + \hat{P}_c(t) \\ &= \sum_{t_i \leq t} \frac{d_{ei}}{n_i} \hat{\bar{\mathsf{F}}}(t_{i-1}) + \sum_{t_i \leq t} \frac{d_{ci}}{n_i} \hat{\bar{\mathsf{F}}}(t_{i-1}) \\ &= \sum_{t_i \leq t} \frac{(d_{ei} + d_{ci}) \hat{\bar{\mathsf{F}}}(t_{i-1})}{n_i} = \sum_{t_i \leq t} \frac{d_i}{n_i} \hat{\bar{\mathsf{F}}}(t_{i-1}). \end{aligned}$$

Thus the probability of all events can be decomposed in the probabilities for each type of event.

Another important method for competing risk is Nelson-Aalen (NA), see e.g. Tsiatis [2005], Njamen [2017]. The NA estimator, as the KM estimator, is a non-parametric estimator. It is used in survival theory, reliability engineering and life insurance to estimate the failure time of an event. For that method the standard mathematical formulation is as follows: let $\mathsf{T} = \{\mathsf{T}_1, ..., \mathsf{T}_N\}$ be the times to event and let $\mathcal{T} = \{\mathcal{T}_1, ..., \mathcal{T}_N\}$ be the censor times. Let F and G be the cdfs for the time to event and the censor time, respectively. The observed times are $\mathsf{Z} = \{\mathsf{Z}_1, ..., \mathsf{Z}_N\}$ where $\mathsf{Z}_j = \min\{T_j, \mathcal{T}_j\}$, $1 \leq j \leq N$. In addition, it is also known what time is really observed; i.e., the final available information are the pairs $\{(\mathsf{Z}_1, \delta_1), ..., (\mathsf{Z}_N, \delta_N)\}$, where $\delta_j = \mathbb{1}_{\{T_j \leq \mathcal{T}_j\}}$, i.e., δ_j takes the value 1 if the time to an event is observed and 0 otherwise. Therefore the Nelson-Aalen (NA) Estimator for the cumulative hazard function is

$$\hat{\Lambda}(t) = \sum_{i=1}^{n} \delta_i c_N(R_i) \mathbb{1}_{\{\mathsf{Z}_i \leq t\}},$$

where R_i is the rank of Z_i among $\mathsf{Z}_1, \mathsf{Z}_2, ..., \mathsf{Z}_N$ and $c_N(i) = \frac{1}{(N-i+1)}$, for $i = 1, 2, ..., N$.

Competing risk are a natural extension of (Markov) phase-type distribution (Ph-distribution), see, e.g., Lindqvist and Kjølen [2018]. In the sequence, we shall describe this last reference. Lindqvist and Kjølen [2018] presented an extension of the phase-type methodology for modeling lifetime distributions to include the case of competing risk. This is done by considering finite state Markov processes in continuous time with more than one absorbing state, letting each absorbing state correspond to a particular risk or cause of failure. First Lindqvist and Kjølen introduce the classical

phase-type distribution. The authors state that a phase-type distribution can be described in terms of a Markov process $\{\mathbb{X}(t);\ t \geq 0\}$, where the system moves through some or all K transient states, or phases, before moving to a single absorbing state $\mathsf{K}+1$. The time of absorption, T, is then said to have a phase-type distribution. The infinitesimal generator $\mathcal{G}$ of the Markov process is a $(\mathsf{K}+1) \times (\mathsf{K}+1)$ matrix given on blocks form as

$$\mathcal{G} = \begin{pmatrix} \Theta & \ell \\ \mathbf{0} & 0 \end{pmatrix}, \tag{1.45}$$

where Θ is the $\mathsf{K} \times \mathsf{K}$ matrix corresponding to the transitions between the transient states, ℓ is the $\mathsf{K} \times 1$ vector defining direct transition from the transient states to the absorbing state, while $\mathbf{0}$ is a $1 \times \mathsf{K}$ vector of zeros. The authors define $\mathbb{P}(t)$ as the matrix of transition probabilities, i.e., $\mathbb{P}_{ij}(t) = \mathbb{P}(\mathbb{X}(t) = j \mid \mathbb{X}(0) = i)$ where

$$\mathbb{P}(t) = e^{\mathcal{G}t} = \sum_{i=0}^{\infty} \mathcal{G}^i \frac{t^i}{i!}.$$

It can be shown that this implies that

$$\mathbb{P}(t) = \begin{pmatrix} e^{\Theta t} & \Theta^{-1}(e^{\Theta t - I})\ell \\ \mathbf{0} & 1 \end{pmatrix}.$$

From this last equation, an expression for the cumulative distribution function of T,

$$\mathsf{F}(t) = \mathbb{P}(\mathsf{T} \leq t) = \mathbb{P}(\mathbb{X}(t) = \mathsf{K}+1) = \alpha\Theta^{-1}(e^{\Theta t - I})\ell,$$

where α is the initial distribution of the Markov process, i.e., $\alpha(i) = \mathbb{P}(\mathbb{X}(0) = i)$, for $i = 1, ..., \mathsf{K}+1$. In standard phase-type distributions, it is considered the time to failure T by a unique cause. In competing risk, it is supposed that the system can experience $m > 1$ competing failure causes. To clarify the idea, suppose the Markov process $\{\mathbb{X}(t);\ t \geq 0\}$ has K transient states and $m > 1$ absorbing states, named $\mathsf{K}+1, \mathsf{K}+2, ..., \mathsf{K}+m$. Letting T be the time of absorption (in any one of the absorbing states), and let C be the cause of failure, which is represented by the state where absorption occurs, i.e., $C = \mathsf{K}+j$ if $\mathbb{X}(\mathsf{T}) = \mathsf{K}+j; j = 1, 2, ..., m$. The pair (T, C) can be viewed as an observation from a classical competing risk process with causes $\mathsf{K}+1, ..., \mathsf{K}+m$. By extending the matrix (1.45) to encompass m absorbing states, we obtain the infinitesimal generating matrix of Markov process to be the $(\mathsf{K}+m) \times (\mathsf{K}+m)$ matrix given on block form as

$$\mathcal{G} = \begin{pmatrix} \Theta & L \\ 0_1 & 0_2 \end{pmatrix}, \tag{1.46}$$

where as before, Θ is the $\mathsf{K} \times \mathsf{K}$ matrix corresponding to the transition between the transient states. The vector ℓ is now replaced by the $\mathsf{K} \times m$ matrix L which contains transitions from the transient states to the absorbing states. Furthermore, 0_1 and 0_2 are, respectively, $m \times \mathsf{K}$ and $m \times m$ matrices of zeros.
It can shown that matrix (1.46) implies that the matrix of transition probabilities

$\mathbb{P}_{ij}(t)$ is given by

$$\mathbb{P}(t) = \begin{pmatrix} e^{\Theta t} & \Theta^{-1}(e^{\Theta t} - I)L \\ 0_1 & I \end{pmatrix}, \tag{1.47}$$

where I is the $\mathsf{K} \times \mathsf{K}$ identity matrix. From (1.47) the authors obtained expressions for the subdistribution functions, given by

$$\mathsf{F}_j(t) := \mathbb{P}(\mathsf{T} \leq t, C = j) = \mathbb{P}(\mathbb{X}(t) = j) = \alpha\Theta^{-1}(e^{\Theta t - I})Lv_j$$

for $j = 1, ..., m$. By differentiation, the authors got the sub-densities

$$\mathsf{f}_j(t) = \mathsf{F}_j'(t) = \alpha e^{\Theta t} L v_j,$$

where α is the initial distribution of the Markov chain and v_j is the m-vector with j-th element equal to 1 and the rest equal to 0. Therefore, the cause-specific hazard rate is given by

$$\lambda_j(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(\mathsf{T} \leq t + \Delta t, C = j \mid \mathsf{T} > t)}{\Delta t} = \frac{\mathsf{F}'(t)}{\mathbb{P}(\mathsf{T} > t)} = \frac{\alpha e^{\Theta t L v_j}}{\alpha e^{\Theta t} \mathbf{1}_k},$$

where $\mathbf{1}_k$ is a k-vector of ones.
One of the aims in this thesis is to extend these results to semi-Markov models which are the generalization of Markov processes see Limnios and Oprişan [2001].

Chapter 2

Identification of Words in Biological Sequences Under the semi-Markov Hypothesis

Genomic sequences [1] are likely to be the most sophisticated information databases created by nature through the evolution process. For this reason, model DNA sequences via mathematical tools are challenged questions for mathematicians and biologists. In most cases DNA sequences are compared to a stochastic process governed by four nitrogenous bases: Adenine (A), Cytosine (C), Thymine (T) and Guanine (G). In probabilistic models the efficiency and precision depend chiefly on the kind of dependency between symbols. For instance, Stefanov et al. [1997], Robin and Daudin [1999] and Chadjiconstantinidis et al. [2000] model DNA sequences derived from independently and identically distributed Bernoulli trials; Antzoulakos [2001], Fu and Chang [2002] between others authors consider DNA sequences are modeled by a Markov chain.

Given a genome sequence it is interesting to recognize a word (pattern), counting the number of or determine the first position that this appears, see, e.g., Abadi and Vergne [2008], Li et al. [2016], Li et al. [2018], Sigwart and Garbett [2018], Hebert et al. [2003] and Robin et al. [2007].

Seek a particular word over a chain formed by a big quantity of nucleotides is a very cumbersome task and in most cases it can not be achieved by simple inspection. There are a number of models and algorithms which propose an automatic treatment for the searching problem in a few seconds see, e.g., Aboluion [2011], Srivastava and Baptista [2016], Stefanov et al. [2011], Picard et al. [2011], Codish et al. [2017], Montemanni [2015], Glaz et al. [2006], Crochemore and Stefanov [2003] and Nuel [2008], Touyar et al. [2008].

[1] This chapter develops the content of an article which appears in the journal 'Computational Biology' put in shape to be inserted in this thesis.
Garcia-Maya, B. I., and Limnios, N. (2020). Identification of Words in Biological Sequences Under the Semi-Markov Hypothesis. Journal of Computational Biology, 27(5), 683-697.

In this chapter we find a pattern in a DNA sequence under the hypothesis that DNA is modeled by a semi-Markov chain. The semi-Markov hypothesis allows us to take into account general distributions in the sojourn time in a state. To achieve our goal we use the prefixes chain. Suppose we search the word (pattern) $w = ACCT$ in a DNA sequence. We construct the word step by step from its first symbol to its last one. The elements of this construction are called prefixes, i.e., consider the following DNA sequence from a bacteriophage:
GGGCGGCGACCTCGCGGGTTTTCGCTATTTATGAAAATTTTCCGGTTTAAG

TTCTTCTTCGTCATAACTTAATGTTTTTATTTAAAATACCCTCTGAAAAG...

and suppose we want to compute the first position of $w = ACCT$, i.e., we want to compute how many nucleotides have to appear in the DNA before w. To do this, we use the prefixes of w, which is the set $\{A, AC, ACC, ACCT\}$. For this example, it is clear that the first appearance of w occurs at position (starting from zero) 11. Using the chain of the longest prefix of the word and all possible backwards times for each prefix, it is computed the distribution for the (first) hitting position of the word in a sequence of letters. To show the applicability of our proposed model, we test it in a bacteriophage DNA sequence. We present, the distribution function, the expected value, the variance and the standard deviation of the random variable which represents the (first) hitting position of the word. The word occurrence rate is also presented.

2.1 Prefix chain of a single word

Let us consider a finite alphabet, say $\mathcal{A} = \{a_1, ..., a_l\}$, $2 \leq l < \infty$. A word formed by the elements of $\mathcal{A}$ is represented by $w := w_1 w_2 \cdots w_h$ where $w_1, w_2, ..., w_h \in \mathcal{A}$. The length of the word is expressed by $|w|$ and represents its number of letters, in this case $|w| = h$, where $h \in \mathbb{N}^*$. We shall denote the set of all words of size $h \in \mathbb{N}^*$ formed by the elements of $\mathcal{A}$ by $\mathcal{A}^h$. As an example, suppose the set of letters is $\mathcal{A} = \{a, b\}$ and $h = 2$ therefore $\mathcal{A}^h = \{aa, ab, ba, bb\}$. Based on the structure of a word $w \in \mathcal{A}^h$, we shall define its prefix set

$$E_w = \{\varepsilon, w_1, w_1 w_2, ..., w_1 w_2 \cdots w_{h-1}, w\}, \tag{2.1}$$

where symbol ε denotes the empty prefix which is used in case of none of the symbols $\{w_1, w_1 w_2, ..., w_1 w_2 \cdots w_{h-1}, w\}$ appears in the sequence. Observe that $|\varepsilon| = 0$.
Let

$$\delta_{E_w} : E_w \times \mathcal{A} \to E_w \tag{2.2}$$

be a mapping such that for a prefix $q \in E_w$ and a letter $a \in \mathcal{A}$, $\delta_{E_w}(q, a)$ is defined as the longest suffix of qa (concatenation of q and a) in the prefix set E_w. To clarify this definition observe the follow examples.

EXAMPLE 1. if $\mathcal{A} = \{a, b\}$, $w = ba$, the prefix set is $E_w = \{\varepsilon, b, ba\}$, then

$$\begin{array}{ll} \delta_{E_w}(\varepsilon, a) = \varepsilon, & \delta_{E_w}(\varepsilon, b) = b, \\ \delta_{E_w}(b, a) = ba, & \delta_{E_w}(b, b) = b, \\ \delta_{E_w}(ba, a) = \varepsilon, & \delta_{E_w}(ba, b) = b, \end{array}$$

In Figure 2.1 we can observe the graphic representation of example 1, where to not confuse letters with prefixes, the prefix set is numerated as follows $\{1 = \varepsilon, 2 = b, 3 = ba\}$.

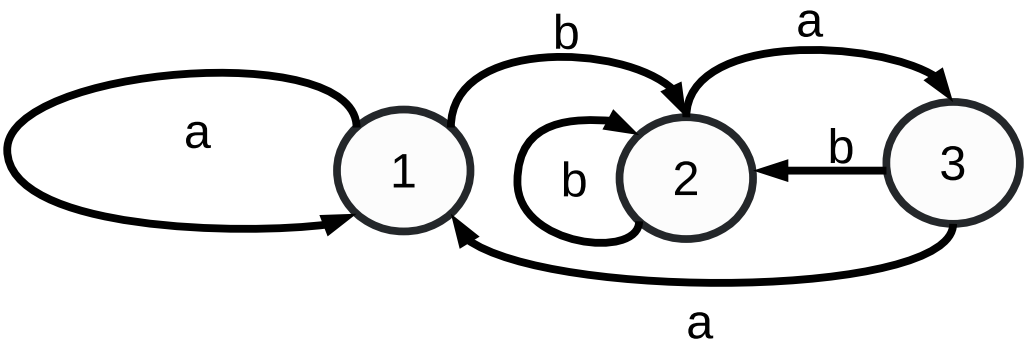


FIGURE 2.1: Graphic representation of example 1

EXAMPLE 2. Suppose we add one letter to the alphabet and we search the same word as in example 1, i.e., $\mathcal{A} = \{a, b, c\}$ and $w = ba$. The prefix set does not change $E_w = \{\varepsilon, b, ba\}$ due to w is the same, nevertheless the results for δ_{E_w} are different due to the alphabet has one letter more. The results for δ_{E_w} are:

$$\begin{array}{lll} \delta_{E_w}(\varepsilon, a) = \varepsilon, & \delta_{E_w}(\varepsilon, b) = b, & \delta_{E_w}(\varepsilon, c) = \varepsilon, \\ \delta_{E_w}(b, a) = ba, & \delta_{E_w}(b, b) = b, & \delta_{E_w}(b, c) = \varepsilon, \\ \delta_{E_w}(ba, a) = \varepsilon, & \delta_{E_w}(ba, b) = b, & \delta_{E_w}(ba, c) = \varepsilon. \end{array}$$

In Figure 2.2 we can observe the graphic representation of example 2, where the prefixes are numerate as in Figure 2.1.

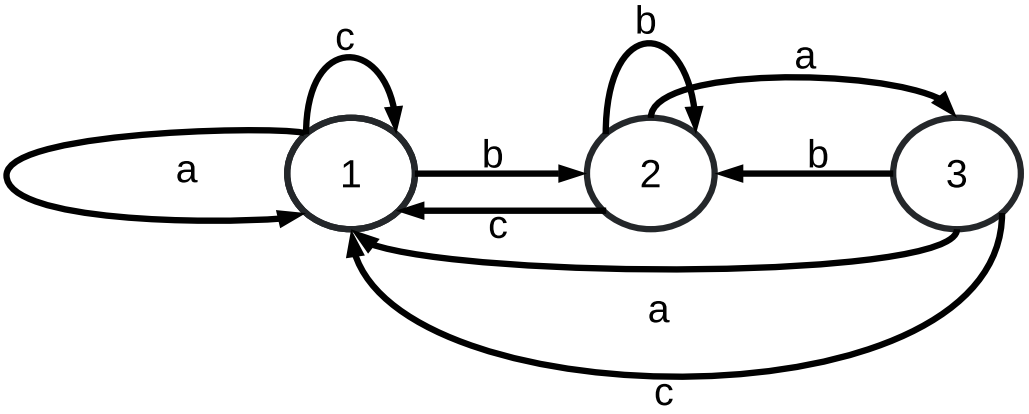


FIGURE 2.2

Observe that, for $p \in E_w$ the set $\{i \in \mathcal{A} : \delta_{E_w}(p,i) = \varepsilon\}$ has more than one element if $|\mathcal{A}| > 2$. Therefore $\delta_{E_w}(p,i)$ is not a one-to-one mapping in general. According to Nicodeme et al. [2002], new elements can be added to E_w such that $\delta_{E_w}(p,i)$ becomes a one-to-one mapping for p fixed. If p and q are two prefixes such that q is different from prefix ε, i.e., $q = w_1 w_2 \cdots w_l$ for $w_1, w_2, \cdots, w_l \in \mathcal{A}$ where $1 \leq l \leq h$ and, if there is some $i \in \mathcal{A}$ such that $\delta_{E_w}(p,i) = q$, therefore i is the last letter of q, i.e., $i = w_l$ for this reason for p fixed only when δ_{E_w} results in ε, i.e., $\delta_{E_w}(p,i) = \varepsilon$, the function δ_{E_w} is not one to one. The empty prefix ε will be labeled according to the letter which is added to p to obtain ε, that means, instead of writing

$$\delta_{E_w}(p,i) = \varepsilon,$$

it will be written

$$\delta_{E_w}(p,i) = \varepsilon_i.$$

Let

$$\mathsf{E} := \cup_{i \in \mathcal{A} \setminus \{w_1\}} \{\varepsilon_i\} \cup \{w_1, w_1 w_2, ..., w_1 w_2 \cdots w_{h-1}, w\}, \tag{2.3}$$

be the extended state space of E_w in which for $p \in \mathsf{E}$ and $i \in \mathcal{A}$, $\delta(p,i)$ is now a one-to-one mapping. The previous definition of δ_{E_w} can be extended as follows, $\delta_{\mathsf{E}} : \mathsf{E} \times \mathcal{A} \to \mathsf{E}$. Henceforth, this last definition will be considered.

The partial inverse of δ_{E} is the function $\delta_{\mathsf{E}}^{-1} : \mathsf{E} \longrightarrow \mathcal{A}$ and it is defined as follows: for all $p \in \mathsf{E}$,

$$\delta_{\mathsf{E}}^{-1}(p) := \{i \in \mathcal{A} \text{ where there exist } q \in \mathsf{E}, \text{ such that } \delta(q,i) = p\}. \tag{2.4}$$

Roughly speaking, the partial inverse of prefix p gives the last letter of p, i.e., if $p = ba$, $\delta_{\mathsf{E}}^{-1}(p) = a$. Observe that the partial inverse defined in E is one-to-one.

2.2 Prefix process in the semi-Markov case

In the considering problem: computing the first hitting position of a word (pattern) in a biological sequence. The biological sequence is modeled by a semi-Markov chain (Z_k), see subsection 1.2, the state space $\mathcal{A}$ is the genomic alphabet: Adenine (A), Cytosine (C), Thymine (T) and Guanine (G); the maximum sojourn time in a state $i \in \mathcal{A}$, see Equation (1.19), represents the maximal number of nucleotides that can be found together through the DNA sequence. The general idea is to compute the first hitting position of the word using the prefix process. In the sequel of this section we shall define the embedded prefix chain in the semi-Markov process.

Let us consider a stochastic process $(Z_k)_{k\in\mathbb{N}}$ which models a sequence of letters taken from a finite alphabet $\mathcal{A}$. If w is a word from $\mathcal{A}$, the embedded prefix chain is defined as follows.

DEFINITION 15. (Garcia-Maya and Limnios [2020]). The prefix chain of $w = w_1 \cdots w_h \in \mathcal{A}^h$ embedded in the SMC (Z_k) and defined in E, see Equation (2.3), is denoted by $Y := (Y_k)_{k\in\mathbb{N}}$ where

$$Y_0 := \begin{cases} w_1 & \text{if } Z_0 = w_1, \\ \varepsilon_i & \text{if } Z_0 = i, \ \ i \in \mathcal{A} \setminus \{w_1\}, \end{cases}$$

and

$$Y_k := \delta_{\mathsf{E}}(Y_{k-1}, Z_k), \quad k \geq 1.$$

EXAMPLE 3. If the alphabet is $\mathcal{A} = \{a, b, c\}$ and the word is $w = ba$ then, if the elements of the prefix set E are listed as follows $\mathsf{E} = \{\varepsilon_a = 0, \varepsilon_c = 1, b = 2, ba = 3\}$. For the following sample

$$Z : aaaccba...$$

we have as result

$$Y : 0001123...$$

We shall say that a word $w = w_1 w_2 \cdots w_h \in \mathcal{A}^h$ occurs at time k in the sequence (Z_k) iff

$$Z_{k-h+1} = w_1, Z_{k-h+2} = w_2, ..., Z_k = w_h.$$

Observe that $w \in \mathcal{W}$ has positive probability to appear in the sequence (Z_k) if

$$\mathbb{P}(Z_{k-h+2} = w_2 \mid Z_{k-h+1} = w_1) \cdots \mathbb{P}(Z_k = w_h \mid Z_{k-1} = w_{h-1}) > 0. \tag{2.5}$$

To avoid trivialities we shall consider Equation (2.5) is always positive.
The number of positions we have to wait for an occurrence of w in (Z_k) correspond to the number of positions we have to wait for an occurrence of w in (Y_k), see Nuel [2008]. Therefore w has positive probability to appear in (Z_k) iff w has positive probability to appear in (Y_k) and the number of positions we have to wait for an occurrence of w in (Z_k) correspond to the number of positions we have to wait for an occurrence of w in (Y_k).

Observe that to embed a prefix process in a semi-Markov chain it is necessary to add the backward recurrence time corresponding to each prefix. Let (Y_k, B_k) be the process of prefixes and backward times. To define the state space of (Y_k, B_k), let us introduce the blocks of the word. If $w = \underbrace{i \cdots i}_{h}$, $i \in \mathcal{A}$ and, $h \in \mathbb{N}^*$, we shall say that w is a block of $i's$ of length h and it will be denoted

$$w = i^{(h)}, \;\; i \in \mathcal{A}, \;\; h \in \mathbb{N}, 1 \leq h < \infty.$$

If w is not a block, it can be obtained by concatenating words which are blocks, see Karaliopoulou [2009]. That is, w can be expressed as

$$w = w(1)w(2)\cdots w(\eta), \tag{2.6}$$

where $w(1) = {i_1}^{(n_1)}$, $w(2) = {i_2}^{(n_2)}$, ..., $w(\eta) = {i_\eta}^{(n_\eta)}$, for $i_1, i_2, ..., i_\eta \in \mathcal{A}$ and $n_1, n_2, ..., n_\eta \in \mathbb{N}^*$, such that $n_1 + n_2 + \cdots + n_\eta = |w|$. Noticing that a prefix is also a word, therefore it can be represented by the concatenation of blocks. We shall introduce the backward time for each prefix

DEFINITION 16. (Garcia-Maya and Limnios [2020]). We shall define the backward position time of a prefix $p = w_1 w_2 \cdots w_{l-1} w_l$ for $1 \leq l \leq h$, as the size of its last block minus one. This definition comes from the time (the number of positions) we have stayed in the last letter of the prefix since the last jump (renewal point). Notice that, in general the backward time of a prefix is different from the backward time B_k of the letters, see Equation (1.18). The backward time of a prefix will be denoted by $B(p)$.

EXAMPLE 4. Let $w = bbaab$ be a word from the alphabet $\mathcal{A} = \{a, b\}$, and let $p = bba$ be a prefix of w. Then, it is clear that, the backward time of p is, $B(p) = 0$.

