Poisson approximation for search of rare words in DNA sequences

Nicolas Vergne, Miguel Abadi

Alea **4**, 223–244 (2008)

Poisson approximation for search of rare words in DNA sequences

Nicolas Vergne and Miguel Abadi

Laboratoire Statistique et Gènome, 523 Place des Terrasses, 91000 Evry, France
E-mail address: nvergne@genopole.cnrs.fr

Instituto de Matemática, Estatística e Computação Científica, Universidade de Campinas, Brasil
E-mail address: miguel@ime.unicamp.br

Abstract. Using recent results on the occurrence times of a string of symbols in a stochastic process with mixing properties, we present a new method for the search of rare words in biological sequences modelled by a Markov chain. We obtain a bound on the error between the distribution of the number of occurrences of a word in a sequence and its Poisson approximation. A global bound is already given by a Chen-Stein method. Our approach, the ψ-mixing method, gives local bounds. Since we only need the error in the tails of distribution, the global uniform bound of Chen-Stein is too large and it is a better way to consider local bounds. It is the first time that local bounds are devised for Poisson approximation. We search for two thresholds on the number of occurrences from which we can regard a studied word as an over-represented or an under-represented one. A biological role is suggested for these over- or under-represented words. Our method gives such thresholds for a panel of words much broader than the Chen-Stein method which cannot give any result in a great number of cases where our method works. Comparing the methods, we observe a better accuracy for the ψ-mixing method for the bound of the tails of distribution. Our method can obviously be used in domains other than biology. We also present the software PANOW (available at http://stat.genopole.cnrs.fr/sg/software/panow/) dedicated to the computation of the error term and the thresholds for a studied word.

1. Introduction

Modelling DNA sequences with stochastic models and developing statistical methods to analyse the enormous set of data that results from the multiple projects of DNA sequencing are challenging questions for statisticians and biologists. Many DNA sequence analysis are based on the distribution of the occurrences of patterns having some special biological function. The most popular model in this domain is the Markov chain model that gives a

Received by the editors November 22, 2007; accepted June 25, 2008.
2000 *Mathematics Subject Classification.* 60F05, 60G10, 92D20 (Primary); 60-04 (Secondary).
Key words and phrases. Poisson approximation, Chen-Stein method, mixing processes, Markov chains, rare words, DNA sequences.

description of the local behaviour of the sequence (see Almagor (1983); Blaisdell (1985); Philips et al. (1987); Gelfand et al. (1992)). An important problem is to determine the statistical significance of a word frequency in a DNA sequence. Nicodème et al. (2002) discuss about this relevance of finding over- or under-represented words. The naive idea is the following: a word may have a significant low frequency in a DNA sequence because it disrupts replication or gene expression, whereas a significantly frequent word may have a fundamental activity with regard to genome stability. Well-known examples of words with exceptional frequencies in DNA sequences are biological palindromes corresponding to restriction sites avoided for instance in *E. coli* (Karlin et al. (1992)), the Cross-over Hotspot Instigator sites in several bacteria (Smith et al. (1981); El Karoui et al. (1999)), and uptake sequences (Smith et al. (1999)) or polyadenylation signals (van Helden et al. (2000)).

The exact distribution of the number of a word occurrences under the Markovian model is known and some softwares are available (Robin and Daudin (1999); Régnier (2000)) but, because of numerical complexity, they are often used to compute expectation and variance of a given count (and thus use, in fact, Gaussian approximations for the distribution). In fact these methods are not efficient for long sequences or if the Markov model order is larger than 2 or 3. For such cases, several approximations are possible: Gaussian approximations (Prum et al. (1995)), Binomial or Poisson approximations (van Helden et al. (1998); Godbole (1991)), compound Poisson approximations (Reinert and Schbath (1998)), or large deviations approach (Nuel (2004)). In this paper we only focus on the Poisson approximation. For the first time, we give a local bound for the Poisson approximation. We approximate $\mathbb{P}(N(A) = k)$ by $\exp(-t\mathbb{P}(A))[t\mathbb{P}(A)]^k(k!)^{-1}$ where $\mathbb{P}(N(A) = k)$ is the stationary probability under the Markov model that the number of occurrences $N(A)$ of word A is equal to k, $\mathbb{P}(A)$ is the probability that word A occurs at a given position, and t is the length of the sequence. Intuitively, a binomial distribution could be used to approximate the distribution of occurrences of a particular word. Length t of the sequence is large, $\mathbb{P}(A)$ is small if A is large. Thus, we use the more numerically convenient Poisson approximation. Our aim is to bound the error between the distribution of the number of occurrences of word A and its Poisson approximation. In Reinert and Schbath (1998), the authors prove an upper bound for a compound Poisson approximation. They use a Chen-Stein method, which is the usual method in this purpose. This method has been developed by Chen on Poisson approximations (Chen (1975)) after a work of Stein on normal approximations (Stein (1972)). Its principle is to bound the difference between the two distributions in total variation distance for all subsets of the definition domain. Since we are interested in under- or over-represented words, we are only interested in this difference for the tails of the distributions. Then, the uniform bound given by the Chen-Stein method is too large for our purpose. We present here a new method, based on the property of mixing processes. Our method has the useful particularity to give a bound on the error at each point of the distribution. More precisely, it offers an error term ϵ, for the number of occurrences k, of word A:

$$\left| \mathbb{P}(N(A) = k) - \frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!} \right| \leq \epsilon(A, k).$$

Moreover, $\epsilon(A, k)$ decays factorially fast with respect to k.

Abadi (2001a, 2004) presents lower and upper bounds for the exponential approximation of the first occurrence time of a rare event, also called *hitting time*, in a stationary stochastic process on a finite alphabet with α- or ϕ-mixing property. (Abadi and Vergne, in

preparation) describe the statistics of *return times* of a string of symbols in such a process. In (Abadi and Vergne, in preparation), the authors prove a Poisson approximation for the distribution of occurrence times of a string of symbols in a ϕ-mixing process. The first part of our present work is to determine some constants not explicitly computed in the results of the above mentioned articles but necessary for the proof of our theorem and moreover for its practical use. Theoritical constants are useless in the way of numerical tests, that is why we have to determine these constants. Our work is complementary to all these articles, in the sense that it relies on them for preliminary results and it adapts them to ψ-mixing processes. Since Markov chains are mixing processes, all these results established for mixing processes also apply to Markov chains which model biological sequences.