Now, we are ready to introduce the backward times for each prefix through the semi-Markov process. Let l_p^*, be the set of backward times which correspond to prefix p. If $i \in \mathcal{A}$ is the partial inverse of p, i.e., $i = \delta_{\mathsf{E}}^{-1}(p)$, see Equation (2.4), and n_1 is the size of the first block of w, see Equation (2.6), then

$$l_p^* = \begin{cases} [\![0, l_i]\!], & \text{if } \; |p| = 0, \\ B(p), & \text{if } \; |p| \neq n_1 \;\text{ and }\; |p| \neq 0, \\ [\![B(p), l_i]\!], & \text{if } \; |p| = n_1. \end{cases} \tag{2.7}$$

where $[\![a, b]\!]$ denotes an interval of integers, i.e., a subset of $\mathbb{N}$. Let

$$K_p(\mathsf{E}) := \{(p, n) : p \in \mathsf{E}, n \in l_p^*\} \tag{2.8}$$

be the set which represents the prefix $p \in \mathsf{E}$ and its backward times, by consequence the set

$$K(\mathsf{E}) := \bigcup_{p \in \mathsf{E}} K_p(\mathsf{E}) \tag{2.9}$$

represents all prefixes in E and their corresponding backward times. Clearly we have: $K_p(\mathsf{E}) \cap K_q(\mathsf{E}) = \emptyset$, if $p \neq q$. The set $K(\mathsf{E})$ is the state space of process (Y_k, B_k). Algorithm 1 proposed here (see Appendix) provide the state space $K(\mathsf{E})$.

PROPOSITION 10. (Garcia-Maya and Limnios [2020]). The process $(Y_k, B_k)_{k\in\mathbb{N}}$ is a Markov chain with state space $K(\mathsf{E})$ and initial distribution $\alpha(i) = \mathbb{P}(Z_0 = i)$, $i \in \mathcal{A}$. Let us consider $p \in \mathsf{E}$, $i = \delta_{\mathsf{E}}^{-1}(p)$. Then the initial distribution of the process (Y_k, B_k) is formally defined by

$$\mathbb{P}(Y_0 = p, B_0 = u) = \begin{cases} \alpha(w_1)\mathbb{1}_{\{u=0\}} & \text{if } p = w_1, \\ \alpha(i)\mathbb{1}_{\{u=0\}} & \text{if } p = \varepsilon_i, \ i \in \mathcal{A} \setminus \{w_1\}, \\ 0 & \text{otherwise,} \end{cases}$$

and its transition probabilities are

$$\mathbb{P}(Y_{k+1} = q, B_{k+1} = v \mid Y_k = p, B_k = u) = \begin{cases} \frac{q_{ia}(u+1)}{1-H_i(u)}\mathbb{1}_{\{\delta_{\mathsf{E}}(p,a)=q\}} & \text{if } i \neq a, \ v = 0, \\ \frac{1-H_i(u+1)}{1-H_i(u)}\mathbb{1}_{\{\delta_{\mathsf{E}}(p,a)=q\}} & \text{if } i = a, \ v = u+1, \\ 0 & \text{otherwise.} \end{cases} \tag{2.10}$$

Proof. The initial distribution comes from the definition of Y_0, see Definition 15, and the backward position time at starting time, see Equation (1.18). We shall use here the fact that (Z_k, U_k) is a Markov chain with known transition probability, see e.g., Barbu and Limnios [2008]. Let $p \in \mathsf{E}$ be a prefix and suppose that there exists $a \in \mathcal{A}$ such that $\delta(p, a) = q$, for some $q \in \mathsf{E}$. If $p = w_1 w_2 \cdots w_{l-1} w_l$ for $1 \leq l \leq h$, then

$$\begin{aligned} &\mathbb{P}(Y_{k+1} = q, B_{k+1} = v \mid Y_k = p, B_k = u) \\ &= \mathbb{P}(Z_{k+1} = a, B_{k+1} = v \mid Z_k = w_l, B_k = u, Z_{k-1} = w_{l-1}, B_{k-1} = \cdot, ..., \\ &\quad Z_{k-l+1} = w_1, B_{k-l+1} = \cdot), \\ &= \mathbb{P}(Z_{k+1} = a, B_{k+1} = v \mid Z_k = w_l, B_k = u). \end{aligned} \tag{2.11}$$

If $p = \varepsilon_j$ for $j \in \mathcal{A} \setminus \{w_1\}$, then

$$\begin{aligned} &\mathbb{P}(Y_{k+1} = q, B_{k+1} = v \mid Y_k = p, B_k = u) \\ &= \mathbb{P}(Z_{k+1} = a, B_{k+1} = v \mid Z_k = j, \cdot, B_{k-1} = \cdot, ..., Z_{k-i+1} = \cdot, B_{k-i+1} = \cdot), \\ &= \mathbb{P}(Z_{k+1} = a, B_{k+1} = v \mid Z_k = j, B_k = u). \end{aligned} \tag{2.12}$$

To denote in a general the transition probability, let $i = \delta_{\mathsf{E}}^{-1}(p)$ be the partial inverse of p. Therefore in Equations (3.4) and (3.5) $Z_k = i$, thus

$$\mathbb{P}(Y_{k+1} = q, B_{k+1} = v \mid Y_k = p, B_k = u) = \mathbb{P}(Z_{k+1} = a, B_{k+1} = v \mid Z_k = i, B_k = u). \tag{2.13}$$

If $k+1$ is a renewal time for (Z_k) then $v = 0$ and $i \neq a$, yields

$$\mathbb{P}(Z_{k+1} = a, B_{k+1} = v \mid Z_k = i, B_k = u) = \frac{q_{ia}(u+1)}{1 - H_i(u)}, \tag{2.14}$$

if $k+1$ is not a renewal time for (Z_k) then $v = u+1$ and $a = i$, yields

$$\mathbb{P}(Z_{k+1} = a, B_{k+1} = v \mid Z_k = i, B_k = u) = \frac{1 - H_i(u+1)}{1 - H_i(u)} \tag{2.15}$$

see Barbu and Limnios [2008], so, the proposition is proved . ■
To simplify the notation, let $\check{P}$ be the transition probability matrix and β the initial distribution of process (Y_k, B_k) respectively, such that

$$\check{P}((p,u),(q,v)) := \mathbb{P}(Y_{k+1} = q, B_{k+1} = v \mid Y_k = p, B_k = u) \tag{2.16}$$

and

$$\beta(p,u) := \mathbb{P}(Y_0 = p, B_0 = u). \tag{2.17}$$

Algorithm 2 proposed here (see appendix B) computes the transition probability matrix $\check{P}$.

If $(J_n, S_n)_{n\in\mathbb{N}}$ is irreducible, then Markov chain (Z_k, B_k) is irreducible too, see Chryssaphinou et al. [2008]. We shall prove, in the next proposition, the same properties for (Y_k, B_k), i.e., the irreducibility and aperiodicity.

PROPOSITION 11. (Garcia-Maya and Limnios [2020]). Let $\mathcal{A}$ be an alphabet, such that $|\mathcal{A}| \geq 2$. If the Markov renewal process (J_n, S_n) with state space $\mathcal{A} \times \mathbb{N}$ is irreducible and aperiodic, then process (Y_k, B_k) with state space $K(\mathsf{E})$ is also irreducible and aperiodic.

Proof. Let be $(p,u),(q,v) \in K(\mathsf{E})$ and $i = \delta_{\mathsf{E}}^{-1}(p)$. If $q = w_1$ or $q = \varepsilon_a$ with $a \in \mathcal{A} \setminus \{w_1\}$ the proof is a direct consequence of the irreducibility of (Z_k, B_k). If $q = w_1 w_2 \cdots w_\ell$, with $1 < \ell \leq h$. Let $q_1 = w_1$, $q_2 = w_1 w_2, ..., q = w_1 w_2 \cdots w_\ell$, be the consecutive prefixes from w_1 to $q = w_1 w_2 \cdots w_\ell$ and (q_1, u_1), $(q_2, u_2), ..., (q, v)$ the consecutive couples in $\mathsf{E} \times \mathbb{N}$, such that for $1 \leq i \leq \ell - 1$

$$u_{i+1} = \begin{cases} 0 & \text{if } w_i \neq w_{i+1} \\ u_i + 1 & \text{elsewhere} \end{cases}$$

due to the irreducibility of (Z_k, B_k), for $(i,u) \in \mathcal{A} \times \mathbb{N}$ there exist $n \in \mathbb{N}^*$ such that, for $m \in \mathbb{N}$

$$\mathbb{P}(Z_{m+n} = w_1, B_{m+n} = u_1 \mid Z_m = i, B_m = u) > 0,$$

by Proposition 10

$$\check{P}((w_1,u_1),(w_1w_2,u_2)) \times \cdots \times \check{P}((w_1 w_2 \cdots w_{\ell-1}, u_{\ell-1}),(w_1 w_2 \cdots w_\ell, u_\ell)) > 0.$$

Therefore there exist $n_1 = n + \ell - 1$ such that

$$\mathbb{P}(Y_{m+n_1} = q, B_{m+n_1} = v \mid Y_m = p, B_m = u) > 0.$$

The aperiodicity of (Y_k) is a direct consequence of the aperiodicity of (Z_k). Let $q = w_1 \in \mathsf{E}$ be a prefix formed by one letter and let d be its period, by the aperiodicity of (Z_k) it is clear that $d = 1$. Therefore (Y_k) is aperiodic. ■

2.3 The hitting time of the word

Let N_w be the number of elements in the sequence of letters before the first hitting position of w, to define the random variable N_w we use the prefix chain and its backward time i.e.,

$$N_w := \min\{k \geq 0 : (Y_k, B_k) = (w, \cdot\,) \in K(\mathsf{E})\}. \tag{2.18}$$

As it has been noted in Section 2.2, an occurrence of w in (Z_k) corresponds to an occurrence of w in (Y_k). The following proposition gives the probability law of N_w.

PROPOSITION 12. (Garcia-Maya and Limnios [2020]). Let $\{W^c, W\}$ be a partition of the state space $K(\mathsf{E})$ such that $W := \{(w, \cdot) \in K(\mathsf{E})\}$ and $W^c := K(\mathsf{E}) \setminus W$. Let $\mathbf{1}$ be a column vector of ones with size $|W^c|$. Let $\check{P}_w = \check{P}|_{W^c \times W^c}$ and β_w be the restrictions respectively, on $W^c \times W^c$ and W^c of the transition matrix $\check{P}$ and the initial distribution β. Then

$$\mathbb{P}(N_w = n) = \begin{cases} 0 & \text{if } n < h-1, \\ \beta_w (\check{P}_w)^{h-1}\mathbf{1} & \text{if } n = h-1, \\ \beta_w \check{P}_w^{n-1}[I - \check{P}_w]\mathbf{1} & \text{if } n \geq h. \end{cases}$$

Proof. For $n < h-1$ it is obvious. For $n = h-1$, let $p_1 = w_1$, $p_2 = w_1w_2$,..., $p_{h-1} = w_1w_1 \cdots w_{h-1}$, $w = w_1w_1 \cdots w_{h-1}w_h$ be the consecutive prefixes from w_1 to w, then

$$\begin{aligned} \mathbb{P}(N_w = h-1) &= \sum_{u_h \in l_w} \sum_{u_{h-1} \in l_{p_{h-1}}} \cdots \sum_{u_2 \in l_{p_2}} \sum_{u_1 \in l_{p_1}} \check{P}((p_{h-1}, u_{h-1}), (w, u_h)) \\ &\cdots \quad \check{P}((p_1, u_1), (p_2, u_2))\mathbb{P}(Y_0 = p_1, B_0 = u_1), \end{aligned}$$

where the set l_p is the set of backward times which corresponds to the prefix p, see Equation (2.7). For $n \geq h$

$$\begin{aligned} \mathbb{P}(N_w > n) &= \mathbb{P}((Y_k, B_k) \in W^c, k \in \{0, 1, ..., n\}) \\ &= \sum_{(q_n,u_n)\in W^c} \sum_{(q_{n-1},u_{n_1})\in W^c} \cdots \sum_{(q_0,u_0)\in W^c} \check{P}((q_{n-1}, u_{n-1}), (q_n, u_n)) \\ &\quad \cdots \check{P}((q_0, u_0), (q_1, u_1))\mathbb{P}(Y_0 = q_0, B_0 = u_0) \\ &= \sum_{(q_n,u_n)\in W^c} \sum_{(q_0,u_0)\in W^c} \check{P}^n((q_0, u_0), (q_n, u_n))\mathbb{P}(Y_0 = q_0, B_0 = u_0), \end{aligned}$$

and therefore

$$\begin{aligned}\mathbb{P}(N_w = n) &= \beta_w(\check{P}_w)^{n-1}\mathbf{1} - \beta_w(\check{P}_w)^{n}\mathbf{1} \\ &= \beta_w(\check{P}_w)^{n-1}[I - \check{P}_w]\mathbf{1}. \ \blacksquare\end{aligned}$$

PROPOSITION 13. (Garcia-Maya and Limnios [2020]). Under the same hypothesis, as in Proposition 3.9, where $f(h-1) = \mathbb{P}(N_w = h-1)$. The generating function of N_w, i.e., $G(s) := \mathbb{E}(s^{N_w})$, for $|s| \leq 1$ is

$$G(s) = s^{h-1}f(h-1)\mathbb{1}_{\{h\geq 1\}} + s\beta_w(s\check{P}_w)^{h-1}(I - s\check{P}_w)^{-1}(I - \check{P}_w)\mathbb{1}_{\{h\geq 1\}}\mathbf{1}.$$

Proof. By definition of the generating function and Proposition 3.9, we write:

$$\begin{aligned}G(s) &= \sum_{k\geq h-1}\mathbb{P}(N_w = k)s^k \\ &= s^{h-1}f(h-1)\mathbb{1}_{\{h\geq 1\}} + \sum_{k\geq h} s^k\beta_w\check{P}_w^{k-1}[I - \check{P}_w]\mathbf{1}. \qquad (2.19)\end{aligned}$$

Due to the fact that W^c is a proper subset of the state space $K(\mathsf{E})$ of an irreducible and aperiodic Markov chain, we write

$$\sum_{k\geq 0}(s\check{P}_w)^k = (I - s\check{P}_w)^{-1} \qquad (2.20)$$

see Neuts [1981b].

Therefore

$$\begin{aligned}\sum_{k\geq h} s^k\beta_w\check{P}_w^{k-1}[I - \check{P}_w]\mathbf{1} &= s^h\beta_w\check{P}_w^{h-1}\sum_{h\geq 0}(s\check{P}_w)^k(I - \check{P}_w)\mathbf{1} \\ &= s\beta_w(s\check{P}_w)^{h-1}(I - s\check{P}_w)^{-1}(I - \check{P}_w)\mathbb{1}_{\{h\geq 1\}}\mathbf{1},\end{aligned}$$

hence

$$G(s) = s^{h-1}f(h-1)\mathbb{1}_{\{h\geq 1\}} + s\beta_w(s\check{P}_w)^{h-1}(I - s\check{P}_w)^{-1}(I - \check{P}_w)\mathbb{1}_{\{h\geq 1\}}\mathbf{1}. \ \blacksquare$$

LEMMA 2. (Garcia-Maya and Limnios [2020]). The n derivate of $(I - s\check{P}_w)$ holds the follow property

$$\frac{d}{ds}(I - s\check{P}_w)^{-n} = n(I - s\check{P}_w)^{-(n+1)}\check{P}_w = n\check{P}_w(I - s\check{P}_w)^{-(n+1)}.$$

PROPOSITION 14. (Garcia-Maya and Limnios [2020]). The mean and variance of N_w are respectively

$$\mathbb{E}(N_w) = (h-1)f(h-1)\mathbb{1}_{\{h\geq 2\}} + \beta_w[Q + h\check{P}_w^{h-2}]\check{P}_w\mathbf{1} \tag{2.21}$$

and for $h \geq 3$

$$\mathbb{V}\text{ar}(N_w) = 2(h-1)f(h-1) + a + b + c + d - (\mathbb{E}(N_w))^2, \tag{2.22}$$

where:

$$\begin{aligned} a &= 2\beta_w\check{P}_w^2Q^2\mathbf{1} \\ b &= 2\beta_w\check{P}_wQ\mathbf{1} \\ c &= h\beta_w\check{P}_w^hQ\mathbf{1} \\ d &= h^2\check{P}_w^{h-1}\mathbf{1} \\ Q &= (I-\check{P}_w)^{-1}. \end{aligned}$$

Proof. Using the generating function of N_w, $G(s)$, and Lemma 2, we get

$$\begin{aligned} \frac{dG(s)}{ds} &= (h-1)s^{h-2}f(h-1)\mathbb{1}_{\{h\geq 2\}} \\ &\quad +\beta_w\left[(I-s\check{P}_w)^{-1} + h(s\check{P}_w)^{h-2}\right](s\check{P})(I-s\check{P}_w)^{-1}(I-\check{P}_w)\mathbf{1}, \end{aligned} \tag{2.23}$$

which yields

$$\frac{dG(s)}{ds}\Big|_{s=1} = \mathbb{E}(N_w) = (h-1)f(h-1)\mathbb{1}_{\{h\geq 2\}} + \beta_w\left[(I-\check{P}_w)^{-1} + h(\check{P}_w)^{h-2}\right]\check{P}\mathbf{1}.$$

For the variance of N_w, the derivate of Equation (2.23) will be computed as follows:

$$\frac{d}{ds}(h-1)s^{h-2}f(h-1) = (h-1)(h-2)s^{h-3}f(h-1)\mathbb{1}_{\{h\geq 3\}}, \tag{2.24}$$

$$\frac{d}{ds}[(I-s\check{P}_w)^{-1} + h(s\check{P}_w)^{h-2}] = (I-s\check{P}_w)^{-2}\check{P}_w + h(h-2)(s\check{P}_w)^{h-3}\check{P}_w\mathbb{1}_{\{h\geq 3\}}, \tag{2.25}$$

and

$$\frac{d}{ds}(s\check{P}_w)(I-s\check{P}_w)^{-1} = (s\check{P}_w)(I-s\check{P}_w)^{-2}\check{P}_w + \check{P}_w(I-s\check{P}_w)^{-1}. \tag{2.26}$$

Then, using Equations (2.25) and (2.26) and simplifying terms, it yields for $h \geq 3$

$$\begin{aligned}\frac{d}{ds}[(I-s\check{P}_w)^{-1}+h(s\check{P}_w)^{h-2}](s\check{P}_w)(I-s\check{P}_w)^{-1} &= \big\{2(s\check{P}_w)(I-s\check{P}_w)^{-2} \\ &+[1+h(s\check{P}_w)^{h-1}](I-s\check{P}_w)^{-1} \\ &+(h-1)h(s\check{P}_w)^{h-2}\big\}(I-s\check{P}_w)^{-1}\check{P}_w.\end{aligned} \tag{2.27}$$

Therefore, using Equations (2.24) and (2.27) for $h \geq 3$:

$$\begin{aligned}\frac{d^2G(s)}{ds^2} &= (h-1)(h-2)s^{h-3}f(h-1)+\beta_w\big\{2(s\check{P}_w)(I-s\check{P}_w)^{-2} \\ &+[1+h(s\check{P}_w)^{h-1}](I-s\check{P}_w)^{-1} \\ &+(h-1)h(s\check{P})^{h-2}\big\}(I-s\check{P}_w)^{-1}\check{P}(I-\check{P})\mathbf{1}.\end{aligned} \tag{2.28}$$

Therefore for $h \geq 3$,

$$\frac{d^2G(s)}{ds^2}\Big|_{s=1}=(h-1)(h-2)f(h-1)+\beta_w\big\{2\check{P}_w^2(I-\check{P}_w)^{-2}+[\check{P}_w+h\check{P}_w^h](I-\check{P}_w)^{-1}+(h-1)h\check{P}_w^{h-1}\big\}\mathbf{1}.$$

Using the expression

$$\mathbb{V}\mathrm{ar}(N_w)=\frac{d^2G(s)}{ds^2}\Big|_{s=1}+\frac{dG(s)}{ds}\Big|_{s=1}-\left(\frac{dG(s)}{ds}\Big|_{s=1}\right)^2,$$

we get for $h \geq 3$,

$$\mathbb{V}\mathrm{ar}(N_w)=(h-1)^2f(h-1)+a+b+c+d-(\mathbb{E}(N_w))^2,$$

where:

$$\begin{aligned}a &= 2\beta_w\check{P}_w^2Q^2\mathbf{1},\\ b &= 2\beta_w\check{P}_wQ\mathbf{1},\\ c &= h\beta_w\check{P}_w^hQ\mathbf{1},\\ d &= h^2\check{P}_w^{h-1}\mathbf{1},\\ Q &= (I-\check{P}_w)^{-1}.\end{aligned}$$

2.4 A genomic application

The mathematical model proposed in this theses can be implemented in any irreducible semi-Markov chain with finite state space, to show one of its applications, in this section we present a genomic example.
Let us consider the DNA sequence of bacteriophage. This DNA includes 48502 nucleotides. In this case the genomic alphabet is $\mathcal{A}=\{A,T,G,C\}$ and consider the pattern $w=CCCGGG$ which is the enzyme *SmaI*. The *SmaI* enzyme is a DNA cutter. That is, when it finds the '$CCCGGG$' fragment it applies and cuts the DNA

in the middle of this fragment, i.e., into '...CCC' and 'GGG...'.
The semi-Markov kernel is estimated according with Equation (1.40). The probability that the *SmaI* enzyme appears after k nucleotides is observed in figure 2.3. The random variable N_w counts the numbers of nucleotides before the apparition of the enzyme and it is defined according with Equation (2.18). The probability function of N_w is denoted $f_w(k) := \mathbb{P}(N_w = k)$, this probability is computed using Preposition 3.9.

In figure 2.4 we can observe that the *SmaI* enzyme does not appears frequently in the DNA. The word occurrence rate for $k \geq 1$ is given by the rate function

$$\check{\lambda}(k) = \begin{cases} 1 - \frac{\bar{F}_w(k)}{\bar{F}_w(k-1)}, & \bar{F}_w(k-1) \neq 0 \\ 0, & \text{otherwise,} \end{cases}$$

where $\bar{F}_w(k) = 1 - F_w(k)$ and $F_w(k) := \sum_{l=0}^{k} f_w(l)$. Figure 2.5 gives the values for the distribution function $F_w(k)$. In the continuous time, the rate function takes values also greater than one, in the discrete-time case it takes values only in the interval $[0, 1]$. For this reason, in an other context Roy and Gupta [1992] proposed another rate function as follows:

$$r(k) = -\ln(1 - \check{\lambda}(k)). \tag{2.29}$$

Nevertheless, when $\check{\lambda}(k)$ is close to 0, we have obviously

$$r(k) \cong \check{\lambda}(k).$$

In the case of the present example, the values $\check{\lambda}(k)$ are very small, so, Figure 2.4 represents also the function $r(k)$.

Considering that DNA sequence has 48502 nucleotides, i.e., it has size $M = 48502$ and using the expressions:

$$\mathbb{E}(N_w \cdot \mathbb{1}_{\{N_w \leq M\}}) = \sum_{k=h-1}^{M} k f_w(k),$$

$$\mathbb{E}(T_w^2 \cdot \mathbb{1}_{\{N_w \leq M\}}) = \sum_{k=h-1}^{M} k^2 f_w(k),$$

where $f_w(k) := \mathbb{P}(N_w = k)$ is obtaining according with Proposition 3.9, the expected value $\mathbb{E}(N_w \cdot \mathbb{1}_{\{N_w \leq M\}})$ and the standard deviation $\sigma(N_w \cdot \mathbb{1}_{\{N_w \leq M\}})$ are computed. Hence, we obtain :

$$\begin{aligned} \mathbb{E}(N_w \cdot \mathbb{1}_{\{N_w \leq M\}}) &= 6335.9, \\ \sigma(N_w \cdot \mathbb{1}_{\{N_w \leq M\}}) &= 6266.7. \end{aligned}$$

Considering different lengths for the DNA sequence, i.e., considering different values for M, the mean value of N_w for each M are observed in Figure 2.6. If we take into account that DNA sequence has infinity length, the mean value of N_w is computed according with Equation (2.21). The variance is computed using Equation (2.22), therefore for the bacteriophage DNA sequence we have the values:

$$\begin{aligned} \mathbb{E}(N_w) &= 6367.6, \\ \sigma(N_w) &= 6354.9. \end{aligned}$$

Notice in Figure 2.6, the value $\mathbb{E}(N_w \cdot \mathbb{1}_{\{N_w \leq M\}})$ reach $\mathbb{E}(N_w)$ as M becomes large enough, as it was expected by the dominated converge theorem. It is worth noticing that the standard deviation here is high. This is due to the fact that the evolution of the rate of occurrence of the word is small. After position 9, it becomes geometric, as we can see in Figure 2.4. According with Geometric distribution with success probability p, the variance is given by formula

$$Var(X) = \frac{1-p}{p^2}.$$

Observing that variance grows if p decrease. In our model the rate of occurrence of the word is tiny, we can see in same Figure 2.4 that after position 9, we have $p = \check{\lambda}(9) = 1.6 \times 10^{-4}$ which means that the probability to have a success is small, this gives the big value for the standard deviation.

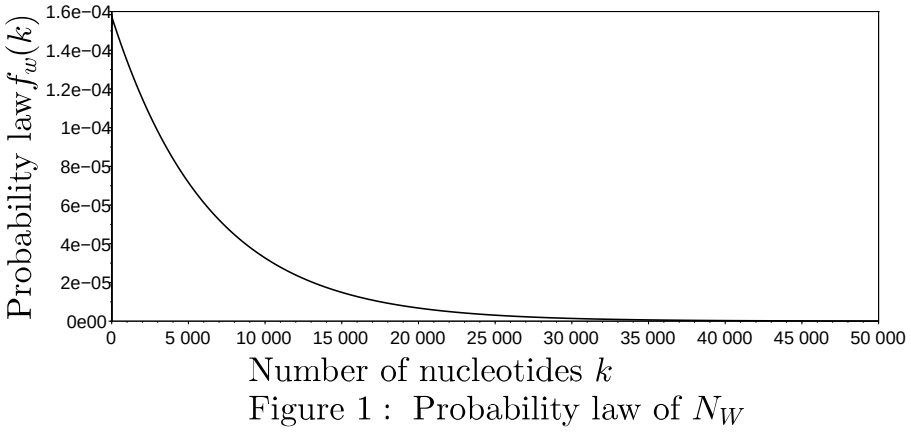


FIGURE 2.3: Probability law of N_w, i.e., $f_w(k) := \mathbb{P}(N_w = k)$

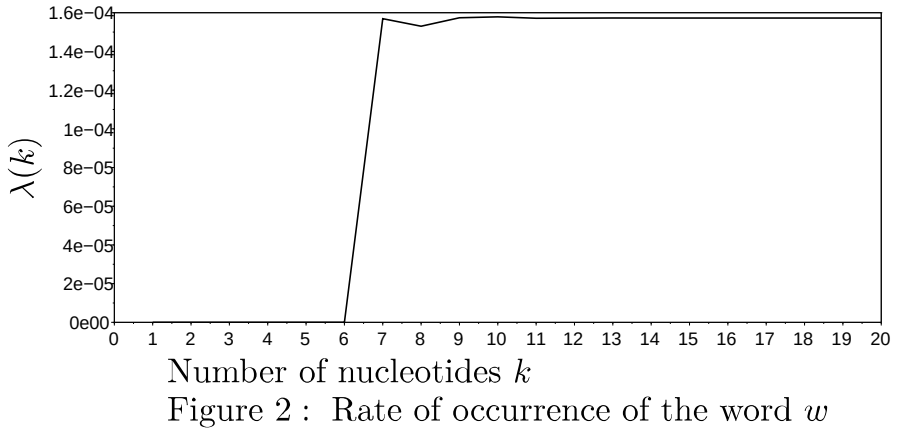


FIGURE 2.4: Rate of occurrence of the word w in the DNA sequence

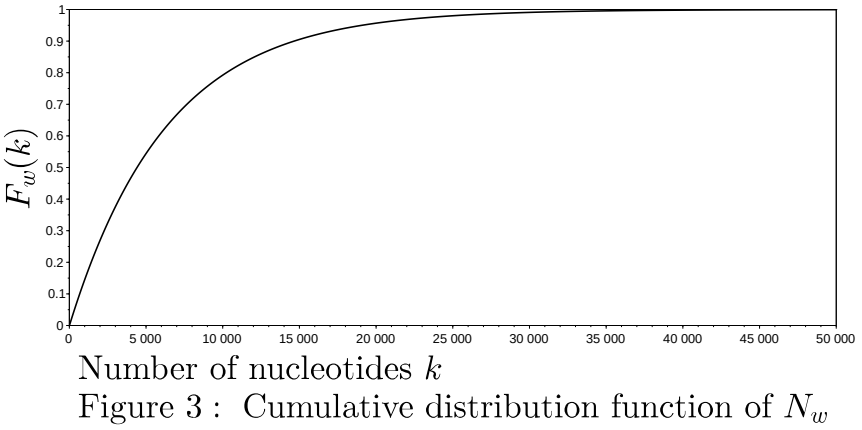


FIGURE 2.5: Cumulative distribution function of N_w, i.e., $F_w(k) := \mathbb{P}(N_w \leq k)$

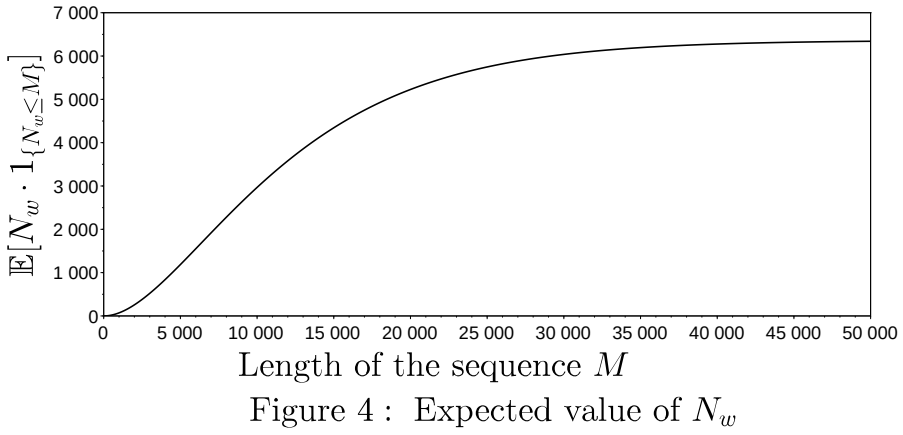


Figure 4 : Expected value of N_w

FIGURE 2.6: Expected value of N_w for different values of M in the DNA

2.5 Concluding remarks

In this chapter we proposed a new model and algorithm that can be implemented in real applications to compute the first hitting position (time) of a word (pattern) in a semi-Markov sequence. Although Chryssaphinou et al. [2008] proposed the only theoretical model before this work, for words occurrence in discrete time semi-Markov chains, the Chryssaphinou method's cannot be implemented because the cardinality of the state space is huge. It has $|\mathcal{A}|^h \times (M-h)$ elements, where M represents the length of the semi-Markov chain (Z_k), $|\mathcal{A}|$ is the alphabet cardinality and h is the length of the pattern. Notice that, if the values of h and/or M grow, the cardinality of the state space becomes enormous, this makes difficult the implementation. By contrast, the model proposed here needs less memory space than the one proposed by Chryssaphinou et al. [2008]. This model is based on prefixes and its extended state space, with these assumptions process (Y_k, B_k) becomes a Markov chain where its transition matrix $\check{P}$ is easily written. For a word w with length $|w| = h$ and maximum value for the backward time of a prefix: $\gamma = \max\{l_p^* : p \in \mathsf{E}\}$, we need a state space of cardinality $(h + |\mathcal{A}| - 1) \times \gamma$ at most.

Moreover, we proposed results for the (first) position time to w, its law, its mean, its variance and its generating function. As we can see in Figure 2.6, the value $\mathbb{E}(N_w \cdot \mathbb{1}_{\{N_w \leq M\}})$ reach $\mathbb{E}(N_w)$, as M becomes large enough, as it was expected by the dominated converge theorem.

It is worth noticing that for words of length one, i.e., a letter, our algorithm recovers the Markov process (Z_k, B_k). Of course, this work (algorithm) could be used for any irreducible semi-Markov sequence with finite state space and any finite word (pattern).

Chapter 3

Asymptotic properties of words in Markov and semi-Markov sequences

Biomolecules [1] have a wide range of sizes and structures and perform a big number of functions. They play a crucial role in bio-informatics and modern biology. Similar to the way the order of letters in an alphabet is used to form a word, the order of nitrogen bases in a DNA sequence forms biomolecules, which in the language of biology, tells cells to do a specific function. In most cases biomolecules have more than one form to be encrypted. For instance, restrictions enzymes that cleaves DNA into fragments are recognized by more than one particular pattern. For instance, the *EcoRI* enzyme is recognized by the sequences $GAATTC$ and $CTTAAG$; the *BamHI* enzyme is recognized by the sequences $GGATCC$ and $CCTAGG$, etc.

In this chapter we compute the average number of times that a biomolecule appears through the DNA by any of its configurations, in other words, we compute the average number of times that the elements from a specific set of words $\mathcal{W}$ appear through a sequence of letters (Z_k). Where the sequence of letters (Z_k) represents the DNA sequence and the biomolecule is represented by the set of words $\mathcal{W}$. To achieve our goal we use the strong law of large numbers. We also provide the central limit theorem for a sequence of patterns. Additionally, we treat the problem of finding a specific biomolecule by any of its configurations. We also compute which of those configurations is more probable to occur at first. In other words, we identify the first hitting time in which the elements from a specific set of words $\mathcal{W}$ appear through a sequence of letters (Z_k). To resolve the problem, we consider two cases: DNA is modeled by an ergodic Markov sequence and DNA is modeled by a semi-Markov

[1] This chapter develops the content of an article submitted in 2020 put in shape to be inserted in this thesis.
Garcia-Maya, B.I., Karaliopoulou, M., and Limnios, N. (2020) Asymptotic properties of words in Markov and semi-Markov sequences.

chain. Even if Markov sequences model properly sequences of letters, see e.g., Nur et al. [2009], Wheeler et al. [2012], Jääskinen et al. [2014], etc. The semi-Markov case is more general, see e.g., Barbu and Limnios [2008], D'Amico et al. [2013], Janssen [2013]. It considers general probabilities laws on $\mathbb{N}$ instead of the geometric only law in the Markov case.

3.1 The prefix chain of a set of words

In Chapter 2 we presented the prefixes of a single word, in this chapter we shall introduce the prefixes of a set of words. First we shall consider that the sequence of letters is modeled by a Markov chain $(\mathbb{X}_k)$ after this, we shall give the analogous results for a semi-Markov chain (Z_k).

Let us consider a particular set of words taken from $\mathcal{A}^h$, we shall denote this set by $\mathcal{W}$. As we can observe in Section 2.1, the set of prefixes for a single word is denoted by E_w, see Equation (2.1), therefore the prefixes of a set of words $\mathcal{W} \subset \mathcal{A}^h$ denoted by $\tilde{E}^*$ is the union $\tilde{E}^* := \bigcup_{w\in\mathcal{W}} E_w$.

Let $\delta_{\tilde{E}^*} : \tilde{E}^* \times \mathcal{A} \to \tilde{E}^*$ be a function analogously defined as Equation (2.2), i.e., it is the longest suffix of $qa \in \tilde{E}^*$ (concatenation of $q \in \tilde{E}^*$ and $a \in \mathcal{A}$) in the prefix set $\tilde{E}^*$. Observe the following examples for $\delta_{\tilde{E}^*}$

EXAMPLE 5. If $\mathcal{A} = \{a, b, c\}$, $\mathcal{W} = \{ab, aa\}$ the prefix set is $\tilde{E}^* = \{\varepsilon, a, ab, aa\}$, then

$$\begin{array}{lll} \delta_{\tilde{E}^*}(\varepsilon, a) = a & \delta_{\tilde{E}^*}(\varepsilon, b) = \varepsilon & \delta_{\tilde{E}^*}(\varepsilon, c) = \varepsilon \\ \delta_{\tilde{E}^*}(a, a) = aa & \delta_{\tilde{E}^*}(a, b) = ab & \delta_{\tilde{E}^*}(a, c) = \varepsilon \\ \delta_{\tilde{E}^*}(ab, a) = a & \delta_{\tilde{E}^*}(ab, b) = \varepsilon & \delta_{\tilde{E}^*}(ab, c) = \varepsilon \\ \delta_{\tilde{E}^*}(aa, a) = aa & \delta_{\tilde{E}^*}(aa, b) = ab & \delta_{\tilde{E}^*}(aa, c) = \varepsilon \end{array}$$

Observe that, for $p \in \tilde{E}^*$ the set $\{i \in \mathcal{A} : \delta_{\tilde{E}^*}(p, i) = \varepsilon\}$ has more than one element. For the same reasons explained in Section 2.1 the prefix ε will be tagged according to the letter which is concatenated to p to results in ε. After redefining the prefix ε we shall introduce the set $\mathscr{F}$. Let us consider a set $\mathscr{F}$ which contains the letters in $\mathcal{A}$ that are different from the first letter of any word $w = w_1 w_2 \cdots w_h \in \mathcal{W}$, i.e.,

$$\mathscr{F} := \{i \in \mathcal{A} : i \neq w_1, \text{ for all } w = w_1 w_2 \cdots w_h \in \mathcal{W}\}. \tag{3.1}$$

For $p \in \tilde{E}^*$ and $i \in \mathcal{A}$ instead of having

$$\delta_{\tilde{E}^*}(p, i) = \varepsilon,$$

we will have

$$\delta_{\tilde{E}^*}(p, i) = \varepsilon_i,$$

where $i \in \mathscr{F}$.
Therefore,

$$E^* := (\cup_{w\in\mathcal{W}} E_w) \bigcup (\cup_{i\in\mathscr{F}} \varepsilon_i) \tag{3.2}$$

is the prefix set of $\mathcal{W}$ in which $\delta_{E^*} : E^* \times \mathcal{A} \to E^*$ is a one to one mapping. Observe that the partial inverse of δ_{E^*}, defined in E^*, i.e., $\delta_{E^*}^{-1} : E^* \to \mathcal{A}$, see Equation(2.4), it is one-to-one. We can notice that $\mathcal{W} \subset \mathcal{A}^h$ and $\mathcal{W} \subset E^*$.
From now we shall denote by ℓ the cardinality of E^*, i.e., $\ell := |E^*|$. Next definition presents the generalization of Definition 15. It introduces the prefix process of a set of words $\mathcal{W}$.

DEFINITION 17. The prefix chain of $\mathcal{W}$ embedded in the Markov chain $(\mathbb{X}_k)$ and defined in E^*, see Equation (3.2), is denoted by $Y^* := (Y_k^*)_{k\in\mathbb{N}}$ where

$$Y_0^* := \begin{cases} w_1 & \text{if } \mathbb{X}_0 = w_1, \text{for some } w = w_1 w_2, \cdots w_h \in \mathcal{W}, \\ \varepsilon_i & \text{if } \mathbb{X}_0 = i \text{ with } i \in \mathscr{F}, \end{cases}$$

and

$$Y_k^* := \delta_{E^*}(Y_{k-1}^*, \mathbb{X}_k), \quad k \geq 1.$$

where $\mathscr{F}$ is defined in Equation (3.1).

PROPOSITION 15. (Nuel [2008]). If the sequence of letters is modeled by a Markov chain $(\mathbb{X}_k)$, the prefix process (Y_k^*) defined in E^* is a Markov chain too with initial distribution

$$\alpha^*(p) = \mathbb{P}(Y_0^* = p) = \begin{cases} \alpha^*(w_1) & \text{if } p = w_1 \text{ for some } w = w_1 w_2 \cdots w_h \in \mathcal{W}, \\ \alpha^*(i) & \text{if } p = \varepsilon_i, \ i \in \mathscr{F}, \\ 0 & \text{otherwise.} \end{cases}$$

and transition probability matrix

$$\tilde{P}(p,q) := \mathbb{P}(Y_{k+1}^* = q \mid Y_k^* = p), \ \ q, p \in E^*;$$

where

$$\tilde{P}(p,q) = \begin{cases} \mathbb{P}(\mathbb{X}_{k+1} = a \mid \mathbb{X}_k = \delta^{-1}(p)) & \text{if } \delta_{E^*}(p,a) = q \\ 0 & \text{elsewhere.} \end{cases} \tag{3.3}$$

Proof. The initial distribution comes from the definition of Y_0^*, see Definition 17. For the transition probability matrix. Consider a prefix $p \in E^*$ and suppose that there exists $a \in \mathcal{A}$ such that $\delta_{E^*}(p,a) = q$, for some $q \in E^*$. If $p = w_1 w_2 \cdots w_{l-1} w_l$ for $1 \leq l \leq h$, then

$$\begin{aligned} \mathbb{P}(Y_{k+1}^* = q, \mid Y_k^* = p,) &= \mathbb{P}(\mathbb{X}_{k+1} = a \mid \mathbb{X}_k = w_l, Z_{k-1} = w_{l-1}, ..., \mathbb{X}_{k-l+1} = w_1), \\ &= \mathbb{P}(\mathbb{X}_{k+1} = a \mid \mathbb{X}_k = w_l). \end{aligned} \tag{3.4}$$

If $p = \varepsilon_j$ for $j \in \mathscr{F}$, then

$$\begin{aligned} \mathbb{P}(Y_{k+1}^* = q \mid Y_k^* = p) &= \mathbb{P}(\mathbb{X}_{k+1} = a, \mid \mathbb{X}_k = j, \mathbb{X}_{k-1} = \cdot, ..., \mathbb{X}_{k-l+1} = \cdot), \\ &= \mathbb{P}(\mathbb{X}_{k+1} = a, \mid \mathbb{X}_k = j). \end{aligned} \tag{3.5}$$

To denote in general the transition probability, let $i = \delta_{E^*}^{-1}(p)$ be the partial inverse of p. Therefore in Equations (3.4) and (3.5) $\mathbb{X}_k = i$, thus

$$\mathbb{P}(Y_{k+1}^* = q, \mid Y_k^* = p) = \mathbb{P}(\mathbb{X}_{k+1} = a \mid \mathbb{X}_k = i),$$

which proves the proposition. ■
Next Proposition proves that if the sequences of letters $(\mathbb{X}_k)$ is an irreducible and aperiodic Markov chain, then prefix chain (Y_k^*) has also the same properties.

PROPOSITION 16. If process of letters $(\mathbb{X}_k)$ is modeled by an irreducible and aperiodic Markov chain then the prefix chain (Y_k^*) described in Definition 17 has the same properties.

Proof. Let be $p, q \in E^*$ and $i = \delta_{E^*}^{-1}(p)$. Suppose $Y_m^* = p$, $m \in \mathbb{N}$. If $q = w_1$ or $q = \varepsilon_a$, $a \in \mathcal{F}$ the proof is a direct consequence of the irreducibility of $(\mathbb{X}_k)$. If $q = w_1 w_2 \cdots w_l$, with $1 < l \leq h$. Let $q_1 = w_1$, $q_2 = w_1 w_2, ..., q = w_1 w_2 \cdots w_l$, be the consecutive prefixes from w_1^j to $q = w_1 w_2 \cdots w_l$. Due to the irreducibility of $(\mathbb{X}_k)$, for $w_1 \in \mathcal{A}$ there exist $n \in \mathbb{N}^*$ such that

$$\mathbb{P}(\mathbb{X}_{m+n} = w_1 \mid \mathbb{X}_m = i) > 0.$$

Therefore

$$\mathbb{P}(Y_{m+1}^* = q_1 \mid Y_m^* = p)\mathbb{P}(Y_{m+2}^* = q_2 \mid Y_{m+1}^* = q_1) \cdots \mathbb{P}(Y_{m+l}^* = q_l \mid Y_{m+l-1}^* = q_{l-1}) > 0$$

Hence there exist $n_1 = n + \ell$ such that

$$\mathbb{P}(Y_{m+n_1}^* = q \mid Y_m^* = p) > 0.$$

The aperiodicity of (Y_k^*) is a direct consequence of the aperiodicity of $(\mathbb{X}_k)$. Let $q = w_1$, $w \in \mathcal{W}$ be a prefix formed only by one letter and let d be its period, by the aperiodicity of $(\mathbb{X}_k)$ it is clear that $d = 1$. Therefore (Y_k^*) is aperiodic. ■

By Proposition 16 we have shown that if the sequence of letters is an ergodic Markov chain then the sequence of prefixes has the same properties. Next Proposition provides the stationary distribution of the prefix process (Y_k^*). We can notice that if $\mathbb{P}$ and $\check{\pi}$ are the transition probability matrix and the stationary distribution (respectively) of the process of letters $(\mathbb{X}_k)$, i.e.,

$$\mathbb{P}^n(i, j) \to \check{\pi}(j), \quad i, j \in \mathcal{A}, \quad n \to \infty \tag{3.6}$$

then, the stationary distribution of the prefix process is a function of the stationary distribution of the sequence of letters.

PROPOSITION 17. The stationary distribution of prefix process (Y_k^*) denoted by $\tilde{\pi}$ is a function of the stationary distribution of the sequence of letters $(\mathbb{X}_k)$ where

$$\tilde{\pi}(p) = \begin{cases} \check{\pi}(j) & \text{if } j = \delta^{-1}(p) \text{ and } |p| = 0 \text{ or } |p| = 1, \\ \check{\pi}(w_1)\mathbb{P}(w_1, w_2) \times \cdots \times \mathbb{P}(w_{l-1}, w_l) & \text{if } p = w_1 w_2 \cdots w_l \text{ for } 1 < l \leq h. \end{cases} \tag{3.7}$$

Proof: If p is a prefix formed by only one letter or it has not letters, i.e., if $p = j$, $j \in \mathcal{A} \setminus \mathscr{F}$ or $p = \varepsilon_j$, $j \in \mathscr{F}$ then, the stationary distribution of p is a direct consequence of the stationary distribution of process $(\mathbb{X}_k)$. In the other hand, if p is formed by more than one letter, i.e., $p = w_1 w_2 \cdots w_l$ for $1 < l \leq h$ and $p_1 = w_1$, $p_2 = w_1 w_2$, ..., $p = w_1 w_2 \cdots w_l$ are the consecutive prefixes from $p_1 = w_1$ to $p = w_1 w_2 \cdots w_l$, then it is clear that the stationary distribution of prefix p is

$$\begin{aligned} \tilde{\pi}(p) &= \check{\pi}(p_1)\tilde{P}(p_1, p_2) \times \cdots \times \tilde{P}(p_{l-1}, p_l) \\ &= \check{\pi}(w_1)\mathbb{P}(w_1, w_2) \times \cdots \times \mathbb{P}(w_{l-1}, w_l). \ \blacksquare \end{aligned}$$

The number of times that the elements from $\mathcal{W}$ are repeated in the sequence $(\mathbb{X}_k)$ correspond to the average number of times that the elements from $\mathcal{W}$ are repeated in prefix process (Y_k^*) see Nuel [2008]. This average number is a function of the transition probability of process $(\mathbb{X}_k)$ as we can observe in the following proposition.

PROPOSITION 18. For the ergodic Markov chain of prefixes (Y_k^*) with stationary distribution $\tilde{\pi}$ (see Equation 3.7), the frequency of words from the set $\mathcal{W}$ is

$$\frac{1}{n}\sum_{k=0}^{n} \mathbb{1}_{\{Y_k^* \in \mathcal{W}\}} \xrightarrow{a.s} \sum_{w \in \mathcal{W}} \check{\pi}(w_1)\mathbb{P}(w_1, w_2) \times \cdots \times \mathbb{P}(w_{l-1}, w_l), \ \ n \longrightarrow \infty.$$

where $w = w_1 w_2 \cdots w_h \in \mathcal{A}^h$.

Proof: The proof is a direct consequence of the strong law of large numbers for an ergodic Markov chain, see e.g., Barbu and Limnios [2008]. ■

The central limit theorem (CLT), for an ergodic Markov chain, establishes the average of the sum of its terms tends toward a normal distribution even if the original variables are not normally distributed, see e.g., Limnios and Oprişan [2001]. The goal of the following Proposition is to describe the central limit theorem for the prefix process.