This paper is organised in the following way. In section 2, we introduce the Chen-Stein method. In section 3, we define a ψ-mixing process and state some preliminary notations, mostly on the properties of a word. We also present in this section the principal result of our work: the Poisson approximation (Theorem 3.3). In section 4, we state preliminary results. Mainly, we recall results of Abadi (2004), but computing all the necessary constants and we present lemmas and propositions necessary for the proof of Theorem 3.3. In section 5, we establish the proof of our main result: Theorem 3.3 on Poisson approximation. Using ψ-mixing properties and preliminary results, we prove an upper bound for the difference between the exact distribution of the number of occurrence of word A and the Poisson distribution of parameter $t\mathbb{P}(A)$. Section 6 is dedicated to numerical results. For the search of over-represented words, we show how our method is better than the Chen-Stein method on both synthetic and biological data. In this section, we also present results obtained by a similar method, the ϕ-mixing method. We end the paper presenting some examples of biological applications, and some conclusions and perspectives of future works.

2. The Chen-Stein method

2.1. *Total variation distance.*

Definition 2.1. For any two random variables X and Y with values in the same discrete space E, the total variation distance between their probability distributions is defined by

$$d_{\mathrm{TV}}(\mathcal{L}(X),\mathcal{L}(Y)) = \frac{1}{2}\sum_{i\in E}\left|\mathbb{P}(X=i)-\mathbb{P}(Y=i)\right|.$$

We remark that for any subset S of E

$$|\mathbb{P}(X\in S)-\mathbb{P}(Y\in S)| \leq d_{\mathrm{TV}}(\mathcal{L}(X),\mathcal{L}(Y)).$$

2.2. *The Chen-Stein method.*

The Chen-Stein method is used to bound the error between the distribution of the number of occurrences of a word A in a sequence X and the Poisson distribution with parameter $t\mathbb{P}(A)$ where t is the length of the sequence and $\mathbb{P}(A)$ the stationary measure of A. The Chen-Stein method for Poisson approximation has been developed by Chen (1975); a friendly exposition is in Arratia et al. (1989) and a description with many examples can be found in Arratia et al. (1990) and Bardour et al. (1992). We will use Theorem 1 in Arratia et al. (1990) with an improved bound by Bardour et al. (1992) (Theorem 1.A and Theorem 10.A).

First, we will fix a few notations. Let $\mathcal{A}$ be a finite set (for example, in the DNA case $\mathcal{A}=\{a,c,g,t\}$). Put $\Omega=\mathcal{A}^{\mathbb{Z}}$. For each $x=(x_m)_{m\in\mathbb{Z}}\in\Omega$, we denote by X_m the m-th coordinate of the sequence x: $X_m(x)=x_m$. We denote by $T:\Omega\to\Omega$ the one-step-left shift operator: so we will have $(T(x))_m = x_{m+1}$. We denote by $\mathcal{F}$ the σ-algebra over Ω

generated by strings and by $\mathcal{F}_I$ the σ-algebra generated by strings with coordinates in I with $I \subseteq \mathbb{Z}$. We consider an invariant probability measure $\mathbb{P}$ over $\mathcal{F}$. Consider a stationary Markov chain $X = (X_i)_{i\in\mathbb{Z}}$ on the finite alphabet $\mathcal{A}$. Let us fix a word $A = (a_1, \ldots, a_n)$. For $i \in \{1, 2, \cdots, t-n+1\}$, let Y_i be the following random variable

$$\begin{aligned} Y_i = Y_i(A) &= \mathbb{1}\{\text{word } A \text{ appears at position } i \text{ in the sequence}\} \\ &= \mathbb{1}\{(X_i, \ldots, X_{i+n-1}) = (a_1, \ldots, a_n)\}, \end{aligned}$$

where $\mathbb{1}\{F\}$ denotes the indicator function of set F. We put $Y = \sum_{i=1}^{t-n+1} Y_i$, the random variable corresponding to the number of occurrences of a word, $\mathbb{E}(Y_i) = m_i$ and $\sum_{i=1}^{t-n+1} m_i = m$. Then, $\mathbb{E}(Y) = m$. Let Z be a Poisson random variable with parameter m: $Z \sim \mathcal{P}(m)$. For each i, we arbitrarily define a set $V(i) \subset \{1, 2, \cdots, t-n+1\}$ containing the point i. The set $V(i)$ will play the role of a neighbourhood of i.

Theorem 2.2 (Arratia et al. (1990); Bardour et al. (1992))**.** *Let I be an index set. For each $i \in I$, let Y_i be a Bernoulli random variable with $p_i = \mathbb{P}(Y_i = 1) > 0$. Suppose that, for each $i \in I$, we have chosen $V(i) \subset I$ with $i \in V(i)$. Let $Z_i, i \in I$, be independent Poisson variables with mean p_i. The total variation distance between the dependent Bernoulli process $\underline{Y} = \{Y_i, i \in I\}$ and the Poisson process $\underline{Z} = \{Z_i, i \in I\}$ satisfies*

$$d_{\mathrm{TV}}(\mathcal{L}(\underline{Y}), \mathcal{L}(\underline{Z})) \leq b_1 + b_2 + b_3$$

where

$$b_1 = \sum_i \sum_{j\in V(i)} \mathbb{E}(Y_i)\mathbb{E}(Y_j),$$

$$b_2 = \sum_i \sum_{j\in V(i), j\neq i} \mathbb{E}(Y_i Y_j),$$

$$b_3 = \sum_i \mathbb{E}\left|\mathbb{E}(Y_i - p_i | Y_j, j \notin V(i))\right|.$$

Moreover, if $W = \sum_{i\in I} Y_i$ and $\lambda = \sum_{i\in I} p_i < \infty$, then

$$d_{TV}(\mathcal{L}(W), \mathcal{P}(\lambda)) \leq \frac{1-e^{-\lambda}}{\lambda}(b_1 + b_2) + \min\left(1, \sqrt{\frac{2}{\lambda e}}\right) b_3.$$

We think of $V(i)$ as a neighbourhood of strong dependence of Y_i. Intuitively, b_1 describes the contribution related to the size of the neighbourhood and the weights of the random variables in that neighbourhood; if all Y_i had the same probability of success, then b_1 would be directly proportional to the neighbourhood size. The term b_2 accounts for the strength of the dependence inside the neighbourhood; as it depends on the second moments, it can be viewed as a "second order interaction" term. Finally, b_3 is related to the strength of dependence of Y_i with random variables outside its neighbourhood. In particular, note that $b_3 = 0$ if Y_i is independent of $\{Y_j | j \notin V(i)\}$.

One consequence of this theorem is that for any indicator function of an event, i.e. for any measurable functional h from Ω to $[0, 1]$, there is an error bound of the form $|\mathbb{E}h(\underline{Y}) - \mathbb{E}h(\underline{Z})| \leq d_{TV}(\mathcal{L}(\underline{Y}), \mathcal{L}(\underline{Z}))$. Thus, if $S(\underline{Y})$ is a test statistic then, for all $t \in \mathbb{R}$,

$$\mathbb{P}(S(\underline{Y}) \geq t) - \mathbb{P}(S(\underline{Z}) \geq t) \leq b_1 + b_2 + b_3,$$

which can be used to construct confidence intervals and to find p-values for tests based on this statistic.

3. Preliminary notations and Poisson Approximation

3.1. *Preliminary notations.* We focus on Markov processes in our biological applications (see 6) but the theorem given in the following subsection is established for more general mixing processes: the so called ψ-mixing processes.

Definition 3.1. Let $\psi = (\psi(\ell))_{\ell \geq 0}$ be a sequence of real numbers decreasing to zero. We say that $(X_m)_{m\in\mathbb{Z}}$ is a ψ-mixing process if for all integers $\ell \geq 0$, the following holds

$$\sup_{n\in\mathbb{N}, B\in\mathcal{F}_{\{0,.,n\}}, C\in\mathcal{F}_{\{n\geq 0\}}} \frac{|\mathbb{P}(B\cap T^{-(n+\ell+1)}(C)) - \mathbb{P}(B)\mathbb{P}(C)|}{\mathbb{P}(B)\mathbb{P}(C)} = \psi(\ell),$$

where the supremum is taken over the sets B and C, such that $\mathbb{P}(B)\mathbb{P}(C) > 0$.

For a word A of Ω, that is to say a measurable subset of Ω, we say that $A \in \mathcal{C}_n$ if and only if

$$A = \{X_0 = a_0, \ldots, X_{n-1} = a_{n-1}\},$$

with $a_i \in \mathcal{A}, i = 1, \ldots, n$. Then, the integer n is the length of word A. For $A \in \mathcal{C}_n$, we define the hitting time $\tau_A : \Omega \to \mathbb{N}\cup\{\infty\}$, as the random variable defined on the probability space (Ω,$\mathcal{F}$,$\mathbb{P}$):

$$\forall x \in \Omega, \quad \tau_A(x) = \inf\{k \geq 1 : T^k(x) \in A\}.$$

τ_A is the first time that the process hits a given measurable A. We also use the classical probabilistic shorthand notations. We write $\{\tau_A = m\}$ instead of $\{x \in \Omega : \tau_A(x) = m\}$, $T^{-k}(A)$ instead of $\{x \in \Omega : T^k(x) \in A\}$ and $\{X_r^s = x_r^s\}$ instead of $\{X_r = x_r, ..., X_s = x_s\}$. Also we write for two measurable subsets A and B of Ω, the conditional probability of B given A as $\mathbb{P}(B|A) = \mathbb{P}_A(B) = \mathbb{P}(B\cap A)/\mathbb{P}(A)$ and the probability of the intersection of A and B by $\mathbb{P}(A\cap B)$ or $\mathbb{P}(A;B)$. For $A = \{X_0^{n-1} = x_0^{n-1}\}$ and $1 \leq w \leq n$, we write $A^{(w)} = \{X_{n-w}^{n-1} = x_{n-w}^{n-1}\}$ for the event consisting of the *last* w symbols of A. We also write $a \vee b$ for the supremum of two real numbers a and b. We define the periodicity p_A of $A \in \mathcal{C}_n$ as follows:

$$p_A = \inf\{k \in \mathbb{N}^* | A \cap T^{-k}(A) \neq \emptyset\}.$$

p_A is called the principal period of word A. Then, we denote by $\mathcal{R}_p = \mathcal{R}_p(n)$ the set of words $A \in \mathcal{C}_n$ with periodicity p and we also define $\mathcal{B}_n$ as the set of words $A \in \mathcal{C}_n$ with periodicity less than $[n/2]$, where $[.]$ defines the integer part of a real number:

$$\mathcal{R}_p = \{A \in \mathcal{C}_n | p_A = p\}, \mathcal{B}_n = \bigcup_{p=1}^{\left[\frac{n}{2}\right]} \mathcal{R}_p.$$

$\mathcal{B}_n$ is the set of words which are self-overlapping before half their length (see Example 3.2). We define $\mathcal{R}(A)$ the set of return times of A which are not a multiple of its periodicity p_A:

$$\mathcal{R}(A) = \left\{k \in \{[n/p_A]p_A + 1, \ldots, n-1\} | A \cap T^{-k}(A) \neq \emptyset\right\}.$$

Let us denote $r_A = \#\mathcal{R}(A)$, the cardinality of the set $\mathcal{R}(A)$. Define also $n_A = \min \mathcal{R}(A)$ if $\mathcal{R}(A) \neq \emptyset$ and $n_A = n$ otherwise. $\mathcal{R}(A)$ is called the set of secondary periods of A and n_A is the smallest secondary period of A. Finally, we introduce the following notation. For an integer $s \in \{0, \ldots, t-1\}$, let $N_s^t = \sum_{i=s}^t \mathbb{1}\{T^{-i}(A)\}$. The random variable N_s^t counts the number of occurrences of A between s and t (we omit the dependence on A). For the sake of simplicity, we also put $N^t = N_0^t$.