PROPOSITION 19. For an ergodic Markov chain (Y_k^*) with stationary distribution $\tilde{\pi}$, we have

$$\sqrt{n}\left(\frac{1}{n}\sum_{k=0}^{n} \mathbb{1}_{\{Y_k^* \in \mathcal{W}\}} - S\right) \xrightarrow{d} \mathcal{N}(0, \sigma_{\mathcal{W}}^2), \ \ n \to \infty$$

where

$$\begin{aligned} \sigma_{\mathcal{W}}^2 \ := \ & \sum_{p_i \in \mathcal{W}} \left(\tilde{\pi}(p_1)[\delta_{p_1 p_i} - \tilde{\pi}(p_i)], ..., \tilde{\pi}(p_{|E^*|})[\delta_{p_{|E^*|} p_i} - \tilde{\pi}(p_i)]\right) \\ & \times (2\check{Z} - I)\left(\tilde{S}(p_1), ..., \tilde{S}(p_{|E^*|})\right)^T. \end{aligned}$$

such that $S := \sum_{p\in\mathcal{W}} \tilde{\pi}(p)$, $\tilde{S}(p) := \mathbb{1}_{\{p\in\mathcal{W}\}} - S$, for $p \in E^*$ and $\check{Z} := (I - \tilde{P} + \Pi)^{-1}$ is the fundamental matrix of $\tilde{P}$ and Π is defined as follows

$$\Pi := \begin{pmatrix} \tilde{\pi}(p_1) & \tilde{\pi}(p_2) & \cdots & \tilde{\pi}(p_{|E|}) \\ \tilde{\pi}(p_1) & \tilde{\pi}(p_2) & \cdots & \tilde{\pi}(p_{|E|}) \\ \vdots & & & \\ \tilde{\pi}(p_1) & \tilde{\pi}(p_2) & \cdots & \tilde{\pi}(p_{|E|}) \end{pmatrix}.$$

Proof:
For an ergodic Markov chain (Y_k^*) with stationary distribution $\tilde{\pi}$ by the CLT, we have

$$\sqrt{n}\left(\frac{1}{n}\sum_{k=0}^{n} \mathbb{1}_{\{Y_k^*\in\mathcal{W}\}} - S\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\mathcal{W}}), \quad n \to \infty$$

where the variance is expressed by $\sigma^2_{\mathcal{W}} = \tilde{S} diag(\tilde{\pi})[2\check{Z} - I]\tilde{S}^T$, see Trevezas and Limnios [2009]. The matrix $\check{Z}$ is the fundamental matrix of process (Y_k^*) and it is defined by $\check{Z} := (I - \tilde{P} + \Pi)^{-1}$, such that I is the identity matrix of size $|E^*| \times |E^*|$, $\tilde{P}$ is the transition matrix of process (Y_k^*). Therefore

$$\begin{aligned}
\sigma^2_{\mathcal{W}} &:= \left(\tilde{S}(p_1), ..., \tilde{S}(p_{|E^*|})\right) \begin{pmatrix} \tilde{\pi}(p_1) & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & \tilde{\pi}(p_{|E^*|}) \end{pmatrix} (2\check{Z} - I) \begin{pmatrix} \tilde{S}(p_1) \\ \vdots \\ \tilde{S}(p_{|E^*|}) \end{pmatrix} \\
&= \left(\tilde{S}(p_1)\tilde{\pi}(p_1), ..., \tilde{S}(p_{|E^*|})\tilde{\pi}(p_{|E^*|})\right) (2\check{Z} - I) \begin{pmatrix} \tilde{S}(p_1) \\ \vdots \\ \tilde{S}(p_{|E^*|}) \end{pmatrix} \\
&= \left(\mathbb{1}_{\{p_1\in\mathcal{W}\}}\tilde{\pi}(p_1) - \sum_{p_i\in\mathcal{W}} \tilde{\pi}(p_i)\tilde{\pi}(p_1), ..., \mathbb{1}_{\{p_{|E^*|}\in\mathcal{W}\}}\tilde{\pi}(p_{|E^*|}) - \sum_{p_i\in\mathcal{W}} \tilde{\pi}(p_i)\tilde{\pi}(p_{|E^*|})\right) \\
&\quad (2\check{Z} - I) \begin{pmatrix} \tilde{S}(p_1) \\ \vdots \\ \tilde{S}(p_{|E^*|}) \end{pmatrix} \\
&= \left(\sum_{p_i\in\mathcal{W}} \tilde{\pi}(p_1)[\delta_{p_1p_i} - \tilde{\pi}(p_i)], ..., \sum_{p_i\in\mathcal{W}} \tilde{\pi}(p_{|E^*|})[\delta_{p_{|E^*|}p_i} - \tilde{\pi}(p_i)]\right) \\
&\quad (2\check{Z} - I) \begin{pmatrix} \tilde{S}(p_1) \\ \vdots \\ \tilde{S}(p_{|E^*|}) \end{pmatrix} \\
&= \sum_{p_i\in\mathcal{W}} \left(\tilde{\pi}(p_1)[\delta_{p_1p_i} - \tilde{\pi}(p_i)], ..., \tilde{\pi}(p_{|E^*|})[\delta_{p_{|E^*|}p_i} - \tilde{\pi}(p_i)]\right) \\
&\quad (2\check{Z} - I) \begin{pmatrix} \tilde{S}(p_1) \\ \vdots \\ \tilde{S}(p_{|E^*|}) \end{pmatrix}. \blacksquare
\end{aligned}$$

Additionally in this chapter we compute the first hitting position in which an element from $\mathcal{W}$ appears in the sequence $(\mathbb{X}_k)$. We consider the words $w \in \mathcal{W}$ like absorbing states. To achieve this goal we make a partition of the state space E^*, such that $E^* = E_0^* \cup E_1^*$ where E_1^* are the prefixes which belong to the set $\mathcal{W}$ and E_0^* are the elements of E^* which are different from the set $\mathcal{W}$, i.e., $E_0^* = E^* \setminus \mathcal{W}$ and $E_1^* = \mathcal{W}$, it is clear that $E_0^* \cap E_1^* = \emptyset$. We also decompose the initial distribution α^*, see Proposition 15, where α_0^* and α_1^* are the restrictions of the initial distributions in states E_0^* and E_1^* and receptively. Let $(\tilde{Y}_k)$ be a prefix process defined as the prefix process (Y_k^*), see Definition 17, but with transition probability matrix

$$\tilde{R} = \begin{pmatrix} \tilde{P}_{00} & \tilde{P}_{01} \\ 0 & I \end{pmatrix}, \tag{3.8}$$

where $\tilde{P}_{ij}$ is the restiction of $\tilde{P}$ in states $E_i^* \times E_j^*$ for $i, j \in \{0, 1\}$. Observe that 0 is the zero matrix of size $E_1^* \times E_0^*$ and I is the identity matrix of size $E_1^* \times E_1^*$.

The time that process $(\tilde{Y}_k)$ has to wait until an element from $\mathcal{W}$ arrives is a random variable and it is defined as follows

$$T := \inf\{k \geq 0 : \tilde{Y}_k \in \mathcal{W}\}. \tag{3.9}$$

We are interested in computing the distribution function of T and also the probability that $(\tilde{Y}_k)$ reaches $\mathcal{W}$ by a specific element $w^j \in \mathcal{W}$. Therefore, let us denote the random variable

$$W := \{w^j \in \mathcal{W} : \tilde{Y}_T = w^j\}. \tag{3.10}$$

The random variable W takes the value w^j if w^j is the first element from $\mathcal{W}$ that appears in the prefixes chain $(\tilde{Y}_k)$. For the bi-dimensional random variable (T, W) we define its distribution function as

$$G_j(k) := \mathbb{P}(T \leq k, W = w^j)$$

and its law by

$$g_j(k) := \mathbb{P}(T = k, W = w^j). \tag{3.11}$$

Last Equation represents the probability that process $(\tilde{Y}_k)$ reaches the set $\mathcal{W}$ at time k by the element $w^j \in \mathcal{W}$. It is easy to observe that for every $w^j \in \mathcal{W}$, Equation (3.11) can be written as

$$g_j(k) = \mathbb{P}(\tilde{Y}_k = w^j, \tilde{Y}_l \in E_0; l = 0, ..., k-1).$$

The following result expresses the probability law and the distribution function of (T, W).

PROPOSITION 20. The law and the distribution function of (T, W) for the above Markov chain $(\tilde{Y}_k)$ are

$$g_j(k) := \mathbb{P}(T = k, W = w^j) = \begin{cases} 0, & k < h-1; \\ \tilde{\alpha}(w_1^j)\tilde{P}(p_1^j, p_2^j)\cdots\tilde{P}(p_{h-1}^j, p_h^j), & k = h-1; \\ \tilde{\alpha}_0\tilde{P}_{00}^{k-1}\tilde{P}_{01}e_j, & k > h-1. \end{cases}$$

and

$$G_j(k) := \mathbb{P}(T \leq k, W = w^j) = \tilde{\alpha}_0(I - \tilde{P}_{00})^{-1}(I - \tilde{P}_{00}^k)\tilde{P}_{01}e_j, \tag{3.12}$$

respectively where e_j is a column vector of size $|E_0^*|$ where all its entries are zeros, except the entry which corresponds to w^j which takes the value one.

Proof. For $k < h-1$ is obvious. For $k = h-1$ suppose $w^j = w_1^j w_2^j \cdots w_h^j$ with $w_1^j, w_2^j, \cdots, w_h^j \in \mathcal{A}$ and let us consider $p_1^j := w_1^j$, $p_2^j := w_1^j w_2^j, ..., p_h^j := w_1^j w_j^2 \cdots w_h^j$ the consecutive prefixes of w^j. Therefore

$$\begin{aligned} g_j(h-1) &= \mathbb{P}(T = h-1, W = w^j) \\ &= \mathbb{P}(\tilde{Y}_0 = p_1^j, \tilde{Y}_1 = p_2^j, ..., \tilde{Y}_{h-1} = p_h^j), \\ &= \tilde{\alpha}(w_1^j)\tilde{P}(p_1^j, p_2^j)\cdots\tilde{P}(p_{h-1}^j, p_h^j). \end{aligned}$$

For $k > h-1$, we have

$$\begin{aligned} g_j(k) &= \mathbb{P}(T = k, W = w^j) \\ &= \mathbb{P}(\tilde{Y}_k = w^j, \tilde{Y}_l \in E_0; l = 0, ..., k-1), \\ &= \sum_{v \in E_0}\sum_{j \in E_0} \mathbb{P}(\tilde{Y}_0 = j)\tilde{P}_{00}^{k-1}(i, v)\tilde{P}_{01}(v, w^j). \end{aligned}$$

Expressing the last equation in matrix form, we have

$$g_j(k) = \tilde{\alpha}_0\tilde{P}_{00}^{k-1}\tilde{P}_{01}e_j.$$

For the distribution function, we have

$$\begin{aligned} G_j(k) &= \sum_{l=0}^{k} \mathbb{P}(T = l, W = w^j) \\ &= \sum_{l=1}^{k} \tilde{\alpha}_0\tilde{P}_{00}^{l-1}\tilde{P}_{01}e_j. \end{aligned}$$

The sum of the above series is

$$\sum_{l=1}^{k} \tilde{P}_{00}^{l-1} = (I - \tilde{P}_{00})^{-1}(I - \tilde{P}_{00}^k),$$

where I is the identity matrix of size $|E_0| \times |E_0|$, therefore we have

$$G_j(k) = \tilde{\alpha}_0 (I - \tilde{P}_{00})^{-1} (I - \tilde{P}_{00}^k) \tilde{P}_{01} e_j. \ \blacksquare$$

Even if Markov processes model properly sequences of symbols. The main drawback of Markov hypothesis is that they cannot take into account general distributions in the sojourn time in a state by contrast discrete-time semi-Markov processes generalize discrete time Markov chains. In semi-Markov processes the distribution function of the sojourn time in a state can be any one. In next section we shall provide the analogues properties if the sequence of letters is modeled by a semi-Markov chain.

3.2 Properties of words in Semi-Markov sequences

Similarly to the Markov case we shall embed the prefix chain in a semi-Markov sequence to present the central limit theorem, the strong law of large numbers and to compute the first hitting position of the elements in $\mathcal{W}$ through the sequence of letters.

Consider the sequence of letters is modeled by a semi-Markov chain (Z_k). Let $(\mathbb{Y}_k)$ be the prefix process defined as in Definition 17 but embedded in the SMC (Z_k) i.e.,

$$\mathbb{Y}_0 := \begin{cases} w_1 & \text{if } Z_0 = w_1, \text{for some } w = w_1 w_2, \cdots w_h \in \mathcal{W}, \\ \varepsilon_i & \text{if } Z_0 = i \text{ with } i \in \mathscr{F}, \end{cases}$$

and

$$\mathbb{Y}_k := \delta_{E^*}(\mathbb{Y}_{k-1}, Z_k), \quad k \geq 1.$$

If we use a semi-Markov sequence (Z_k) we need to introduce the backward time process for each prefix, let us denote by $(\mathbb{Y}_k, B_k)$ this process. In the sequel we shall introduce the state space of this process.

Consider the prefixes of the set of words $\mathcal{W}$: E^*, see Equation (3.2). Let

$$K_p(E^*) := \{(p, n) : n \in l_p^*\}, \ \ p \in E^* \tag{3.13}$$

be the set which represents the prefix $p \in E^*$ and its backward times where l_p^* is defined according with Equation (2.7). The set

$$K(E^*) := \bigcup_{p \in E^*} K_p(E^*) \tag{3.14}$$

represents all prefixes in E^* and their corresponding backward times. Clearly we have: $K_p(E^*) \cap K_q(E^*) = \emptyset$, if $p \neq q$. The set $K(E^*)$ is the state space of process $(\mathbb{Y}_k, B_k)$.

PROPOSITION 21. (Garcia-Maya et al. [Submitted in 2020]). The process $(\mathbb{Y}_k, B_k)_{k\in\mathbb{N}}$ is a Markov chain with state space $K(E^*)$. Let us consider $p \in E^*$, $i = \delta_{E^*}^{-1}(p)$ and $\alpha(i) = \mathbb{P}(Z_0 = i)$. Then the initial distribution of the process $(\mathbb{Y}_k, B_k)$ is

$$\mathbb{P}(\mathbb{Y}_0 = p, B_0 = u) = \begin{cases} \alpha^*(w_1)\mathbb{1}_{\{u=0\}} & \text{if } p = w_1, \text{ for some } w = w_1w_2\cdots w_h \in \mathcal{W} \\ \alpha^*(i)\mathbb{1}_{\{u=0\}} & \text{if } p = \varepsilon_i, \ i \in \mathscr{F}, \\ 0 & \text{otherwise,} \end{cases}$$

and its transition probabilities are

$$\mathbb{P}(\mathbb{Y}_{k+1} = q, B_{k+1} = v \mid \mathbb{Y}_k = p, B_k = u) = \begin{cases} \frac{q_{ia}(u+1)}{1-H_i(u)}\mathbb{1}_{\{\delta_{E^*}(p,a)=q\}} & \text{if } i \neq a, \ v = 0, \\ \frac{1-H_i(u+1)}{1-H_i(u)}\mathbb{1}_{\{\delta_{E^*}(p,a)=q\}} & \text{if } i = a, \ v = u+1, \\ 0 & \text{otherwise,} \end{cases} \tag{3.15}$$

where $\delta_{E^*}^{-1}(p)$ is the partial inverse of prefix p, $\mathscr{F}$ is a set defined as in Equation (3.1) and $\mathbb{1}_A$ is the indicator function of A.

Proof. The proof is analogous to the proof in Proposition 10. ■

To simplify the notation let us denote by $\overline{P}$ the transition probability matrix of process $(\mathbb{Y}_k, B_k)$, i.e.,

$$\overline{P}((q,v),(p,u)) := \mathbb{P}(\mathbb{Y}_{k+1} = p, B_{k+1} = v \mid \mathbb{Y}_k = p, B_k = u) \tag{3.16}$$

and

$$\overline{\alpha}(p,u) := \mathbb{P}((\mathbb{Y}_0, B_0) = (p,u)) \tag{3.17}$$

its initial distribution.
It has been proved in Proposition 11 that if (Z_k, B_k) is an irreducible and aperiodic Markov chain, then the sequence $(\mathbb{Y}_k, B_k)$ has the same properties. If π is the stationary distribution of Markov process (Z_k, B_k). Then the stationary distribution of $(\mathbb{Y}_k, B_k)$ denoted by $\overline{\pi}$ is a function of π, where for $m \in l_p^*$

$$\overline{\pi}(p,m) = \begin{cases} \pi(\delta^{-1}(p), m) & \text{if } |p| = 0 \text{ or } |p| = 1; \\ \sum_{u\in l_p}\sum_{u_{l-1}\in l_{p_{l-1}}}\cdots\sum_{u_2\in l_{p_2}}\sum_{u_1\in l_{p_1}} & \\ \overline{P}((p_{l-1},u_{l-1}),(p,u))\cdots\overline{P}((p_1,u_1),(p_2,u_2))\pi(p_1,u_1) & \text{if } p = w_1\cdots w_\ell, \\ & 1 < \ell \leq h. \end{cases} \tag{3.18}$$

If prefix p has the expression $p = w_1w_2\cdots w_\ell$, $1 < \ell \leq h$ then the prefixes $p_1 = w_1^i$, $p_2 = w_1^iw_2^i, ..., p = w_1^iw_2^i\cdots w_\ell^i$ are the consecutive prefixes from $p_1 = w_1$ to $p = w_1w_2\cdots w_\ell$ and (p_1,u_1), (p_2,u_2), ..., (p,u) are the consecutive couples in $E^* \times l_p$, such that for $1 \leq j \leq \ell - 1$

$$u_{j+1} = \begin{cases} 0 & \text{if } w_j \neq w_{j+1}, \\ u_j + 1 & \text{elsewhere .} \end{cases}$$

The average number of times that the elements $\{(p,u):(p,u)\in(\mathcal{W},\cdot)\}$ are repeated through the sequence $(\mathbb{Y}_k,B_k)$ is a function of the transition probability matrix of process (Z_k,B_k) as we observe in the following proposition.

PROPOSITION 22. (Garcia-Maya et al. [Submitted in 2020]). For an ergodic Markov chain $(\mathbb{Y}_k,B_k)$ with stationary distribution $\overline{\pi}$ (see Equation 3.18) the average number of times that the elements from $\mathcal{W}$ appear through process $(\mathbb{Y}_k,B_k)$ is

$$\frac{1}{n}\sum_{k=0}^{n}\mathbb{1}_{\{(\mathbb{Y}_k,B_k)\in(\mathcal{W},\cdot)\}}\xrightarrow{a.s.}\sum_{p\in\mathcal{W}}\sum_{u\in l_p}\sum_{u_{l-1}\in l_{p_{l-1}}}\cdots\sum_{u_2\in l_{p_2}}\sum_{u_1\in l_{p_1}}\overline{P}((p_{l-1},u_{l-1}),(p,u))\cdots\overline{P}((p_1,u_1),(p_2,u_2))\pi(p_1,u_1),n\longrightarrow\infty.$$

where $p_1=w_1$, $p_2=w_1w_2,...,p=w=w_1w_2\cdots w_h$ are the consecutive prefixes from $p_1=w_1$ to $w=w_1w_2\cdots w_h$.

Proof: By the strong law of large numbers for an ergodic Markov chain

$$\frac{1}{n}\sum_{k=0}^{n}\mathbb{1}_{\{(\mathbb{Y}_k,B_k)\in(\mathcal{W},\cdot)\}}\xrightarrow{n\to\infty}\sum_{(p,u)\in K(E^*)}\overline{\pi}(p,u)\mathbb{1}_{\{(p,u)\in(\mathcal{W},\cdot)\}}$$

where

$$\begin{aligned}\sum_{(p,u)\in K(E^*)}\overline{\pi}(p,u)\mathbb{1}_{\{(p,u)\in(\mathcal{W},\cdot)\}} &= \sum_{(p,u)\in(\mathcal{W},\cdot)}\overline{\pi}(p,u)\\ &= \sum_{p\in\mathcal{W}}\sum_{u\in l_p}\sum_{u_{l-1}\in l_{p_{l-1}}}\cdots\sum_{u_2\in l_{p_2}}\sum_{u_1\in l_{p_1}}\overline{P}((p_{l-1},u_{l-1}),(p,u))\\ &\quad\cdots\overline{P}((p_1,u_1),(p_2,u_2))\pi(p_1,u_1).\end{aligned}$$

The following proposition describes the word frequencies using prefix and backward process.

PROPOSITION 23. (Garcia-Maya et al. [Submitted in 2020]). For an ergodic Markov chain $(\mathbb{Y}_k,B_k)$ with stationary distribution $\overline{\pi}$, we have

$$\sqrt{n}\left(\frac{1}{n}\sum_{k=0}^{n}\mathbb{1}_{\{(\mathbb{Y}_k,B_k)\in(\mathcal{W},\cdot)\}}-S\right)\xrightarrow{d}\mathcal{N}(0,\sigma^2_{\mathcal{W}}),n\longrightarrow\infty$$

where

$$\begin{aligned}\sigma^2_{\mathcal{W}} &= \sum_{(p_i,u_i)\in(\mathcal{W},\cdot)}\\ &\quad\left(\overline{\pi}(p_1,u_1)[\delta_{(p_1,u_1)(p_i,u_i)}-\overline{\pi}(p_i,u_i)],...,\overline{\pi}(p_{|K(E^*)|},u_{|K(E^*)|})[\delta_{(p_{|K(E^*)|},u_{|K(E^*)|})(p_i,u_i)}-\overline{\pi}(p_i,u_i)]\right)\\ &\quad\times(2Z-I)\begin{pmatrix}\overline{S}(p_1,u_1)\\ \vdots\\ \overline{S}(p_{|K(E^*)|},u_{|K(E^*)|})\end{pmatrix}.\end{aligned}\tag{3.19}$$

such that $\overline{S}(p,u) := \mathbb{1}_{\{(p,u)\in(\mathcal{W},\cdot)\}} - S$ and $S := \sum_{(p,u)\in(\mathcal{W},\cdot)} \overline{\pi}(p,u)$.

Proof: For an ergodic Markov chain $(\mathbb{Y}_k, B_k)$ with state space $K(E^*)$ and stationary distribution $\overline{\pi}$ by the CLT, we have

$$\sqrt{n}\left(\frac{1}{n}\sum_{k=0}^{n}\mathbb{1}_{\{(\mathbb{Y}_k,B_k)\in(\mathcal{W},\cdot)\}} - \sum_{(p,u)\in(\mathcal{W},\cdot)}\overline{\pi}(p,u)\right) \xrightarrow{n\to\infty} \mathcal{N}(0,\overline{\sigma}^2),$$

where $\overline{\sigma}^2 = \overline{S}diag(\overline{\pi})[2\overline{Z} - I]\overline{S}^T$ such that $\overline{S}(p,u) := \mathbb{1}_{\{(p,u)\in(\mathcal{W},\cdot)\}} - S$ and $S := \sum_{(p,u)\in(\mathcal{W},\cdot)} \overline{\pi}(p,u)$. The matrix $\overline{Z}$ is the fundamental matrix of process $(\mathbb{Y}_k, B_k)$ and it is defined by $\overline{Z} := (I - \overline{P} + \overline{\Pi})^{-1}$, such that I is the identity matrix of size $|K(E^*)| \times |K(E^*)|$, $\overline{P}$ is the transition matrix of process $(\mathbb{Y}_k, B_k)$ and $\overline{\Pi} = \lim_{n\to\infty} \overline{P}$.

Therefore

$$
\begin{aligned}
\overline{\sigma}^2_{\mathcal{W}} &:= \left(\overline{S}(p_1,u_1),...,\overline{S}(p_{|K(E^*)|},u_{|K(E^*)|})\right)
\begin{pmatrix} \overline{\pi}(p_1,u_1) & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & \overline{\pi}(p_{|K(E^*)|},u_{|K(E^*)|}) \end{pmatrix} \\
&\quad (2\overline{Z}-I)\begin{pmatrix} \overline{S}(p_1,u_1) \\ \vdots \\ \overline{S}(p_{|K(E^*)|},u_{|K(E^*)|}) \end{pmatrix} \\
&= \left(\overline{S}(p_1,u_1)\overline{\pi}(p_1,u_1),...,\overline{S}(p_{|K(E^*)|},u_{|K(E^*)|})\overline{\pi}(p_{|K(E^*)|},u_{|K(E^*)|})\right) \\
&\quad (2\overline{Z}-I)\begin{pmatrix} \overline{S}(p_1,u_1) \\ \vdots \\ \overline{S}(p_{|K(E^*)|},u_{|K(E^*)|}) \end{pmatrix} \\
&= \Bigg(\mathbb{1}_{\{(p_1,u_1)\in(\mathcal{W},\cdot)\}}\overline{\pi}(p_1,u_1) - \sum_{(p_i,u_i)\in(\mathcal{W},\cdot)} \overline{\pi}(p_i,u_i)\overline{\pi}(p_1,u_1),..., \\
&\quad \mathbb{1}_{\{(p_{|K(E^*)|},u_{|K(E^*)|})\in(\mathcal{W},\cdot)\}}\overline{\pi}(p_{|K(E^*)|},u_{|K(E^*)|}) - \sum_{(p_i,u_i)\in(\mathcal{W},\cdot)} \overline{\pi}(p_i,u_i)\overline{\pi}(p_{|K(E^*)|},u_{|K(E^*)|})\Bigg) \\
&\quad (2\overline{Z}-I)\begin{pmatrix} \overline{S}(p_1,u_1) \\ \vdots \\ \overline{S}(p_{|K(E^*)|},u_{|K(E^*)|}) \end{pmatrix} \\
&= \Bigg(\sum_{(p_i,u_i)\in(\mathcal{W},\cdot)} \overline{\pi}(p_1,u_1)[\delta_{(p_1,u_1)(p_i,u_i)} - \overline{\pi}(p_i,u_i)],..., \\
&\quad \sum_{(p_i,u_i)\in(\mathcal{W},\cdot)} \overline{\pi}(p_{|K(E^*)|},u_{|K(E^*)|})[\delta_{(p_{|K(E^*)|},u_{|K(E^*)|})(p_i,u_i)} - \overline{\pi}(p_i,u_i)]\Bigg) \\
&\quad (2\overline{Z}-I)\begin{pmatrix} \overline{S}(p_1,u_1) \\ \vdots \\ \overline{S}(p_{|K(E^*)|},u_{|K(E^*)|}) \end{pmatrix} \\
&= \sum_{(p_i,u_i)\in(\mathcal{W},\cdot)} \\
&\quad \Big(\overline{\pi}(p_1,u_1)[\delta_{(p_1,u_1)(p_i,u_i)} - \overline{\pi}(p_i,u_i)],...,\overline{\pi}(p_{|K(E^*)|},u_{|K(E^*)|})[\delta_{(p_{|K(E^*)|},u_{|K(E^*)|})(p_i,u_i)} - \overline{\pi}(p_i,u_i)]\Big) \\
&\quad (2\overline{Z}-I)\begin{pmatrix} \overline{S}(p_1,u_1) \\ \vdots \\ \overline{S}(p_{|K(E^*)|},u_{|K(E^*)|}) \end{pmatrix}. \ \blacksquare
\end{aligned}
$$