TABLE 3.1. Periods and overlapping in `aaataaataaa`

0	1	2	3	**4**	5	6	7	**8**	**9**	**10**										
a	a	a	t	a	a	a	t	a	a	a										
				a	a	a	t	a	a	a	t	a	a	a						
								a	a	a	t	a	a	a	t	a	a	a		
									a	a	a	t	a	a	a	t	a	a	a	
										a	a	a	t	a	a	a	t	a	a	a

Example 3.2. Consider the word $A = aaataaataaa$. Since $p_A = 4$, we have $A \in \mathcal{B}_n$ where $n = 11$. See TABLE 3.1 to note that $\mathcal{R}(A) = \{9; 10\}$, $r_A = 2$ and $n_A = 9$.

3.2. *The mixing method.* We present a theorem that gives an error bound for the Poisson approximation. Compared to the Chen-Stein method, it has the advantage to present non uniform bounds that strongly control the decay of the tail distribution of N^t.

Theorem 3.3 (ψ-mixing approximation)**.** *Let $(X_m)_{m\in\mathbb{Z}}$ be a ψ-mixing process. There exists a constant $C_\psi = 254$, such that for all $A \in \mathcal{C}_n \setminus \mathcal{B}_n$ and all non negative integers k and t, the following inequality holds:*

$$\left|\mathbb{P}(N^t = k) - \frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!}\right| \leq C_\psi e_\psi(A) e^{-(t-(3k+1)n)\mathbb{P}(A)} g_\psi(A,k)$$

$$\text{where } g_\psi(A,k) = \begin{cases} \dfrac{(2\lambda)^{k-1}}{(k-1)!} & k \notin \{\frac{\lambda}{e_\psi(A)}, ..., \frac{2t}{n}\} \\ \dfrac{(2\lambda)^{k-1}}{\left(\frac{\lambda}{e_\psi(A)}\right)!\left(\frac{1}{e_\psi(A)}\right)^{k-\frac{\lambda}{e_\psi(A)}-1}} & k \in \{\frac{\lambda}{e_\psi(A)}, ..., \frac{2t}{n}\} \end{cases},$$

$$e_\psi(A) = \inf_{1\leq w\leq n_A}\left[(r_A+n)\mathbb{P}\left(A^{(w)}\right)(1+\psi(n_A-w))\right],$$

$$\text{and } \lambda = t\mathbb{P}(A)(1+\psi(n)).$$

This result is at the core of our study. It shows an upper bound for the difference between the distribution of the number of occurrences of word A in a sequence of length t and the Poisson distribution of parameter $t\mathbb{P}(A)$. Proof is postponed in Section 5.

4. Calculation of the constants

Our goal is to compute a bound as small as possible to control the error between the Poisson distribution and the distribution of the number of occurrences of a word. Thus, we determine the global constant C_ψ appearing in Theorem 3.3 by means of intermediary bounds appearing in the proof. General bounds are interesting asymptotically in n, but for biological applications, n is approximately between 10 or 20, which is too small. Then along the proof, we will indicate the intermediary bounds that we compute. Before establishing the proof of that Theorem 3.3, we point out here, for easy references, some results of Abadi (2004), and some other useful results. In Abadi (2004), these results are given only in the ϕ-mixing context. Moreover exact values of the constants are not given, while these are necessary for practical use of these methods. We provide the values of all the constants appearing in the proofs of these results.

Proposition 4.1 (Proposition 11 in Abadi (2004))**.** *Let $(X_m)_{m\in\mathbb{Z}}$ be a ψ-mixing process. There exist two finite constants $C_a > 0$ and $C_b > 0$, such that for any n, any word $A \in \mathcal{C}_n$, and any $c \in \left[4n, \frac{1}{2\mathbb{P}(A)}\right]$ satisfying*

$$\psi(c/4) \leq \mathbb{P}\left(\{\tau_A \leq c/4\} \cap \{\tau_A \circ T^{c/4} > c/2\}\right),$$

there exists Δ, with $n < \Delta \leq c/4$, such that for all positive integers k, the following inequalities hold:

$$\left|\mathbb{P}(\tau_A > kc) - \mathbb{P}(\tau_A > c - 2\Delta)^k\right| \leq C_a\varepsilon(A)\, k\mathbb{P}(\tau_A > c - 2\Delta)^k, \quad (4.1)$$

$$\left|\mathbb{P}(\tau_A > kc) - \mathbb{P}(\tau_A > c)^k\right| \leq C_b\varepsilon(A)\, k\mathbb{P}(\tau_A > c - 2\Delta)^k, \quad (4.2)$$

$$\text{with } \varepsilon(A) = \inf_{n\leq\ell\leq\frac{1}{\mathbb{P}(A)}} [\ell\mathbb{P}(A) + \psi(\ell)].$$

Both inequalities provide an approximation of the hitting time distribution by a geometric distribution at any point t of the form $t = kc$. The difference between these distributions is that in 4.1, the geometric term inside the modulus is the same as in the upper bound, while in 4.2, the geometric term inside the modulus is larger than the one in the upper bound. That is, the second bound gives a larger error. We will use both in the proof of Theorem 4.3.

Proposition 4.2. *We have $C_a = 24$ and $C_b = 25$.*

Proof. For the details of the proof of Proposition 4.1, we refer to Proposition 11 in Abadi (2004).

For any $c \in \left[4n, \frac{1}{2\mathbb{P}(A)}\right]$ and $\Delta \in [n, c/4]$, we denote $\mathcal{N}_j^i = \left\{\tau_A \circ T^{ic+j\Delta} > c - j\Delta\right\}$ and $\mathcal{N} = \{\tau_A > c - 2\Delta\}$ for the sake of simplicity. Abadi (2004) obtains the following bound:

$$\forall k \geq 2, \left|\mathbb{P}(\tau_A > kc) - \mathbb{P}(\mathcal{N})^k\right| \leq (a) + (b) + (c), \text{ with}$$

$$(a) = \sum_{j=0}^{k-2} \mathbb{P}(\mathcal{N})^j \left|\mathbb{P}((k-j)c) - \mathbb{P}\left(\tau_A > (k-j-1)c; \mathcal{N}_2^{k-j-1}\right)\right|,$$

$$(b) = \sum_{j=0}^{k-2} \mathbb{P}(\mathcal{N})^j \left|\mathbb{P}\left(\tau_A > (k-j-1)c; \mathcal{N}_2^{k-j-1}\right) - \mathbb{P}(\tau_A > (k-j-1)c)\,\mathbb{P}\left(\mathcal{N}_2^0\right)\right|,$$

$$(c) = \mathbb{P}(\mathcal{N})^{(k-1)} \left|\mathbb{P}(\tau_A > c) - \mathbb{P}(\mathcal{N})\right|.$$

First, for any measurable $B \in \mathcal{F}_{\{(\ell+1)c,(\ell+2)c+n-1\}}$, we have $\mathbb{P}(B) + \psi(\Delta) \leq 3\psi(\Delta) \leq \frac{3}{2}\varepsilon(A)$. We can also remark that $\mathbb{P}(\mathcal{N}) \geq \mathbb{P}(\tau_A > c) \geq \mathbb{P}\tau_A > 1/(2\mathbb{P}(A)) \geq 1/2$. Then, by iteration of the mixing property, we have the following inequality for all $\ell \in \mathbb{N}$:

$$\mathbb{P}\left(\bigcap_{i=0}^{\ell} \mathcal{N}_1^i; B\right) \leq 6\mathbb{P}(\mathcal{N})^{\ell+1}\varepsilon(A).$$

We apply this bound in the inequalities (14) and (15) of Abadi (2004) to get

$$(a) \leq \sum_{j=0}^{k-2} \mathbb{P}(\mathcal{N})^j \left(6\mathbb{P}(\mathcal{N})^{k-j-2+1}\varepsilon(A)\right) = 6(k-1)\varepsilon(A)\,\mathbb{P}(\mathcal{N})^{(k-1)},$$

$$(b) \leq \sum_{j=0}^{k-2} \mathbb{P}(\mathcal{N})^j \left(6\mathbb{P}(\mathcal{N})^{k-j-2+1}\varepsilon(A)\right) = 6(k-1)\varepsilon(A)\,\mathbb{P}(\mathcal{N})^{(k-1)}.$$

We also have $(c) \leq \mathbb{P}(\mathcal{N})^{k-1} \mathbb{P}\left(\mathcal{N}; \tau_A \circ T^{c-2\Delta} \leq 2\Delta\right) \leq \varepsilon(A) \mathbb{P}(\mathcal{N})^{k-1}$.
We obtain (4.1): $\left|\mathbb{P}(\tau_A > kc) - \mathbb{P}(\mathcal{N})^k\right| \leq 24k\varepsilon(A)\mathbb{P}(\mathcal{N})^k$.
We deduce (4.2): $\left|\mathbb{P}(\tau_A > kc) - \mathbb{P}(\tau_A > c)^k\right| \leq 25k\varepsilon(A)\mathbb{P}(\mathcal{N})^k$.
Then, $C_a = 24$ and $C_b = 25$. □

Theorem 4.3 (Theorem 1 in Abadi (2004))**.** *Let $(X_m)_{m\in\mathbb{Z}}$ be a ψ-mixing process. Then, there exist constants $C_h > 0$ and $0 < \Xi_1 < 1 \leq \Xi_2 < \infty$, such that for all $n \in \mathbb{N}$ and any $A \in \mathcal{C}_n$, there exists $\xi_A \in [\Xi_1, \Xi_2]$, for which the following inequality holds for all $t > 0$:*

$$\left|\mathbb{P}\left(\tau_A > \frac{t}{\xi_A}\right) - e^{-t\mathbb{P}(A)}\right| \leq C_h \varepsilon(A) f_1(A,t),$$

with $\varepsilon(A) = \inf_{n \leq \ell \leq \frac{1}{\mathbb{P}(A)}} [\ell\mathbb{P}(A) + \psi(\ell)]$ *and* $f_1(A,t) = (t\mathbb{P}(A) \vee 1)e^{-t\mathbb{P}(A)}$.

We prove an upper bound for the distance between the rescaled hitting time and the exponential law of expectation equal to one. The factor $\varepsilon(A)$ in the upper bound shows that the rate of convergence to the exponential law is given by a trade off between the length of this time and the velocity of loosing memory of the process.

Proposition 4.4. *We have $C_h = 105$.*

Proof. We fix $c = \frac{1}{2\mathbb{P}(A)}$ and Δ given by Proposition 4.1. We define

$$\xi_A = \frac{-\log \mathbb{P}(\tau_A > c - 2\Delta)}{c\mathbb{P}(A)}.$$

There are three steps in the proof of the theorem. First, we consider t of the form $t = kc$ with k a positive integer. Secondly, we prove the theorem for any t of the form $t = (k + p/q)c$ with k, p positive integers and $1 \leq p \leq q$ with $q = \left[\frac{1}{2\varepsilon(A)}\right]$, where $[.]$ defines the integer part of a real number. Finally, we consider the remaining cases. Here, for the sake of simplicity , we do not detail the two first steps (for that, see Abadi (2004)), but only the last one. Let t be any positive real number. We write $t = kc + r$, with k a positive integer and r such that $0 \leq r < c$. We can choose a $\bar{t}$ such that $\bar{t} < t$ and $\bar{t} = (k + p/q)c$ with p, q as before. Abadi (2004) obtains the following bound:

$$\begin{aligned}\left|\mathbb{P}(\tau_A > t) - e^{-\xi_A\mathbb{P}(A)t}\right| &\leq |\mathbb{P}(\tau_A > t) - \mathbb{P}(\tau_A > \bar{t})| + \left|\mathbb{P}(\tau_A > \bar{t}) - e^{-\xi_A\mathbb{P}(A)\bar{t}}\right| \\ &+ \left|e^{-\xi_A\mathbb{P}(A)\bar{t}} - e^{-\xi_A\mathbb{P}(A)t}\right|.\end{aligned}$$