Let $(\tilde{\mathbb{Y}}_k, B_k)$ be a process defined as process $(\mathbb{Y}_k, B_k)$ but with transition probability matrix $\tilde{R}$ where

$$\tilde{R} = \mathbb{P}(\tilde{\mathbb{Y}}_{k+1} = q, B_{k+1} = v \mid \tilde{\mathbb{Y}}_k = p, B_k = u) = \begin{cases} \frac{q_{ia}(u+1)}{1-H_i(u)} \mathbb{1}_{\{\delta_{E^*}(p,a)=q\}} & \text{if } i \neq a, \ v = 0, \\ \frac{1-H_i(u+1)}{1-H_i(u)} \mathbb{1}_{\{\delta_{E^*}(p,a)=q\}} & \text{if } i = a, \ v = u+1, \\ & q \neq p \\ 1 & \text{if } p = q \\ 0 & \text{otherwise.} \end{cases} \tag{3.20}$$

The time that process $(\tilde{\mathbb{Y}}_k, B_k)$ has to wait until an element from $\mathcal{W}$ arrives is a random variable and it is defined as follows

$$\tilde{T} := \inf\{k \geq 0 : (\tilde{\mathbb{Y}}_k, B_k) \in (\mathcal{W}, \cdot)\}. \tag{3.21}$$

We are interested in computing the distribution function of $\tilde{T}$ and also the probability that $(\tilde{\mathbb{Y}}_k, B_k)$ reaches $(\mathcal{W}, \cdot)$ by a specific element $w^j \in \mathcal{W}$. Therefore, let us denote the random variable

$$\tilde{W} := \{w^j \in \mathcal{W} : \tilde{\mathbb{Y}}_T = w^j, B_k = \cdot\}. \tag{3.22}$$

The random variable $\tilde{W}$ takes the value w^j if w^j is the first element from $\mathcal{W}$ that appears in the chain $(\tilde{\mathbb{Y}}_k, B_k)$. To provide the distribution function of $(\tilde{T}, \tilde{W})$ we shall decompose the state space $K(E^*)$ according with states K_0 and K_1 where $K_0 := \{(\tilde{\mathbb{Y}}_k, B_k) \in K(E^*) : \tilde{\mathbb{Y}}_k \notin \mathcal{W}\}$ and $K_1 := \{(\tilde{\mathbb{Y}}_k, B_k) \in K(E^*) : \tilde{\mathbb{Y}}_k \in \mathcal{W}\}$. We shall consider $\tilde{R}$ the transition probability matrix of process $(\tilde{Y}_k, B_k)$, where $\tilde{R}$ is defined as follows

$$\tilde{\mathbb{R}} = \mathbb{P}(\tilde{\mathbb{Y}}_{k+1} = q, B_{k+1} = v \mid \tilde{\mathbb{Y}}_k = p, B_k = u) = \begin{cases} \frac{q_{ia}(u+1)}{1-H_i(u)} \mathbb{1}_{\{\delta_{E^*}(p,a)=q\}} & \text{if } i \neq a, \ v = 0, \\ \frac{1-H_i(u+1)}{1-H_i(u)} \mathbb{1}_{\{\delta_{E^*}(p,a)=q\}} & \text{if } i = a, \ v = u+1 \\ & \text{and } q \neq p \\ 1 & p = q \\ 0 & \text{otherwise.} \end{cases} \tag{3.23}$$

PROPOSITION 24. The law and the distribution function of $(\tilde{T}, \tilde{W})$ for the Markov chain $(\tilde{\mathbb{Y}}_k, B_k)$ are:

$$\begin{aligned} \overline{g}_i(k) &:= \mathbb{P}(\tilde{T} = k, \tilde{W} = w^i) \\ &= \begin{cases} 0, & k < h-1; \\ \overline{\alpha}(w_1^i, 0)\overline{P}((p_1^i, u_1), (p_2^i, u_2)) \cdots \overline{P}((p_{h-1}^i, u_{h-1}), (p_h^i, u_h), & k = h-1; \\ \overline{\alpha}_0 \overline{P}_{00}^{k-1} \overline{P}_{01} e_i, & k > h-1; \end{cases} \end{aligned}$$

therefore its distribution function is

$$\overline{G}_i(k) := \mathbb{P}(\tilde{T} \leq k, \tilde{W} = w^i) = \overline{\alpha}_0 (I - \overline{P}_{00})^{-1} (I - \overline{P}_{00}^k) \overline{P}_{10} e_j, \tag{3.24}$$

where e_i is a column vector of size $|K_1|$ where all its entries are zeros, except the entry which corresponds to w^i which takes the value one.

Proof. For $k < h-1$ it is obvious. For $k = h-1$, let $p_1^i := w_1^i$, $p_2^i := w_1^i w_2^i$, $p_{h-1}^i := w_1^i w_1^i \cdots w_{h-1}^i$, $p_h^i := w^i = w_1^i w_1^i \cdots w_{h-1}^i w_h^i$ be the consecutive prefixes from w_1^i to

w^i, then

$$\begin{aligned}\overline{g}_i(k=h-1) &= \mathbb{P}(\tilde{T}=h-1,\tilde{W}=w^i)\\ &= \mathbb{P}(\tilde{\mathbb{Y}}_0=p_1^i,B_0=0)\mathbb{P}(\tilde{\mathbb{Y}}_1=p_2^i,B_1=u_2\mid\tilde{\mathbb{Y}}_0=p_1^i,B_0=0),\\ &\quad\cdots\mathbb{P}(\tilde{\mathbb{Y}}_{h-1}=p_h^i,B_{h-1}=u_h\mid\tilde{\mathbb{Y}}_{h-2}=p_{h-1}^i,B_{h-2}=u_{h-1})\\ &= \overline{\alpha}(w_1^i,0)\tilde{\mathbb{R}}((p_1^i,0),(p_2^i,u_2))\cdots\tilde{\mathbb{R}}((p_{h-1}^i,u_{h-1}),(p_h^i,u_h)),\end{aligned}$$

where for $1\le j\le h-1$, u_{j+1} are the elements in l_p^* such that for $p_j^i=w_1^i\cdots w_j^i$ and $p_{j+1}^i=w_1^i\cdots w_{j+1}^i$ we have

$$u_{j+1}=\begin{cases}0 & \text{if } w_j^i\neq w_{j+1}^i\\ u_j+1 & \text{elsewhere .}\end{cases}$$

For $k>h-1$, we have

$$\begin{aligned}\bar{g}_j(k) &= \mathbb{P}(\tilde{T}=k,\tilde{W}=w^j) \qquad (3.25)\\ &= \mathbb{P}(\tilde{\mathbb{Y}}_k=w^j,B_k=\cdot \text{ such that } (\tilde{\mathbb{Y}}_l,B_l)\in K_0;l=0,...,k-1),\\ &= \sum_{u_h\in l^*_{w^j}}\ \sum_{(p_{h-1}^i,u_{h-1})\in K_0}\ \sum_{(p,u_0)\in K_0}\mathbb{P}(\tilde{\mathbb{Y}}_0=p,B_0=u_0)\\ &\quad\times\tilde{\mathbb{R}}_{00}^{k-1}((p,u_0),(p_{h-1}^i,u_{h-1}))\tilde{\mathbb{R}}_{01}((p_{h-1}^i,u_{h-1}),(w^j,u_h)).\end{aligned}$$

Expressing the last equation in matrix form, we have

$$\bar{g}_j(k)=\tilde{\alpha}_0\tilde{\mathbb{R}}_{00}^{k-1}\tilde{\mathbb{R}}_{01}e_j.$$

For the distribution function, we have

$$\begin{aligned}\overline{G}_j(k) &= \sum_{l=0}^{k}\mathbb{P}(T=l,W=w^j)\\ &= \sum_{l=1}^{k}\tilde{\alpha}_0\tilde{\mathbb{R}}_{00}^{l-1}\tilde{\mathbb{R}}_{01}e_j.\end{aligned}$$

The sum of the above series is

$$\sum_{l=1}^{k}\tilde{\mathbb{R}}_{00}^{l-1}=(I-\tilde{\mathbb{R}}_{00})^{-1}(I-\tilde{\mathbb{R}}_{00}^{k}),$$

where I is the identity matrix of size $|K_0|\times|K_0|$, therefore we have

$$\overline{G}_j(k)=\tilde{\alpha}_0(I-\tilde{\mathbb{R}}_{00})^{-1}(I-\tilde{\mathbb{R}}_{00}^{k})\tilde{\mathbb{R}}_{01}e_j.\ \blacksquare$$

3.3 Example

To show one of the implementations of the proposed model, in this section we present a genomic example where DNA is modeled by a Markov and semi-Markov process and we

present the number of times that the word CG is repeated through the DNA sequence. We also searched the enzyme SmaI by its two possible configurations.

Let us consider a bacteriophage DNA sequence. The genomic alphabet is $\mathcal{A} = \{A, T, G, C\}$ where the letters A, T, G, C represent the nucleotides adenine, thymine, guanine and cytosine respectively. In this application, we count the average number of times that the word CG is repeated through the DNA. For this word in the Markov case the main and variance computed according with Propositions 18 and 19 are 0.0869 and 0.0709 respectively. The analogous results for the semi-Markov case according with Propositions 22 and 23 are 0.0500723 and 0.052394. Additionally we search the *SmaI* enzyme by any of its two configuration, i.e., '$CCCGGG$' and '$GGGCCC$'. Considering DNA is modeled by a Markov chain, the distribution function of the first hitting position of the *SmaI* enzyme is computing using Equation (3.12). In figure 3.1 we can observe the results for both words. Similarly, we also consider DNA sequence is modeled by a semi-Markov chain. Under this hypothesis the distribution function of the first hitting position of the *SmaI* enzyme is computing using Equation (3.24). In figure 3.2 we can observe which pattern of the enzyme is more probable to appear at first under the semi-Markov hypothesis. Analyzing figures 3.1 and 3.2 we can observe that the configuration $CCCGGG$ has a big probability to appear at first time even if DNA is modeled by Markov or semi-Markov chain.

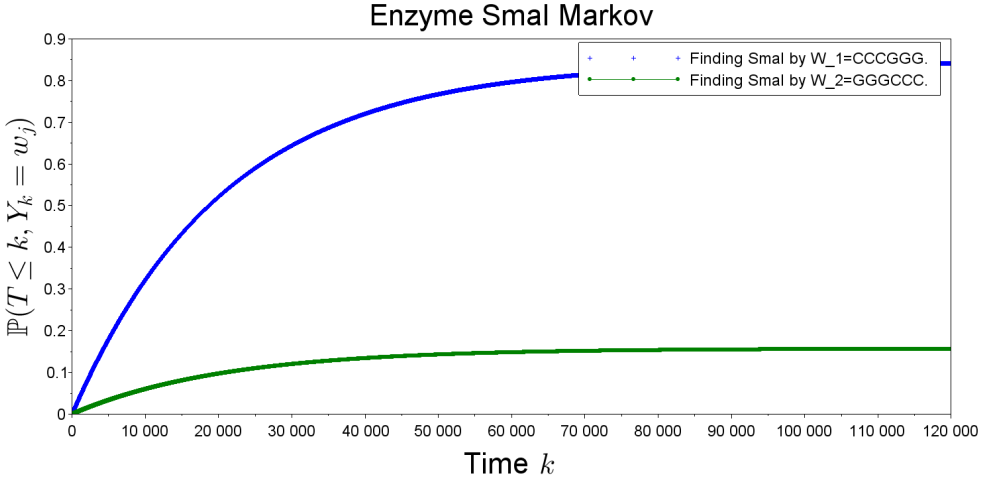


FIGURE 3.1: Probability to reach SmaI by any of its configuration under Markov hypothesis

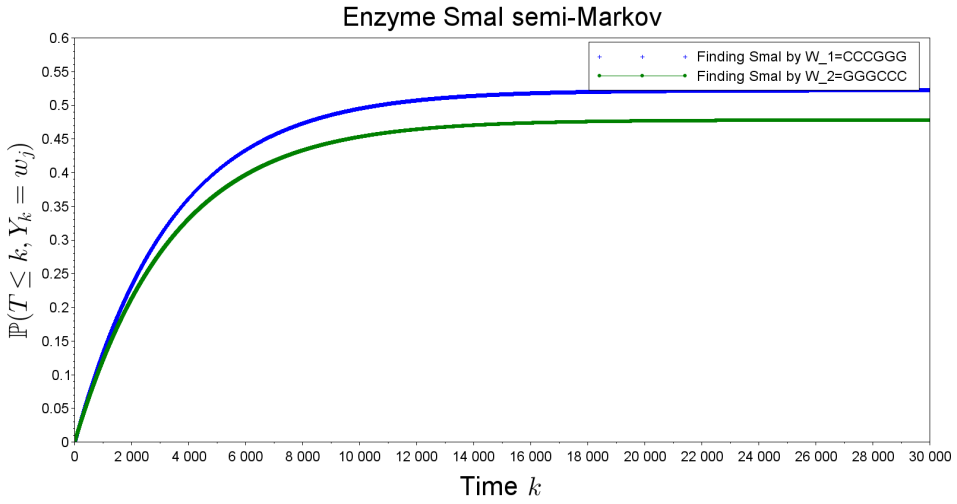


FIGURE 3.2: Probability to reach SmaI by any of its configuration under semi-Markov hypothesis

3.4 Concluding remarks

In this chapter we count the number of times that the elements from a set $\mathcal{W}$ are repeated through the DNA, i.e., we provide the strong law of large numbers for a set $\mathcal{W}$. We considered two possibilities to model DNA sequences. We consider DNA is modeled by an ergodic Markov sequence and DNA is modeled by a semi-Markov chain. For both hypothesis the Central Limit Theorem has been presented. The results that appear in the semi-Markov case can be deduced to the results in the Markov case when we have a geometric distribution. We also computed the first hitting position of the elements from the set of words $\mathcal{W}$ through the DNA sequence. To show one of its applications we computed the first hitting position of an enzyme through a DNA sequence.

Chapter 4

Phase-type Semi-Markov Distributions and Competing Risks

We present [1] here competing risks models within a semi-Markov process framework via the semi-Markov phase-type distribution. We consider semi-Markov processes in continuous and discrete time with a finite number of transient states and a finite number of absorbing states. Each absorbing state represents a failure mode (in reliability of a system) or a cause of death of an individual (in survival analysis). We express the probability a failure occurs at a certain time due to a unique cause. This is an extension of the continuous-time Markov competing risks model presented in Lindqvist and Kjølen [2018]. We give the joint distribution of the lifetime and the failure cause via the transition function of semi-Markov process in continuous and discrete-time cases. Some examples are given for illustration.

4.1 Introduction

In competing risks there are two random variables of interest T is the time to failure, and C is the cause of failure, see, e.g., Crowder [2001], Aalen [1995], Lindqvist and Kjølen [2018]. For instance, we can consider that a person could die for different causes, lung cancer, heart attack, HIV, etc. If we are interested in knowing the time to death and the cause of death, the model therefore has to include more than one absorbing state (failure state), see e.g., Crowder [2001],Crowder [2012]. Thus, if the interest is focused on a specific cause of failure in presence of different causes, we are in the case of a competing risks models. In engineering, competing risks refer to the lifetime of a machine and its cause of breakdown. For instance, if we consider a car, it can stop working because of electrical problems, dead battery, malfunctioning sensors, etc. The idea of competing risks is to model a process

[1] This chapter develops the content of an article submitted in 2020 put in shape to be inserted in this thesis.
Garcia-Maya, B.I., Limnios, and N., Lindqvist, B. (2019). Competing Risks Modeling by Extended Phase-type Semi-Markov Distributions.

where the system is exposed to several causes of failure and its eventual failure is attributed exactly to only one of them.

A natural extension of (Markov) phase-type distribution (Ph-distribution), see, e.g., Neuts [1981a], Aalen [1995], Asmussen and O'Cinneide [2006], is the semi-Markov Ph-distribution in continuous or discrete-time. See, e.g., Limnios [2012a], where the Ph-distribution is defined in semi-Markov processes for both continuous and discrete time. The aim here is to extend semi-Markov processes to competing risks models (see, e.g., Crowder [2001], Beyersmann et al. [2011], Crowder [2012]).

4.2 Semi-Markov process and extended ph-type distributions

Let us consider a semi-Markov process $Z = (Z_k)_{k\in\mathbb{N}}$ with state space $E = \{1, 2, ..., r+1\}$, where states $E_0 := \{1, 2, ..., r\}$ are the transient states and state $\{r+1\}$ is an absorbing state. Let $(J_n, S_n), n \geq 0$, be the (embedded) Markov Renewal Process (MRP) of Z. Let i, j be two elements of E. Then the semi-Markov kernel $Q(t)$ is defined as follows,

$$Q_{ij}(t) := \mathbb{P}(J_{n+1} = j, S_{n+1} - S_n \leq t \mid J_n = i), \quad n \geq 0, t \in \mathbb{R}_+. \tag{4.1}$$

Let α be the initial distribution of the semi-Markov process Z, i.e., $\alpha(i) := \mathbb{P}(Z_0 = i) = \mathbb{P}(J_0 = i)$, $i \in E$. Let $P_{ij}(t) := \mathbb{P}(Z_t = j \mid Z_0 = i)$, for $i \in E_0$, $j \in E$ be the transition function of the semi-Markov (Z_k). Of course, we have $P_{r+1,j}(t) = 0$, $j \in E_0$ and $P_{r+1,r+1}(t) = 1$, for $t \geq 0$, see section 1.2.

Consider now a partition of the semi-Markov kernel and the initial law, following sets E_0 and $\{r+1\}$, as follows:

$$Q(t) = \begin{bmatrix} Q_0(t) & L(t) \\ 0_{1\times r} & 0 \end{bmatrix} \tag{4.2}$$

and $\alpha = (\alpha_0, 0)$, where α_0 is the sub-vector corresponding to transient states E_0. The matrix $Q_0(t)$ is the restriction of the semi-Markov kernel over the transient states $E_0 \times E_0$, an $r \times r$ matrix function, and $L(t)$ is an $r \times 1$ column vector function.

Consider also the matrix

$$\overline{H} := \begin{bmatrix} \overline{H}_0(t) & 0 \\ 0 & \overline{H}_1(t) \end{bmatrix} \tag{4.3}$$

where $\overline{H}_0(t) := diag(\overline{H}_i(t), i = 1, ..., r)$ is the restriction of the sojourn times survival functions on the transient states, i.e., $\overline{H}_i(t) := 1 - \sum_{j\in E} Q_{ij}(t)$ and
$H_1(t) := diag(\overline{H}_i(t), i = r+1, ..., r+m)$ is the restriction of the sojourn times survival functions on the absorbing states.
The closed form solution of a semi-Markov phase-type distribution, say F on $[0, \infty)$, is (see, e.g., **?**Limnios [2012a]),

$$\overline{F}(t) := 1 - F(t) = \alpha_0 (I - Q_0(t))^{(-1)} * \overline{H}_0(t) \tag{4.4}$$

where I is the identity matrix for $t \geq 0$, and the zero matrix for $t < 0$, and $(I - Q_0(t))^{(-1)} = \sum_{n\geq 0} Q_0^{(n)}(t)$ where $Q_0^{(n)}$ is the n-fold convolution of Q_0 (see, e.g., **?**), i.e.,

$$Q_{0ij}^{(n)}(t) = \begin{cases} \delta_{ij}\mathbf{1}_{\{t\geq 0\}} & n = 0 \\ Q_{0ij}(t) & n = 1 \\ \sum_{k\in E} \int_0^t Q_{0ik}(ds) Q_{0kj}^{(n-1)}(t-s) & n \geq 2. \end{cases} \tag{4.5}$$

For the non singularity of this matrix see Section 4.3.

It is worth noticing here that the semi-Markov Ph-distributions on $[0, \infty)$, given by (4.4), is a dense class for the weak topology, in the set of all probability distributions on $[0, \infty)$, since this class includes as a particular case the dense class of Markov Ph-distributions (e.g., Neuts [1981a]).

4.3 Semi-Markov process and competing risks

In this section we are going to extend the semi-Markov Ph-distributions to the competing risks setting, as it has been done for the Markov case by Lindqvist and Kjølen [2018]. Let us consider a continuous-time semi-Markov process $(Z_t, t \in \mathbb{R}_+)$, with state space $E = \{1, 2, ..., r+1\}$, where states $E_0 := \{1, 2, ..., r\}$ are the transient states and state $\{r+1\}$ is an absorbing state and initial distribution α. We shall decompose the state space E in transient E_0 (good performance states) and absorbing states E_1 (failures states), i.e., $E = E_0 \cup E_1$. We shall consider $r \geq 1$ transient states and $m \geq 2$ absorbing states. Under these conditions, we shall give the main results for the extended semi-Markov continuous time Ph-distribution. The time that the process has to wait until it arrives to a failure state (an absorbing state) is a random variable. It is known as time to absorption T. We define this time as follows,

$$\mathsf{T} := \inf\{k \geq 0 : Z_k \in E_1\}. \tag{4.6}$$

The lifetime T and the cause of failure, C, with values in the set $\{1, ..., m\}$, depend on Z_t. More precisely, we have $\{\mathsf{T} \leq t, C = j\} = \{Z_t = r + j\}$. This is the key relation of the connection between competing risks and the extended Semi-Markov Ph-distributions.

Consider now the partition of the semi-Markov kernel Q, and the initial law, in this new situation following the partition E_0, E_1 of E, as follows:

$$Q(t) = \begin{bmatrix} Q_0(t) & L(t) \\ 0_{m\times r} & 0_{m\times m} \end{bmatrix} \tag{4.7}$$

and $\alpha = (\alpha_0, \alpha_1)$, notice that, in this particular case α_1 is the zero vector of dimension $1 \times r$. The function $L(t)$ is now an $r \times m$ matrix.

Consider also the diagonal matrix $\overline{H}_0(t) := diag(\overline{H}_i(t), i = 1, ..., r)$ and $H_1(t) = 0$, so $\overline{H}_1(t) = I$ the identity matrix, for $t \geq 0$, and $0_{m\times m}$ otherwise.

PROPOSITION 25. (Limnios and Oprişan [2001]). For an absorbing semi-Markov process as described above, the transition function is given by

$$P(t) = \begin{bmatrix} (I - Q_0)^{(-1)} * \overline{H}_0(t) & (I - Q_0)^{(-1)} * L(t) \\ 0_{m\times r} & I(t) \end{bmatrix}$$

Proof. For any fixed $t \geq 0$, the matrix $Q_0(t)$ is sub-stochastic, i.e., there is at least an index $i \in E_0$ such that $\sum_{j\in E_0} Q_0(i,j) < 1$. So, $(Q_0(t))^n$ goes to zero, as n goes to infinity (see, e.g., Theorem 3.65 in Girardin and Limnios [2018]). But, from Lebesgue-Stieltjes integral and Equation (4.5) we have $Q_0^{(n)}(t) \leq (Q_0(t))^n$, so $Q_0^{(n)}(t)$ goes to zero, as n goes to infinity. Consequently, the matrix $I - Q_0(t)$ is non-singular.

The transition function $P(t)$ of the semi-Markov process satisfies the following Markov renewal equation (see, e.g., **?**)

$$P(t) = \overline{H}(t) + Q * P(t),$$

where $\overline{H}(t)$ is defined as in Equation (4.3). Now, for the same reasons as previously, the matrix $I - Q(t)$ is non singular for any fixed t in the interior of the support of Q, and then via the convolution algebra, we get

$$P(t) = (I - Q)^{(-1)} * \overline{H}(t).$$

Let us consider the inversion formula of a bloc matrix, (see Lu and Shiou [2002]), i.e., in case where A and D are square and non singular matrices, we have

$$\begin{pmatrix} A & B \\ 0 & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} & -A^{-1}BD^{-1} \\ 0 & D^{-1} \end{pmatrix}.$$

Now, from the partition of the semi-Markov kernel matrix (4.2), and that of the diagonal matrix $\overline{H}(t)$, and the non singularity of $(I - Q_0)(t)$ and $I(t)$ for any $t \geq 0$, and the linearity of the convolution operation, we get, straightforwardly, the desired results.

Then the probability that the absorbing state is the state $j \in E_1$, starting from a state $i \in E_0$, is given by the (i,j) entry of the matrix $(I - Q_0)^{(-1)} * L(t)$. ■

Let us denote the distribution function of (T, C) (i.e. the cumulative incidence function) by $\mathsf{F}_{ij}(t) := \mathbb{P}_i(\mathsf{T} \leq t, C = j)$ and the corresponding failure rate $\lambda_{ij}(t)$, for initial state $i \in E_0$, and cause $j \in E_1$, conditional on survival up to time t, (which is the cause-specific hazard in the competing risks terminology,)

$$\lambda_{ij}(t) := \lim_{h\downarrow 0} \frac{\mathbb{P}_i(t < \mathsf{T} \leq t + h, C = j \mid \mathsf{T} > t)}{h}.$$

It is worth noticing here that for fixed $i \in E_0$, $j \in E_1$, $\mathsf{F}_{ij}(t)$ is a sub-distribution function.

Let us define the matrix functions $\mathsf{F}(t) := (\mathsf{F}_{ij}(t); i \in E_0, j \in E_1)$ and $\lambda(t) := (\lambda_{ij}(t); i \in E_0, j \in E_1)$.