The first term in the triangular inequality is bounded in the following way:

$$\begin{aligned}|\mathbb{P}(\tau_A > t) - \mathbb{P}(\tau_A > \bar{t})| &= \mathbb{P}\left(\tau_A > \bar{t}; \tau_A \circ T^{\bar{t}} \leq t - \bar{t}\right) \\ &\leq \mathbb{P}\left(\tau_A > kc; \tau_A \circ T^{\bar{t}} \leq \Delta\right) \\ &\leq \mathbb{P}(\mathcal{N})^{k-2}(\Delta\mathbb{P}(A) + \psi(\Delta))) \\ &\leq 4\mathbb{P}(\mathcal{N})^k\varepsilon(A) \\ &\leq 4\varepsilon(A)e^{-\xi_A\mathbb{P}(A)t}.\end{aligned}$$

The second term is bounded like in the two first steps of the proof in Abadi (2004). We apply inequalities (4.1) and (4.2) to obtain

$$\left|\mathbb{P}(\tau_A > \bar{t}) - e^{-\xi_A\mathbb{P}(A)\bar{t}}\right| \leq (3 + C_a t\mathbb{P}(A) + C_a + 2C_b)\varepsilon(A)e^{-\xi_A\mathbb{P}(A)t}.$$

Finally, with the definition of ξ_A and knowing thaht $0 \leq r < c$, the third term is bounded using the Mean Value Theorem (see for example Douglass (1996))

$$\left| e^{-\xi_A \mathbb{P}(A)\tilde{t}} - e^{-\xi_A \mathbb{P}(A)t} \right| \leq \xi_A \mathbb{P}(A) \left(r - \frac{p}{q} c \right) e^{-\xi_A \mathbb{P}(A)\tilde{t}} \leq \varepsilon(A) e^{-\xi_A \mathbb{P}(A)t}.$$

Thus we have $\left| \mathbb{P}\left(\tau_A > t\right) - e^{-\xi_A \mathbb{P}(A)t} \right| \leq 105 \varepsilon(A) f_1(A, \xi_A t)$ and the theorem follows by the change of variables $\tilde{t} = \xi_A t$. Then $C_h = 105$. □

Lemma 4.5. *$(X_m)_{m\in\mathbb{Z}}$ be a ψ-mixing process. Suppose that $B \subseteq A \in \mathcal{F}_{\{0,\ldots,b\}}$, $C \in \mathcal{F}_{\{b+g,\ldots,\infty\}}$ with $b, g \in \mathbb{N}$. The following inequality holds:*

$$\mathbb{P}_A(B \cap C) \leq \mathbb{P}_A(B)\mathbb{P}(C)(1 + \psi(g)).$$

Proof. Since $B \subseteq A$, obviously $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(B \cap C)$. By the ψ-mixing property $\mathbb{P}(B \cap C) \leq \mathbb{P}(B)(\mathbb{P}(C) + \psi(g))$. We divide the above inequality by $\mathbb{P}(A)$ and the lemma follows. □

For all the following propositions and lemmas, we recall that

$$e_\psi(A) = \inf_{1 \leq w \leq n_A} \left[(r_A + n) \mathbb{P}\left(A^{(w)} \right) \left(1 + \psi\left(n_A - w \right) \right) \right].$$

Proposition 4.6. *Let $(X_m)_{m\in\mathbb{Z}}$ be a ψ-mixing process. Let $A \in \mathcal{R}_p(n)$. We recall that p_A is the principal period of word A. Then the following holds:*

(a) *For all $M, M' \geq g \geq n$,*

$$\begin{aligned} & \left| \mathbb{P}_A\left(\tau_A > M + M'\right) - \mathbb{P}_A\left(\tau_A > M\right) \mathbb{P}\left(\tau_A > M'\right) \right| \\ \leq \quad & \mathbb{P}_A\left(\tau_A > M - g\right) 2g\mathbb{P}(A) \left[1 + \psi(g) \right], \end{aligned}$$

and similarly

$$\begin{aligned} & \left| \mathbb{P}_A\left(\tau_A > M + M'\right) - \mathbb{P}_A\left(\tau_A > M\right) \mathbb{P}\left(\tau_A > M' - g\right) \right| \\ \leq \quad & \mathbb{P}_A\left(\tau_A > M - g\right) \left[g\mathbb{P}(A) + 2\psi(g) \right]. \end{aligned}$$

(b) *For all $t \geq p_A \in \mathbb{N}$, with $\zeta_A = \mathbb{P}_A(\tau_A > p_A)$,*

$$\left| \mathbb{P}_A\left(\tau_A > t\right) - \zeta_A \mathbb{P}\left(\tau_A > t\right) \right| \leq 2e_\psi(A).$$

The above proposition establishes a relation between hitting and return times with an error bound uniform with respect to t. In particular, (b) says that these times coincide if and only if $\zeta_A = 1$, namely, the string A is non-self-overlapping.

Proof. In order to simplify notation, for $t \in \mathbb{Z}$, $\tau_A^{[t]}$ stands for $\tau_A \circ T^t$. We introduce a gap of length g after coordinate M to construct the following triangular inequality

$$\begin{aligned} & \left| \mathbb{P}_A\left(\tau_A > M + M'\right) - \mathbb{P}_A\left(\tau_A > M\right) \mathbb{P}\left(\tau_A > M'\right) \right| \\ \leq \quad & \left| \mathbb{P}_A\left(\tau_A > M + M'\right) - \mathbb{P}_A\left(\tau_A > M; \tau_A^{[M+g]} > M' - g\right) \right| \qquad (4.3) \\ + \quad & \left| \mathbb{P}_A\left(\tau_A > M; \tau_A^{[M+g]} > M' - g\right) - \mathbb{P}_A\left(\tau_A > M\right) \mathbb{P}\left(\tau_A > M' - g\right) \right| \qquad (4.4) \\ + \quad & \mathbb{P}_A\left(\tau_A > M\right) \left| \mathbb{P}\left(\tau_A > M' - g\right) - \mathbb{P}\left(\tau_A > M'\right) \right|. \qquad (4.5) \end{aligned}$$

Term (4.3) is bounded with Lemma 4.5 by

$$\mathbb{P}_A\left(\tau_A > M; \tau_A^{[M]} \leq g\right) \leq \mathbb{P}_A\left(\tau_A > M - g\right) g\mathbb{P}(A) \left[1 + \psi(g) \right].$$