PROPOSITION 26. (Limnios and Oprişan [2001]). Suppose that the entries of the matrix function L, in the semi-Markov (cumulative) kernel (4.2), have Radon-Nikodym derivatives.

Then the distribution function matrix $\mathsf{F}(t)$ and the conditional, on survival up to time t, failure rate matrix $\lambda(t)$, are given by

$$\mathsf{F}(t) = (I - Q_0)^{(-1)} * L(t)$$

and

$$\lambda_{ij}(t) = \frac{e_i(I - Q_0)^{(-1)} * \ell(j)}{e_i(I - Q_0)^{(-1)} * \overline{H}_0(t)\mathbf{1}_r},$$

where $\ell(t) := L'(t)$, the pointwise derivatives of L with respect to t and $e_i := (0, ..., 0, 1, 0, ..., 0)$ with 1 in the i-th entry.

Remark. In the case when we consider a general initial distribution α_0 on E_0, then the above formula can be written as $\mathsf{F}_j(t) := \mathbb{P}(\mathsf{T} \leq t, C = j) = \alpha_0(I - Q_0)^{(-1)} * L(j)$ and

$$\lambda_j(t) = \lim_{h \downarrow 0} \frac{\mathbb{P}(t < \mathsf{T} \leq t + h, C = j \mid \mathsf{T} > t)}{h} = \frac{\alpha_0(I - Q_0)^{(-1)} * \ell(j)}{\alpha_0(I - Q_0)^{(-1)} * \overline{H}_0(t)\mathbf{1}_r}.$$

Proof. We have $\mathsf{F}_j(t) := \mathbb{P}(\mathsf{T} \leq t, C = j) = \mathbb{P}(Z_t = j)$. So, by Proposition 1, we get $\mathsf{F}(t) = \alpha_0 P_{12}(t)$, and the result follows.

Let us now consider the probabilities $R_{ik} = \mathbb{P}_i(Z_{\mathsf{T}} = r + k) = \mathbb{P}_i(C = k)$, for starting in state $i \in E_0$ and being absorbed in state $k = 1, ..., m$, and define the matrix $R := (R_{ik}; i = 1, ..., r; k = 1, ..., m)$. Consider also the transition probability matrix P of the embedded Markov chain (J_n) of the semi-Markov process (Z_t) (see, e.g., **?**) and its partition following sets E_0, E_1, i.e.,

$$P = \begin{bmatrix} P_0 & P_1 \\ 0_{m \times r} & I \end{bmatrix}.$$

PROPOSITION 27. (Limnios and Oprişan [2001]). We have

$$R = (I - P_0)^{-1} P_1.$$

Proof. Since this probability depends only on the transition probabilities of the embedded Markov chain, the proof of the result is straightforward by Markov chain theory (see, e.g., Girardin and Limnios [2018]).

Examples in continuous time Let us consider a four states semi-Markov process, i.e., let $E = \{1, 2, 3, 4\}$, where states $1, 2$ are transient and states $3, 4$ are absorbing states, i.e., $E_0 = \{1, 2\}$ and $E_1 = \{3, 4\}$. The semi-Markov kernel of this process is $Q(t)$,

$$Q(t) = \begin{bmatrix} 0 & Q_{12}(t) & Q_{13}(t) & 0 \\ Q_{21}(t) & 0 & 0 & Q_{24}(t) \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

with the following blocks of its partition:

$$Q_0(t) = \begin{bmatrix} 0 & Q_{12}(t) \\ Q_{21}(t) & 0 \end{bmatrix}, \qquad L(t) = \begin{bmatrix} Q_{13}(t) & 0 \\ 0 & Q_{24}(t) \end{bmatrix}.$$

Now we have the following block matrix of the transition function

$$P_{12}(t) = (I - Q_0)^{(-1)} * L(t) = M * \begin{bmatrix} Q_{13}(t) & Q_{12} * Q_{24}(t) \\ Q_{21} * Q_{13}(t) & Q_{24}(t) \end{bmatrix}$$

where $M(t) := (1 - Q_{21} * Q_{13})^{(-1)}(t) = 1 + \sum_{k=1}^{\infty}(Q_{21} * Q_{13})^{(k)}(t)$. This is a usual renewal type function.

So, we have

$$F_1(t) = \alpha_0 P_{12}(t) e_1 = \alpha(1) M * Q_{13}(t) + \alpha(2) M * Q_{21} * Q_{13}(t)$$

and

$$F_2(t) = \alpha_0 P_{12}(t) e_1 = \alpha(1) M * Q_{12} * Q_{24}(t) + \alpha(2) M * Q_{24}(t).$$

The primes here mean derivatives with respect to t.

We also calculate the cause specific failure rates $\lambda_j(t)$, for $j = 1, 2$, as follows:

$$\lambda_1(t) = \frac{\alpha(1) M * Q'_{13}(t) + \alpha(2) M * Q_{21} * Q'_{13}(t)}{M * [\alpha(1) Q_{13} * \overline{H}_1(t) + \alpha(1) Q_{12} * Q_{24} \overline{H}_2(t) + \alpha(2) Q_{12} * Q_{13} * \overline{H}_1(t) + \alpha(2) Q_{24} * \overline{H}_2(t)]}$$

$$\lambda_2(t) = \frac{\alpha(1) M * Q_{12} * Q_{24}(t)' + \alpha(2) M * Q_{24}(t)'}{M * [\alpha(1) Q_{13} * \overline{H}_1(t) + \alpha(1) Q_{12} * Q_{24} \overline{H}_2(t) + \alpha(2) Q_{12} * Q_{13} * \overline{H}_1(t) + \alpha(2) Q_{24} * \overline{H}_2(t)]}$$

where $\overline{H}_1(t) = 1 - (Q_{12}(t) + Q_{13}(t))$ and $\overline{H}_2(t) = 1 - (Q_{21}(t) + Q_{24}(t))$, for $t \geq 0$, and $\alpha(i) := \mathbb{P}(Z_0 = i)$, for $i = 1, 2$.

Finally, the matrix R is

$$R = (1 - p_{12} p_{21})^{-1} \begin{bmatrix} p_{13} & p_{12} p_{24} \\ p_{21} & p_{24} \end{bmatrix}$$

where $p_{ij} := Q_{ij}(\infty)$, for $i, j \in E_0$ and $p_{ij} := \delta_{ij}$ for $i, j \in E_1$ (Kronecker's δ).

It is worth noticing that from $p_{12} + p_{13} = 1$ and $p_{21} + p_{24} = 1$, we can see that R is a stochastic matrix.

4.4 The discrete-time competing risk

Discrete-time semi-Markov setting. Let $(Z_k), k \in \mathbb{N}$ be a semi-Markov discrete-time process, i.e., a semi-Markov chain (SMC) with state space E, and $(J_n, S_n), n \in \mathbb{N}$, its embedded Markov renewal chain, see section 1.2.

We shall make the same considerations for the semi-Markov chain $(Z_k)_{k \in \mathbb{N}}$ as in continuous time, i.e., we shall decompose the state space in transient (good performance states) and absorbing states (failures states), i.e., $E = E_0 \cup E_1$. We shall consider $r \geq 1$ transient states and $m \geq 2$ absorbing states. We shall also make a partition the semi-Markov kernel following the states E_0 and E_1, i.e.,

$$q(k) = \begin{pmatrix} q_0(k) & q_1(k) \\ \mathbf{0} & \mathbf{0} \end{pmatrix}. \tag{4.8}$$

Observe that the first zero in the second line is the $m \times r$ zero matrix and the second one is the $m \times m$ matrix; $q_0(k)$ and $q_1(k)$ are the restriction of $q(k)$ on $E_0 \times E_0$ and $E_0 \times E_1$ respectively. The next proposition gives the main result for the extended Semi-Markov Ph-distribution in discrete time.

PROPOSITION 28. (Garcia-Maya et al. [Submitted in 2020]). For a semi-Markov discrete-time process (Z_k), $k \in \mathbb{N}$ with state space E and initial distribution α as described above,

$$\mathsf{g}_j(k) := \mathbb{P}(\mathsf{T} = k, C = j) = \begin{cases} 0, & k = 0; \\ \alpha_0 (I - q_0)^{(-1)} * q_1(k) e_j, & k \in \mathbb{N}^*; \end{cases}$$

Therefore

$$\mathbb{G}_j(k) = \mathbb{P}(\mathsf{T} \leq k, C = j) = \sum_{l=0}^{k} \alpha_0 (I - q_0)^{(-1)} * q_1(l) e_j,$$

where e_j is a column vector of size $|E_1|$ where all its coordinates are zero except the coordinate which correspond to state j.

Proof: Set

$$\mathsf{g}_{ij}(k) = \mathbb{P}_i(\mathsf{T} = k, C = j), \;\; i \in E_0, j + r \in E_1.$$

Obviously, we have:

$$\mathsf{g}_j(k) = \sum_{i \in E_0} \alpha_i \mathsf{g}_{ij}(k). \tag{4.9}$$

Now, we can write:

$$\begin{aligned} \mathsf{g}_{ij}(k) &= \mathbb{P}_i(\mathsf{T} = k, C = j, S_1 \geq k) + \mathbb{P}_i(\mathsf{T} = k, C = j, S_1 < k) \\ &= \mathbb{P}_i(J_1 = r + j, S_1 = k) + \sum_{l \leq k-1} \sum_{p \in E_0} \mathbb{P}_i(\mathsf{T} = k, Z_k = j, S_1 = l, J_1 = p). \end{aligned}$$

Then:

$$\mathsf{g}_{ij}(k) = q_{i,r+j}(t) + \sum_{l=0}^{k} \sum_{p \in E_0} q_{ip}(l) \mathsf{g}_{pj}(k - l). \tag{4.10}$$

This is a discrete-time Markov renewal equation which in matrix form gives:

$$\mathsf{g}(k) = q_1(k) + q_0 * \mathsf{g}(k).$$

Its solution is

$$\mathsf{g}(k) = (I - q_0)^{(-1)} * q_1(k) \tag{4.11}$$

see e.g., Barbu and Limnios [2008]. Equation (4.11) combined with Equation (4.9) gives the desired result. ■

Examples in discrete time

We present an example of semi-Markov chain with two absorbing states, i.e., we consider two causes of failure. For these examples the state space is $E = \{1, 2, 3, 4\}$, with up states: 1 and 2; and down states: 3 and 4, i.e., the first cause of failure is the state 3, and the

second cause of failure is the state 4.

As it was mentioned in the previous sections, for a Markov process in discrete time, the sojourn time in a state, see Equation (1.14), always has geometric distribution on $\mathbb{N}^*$ i.e., for all $i, j \in E$,

$$f_{ij}(0) := 0$$

and

$$f_{ij}(k) := p(1-p)^{k-1}, \qquad p \in [0,1], k \geq 1.$$

In a semi-Markov chain, $f_{ij}(k)$ could be any distribution. In this example every $f_{ij}(\cdot)$ is a discrete-time Weibull distribution, i.e.,

$$f_{ij}(k) := W_{ai,bj}(k)$$

where

$$W_{a,b}(0) := 0$$

and

$$W_{a,b}(k) := a^{(k-1)^b} - a^{k^b}, \qquad k \geq 1.$$

see Nakagawa and Osaki [1975]. For this particular example

$$q(k) = \big(q_{ij}(k)\big)_{1\leq i,j\leq 4} = \begin{pmatrix} 0 & p_{12}f_{12}(k) & p_{13}f_{13}(k) & p_{14}f_{14}(k) \\ p_{21}f_{21}(k) & 0 & 0 & p_{24}f_{24}(k) \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad k \in \mathbb{N},$$

where $p_{ij} := \mathbb{P}(J_{n+1} = j \mid J_n = i)$, $i, j \in E$, $n \in \mathbb{N}$. In the next figure we show the extended Phase-type distribution of the random pair (T, C) for a semi-Markov chain. In figure 4.1, the system is modeled by a semi-Markov discrete-time process. In the figure we can observe that the process enter to the absorbing state three (first cause of failure) and, the absorbing state four (second cause of failure).

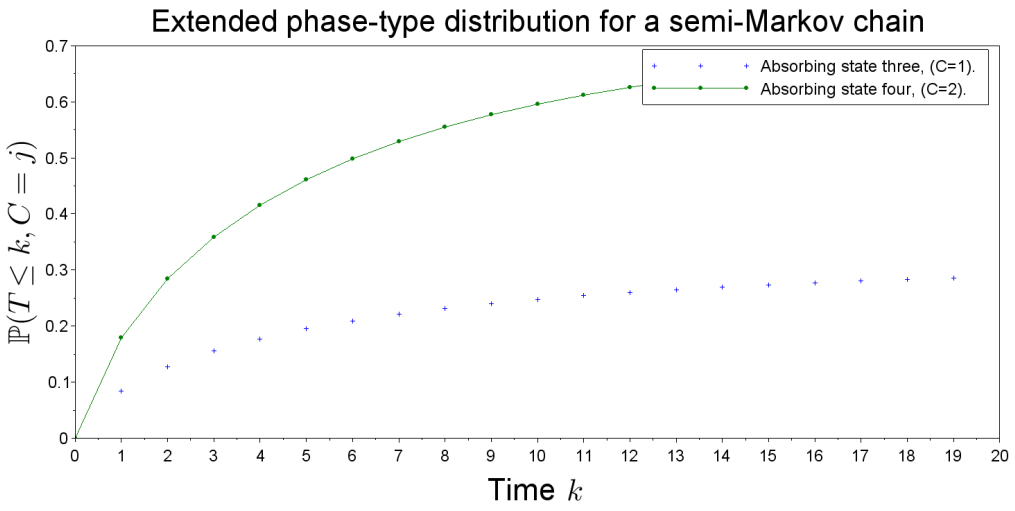


FIGURE 4.1: Probability of absorption before time k for states 3 and 4 for a semi-Markov chain

4.5 Concluding remarks

We have presented competing risks models for semi-Markov process in discrete and continuous time via phase-type distributions. We considered stochastic systems which always start in a functional state and for which there are a finite number of absorbing states. Each absorbing state may represents a failure mode in reliability applications and a cause of death of an individual in survival analysis. We derived expressions for the probability that failures occur at a certain time due to a given cause. We also gave the joint distribution of the lifetime and the cause of failure. In further work we aim at using the model as a basis for statistical estimation problems under semi-Markov assumptions.

Chapter 5

Using semi-Markov chains to solve semi-Markov processes

5.1 Introduction

Even though SMCs and SMPs [1] model properly a vast quantity of real problems most of them requires the Markov renewal equation (MRE) to be solved. The MRE finds applications in many areas of applied probability including reliability, queueing systems, inventory management, risk theory and decision analysis, see e.g., Hou et al. [2017], Asmussen et al. [2016], Dhulipala and Flint [2020],Wang [2017], etc. When the Markov renewal equation (MRE) is required in the solution of these problems, the semi-Markov hypothesis generates some difficulties. This is due to the Markov renewal equation is a function of convolutions products. The convolution of a function is a recursive method that demands a big computational memory to be implemented and even more in continuous time semi-Markov process. In a discrete-time semi-Markov process the Markov renewal function is expressed as a finite series of the semi-Markov kernel convolution product, instead in the continuous case, it is expressed in terms of an infinite series. This give an important advantage to discrete-time semi-Markov processes in numerical calculus.

A number of methods have studied the discretization of the MRE. For instance, in Barbu et al. [2004] the authors proposed a computation procedure for solving the corresponding Markov renewal equation, necessary for reliability measurements. In Barbu and Limnios [2006] the authors obtained empirical estimators for the semi-Markov kernel and the semi-Markov transition function, which allows to present a discretization of then MRE. In Elkins and Wortman [2001] it was developed tight bounds and an algorithm to compute the Markov renewal kernel. Knowledge of the kernel allows to solve Markov renewal equations numerically. In Li and Luo [2005] upper and lower bounds are studied for the solutions of Markov renewal equations. These bounds are applied to a shock model and an

[1] This chapter develops the content of an article submitted in 2020 put in shape to be inserted in this thesis.
Wu, B., Garcia-Maya, B.I., and Limnios, N., (2020). Using semi-Markov chains to solve semi-Markov processes.

age-dependent branching process under Markovian environment, etc.

The main idea in this chapter is to use SMCs to handle SMPs, whith this idea MRE can be expressed as a finite series of semi-Markov kernel convolution product, instead of an infinite series in SMPs. This good property of SMCs facilitates the computational cost making possible the implementation of the theoretical models. In order to illustrate our method, we present an example concerning cyber-attacks where it is evaluated the system availability.

We want to emphasize that this model and all mathematical results shown in this chapter were taken from the article Wu et al. [Submitted in 2020].

5.2 Continuous-time MRE solution given by discrete-time method

In this section we shall present an algorithm which computes the MRE of SMPs, see Equation (1.6). To achieve our goal first we shall present a discretization of the semi-Markov kernel.
For a given SMP $Z(t)$ with continuous semi-Markov kernel $\mathbf{Q}(t)$ and state space $E := \{1, 2, \ldots, s\}$, we define an SMC $Z_h(k)$, where $k \in \mathbb{N}$ and $h > 0$, such that the semi-Markov kernel $\mathbf{q}_h(k)$ is given by the equation

$$\mathbf{q}_h(k) := \mathbf{Q}(kh) - \mathbf{Q}((k-1)h) \text{ for } \quad k \geq 1, \text{ and } \mathbf{q}_h(0) := 0.$$

Then, the transition function matrix of the SMP, denoted by $\mathbf{P}(t)$, which satisfies Equation (1.8), has an approximate solution

$$\mathbf{P}_h(k) = \boldsymbol{\psi}_h * (\mathbf{I} - \mathbf{H}_h)(k), \quad k \in \mathbb{N}, \tag{5.1}$$

where $\boldsymbol{\psi}_h$ can be calculated by Equation (1.6) based on $\mathbf{q}_h$.

In the remainder of this chapter, let h denote the step size of discretization. Note that the matrix functions which refer to SMC $Z_h(k)$ are denoted with index h, for example, $\mathbf{P}_h$. Instead, matrix functions which refer to SMP $Z(t)$ are denoted without index h, for example, $\mathbf{P}$.

Let us denote the matrix norm by $\|\cdot\|$, defined as

$$\|\mathbf{A}\|(k) := \max_{i,j \in E} |A_{ij}(k)|,$$

where $\mathbf{A}$ is a matrix-valued function. Consider now the function $A_{ij}(t)$ defined on $\mathcal{M}_E(\mathbb{N})$ such that $A_{ij}(t) = 0$ for all $i, j \in E$ and $t < 0$, and define the following norm on $[0, t]$:

$$|||\mathbf{A}|||(t) = \sup_{0 \leq u \leq t} \|\mathbf{A}\|(u).$$

Now the continuous-time semi-Markov kernel $\mathbf{Q}(t)$ can be approximated by $\tilde{\mathbf{Q}}_h(t)$, which is defined by

$$\tilde{\mathbf{Q}}_h(t) \equiv \mathbf{Q}_h(k-1)$$

when $(k-1)h \leq t < kh$. Meanwhile, the transition function matrix of the SMP $\mathbf{P}(t)$ can be correspondingly approximated by $\tilde{\mathbf{P}}_h(t)$, which is stated in the following proposition.

PROPOSITION 29. (Wu et al. [Submitted in 2020]). If for any fixed $t > 0$, $kh \to t$ as $k \to \infty$ $(h \downarrow 0)$, then

$$\|\tilde{\mathbf{P}}_h - \mathbf{P}\|(t) \to 0, \quad \text{as} \quad h \downarrow 0.$$

Proof. The cumulative semi-Markov kernel for the SMC $Z_h(k)$ is given by

$$\mathbf{Q}_h(k) = \sum_{l \leq k} \mathbf{q}_h(l) = \mathbf{Q}(kh),$$

which implies that the cumulative distribution function of the sojourn time for the SMC $Z_h(k)$ can be written as

$$\mathbf{H}_h(k) = \mathbf{H}(kh).$$

If we consider a bounded matrix $\mathbf{A}$ of dimensions $s \times s$ and its corresponding pointwise discrete version $\mathbf{A}_h$, we get easily that

$$\mathbf{q}_h * \mathbf{A}_h(k) = \mathbf{Q} * \mathbf{A}(kh). \tag{5.2}$$

And then, for any $n \geq 1$, we get, by induction from Equation (5.2), that

$$\mathbf{q}_h^{(n)}(k) = \mathbf{Q}^{(n)}(kh),$$

which follows that, for any $k \in \mathbb{N}$,

$$\boldsymbol{\psi}_h(k) = \boldsymbol{\psi}(kh).$$

Further, we have

$$\mathbf{P}_h(k) = \boldsymbol{\psi}_h * (\mathbf{I} - \mathbf{H}_h)(k) = \boldsymbol{\psi} \star (\mathbf{I} - \mathbf{H})(kh) = \mathbf{P}(kh).$$

And then, if $kh \to t$ as $k \to \infty$ $(h \downarrow 0)$, we get by continuity that $\boldsymbol{\psi} \star (\mathbf{I} - \mathbf{H})(kh) \to \boldsymbol{\psi} \star (\mathbf{I} - \mathbf{H})(t)$, namely,

$$\mathbf{P}_h(k) \to \mathbf{P}(t) \quad \text{as} \quad k \to \infty \quad (h \downarrow 0),$$

element-wise, and then

$$\max_{i,j \in E} \left| \tilde{\mathbf{P}}_{h;ij}(t) - \mathbf{P}_{ij}(t) \right| = \max_{i,i \in E} |\mathbf{P}_{h;ij}(k-1) - \mathbf{P}_{ij}(t)| \to 0, \quad \text{as} \quad h \downarrow 0. \ \blacksquare$$

In the following, two useful propositions are presented for bounding the error for the transition function matrix due to discretization.

Let us define

$$\delta_h := \max_{\{i,j\in E\},\{k:kh<t\}} q_{ij}^h(k),$$

where $\mathbf{q}_h(k) = [q_{ij}^h(k)]$. Obviously, when $k \to \infty$ or $h \to 0$, $\delta_h \to 0$.

Recall that $\tilde{\mathbf{Q}}_h(t) \equiv \mathbf{Q}_h(k-1)$ when $(k-1)h < t < kh$, which is an extension of the discrete-time semi-Markov kernel to the continuous-time case. Based on Proposition 4.3 in Limnios and Oprisan (2012), we can obtain the distance between transition function matrices based on semi-Markov kernels $\tilde{\mathbf{Q}}_h(t)$ and $\mathbf{Q}(t)$ in the following proposition.

PROPOSITION 30. (Wu et al. [Submitted in 2020]). The distance between transition function matrices, $\tilde{\mathbf{P}}_h(t)$ and $\mathbf{P}(t)$, verifies the following inequality

$$\left|\left|\left|\tilde{\mathbf{P}}_h - \mathbf{P}\right|\right|\right|(t) \leqslant \min\{\kappa(\tilde{\mathbf{Q}}_h, \mathbf{Q})(t), \kappa(\mathbf{Q}, \tilde{\mathbf{Q}}_h)(t)\},$$

where

$$\kappa(\tilde{\mathbf{Q}}_h, \mathbf{Q})(t) = s^2 \cdot \|\psi_h\|(t) \cdot \left|\left|\left|\tilde{\mathbf{Q}}_h - \mathbf{Q}\right|\right|\right|(t) \cdot (2\|\psi\|(t) + 1).$$

Moreover, we have the following proposition which can avoid the inverse in the convolution seance.

PROPOSITION 31. (Wu et al. [Submitted in 2020]). If for a fixed time $t > 0$, the matrices $(\mathbf{I} - \tilde{\mathbf{P}}_h(t))$ and $(\mathbf{I} - \mathbf{P}(t))$ are non-singular, then

$$\left|\left|\left|\tilde{\mathbf{P}}_h - \mathbf{P}\right|\right|\right|(t) \leqslant \min\{\lambda(\tilde{\mathbf{Q}}_h, \mathbf{Q})(t), \lambda(\mathbf{Q}, \tilde{\mathbf{Q}}_h)(t)\}$$

where

$$\lambda(\tilde{\mathbf{Q}}_h, \mathbf{Q})(t) = \delta_h \cdot s^2 \cdot \max_{i,j}[\mathbf{I} - \mathbf{Q}(t)]^{-1}(i,j) \cdot (2 \max_{i,j}[\mathbf{I} - \tilde{\mathbf{Q}}_h(t)]^{-1}(i,j) + 1).$$

Meanwhile, let us denote by $\mathbf{P}_{h,i}$ the probability distribution of the SMC Z_h, and $\mathbf{P}_i$ the probability distribution of the SMP Z, conditional on starting from state $i \in E$. According to Karr's theorem (Karr 1975) (see also Limnios and Oprisan 2012), we get the following result.

PROPOSITION 32. (Wu et al. [Submitted in 2020]). For any $i \in E$, the following weak convergence holds true

$$\mathbf{P}_{h,i} \Rightarrow \mathbf{P}_i \quad \text{as} \quad h \downarrow 0.$$

In the following subsection, we shall give some numerical examples to illustrate our proposed model.