Term (4.4) is bounded using the ψ-mixing property by $\mathbb{P}_A\left(\tau_A > M\right) \psi(g)$. The modulus in (4.5) is bounded using stationarity by $\mathbb{P}\left(\tau_A \leq g\right) \leq g\mathbb{P}(A)$. This ends the proof of both

inequalities of item (a).
Item (b) for $t \geq 2n$ is proven similarly to item (a) with $t = M + M'$, $M = p_A$, and $g = w$ with $1 \leq w \leq n_A$. Consider now $p_A \leq t < 2n$.

$$\zeta_A - \mathbb{P}_A(\tau_A > t) = \mathbb{P}_A(p < \tau_A \leq t) = \mathbb{P}_A(\tau_A \in \mathcal{R}(A) \cup (n \leq \tau_A \leq t)) \leq e_\psi(A).$$

The first equality follows directly by definition of p_A. The second one follows by definition of $\mathcal{R}(A)$ and the commentaries previous to Example 3.2. The inequality follows by an application of Lemma 4.5 with $B = A$, $C = \cup_{i \in \mathcal{R}(A) \cap \{n,\ldots,t\}} T^{-i} A^{(w)}$ and $g = n_A - w$. □

Let $\zeta_A = \mathbb{P}_A(\tau_A > p_A)$ and $h = 1/(2\mathbb{P}(A)) - 2\Delta$, then $\xi_A = -2\log \mathbb{P}(\tau_A > h)$.

Lemma 4.7. *Let $(X_m)_{m \in \mathbb{Z}}$ be a ψ-mixing process. Then the following inequality holds:*

$$|\xi_A - \zeta_A| \leq 11 e_\psi(A).$$

Hence, we have

$$\zeta_A - 11 e_\psi(A) \leq \xi_A \leq \zeta_A + 11 e_\psi(A).$$

Proof.

$$\begin{aligned}
\mathbb{P}(\tau_A > h) &= \prod_{i=1}^{h} \mathbb{P}(\tau_A > i | \tau_A > i-1) = \prod_{i=1}^{h} (1 - \mathbb{P}\left(T^{-i}(A) | \tau_A > i-1\right)) \\
&= \prod_{i=1}^{h} (1 - \rho_i \mathbb{P}(A)),
\end{aligned}$$

where $\rho_i \overset{def}{=} \dfrac{\mathbb{P}_A(\tau_A > i-1)}{\mathbb{P}(\tau_A > i-1)}$. Therefore

$$\begin{aligned}
&\left| \xi_A + 2 \sum_{i=1}^{p_A} \log(1 - \rho_i \mathbb{P}(A)) - 2 \sum_{i=p_A+1}^{h} \zeta_A \mathbb{P}(A) \right| \\
\leq\ & 2 \sum_{i=p_A+1}^{h} |-\log(1 - \rho_i \mathbb{P}(A)) - \zeta_A \mathbb{P}(A)|.
\end{aligned}$$

The above modulus is bounded by

$$|-\log(1 - \rho_i \mathbb{P}(A)) - \rho_i \mathbb{P}(A)| + |\rho_i - \zeta_A| \, \mathbb{P}(A).$$

Now note that $|y - (1 - e^{-y})| \leq (1 - e^{-y})^2$ for $y > 0$ small enough. Apply it with $y = -\log(1 - \rho_i \mathbb{P}(A))$ to bound the most left term of the above expression by $(\rho_i \mathbb{P}(A))^2$. Further by Proposition 4.6 (b) and the fact that $\mathbb{P}(\tau_A > h) \geq 1/2$ (see in Proposition 4.2 that $\mathbb{P}(\mathcal{N}) \geq 1/2$) we have

$$|\rho_i - \zeta_A| \leq \frac{2 e_\psi(A)}{\mathbb{P}(\tau_A > h)} \leq 4 e_\psi(A).$$

for all $i = p_A + 1, \ldots, h$. Yet as before

$$-\sum_{i=1}^{p_A} \log(1 - \rho_i \mathbb{P}(A)) \leq p_A \left(\rho_i \mathbb{P}(A) + (\rho_i \mathbb{P}(A))^2\right) \leq e_\psi(A).$$

Finally, by definition of h

$$\left| 2 \sum_{i=p_A+1}^{h} \zeta_A \mathbb{P}(A) - \zeta_A \right| \leq 4\Delta\mathbb{P}(A) + 2p_A\mathbb{P}(A) \leq 6e_\psi(A).$$

This ends the proof of the lemma. □

Proposition 4.8. *Let $(X_m)_{m\in\mathbb{Z}}$ be a ψ-mixing process. Then the following inequality holds:*

$$|\mathbb{P}(\tau_A > t) - e^{-t\mathbb{P}(A)}| \leq C_p e_\psi(A)(t\mathbb{P}(A) \vee 1)e^{-(\zeta_A - 11e_\psi(A))t\mathbb{P}(A)}.$$

Proof. We bound the first term with Theorem 4.3 and the second with Lemma 4.7 :

$$\begin{array}{rcl} |\mathbb{P}(\tau_A > t) - e^{-t\mathbb{P}(A)}| & \leq & |\mathbb{P}(\tau_A > t) - e^{-\xi_A t\mathbb{P}(A)}| + |e^{-\xi_A t\mathbb{P}(A)} - e^{-t\mathbb{P}(A)}| \\ |\mathbb{P}(\tau_A > t) - e^{-\xi_A t\mathbb{P}(A)}| & \leq & C_h\varepsilon(A)e^{-\xi_A t\mathbb{P}(A)} \leq C_h e_\psi(A)e^{-(\zeta_A - 11e_\psi(A))t\mathbb{P}(A)} \\ |e^{-\xi_A t\mathbb{P}(A)} - e^{-t\mathbb{P}(A)}| & \leq & t\mathbb{P}(A)|\xi_A - 1|e^{-\min\{1,\xi_A\}t\mathbb{P}(A)} \\ & \leq & 11t\mathbb{P}(A)e_\psi(A)e^{-(\zeta_A - 11e_\psi(A))t\mathbb{P}(A)}. \end{array}$$