5.3 Numerical examples in reliability problem

In order to illustrate our mathematical model, we consider a reliability application where it is analyzed the sequential cyber-attacks that was explored in Liu et al. [2019]. In this last article it was explored the Trojan attacks. A Trojan horse or Trojan is a type of malware that is often disguised as legitimate software. Trojans are employed by hackers trying to

get access to users' systems. The trojan attacks were designed to commit crimes one of the most important of them is stealing identity. The way trojan horses operate is by tricking cyber users. Users are typically tricked by some form of social networks into loading and executing Trojans on their systems. We might think we have received an email from someone we know and we click on what looks like a legitimate attachment. But we have been tricked. Once activated, Trojans cyber-criminals can get access to the computational system.

A system which is exposed to Trojan attack is assumed to evolve according to an homogeneous SMP with four states, which are

- state 0: when systems work normally.
- state 1: when users receive Trojan virus links.
- state 2: when malicious links are clicked.
- state 3: when users make payment according to links.

The state transition diagram is illustrated by Figure 5.1.

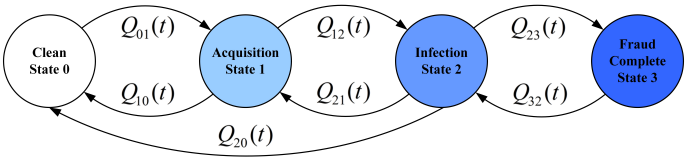


FIGURE 5.1: State transition diagram of systems subject to Trojan attacks

Hence, the system works when it is in states 0, 1 and 2, and fails due to the completed fraud in state 3. The initial distribution is assumed to be $\boldsymbol{\alpha} = (\ 1 \quad 0 \quad 0 \quad 0\ )$.

We aim to calculate the instantaneous availability $A(t)$ of the system. The instantaneous availability of a system S at time $t \in \mathbb{R}_+$ is the probability that the system is in an operational state at time t (independently that the system has fail or not in $[0, t)$). To compute the availability of the system the state space is partitioned in two groups of states: the operational or functional states, i.e., the up states which will be denoted by the letter U and the failure states, i.e., the down states which will be denoted by the letter D. For this particular example the up states is $U = \{0, 1, 2\}$ and the down states are $D = \{3\}$. Therefore the instantaneous availability of a semi-Markov process at time $t \in \mathbb{R}_+$ is

$$A(t) := \mathbb{P}(Z_t \in U).$$

It is well-known that if the system state evolution is governed by an SMP, the availability $A(t)$ is expressed by the following equation

$$A(t) = \boldsymbol{\alpha}\mathbf{P}(t)\mathbf{1}_{s,r},$$

where $\mathbf{1}_{s,r}$ is an s-dimensional vector with 1's as first r components and 0's as last $s - r$ ones. In this example, we have $s = 4$ and $r = 3$ and $\mathbf{P}(t)$ is the transition matrix of the

SMP (Z_t), see Equation (1.3). Then, we can calculate the point-wise availability of the system by computing the transition function matrix $\mathbf{P}(t)$.

5.3.1 Exponentially distributed sojourn times - Markov case

In order to verify the effectiveness of the proposed method, let us first consider a Markov case where the sojourn time in each state is exponentially distributed. As it is well known the transition matrix of a Markov process at continuous time is a function of its infinitesimal generator. In this case the infinitesimal generator is given by

$$\mathcal{G} = \begin{pmatrix} -0.2 & 0.2 & 0 & 0 \\ 0.01 & -0.11 & 0.1 & 0 \\ 0.15 & 0.3 & -0.85 & 0.4 \\ 0 & 0 & 0.5 & -0.5 \end{pmatrix},$$

then the transition probability matrix in the Markov process is

$$P(t) = \begin{pmatrix} 0 & 1-e^{-0.2t} & 0 & 0 \\ \frac{1}{11}\left(1-e^{-0.11t}\right) & 0 & \frac{10}{11}\left(1-e^{-0.11t}\right) & 0 \\ \frac{3}{17}\left(1-e^{-0.85t}\right) & \frac{6}{17}\left(1-e^{-0.85t}\right) & 0 & \frac{8}{17}\left(1-e^{-0.85t}\right) \\ 0 & 0 & 1-e^{-0.5t} & 0 \end{pmatrix}.$$

Therefore if the state evolution of the system is governed by a Markov process, the system availability can be calculated by

$$A(t) = \boldsymbol{\alpha}\mathbf{e}^{\mathcal{G}t}\mathbf{1}_{s,r}. \tag{5.3}$$

To compute the availability of the continuous system we shall make a discretization of the time. We shall consider three different approximations where the step time are $h = 0.05$, $h = 0.01$ and $h = 0.001$. We consider the approximate results when $t = 1$. The exact value calculated by Equation (5.3) and the three consider approximations are illustrated the following figure.

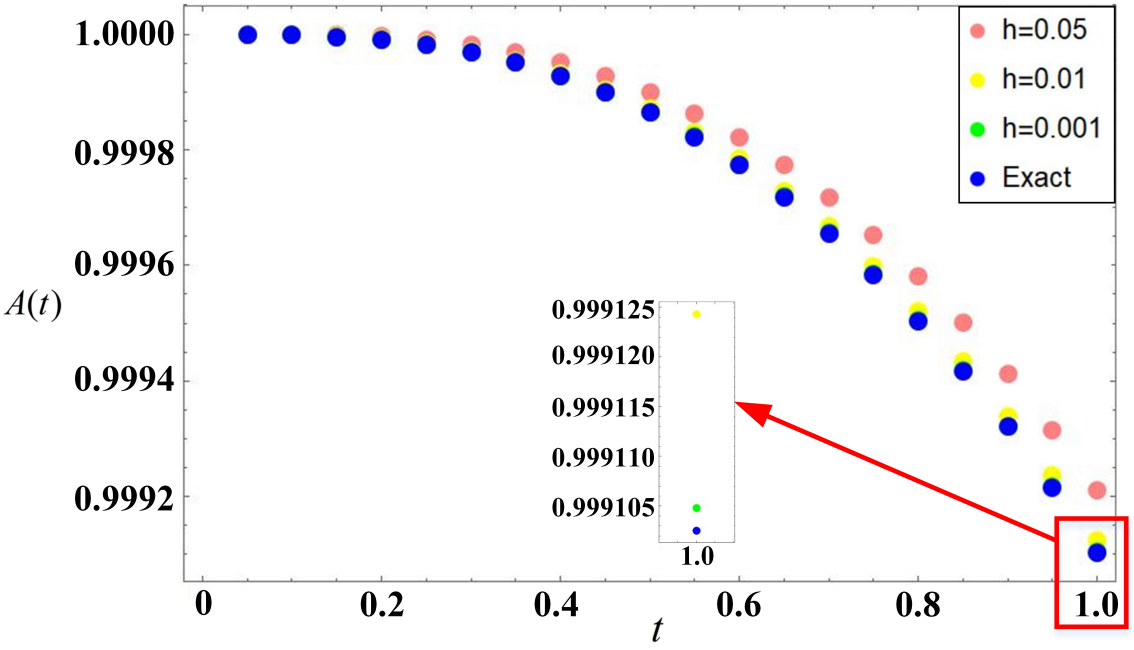


FIGURE 5.2: Exact and approximate values of the availability $A(t)$ in Markov case.

It can be seen from Figure 5.2 that the smaller h is, the closer the approximation is to the exact value, which is consistent with Proposition 29. When $h = 0.001$, the approximate value of $A(t)$ nearly coincides with the exact value.

Meanwhile, we can estimate approximation errors based on Propositions 30 and 31, whose results are shown in Table 5.1. Note that in Table 5.1, $\bar{A}(1)$ represents the approximate result of the system availability at time $t = 1$ obtained by employing the proposed method in subsection 5.2, where the discretization for the transition probability matrix in the Markov case is given by the expression

$$P_h(k) = e^{\mathcal{G}kh} - e^{\mathcal{G}(k-1)h}, \quad k \geq 1 \text{ and } P_h(0) := 0.$$

.

TABLE 5.1: Computation errors of Markov case.

h	0.1	0.05	0.01	0.001
$\bar{A}(1)$	0.999315	0.999210	0.999124	0.999105
$\bar{A}(1) - A(1)$	0.000212	0.000108	0.000022	0.000002
Estimated error by Proposition 30	0.037180	0.009013	0.000352	0.000005
Estimated error by Proposition 31	1.34437	0.654051	0.128003	0.01823

From Table 5.1, we can see that a decreased h reduces the estimated errors obtained by Propositions 30 and 31. Meanwhile, the effect of Proposition 30 is better than the estimation effect of Proposition 31, which is the penalty of Propositions 31 by reducing the

computational complexity.

In next subsection we shall apply the mathematic discretization in a semi-Markov process. For this end, we shall consider sojourn times in a states are governed by a Weibull distribution function.

5.3.2 Weibull distributed sojourn times - semi-Markov case

In this section, the sojourn time in each state is assumed to follow the Weibull distribution with scale and shape parameters $(\alpha_{ij}, \beta_{ij})$ as Table 5.2 shows, which is identical to those in Liu et al. (1990). The cumulative distribution function of the Weibull distribution is $F_{ij}\left(t; \alpha_{ij}, \beta_{ij}\right) = 1 - \exp\left[-\left(\frac{t}{\alpha_{ij}}\right)^{\beta_{ij}}\right]$.

TABLE 5.2: Baseline values of model parameters of Weibull distribution.

CDF	Parameter values
F_{01}	$\alpha_{01} = 1/0.034, \beta_{01} = 0.54$
F_{10}	$\alpha_{10} = 1/0.0125, \beta_{10} = 0.86$
F_{12}	$\alpha_{12} = 1/0.106, \beta_{12} = 2$
F_{20}	$\alpha_{20} = 1/0.15, \beta_{20} = 1$
F_{21}	$\alpha_{21} = 1/0.2, \beta_{21} = 1$
F_{23}	$\alpha_{23} = 1/0.0023, \beta_{23} = 0.072$
F_{32}	$\alpha_{32} = 1/0.39, \beta_{32} = 1$

Based on Equation (5.3), we set $h = 0.1$, $h = 0.05$, $h = 0.01$ and $h = 0.001$. We consider the approximation when $t = 2$. We can draw the curves of system point-wise availabilities as Figure 5.3 shows.

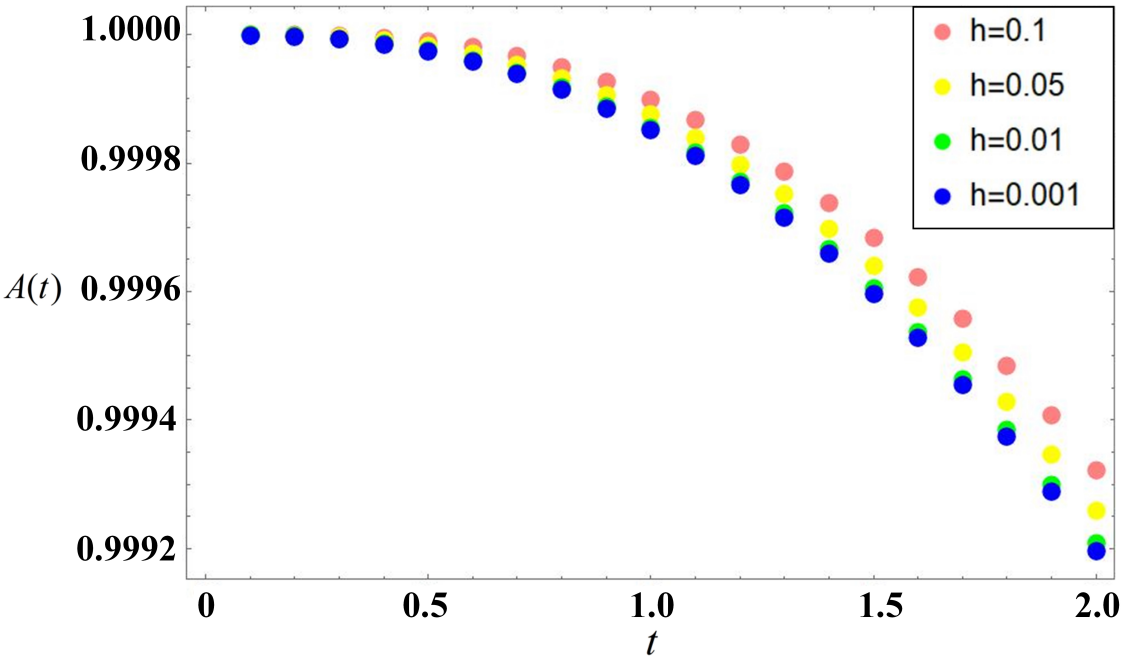


FIGURE 5.3: Approximate values of $A(t)$ under the semi-Markov hypothesis.

5.4 Conclusions

In this chapter, we presented a novel method to solve the continuous-time MRE based on the algorithm from discrete-time case. This method sheds new light on handling continuous-time SMPs which has versatility and flexibility in distributions of sojourn times with good approximate results. The error bounds caused by discretization for transition function matrices of continuous-time SMPs are studied, which provide an efficient way to decide the step size of discretization.

The proposed method is applied to any problem modeled by finite state space continuous-time semi-Markov processes. The proposed model has many applications for instance in reliability problems (as above) for availability, reliability, maintainability, etc. The effectiveness of our method is verified under the Markov case where the exact value of the system availability can be obtained in order to make comparisons with our results. Meanwhile, the case where sojourn times follow Weibull distributions is considered and computed to illustrate the applicability of our method in SMPs.

Chapter 6

Conclusions and Perspectives

SMPs have become increasingly important in probability and statistical modeling because they have a lot of applications. The popularity of SMPs is because they allow to model the sojourn time in a state by any distribution function. In the Markov context, the waiting times between states are geometric distributed (in discrete time) or exponentially distributed (in continuous time). This is the reason why SMPs fit better than Markov hypothesis for real problems.

In this thesis, we tackled DNA analysis and competing risk problems from a point of view of SMPs. We proposed a model and an algorithm that can be implemented in real applications to compute the first hitting position (time) of a set of words (patterns) in a semi-Markov sequence. This model is based on prefixes and its extended state space. For a word w with length $|w| = h$ and maximum value for the backward time of a prefix: $\gamma = \max\{l_p^* : p \in E^*\}$, we need a state space of cardinality $(h + |\mathcal{A}| - 1) \times \gamma$ at most. We also estimate the number of times that a word w, from a specific set of words $\mathcal{W}$, is repeated through out the DNA by any of its configurations, i.e., we provide the strong law of large numbers for a word sequence. To this problem, we consider two cases: DNA is modeled by an ergodic Markov sequence, and DNA is modeled by a semi-Markov chain. For both hypothesis we presented the Central Limit Theorem.

We have also tackle competing risks models from the point of view of semi-Markov processes in discrete and continuous time via phase-type distributions. We considered stochastic systems which always start in a functional state and for which there are a finite number of absorbing states. Each absorbing state represents a failure mode in reliability applications and a cause of death of an individual in survival analysis. We derived expressions for the probability that failures occur at certain time due to a given cause. We also gave the joint distribution of the lifetime and the cause of failure.

Continuous-time MRE are difficult to solve because they are expressed in terms of infinite series of the convolution kernel. In this thesis we presented an algorithm to express MRE at continous time using SMCs. The error bounds caused by discretization were also obtained. These bounds provide the efficiency of the proposed algorithm. The proposed method is applied to any problem modeled by finite state space continuous-time semi-Markov process.

6.1 Perspectives for extension of the present work

During the development of this thesis many ideas have been appeared which for lacking of time were not developed through this work but they can be considered as a future work. In the following paragraphs we would like to mention some of them:

- Firstly, identify how many different patterns of particular length are presented in a random finite sequence has been of interest in recent years. Trifonov [1990] named this problem as: the complexity of a sequence. It is well known that patterns are not always uniformly distributed and not all patterns appear in a finite stochastic sequence. To mention an example, crowd of scenarios suggest that the number of patterns of particular length through the DNA sequence is far less than the total number of patterns, see, e.g., Manfred and Winkler-Oswatitsch [1996]; Kauffman [1993]; Yockey [1992]. This means that every sub-sequence in the DNA has a specific function. Identify how many different patterns (words or motifs) of particular length are presented in a DNA sequence is fundamental to understand the structure and function of organisms, see, e.g., Badis et al. [2009], Taft et al. [2007], Blin et al. [2018].
  After Trifonov the problem of determining the complexity function for finite sequence has been considered by several authors. Like a future work we would like to compute the probability that a stochastic finite sequence reaches $c \in \mathbb{N}$ different patterns of size $h \in \mathbb{N}$, after $k \in \mathbb{N}$ positions (from the beginning of the sequence). We shall consider the hypothesis that the stochastic sequence is modeled by a semi-Markov chain. To exemplify the problem we give an example. Considering that genomic sequences are the result of a certain stochastic process governed by four nucleotides represented by the following set $\mathcal{A} = \{A, T, G, C\}$, until the position 5 (starting from 0) the number of different patterns of size $h = 3$ in the following DNA sequence taken from a bacteriophage:
  GGGCGGCGACCTCGCGGGTTTTCGCTATTTATGAAAATTTTCCGGTTTAAG
  TTTCCGTTCTTCTTCGTCATAACTTAATGTTTTTATTTAAAATACCCTCTG...
  is 4 and are the patterns: $\{GGG, GGC, GCG, CGG\}$. This problem has been encountered in biology, but it can also be encountered in other fields.

- Another point of improvement is in competing risk models. The states of semi-Markov systems can be divided into three categories: a normal working subset, a defective working subset and a breakdown subset which contains an absorbing state where the system cannot escape once entering it. If the number of transitions between the normal and defective working subsets exceeds a given value, the system will be abandoned due to the high maintenance costs. So, we are interested in computing the number of transitions between normal and effective working during a time interval.

- Speech recognition works using algorithms through acoustic and language modeling. Acoustic modeling represents the relationship between linguistic units of speech and audio signals; language modeling matches sounds with word sequences to help distinguish between words that sound similar. It is well known that hidden Markov models (HMM) are used to make speech recognition. HMM is a stochastic process

that is not directly observable, but it can be observed through another set of stochastic processes that produce the sequence of observations Van der Hoek and Elliott [2019]. The five components that characterize Hidden Markov Models are: number of hidden states in HMM, number of observation symbols per state, state transition probability distribution, observation symbol probability distribution in each state and initial state probability distribution. Nevertheless, these models have a number of limitations. The major of them, is the duration of conversations which should be exponentially distributed. For this reason we propose a hidden semi-Markov process for modeling speech recognizing.

- We proposed like a future work to develop a Randomly Stitched System (RSS) based in a continuous time SMP with discrete state space. The idea is to propose a controller for a Unmanned Aerial Vehicles (UAV) taking into acount that the availability of the position is modeled by a switching process where the state space makes reference to the GPS signal quality and which is modeled by a SMP. The GPS quality will be classify in: good, bad and fair. Depending of the state of the GPS signal the UAV determines if its position measure is the real position or the UAV is lost.

Appendix A

Publications

This PhD work has resulted in the following publications:

A.1 Book Chapters

Ch1 **B. Garcia-Maya** and N. Limnios, *Chapter: "Identification of patterns in a semi-Markov chain", Book: Statistical Topics and Stochastic Models for Dependent Data - Applications in Reliability, Survival Analysis and Related Fields, First Edition*, V.S. BARBU and N. VERGNE, ISTE, Wiley, 2020.

A.2 Published articles

B. Garcia-Maya, N. Limnios, (2019). *Identification of words in biological sequences*, Journal of Computational Biology. http://doi.org/10.1089/cmb.2019.0253

A.3 Submitted articles

Garcia-Maya, B.I., Limnios, and N., Lindqvist, B. (2019). Competing Risks Modeling by Extended Phase-type Semi-Markov Distributions.

Wu, B.,**Garcia-Maya, B.I.**, and Limnios, N., (2020). Using semi-Markov chains to solve semi-Markov processes.

Garcia-Maya, B.I., Karaliopoulou, M., and Limnios, N. (2020) Asymptotic properties of words in Markov and semi-Markov sequences.

A.4 International conferences

SMATDA 2020. Stochastic Modeling Techniques and Data Analysis International Conference and Demographics Workshop , Barcelona, Spain.

ASMADA 2019. International and SMTDA conference; Applied Stochastic Models and data analysis, Florence, Italy.

STODEP 2018. Statistical topics and stochastic models for dependent data; applications in reliability, survival analysis and related fields; Rouen, France.

Appendix B

Algorithm

We present here two algorithms that are needed in order to work with our model (proposed in the chapter 2) in practical problems. Where $\mathcal{A}$ is the alphabet, w is the word, $\mathbf{q}$ is the semi-Markov kernel of the SMC (Z_k), E is the extended state space of E_w, n_1 is the size of the first block of w and $K(\mathsf{E})$ is the state space of process (Y_k, B_k).

Algorithm 1 State space $K(\mathsf{E})$

. Require: $\mathcal{A}$, $\mathbf{q}$, E, n_1
. Initialization
for every $a \in \mathcal{A}$ **do**
  . Compute r_a according with Equation (1.15)
end for
for every $p \in \mathsf{E}$ **do**
  if $|p| == 0$ **then**
    for $i = 1$ until $i = r_{\delta^{-1}(p)}$ **do**
      . add state $(p, i-1)$ to $K(\mathsf{E})$
    end for
  end if
  if $|p| \neq 0$ and $|p| \neq n_1$, i.e., $p = w_1 w_2 \cdots w_l$, $l \neq n_1$ and $1 \leq l \leq h$ **then**
    . Let $\delta_{\mathsf{E}}^{-1}(p) = w_l$,
    . i=1,
    . backward-time $= 0$
    while $\delta_{\mathsf{E}}^{-1}(p) == w_{l-i}$ **do**
      .backward-time = backward-time $+ 1$
      . $i = i + 1$
    end while
    . add state $(p, \text{backward-time})$ to $K(\mathsf{E})$
  end if
  if $|p| = n_1$ **then**
    for $i = n_1$ until $i = r_{\delta_{\mathsf{E}}^{-1}(p)}$ **do**
      . add the state (p, i) to $K(\mathsf{E})$
    end for
  end if
end for

Algorithm 2 Transition probability matrix $\check{P}$

. Require: $\mathcal{A}$, $\mathbf{q}$, E, $K(\mathsf{E})$
. Initialization
for every $(p,u),(q,v) \in K(\mathsf{E})$ **do**
  for every $a \in \mathcal{A}$ **do**
    if $\delta_{\mathsf{E}}(p,a) == q$ **then**
      if $\delta_{\mathsf{E}}^{-1}(p) == a$ and $u+1=v$ **then**
        $\check{P}((p,u),(q,v)) =$ **Equation**(2.15)
      else if $\delta_{\mathsf{E}}^{-1}(p) \neq a$ and $v == 0$ **then**
        $\check{P}((p,u),(q,v)) =$ **Equation(2.14)**
      end if
    else
      $\check{P}((p,u),(q,v)) = 0$
    end if
  end for
end for

Appendix C

Notation and abbreviations

SMPs	semi-Markov processes
MC	Markov chain
MP	Markov process
MRP	Markov renewal process
a.s.	almost sure
EMC	embedded Markov chain
MREs	Markov renewal equations
SMCs	semi-Markov chains
MRC	Markov renewal chain
EMC	embedded Markov chain
MRF	Markov renewal function
DTMRE	discrete-time Markov renewal equation
r.v.	random variable
i.i.d.	independent and identically distributed
DFA	determinism finite automata
FMCE	finite Markov chain embedding
cdf	cumulative distribution function
KM	Kaplan Meier
CR	competing risk
NA	Nelson-Aalen

E	finite state space of a SMC and/or SMP
$\mathbb{N}$	natural numbers $\{0, 1, 2, ..\}$
$\mathbb{N}^*$	set of numbers $\{1, 2, 3, ...\}$
$\mathbb{R}_+$	set of positive real numbers
$\mathbb{P}$	probability measure
$(Z_t)_{t\in\mathbb{R}_+}$	continuous time semi-Markov process
$n \in \mathbb{N}$	renewal time
$S = (S_n)_{n\in\mathbb{N}}$	jump times of process (Z_t) or (Z_k)
$J = (J_n)_{n\in\mathbb{N}}$	embedded Markov chain (EMC)
$(J_n, S_n)_{n\in\mathbb{N}}$	Markov renewal process (MRP)
$Q_{ij}(t),\ \ i, j \in E; t \in \mathbb{R}_+$	transition kernel of semi-Markov process (Z_t)
$\psi(t), t \in \mathbb{R}$	$\psi(t) = \sum_{n=0}^{\infty} Q^{(n)}(t)$
$N(t)$	number of jumps of SMP (Z_t) in the interval time $[0, t]$
$\alpha(i), i \in E$	initial distribution of SMP (Z_t) and/or SMC (Z_k)
$H_i(\cdot)$	cumulative distribution function of the sojourn time in state $i \in E$
$H(t)$	diagonal matrix $H(t) := \mathrm{diag}(H_i(t))$
$F_{ij}(\cdot)\ \ i, j \in E$	conditional distribution function of the holding time in state i before visiting state j
$\mathbf{p} = (p_{ij})_{i,j\in E}$	transition function of EMC (J_n)
$\mathbf{P}(t) = (P_{ij}(t); i, j \in E, t \in \mathbb{R}_+)$	transition function of (Z_t)
ν	stationary distribution of the EMC (J_n)
m_i	mean sojourn time in state $i \in E$
μ^*_{ii}	mean passage time from state i to state j
π	stationary distribution of SMP (Z_t)
B_t	backward recurrence time of SMP (Z_t)
V_t	forward recurrence time of SMP (Z_t)
$k \in \mathbb{N}$	calendar time
$(Z_k)_{k\in\mathbb{N}}$	discrete time semi-Markov process i.e., SMC
$(X_n)_{n\in\mathbb{N}}$	successive sojourn times between successive jumps
$\mathbf{q} = (q_{ij}(k); i, j \in E, k \in \mathbb{N})$	discrete-time semi-Markov kernel
$Q_{ij}(k),\ \ i, j \in E; k \in \mathbb{N}$	cumulative semi-Markov kernel of SMC (Z_k)
$f_{ij}(\cdot)$	conditional sojourn time distribution conditioned by next state to be visited
$h_i(\cdot)$	sojourn time distribution in state i
C_i	support of h_i
r_i	maximum sojourn time in state i
l_i	maximum value for the backward time in state i
$\mathcal{M}_E$	real matrice on $E \times E$
$\mathcal{M}_E(\mathbb{N})$	real matrice on $E \times E$ which evolves in a discrete time $k \in \mathbb{N}$
$I(t) = I$	identity matrix in $\mathcal{M}_E(\mathbb{N})$
$A * B$	discrete-time matrix convolution product of $A, B \in \mathcal{M}_E(\mathbb{N})$
$A^{(-1)}$	left inverse in the convolution sense of $A \in \mathcal{M}_E(\mathbb{N})$
$A^{(n)}$	n-fold convolution of $A \in \mathcal{M}_E(\mathbb{N})$
$\psi(k),\ \ k \in \mathbb{N}$	$\psi(k) = \sum_{n=0}^{k} q^{(n)}(k)$
$\mathbf{\Psi} = (\Psi_{i,j}(k), i, j \in E, k \geq 0)$	Markov renewal function
$\hat{q}$	estimator of $\mathbf{q}$

$\mathcal{A}$	finite alphabet
$w = w_1 w_2 \cdots w_h$	word of size h formed by the letters $w_1, w_2, .., w_h \in \mathcal{A}$
$\|w\|$	number of symbols of the word w
$\mathcal{A}^h$	the set of words of size h formed by the elements in $\mathcal{A}$
$\overline{z}_j$	an element from $\mathcal{A}^h$
$(Z_k, B_k)_{k\in\mathbb{N}}$	MC with state $E \times \mathbb{N}$
$\tilde{p}$	transition probability matrix of MC (Z_k, B_k)
$(\overline{Z}_k, \overline{B}_k)_{k\in\mathbb{N}}$	h-dimensional Markov process where $\overline{Z}_k := (Z_k, .., Z_{k+h-1})$ and $\overline{B}_k := (B_k, ..., B_{k+h-1})$
K_j	subset of $\mathcal{A}^h \times \mathbb{N}^h$ which represents the word $\bar{z}_j$ and all possible backwards time for each letter in the word
K	state space of Markov process $(\overline{Z}_k, \overline{B}_k)$
$\alpha(\overline{z}, \overline{u})$	initial distribution of MC $(\overline{Z}_k, \overline{B}_k)$
$\widetilde{P}$	transition probability of Markov process $(\overline{Z}_k, \overline{B}_k)$
$\mathbb{1}_A$	indication function of A
$T_{\mathcal{W}}$	r.v. which determines the first hitting time of an element form the set $\mathcal{W}$
$\mathbf{1}_{\|A\|}$	column vector of ones with size cardinality of A
$\mathbb{X} = (\mathbb{X}_k)_{k\in\mathbb{N}}$	Markov chain (different from the MCE (J_n))
$\mathbb{P}$	transition probability matrix of MC $\mathbb{X}$
$\check{\pi}$	stationary distribution function of MC $\mathbb{X}$
$(\mathcal{A}, \mathcal{Q}, s, \mathcal{F}, \delta)$	determinism finite automata (DFA) where $\mathcal{A}$ is a finite set of symbols, $\mathcal{Q}$ is a set of states, $s \in \mathcal{Q}$ is the initial state, $\mathcal{F} \subset \mathcal{Q}$ is a set of accepted states and $\delta : \mathcal{Q} \times \mathcal{A} \to \mathcal{Q}$ is a transition function
ε	empty prefix
E_w	prefix set of the word w
$\delta_{E_w} : E_w \times \mathcal{A} \to E_w$	function which is defined as the longest prefix in E_w that can be formed with the concatenation between a prefix in E_w and an element from $\mathcal{A}$
$\delta^{-1}_{E_w}(\cdot)$	partial inverse of δ_{E_w}
E	extended state space of E_w
$Y = (Y_k)$	chain of prefixes embedded in a SMC (Z_k) and defined in E
$w(\eta)$	$\eta \in \mathbb{N}$ block of the word w
l^*_p	backward times for prefix p through the SMC (Z_k)
$K_p(\mathsf{E})$	prefix $p \in \mathsf{E}$ and its corresponding backward times
$K(\mathsf{E})$	all prefixes in E and its backward times
(Y_k, B_k)	chain of prefixes and backward times defined in E
$\check{P}$	transition matrix of the Markov chain (Y_k, B_k)
β	initial distribution of the Markov chain (Y_k, B_k)
N_w	number of elements in (Z_k) before the first position of w
$\mathbb{E}(T)$	expected value of the r.v. T
$\mathrm{Var}(T)$	variance of the r.v. T
$\sigma(T)$	standard deviation of the r.v. T
$\check{\lambda}$	word occurrence rate

$\mathcal{W}$	subset of $\mathcal{A}^h$
$\mathscr{F}$	set of letters in $\mathcal{A}$ that are different from the first letter of any word $w \in \mathcal{W}$
$\tilde{E}^*$	prefix set of $\mathcal{W}$
E^*	extended state space of $\tilde{E}^*$
$Y^* := (Y_k^*)_{k\in\mathbb{N}}$	prefix chain of $\mathcal{W}$ embedded in MC $(\mathbb{X}_k)$ and defined in E^*
α^*	initial distribution of prefix chain Y^*
$\tilde{P}$	transition probability matrix of prefix chain Y^*
$\tilde{\pi}$	stationary distribution of prefix process Y^*
$\tilde{Y} = (\tilde{Y}_k)$	prefix process defined as prefix process Y^* but where the elements in $\mathcal{W}$ are consider absorbing states
$\tilde{R}$	transition probability matrix of prefix process $\tilde{Y}$
T	r.v. which represents the time that process $(\tilde{Y}_k)$ has to wait until an element from $\mathcal{W}$ arrives
W	r.v. which takes the value w^j if w^j is the first element from $\mathcal{W}$ that appears in the prefixes chain $(\tilde{Y}_k)$
$G_j(k)$	cdf of the r.v. (T, W)
$g_j(k)$	pdf of the r.v. (T, W)
$(\mathbb{Y}_k)$	prefix chain embedded in SMC (Z_k) and defined in E^*
$(\mathbb{Y}_k, B_k)_{k\in\mathbb{N}}$	chain of prefixes and backward times defined in $K(E^*)$
$\overline{P}$	transition probability matrix of process $(\mathbb{Y}_k, B_k)$
$\overline{\alpha}$	initial distribution of process $(\mathbb{Y}_k, B_k)$
$\overline{\pi}$	stationary distribution of process $(\mathbb{Y}_k, B_k)$
$(\tilde{\mathbb{Y}}_k, B_k)$	process of prefixes and backward times defined as process $(\mathbb{Y}_k, B_k)$ but where the elements in $\mathcal{W}$ are considered absorbing states
$\tilde{\mathbb{R}}$	transition probability matrix of process $(\tilde{\mathbb{Y}}_k, B_k)$
$\tilde{T}$	r.v. which represents the time that process $(\tilde{\mathbb{Y}}_k, B_k)$ has to wait until an element from $\mathcal{W}$ arrives
$\tilde{W}$	r.v. that takes the value w^j if w^j is the first element from $\mathcal{W}$ that appears in the chain $(\tilde{\mathbb{Y}}_k, B_k)$
$\overline{g}_i(k)$	cdf of the r.v. $(\tilde{T}, \tilde{W})$
$\overline{G}_i(k)$	pdf of the r.v. $(\tilde{T}, \tilde{W})$
$\|\cdot\|$	matrix norm
$A(t)$	instantaneous availability of a system
$\mathcal{G}$	infinitesimal generator of a Markov process
T	time to failure
C	cause of failure
$\mathsf{g}_j(k)$	pdf of a r.v. (T, C)
$\mathsf{G}_j(k)$	cdf of a r.v. (T, C)
λ	hazard function
Λ	cumulative hazard function
$\mathsf{F}_j(t)$	sub-distribution function $\mathsf{F}_j(t) := \mathbb{P}(\mathsf{T} \leq t, C = j)$

Bibliography

O. O. Aalen. Phase type distributions in survival analysis. *Scandinavian Journal of Statistics*, pages 447–463, 1995.

M. Abadi and N. Vergne. Poisson approximation for search of rare words in dna sequences. *Am. J. Prob. Math. Stat*, 4:233–44, 2008.

N. A. Aboluion. *The construction of DNA codes using a computer algebra system.* PhD thesis, University of Glamorgan, 2011.

D. L. Antzoulakos. Waiting times for patterns in a sequence of multistate trials. *Journal of Applied Probability*, 38(2):508–518, 2001.

S. Asmussen and C. O'Cinneide. Matrix-exponential distributions. *Encyclopedia of Statistical Sciences*. John Wiley and Sons, 3, 2006.

S. Asmussen, L. Lipsky, and S. Thompson. Markov renewal methods in restart problems in complex systems. In *The Fascination of Probability, Statistics and their Applications*, pages 501–527. Springer, 2016.

K. Athreya and P. Ney. Limit theorems for semi-Markov processes. *Bulletin of the Australian Mathematical Society*, 19(2):283–294, 1978.

P. C. Austin, D. S. Lee, and J. P. Fine. Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, 133(6):601–609, 2016.

G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, et al. Diversity and complexity in dna recognition by transcription factors. *Science*, 324(5935):1720–1723, 2009.

C. Bailey, N. Poole, and D. J. Blackburn. Identifying patterns of communication in patients attending memory clinics: a systematic review of observations and signs with potential diagnostic utility. *Br J Gen Pract*, 68(667):123–138, 2018.

V. Barbu and N. Limnios. Empirical estimation for discrete-time semi-Markov processes with applications in reliability. *Nonparametric Statistics*, 18(7-8):483–498, 2006.

V. Barbu, M. Boussemart, and N. Limnios. Discrete-time semi-Markov model for reliability and survival analysis. *Communications in Statistics-Theory and Methods*, 33(11):2833–2868, 2004.

V. S. Barbu and N. Limnios. *Semi-Markov chains and hidden semi-Markov models toward applications: their use in reliability and DNA analysis*, volume 191. Springer, 2008.

J. Beyersmann, A. Allignol, and M. Schumacher. *Competing risks and multistate models with R.* Springer Science & Business Media, 2011.

K. Blin, W. Wohlleben, and T. Weber. Patscanui: An intuitive web interface for searching patterns in dna and protein data. *Nucleic acids research*, 46(W1):W205–W208, 2018.

R. A. Bobbitt, V. P. Gourevitch, L. E. Miller, and G. D. Jensen. Dynamics of social interactive behavior: A computerized procedure for analyzing trends, patterns, and sequences. *Psychological Bulletin*, 71(2):110, 1969.

S. Chadjiconstantinidis, D. Antzoulakos, and M. Koutras. Joint distributions of successes, failures and patterns in enumeration problems. *Advances in Applied Probability*, 32(3): 866–884, 2000.

O. Chryssaphinou, M. Karaliopoulou, and N. Limnios. On discrete time semi-Markov chains and applications in words occurrences. *Communications in Statistics-Theory and Methods*, pages 1306–1322, 2008.

E. Çinlar. Introduction to stochastic processes. Prentice Hill, 1975, N.J

M. Codish, M. Frank, and V. Lagoon. The dna word design problem: A new constraint model and new results. In *IJCAI*, pages 585–591, 2017.

M. Crochemore and V. T. Stefanov. Waiting time and complexity for matching patterns with automata. *Information Processing Letters*, 87(3):119–125, 2003.

M. Crowder. *Classical competing risks.* Chapman & Hall/CRC, 2001.

M. J. Crowder. *Multivariate survival analysis and competing risks.* Chapman and Hall/CRC, 2012.

G. D'Amico, F. Petroni, and F. Prattico. First and second order semi-Markov chains for wind speed modeling. *Physica A: Statistical Mechanics and its Applications*, 392(5): 1194–1201, 2013.

J. H. de Haan and J. Rotmans. Patterns in transitions: understanding complex chains of change. *Technological Forecasting and Social Change*, 78(1):90–102, 2011.

S. L. Dhulipala and M. M. Flint. Series of semi-Markov processes to model infrastructure resilience under multihazards. *Reliability Engineering & System Safety*, 193:106659, 2020.

D. Elkins and M. Wortman. On numerical solution of the Markov renewal equation: tight upper and lower kernel bounds. *Methodology and Computing in Applied Probability*, 3 (3):239–253, 2001.

D. Farré, R. Roset, M. Huerta, J. E. Adsuara, L. Roselló, M. M. Albà, and X. Messeguer. Identification of patterns in biological sequences at the alggen server: Promo and malgen. *Nucleic acids research*, 31(13):3651–3653, 2003.

W. Feller. On semi-Markov processes. *Proceedings of the National Academy of Sciences of the United States of America*, 51(4):653, 1964.

J. C. Fu and Y. Chang. On probability generating functions for waiting time distributions of compound patterns in a sequence of multistate trials. *Journal of Applied Probability*, 39(1):70–80, 2002.

B. I. Garcia-Maya and N. Limnios. Identification of words in biological sequences under the semi-Markov hypothesis. *Journal of Computational Biology*, 27(5):683–697, 2020.

B. I. Garcia-Maya, M. Karaliopoulou, and N. Limnios. Asymptotic properties of words in semi-Markov sequences. Submitted in 2020.

B. I. Garcia-Maya, N. Limnios, and B. H. Lindqvist. Competing risk modeling by extended phase-type semi-Markov distributions. Submitted in 2020.

V. Girardin and N. Limnios. *Applied Probability: From Random Sequences to Stochastic Processes*. Springer, 2018.

J. Glaz, M. Kulldorff, V. Pozdnyakov, and J. M. Steele. Gambling teams and waiting times for patterns in two-state Markov chains. *Journal of Applied Probability*, 43(1):127–140, 2006.

S. Grigorescu and G. Oprişan. Limit theorems for J- X processes with a general state space. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 35(1):65–73, 1976.

P. D. Hebert, A. Cywinska, S. L. Ball, and J. R. Dewaard. Biological identifications through dna barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1512):313–321, 2003.

J. E. Hopcroft, R. Motwani, and J. D. Ullman. Introduction to automata theory, languages, and computation. *Acm Sigact News*, 32(1):60–65, 2001.

Y. Hou, N. Limnios, and W. Schön. On the existence and uniqueness of solution of Markov renewal equations and applications. *Methodology and Computing in Applied Probability*, 19(4):1241–1250, 2017.

V. Jääskinen, V. Parkkinen, L. Cheng, and J. Corander. Bayesian clustering of dna sequences using Markov chains and a stochastic partition model. *Statistical Applications in Genetics and Molecular Biology*, 13(1):105–121, 2014.

J. M. Jaffe, Y.-E. Lee, L.-N. Huang, and H. Oshagan. Gender identification, interdependence, and pseudonyms in cmc: Language patterns in an electronic conference. *The Information Society*, 15(4):221–234, 1999.

B. J. Jansen, M. Zhang, and A. Spink. Patterns and transitions of query reformulation during web searching. *IJWIS*, 3(4):328–340, 2007.

J. Janssen. (Ed.). *Semi-Markov models: theory and applications*. Springer Science & Business Media, 2013.

P. J. Jungck and D. Helms. Identification of patterns in stateful transactions, 2013.

E. Kaplan and D. Sil'vestrov. Theorems of the invariance principle type for recurrent semi-Markov processes with arbitrary phase space. *Theory of Probability and Its Applications*, 24(3):536–547, 1980.

M. Karaliopoulou. On the number of word occurrences in a semi-Markov sequence of letters. *ESAIM: Probability and Statistics*, 13:328–342, 2009.

S. A. Kauffman. *The origins of order: Self-organization and selection in evolution*. Oxford University Press, USA, 1993.

H. Kitano. Computational systems biology. *Nature*, 420(6912):206–210, 2002.

S. Lacny, T. Wilson, F. Clement, D. J. Roberts, P. Faris, W. A. Ghali, and D. A. Marshall. Kaplan–Meier survival analysis overestimates cumulative incidence of health-related events in competing risk settings: a meta-analysis. *Journal of clinical epidemiology*, 93: 25–35, 2018.

P. Levy. Processus semi-markoviens. In *Proc. Int. Congress. Math. III, Amsterdam, 1954*, 1954.

F. Li, M. Zhang, B. Tian, B. Chen, G. Fu, and D. Ji. Recognizing irregular entities in biomedical text via deep neural networks. *Pattern Recognition Letters*, 105:105–113, 2018.

G. Li and J. Luo. Upper and lower bounds for the solutions of Markov renewal equations. *Mathematical Methods of Operations Research*, 62(2):243–253, 2005.

Z. Li, H. Cao, Y. Cui, and Y. Zhang. Extracting dna words based on the sequence features: non-uniform distribution and integrity. *Theoretical Biology and Medical Modelling*, 13 (1):2, 2016.

N. Limnios. Reliability measures of semi-Markov systems with general state space. *Methodology and Computing in Applied Probability*, 14(4):895–917, 2012a.

N. Limnios. Reliability measures of semi-Markov systems with general state space. *Methodology and Computing in Applied Probability*, 14(4):895–917, 2012b.

N. Limnios and G. Oprişan. *Semi-Markov processes and reliability*. Springer Science & Business Media, 2001.

B. H. Lindqvist and S. H. Kjølen. Phase-type models and their extension to competing risks. In *Recent Advances in Multi-state Systems Reliability*, pages 107–120. Springer, 2018.

Q. Liu, L. Xing, and C. Zhou. Probabilistic modeling and analysis of sequential cyber-attacks. *Engineering Reports*, 1(4):e12065, 2019.

T.-T. Lu and S.-H. Shiou. Inverses of 2× 2 block matrices. *Computers & Mathematics with Applications*, 43(1-2):119–129, 2002.

V. Malinovskii. Limit theorems for recurrent semi-Markov processes and Markov renewal processes. *Journal of Soviet Mathematics*, 36(4):493–502, 1987.

E. Manfred and R. Winkler-Oswatitsch. Steps towards life. a perspective on evolution, 1996.

R. Montemanni. Combinatorial optimization algorithms for the design of codes: a survey. *Journal of Applied Operational Research Vol*, 7(1):37, 2015.

T. Nakagawa and S. Osaki. The discrete Weibull distribution. *IEEE Transactions on Reliability*, 24(5):300–301, 1975.

M. Neuts. *Matrix-geometric solutions an algorithmic approach.* The Johns Hopkins University Press, Baltimore, MD, 1981a.

M. Neuts. Matrix-geometric solutions–an algorithmic approach, 1981b.

P. Nicodeme, B. Salvy, and P. Flajolet. Motif statistics. *Theoretical Computer Science*, 287(2):593–617, 2002.

D. Njamen. Convergence of the Nelson-Aalen estimator in competing risks. *International Journal of Statistics and Probability*, 6(3):9–23, 2017.

G. Nuel. Pattern Markov chains: optimal Markov chain embedding through deterministic finite automata. *Journal of Applied Probability*, 45(1):226–243, 2008.

E. Nummelin. Uniform and ratio limit theorems for Markov renewal and semi-regenerative processes on a general state space. In *Annales de l'IHP Probabilités et statistiques*, volume 14, pages 119–143, 1978.

D. Nur, D. Allingham, J. Rousseau, K. L. Mengersen, and R. McVinish. Bayesian hidden Markov model for DNA sequence segmentation: A prior sensitivity analysis. *Computational Statistics & Data Analysis*, 53(5):1873–1882, 2009.

A. A. Papadopoulou. Some results on modeling biological sequences and web navigation with a semi Markov chain. *Communications in statistics-Theory and Methods*, 42(16): 2853–2871, 2013.

P. Paraskevopoulos, T.-C. Dinh, Z. Dashdorj, T. Palpanas, and L. Serafini. Identification and characterization of human behavior patterns from mobile phone data. *D4D Challenge session, NetMob*, 2013.

J. C. Pederson. Intelligent observation and identification database system, 2016.

F. Picard, S. Schbath, E. Lebarbier, P. Neuvial, and J. Chiquet. Statistiques et génome. *La Gazette des Mathématiciens*, 130:51–82, 2011.

M. Pintilie. An introduction to competing risks analysis. *Revista Española de Cardiología (English Edition)*, 64(7):599–605, 2011.

R. L. Prentice, J. D. Kalbfleisch, A. V. Peterson Jr, N. Flournoy, V. T. Farewell, and N. E. Breslow. The analysis of failure times in the presence of competing risks. *Biometrics*, pages 541–554, 1978.

E. Rachelson, G. Quesnel, F. Garcia, and P. Fabiani. A simulation-based approach for solving generalized semi-Markov decision processes. In *ECAI*, pages 583–587, 2008.

S. Robin and J.-J. Daudin. Exact distribution of word occurrences in a random sequence of letters. *Journal of Applied Probability*, 36(1):179–193, 1999.

S. Robin, S. Schbath, and V. Vandewalle. Statistical tests to compare motif count exceptionalities. *BMC bioinformatics*, 8(1):84, 2007.

S. M. Ross. *Applied probability models with optimization applications*. Courier Corporation, 2013.

D. Roy and R. Gupta. Classifications of discrete lives. *Microelectronics Reliability*, 32(10): 1459–1473, 1992.

S. Schbath. An overview on the distribution of word counts in Markov chains. *Journal of Computational Biology*, 7(1-2):193–201, 2000.

V. Shurenkov and Y. I. Eleiko. Limit distributions of time averages for a semi-Markov process with finite number of states. *Ukrainian Mathematical Journal*, 31(5):475–479, 1979.

J. D. Sigwart and A. Garbett. Biodiversity assessment, dna barcoding, and the minority majority. *Integrative and comparative biology*, 58(6):1146–1156, 2018.

W. L. Smith. Regenerative stochastic processes. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 232(1188):6–31, 1955.

S. Srivastava and M. S. Baptista. Markovian language model of the dna and its information content. *Royal Society open science*, 3(1):150527, 2016.

V. Stefanov, A. G. Pakes, et al. Explicit distributional results in pattern formation. *The Annals of Applied Probability*, 7(3):666–678, 1997.

V. T. Stefanov, S. Robin, and S. Schbath. Occurrence of structured motifs in random sequences: Arbitrary number of boxes. *Discrete Applied Mathematics*, 159(8):826–831, 2011.

R. J. Taft, M. Pheasant, and J. S. Mattick. The relationship between non-protein-coding dna and eukaryotic complexity. *Bioessays*, 29(3):288–299, 2007.

L. Takács. Some investigations concerning recurrent stochastic processes of a certain type. *Maygyar Tud. Akad. Mad. Kutato int. Kozl*, 3:115–128, 1954.

N. Touyar, S. Schbath, D. Cellier, and H. Dauchel. Poisson approximation for the number of repeats in a stationary Markov chain. *Journal of Applied Probability*, 45(2):440–455, 2008.

S. Trevezas and N. Limnios. Variance estimation in the central limit theorem for Markov chains. *Journal of Statistical Planning and Inference*, 139(7):2242–2253, 2009.

E. N. Trifonov. Making sense of the human genome. *Structure and methods: proceedings of the Sixth Conversation in the Discipline Biomolecular Stereodynamics held at the State University of New York at Albany, June 6-10, 1989/edited by RH Sarma & MH Sarma*, 1990.

A. Tsiatis. Competing risks. *Encyclopedia of biostatistics*, J. Wiley, 2, 2005.

J. Van der Hoek and R. J. Elliott. *Introduction to Hidden Semi-Markov Models*, volume 445. Cambridge University Press, 2019.

Z. Wang. Variance and volatility swaps and futures pricing under geometric Markov renewal processes and stochastic volatility models. Master's thesis, Graduate Studies, 2017.

T. J. Wheeler, J. Clements, S. R. Eddy, R. Hubley, T. A. Jones, J. Jurka, A. F. Smit, and R. D. Finn. Dfam: a database of repetitive dna based on profile hidden Markov models. *Nucleic Acids Research*, 41(D1):D70–D82, 2012.

P. Willett. Searching for pharmacophoric patterns in databases of three-dimensional chemical structures. *Journal of Molecular Recognition*, 8(5):290–303, 1995.

B. Wu, B. I. Garcia-Maya, and N. Limnios. Using semi-Markov chains to solve semi-Markov processes. Submitted in 2020.

J. Yackel. Limit theorems for semi-Markov processes. *Transactions of the American Mathematical Society*, 123(2):402–424, 1966.

H. P. Yockey. *Information theory & molecular biology.* Cambridge University Press, 1992.