This ends the proof of the proposition with $C_p = C_h + 11$. □

Definition 4.9. Given $A \in \mathcal{C}_n$, we define for $j \in \mathbb{N}$, the j-th occurrence time of A as the random variable $\tau_A^{(j)} : \Omega \to \mathbb{N} \cup \{\infty\}$, defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ as follows: for any $x \in \Omega$, $\tau_A^{(1)}(x) = \tau_A(x)$ and for $j \geq 2$,

$$\tau_A^{(j)}(x) = \inf\{k > \tau_A^{(j-1)}(\omega) : T^k(x) \in A\}.$$

Proposition 4.10. *Let $(X_m)_{m\in\mathbb{Z}}$ be a ψ-mixing process. Then, for all $A \notin \mathcal{B}_n$, all $k \in \mathbb{N}$, and all $0 \leq t_1 < t_2 < ... < t_k \leq t$ for which $\min_{2\leq j\leq k}\{t_j - t_{j-1}\} > 2n$, there exists a positive constant C_1 independent of A, n, t and k such that*

$$\begin{array}{rcl} & & \left| \mathbb{P}\left(\bigcap_{j=1}^{k} \left(\tau_A^{(j)} = t_j\right); \tau_A^{(k+1)} > t \right) - \mathbb{P}(A)^k \prod_{j=1}^{k+1} \mathcal{P}_j \right| \\ & \leq & C_1 k(\mathbb{P}(A)(1 + \psi(n)))^k e_\psi(A)e^{-(t-(3k+1)n)\mathbb{P}(A)} \end{array}$$

where $\mathcal{P}_j = \mathbb{P}(\tau_A > (t_j - t_{j-1}) - 2n)$.

Proof. We will show this proposition by induction on k. We put $\Delta_j = t_j - t_{j-1}$ for $j = 2, ..., k$, $\Delta_1 = t_1$ and $\Delta_{k+1} = t - t_k$. Firstly, we note that by stationarity

$$\mathbb{P}(\tau_A = t) = \mathbb{P}(A; \tau_A > t - 1).$$

For $k = 1$, by a triangular inequality we obtain

$$\begin{array}{rcll} & & \left| \mathbb{P}\left(\tau_A = t_1; \tau_A^{(2)} > t\right) - \mathbb{P}(A) \prod_{j=1}^{2} \mathcal{P}_j \right| & \\ & \leq & \left| \mathbb{P}\left(\tau_A = t_1; \tau_A^{(2)} > t\right) - \mathbb{P}\left(\tau_A = t_1; N_{t_1+2n}^t = 0\right) \right| & (4.6) \\ & + & \left| \mathbb{P}\left(\tau_A = t_1; N_{t_1+2n}^t = 0\right) - \mathbb{P}\left(\tau_A = t_1\right)\mathcal{P}_2 \right| & (4.7) \\ & + & \left| \mathbb{P}(A; \tau > t_1 - 1) - \mathbb{P}\left(A; N_{2n}^{t_1-1} = 0\right) \right| \mathcal{P}_2 & (4.8) \\ & + & \left| \mathbb{P}\left(A; N_{2n}^{t_1-1} = 0\right)\mathcal{P}_2 - \mathbb{P}(A) \prod_{j=1}^{2} \mathcal{P}_j \right|. & (4.9) \end{array}$$

Term (4.6) is equal to $\mathbb{P}\left(\tau_A = t_1; \bigcup_{i=t_1+1}^{t_1+2n} T^{-i}(A); N_{t_1+2n}^t = 0\right)$ and then

$$(4.6) = \mathbb{P}\left(A; \bigcup_{i\in\mathcal{R}(A)\cup i=1}^{2n} T^{-i}(A); N_{2n}^t = 0\right).$$

Since $A \notin \mathcal{B}_n$, for $1 \le i < p_A$, the above probability is zero. Thus, using mixing property

$$\begin{aligned}(4.6) &\le \mathbb{P}\left(A; \bigcup_{i\in\mathcal{R}(A)\cup i=p_A}^{2n} T^{-i}(A); N_{2n}^t = 0\right)\\ &\le 2\mathbb{P}(A)\mathbb{P}(A)(r_A+n)(1+\psi(n))\mathbb{P}\left(N_{2n}^t = 0\right)\\ &\le 2\mathbb{P}(A)e_\psi(A)e^{-(t-(3k+1)n)\mathbb{P}(A)}.\end{aligned}$$

Term (4.7) is bounded using ψ-mixing property

$$\begin{aligned}(4.7) &\le \psi(n)(1+\psi(n))\mathbb{P}(A)\mathcal{P}_1\mathcal{P}_2\\ &\le \psi(n)\mathbb{P}(A)e_\psi(A)e^{-(t-(3k+1)n)\mathbb{P}(A)}.\end{aligned}$$

Analogous computations are used to bound terms (4.8) and (4.9).

Now, let us suppose that the proposition holds for $k-1$ and let us prove it for k. We put $\mathcal{S}_i = \{\tau_A^{(i)} = t_i\}$. We use a triangular inequality again to bound the term in the left hand side of the inequality of the proposition by a sum of five terms:

$$\left|\mathbb{P}\left(\bigcap_{j=1}^{k}\left(\tau_A^{(j)} = t_j\right); \tau_A^{(k+1)} > t\right) - \mathbb{P}(A)^k\prod_{j=1}^{k+1}\mathcal{P}_j\right| \le I + II + III + IV + V.$$

$$\begin{aligned}I &= \left|\mathbb{P}\left(\bigcap_{j=1}^{k}\mathcal{S}_j; \tau_A^{(k+1)} > t\right) - \mathbb{P}\left(\bigcap_{j=1}^{k-1}\mathcal{S}_j; N_{t_{k-1}+1}^{t_k-2n} = 0; T^{-t_k}(A); N_{t_k+1}^t = 0\right)\right|\\ &= \mathbb{P}\left(\bigcap_{j=1}^{k-1}\mathcal{S}_j; N_{t_{k-1}+1}^{t_k-2n} = 0; \bigcup_{i=t_k-2n+1}^{t_k-1} T^{-i}(A); T^{-t_k}(A); N_{t_k+1}^t = 0\right)\\ &\le \left(\mathbb{P}(A)(1+\psi(n))\right)^k(1-\psi(n))\left(np_A + (r_A+n)\mathbb{P}(A^{(w)})\right)e^{-(t-(3k+1)n)\mathbb{P}(A)},\\ II &= \left|\mathbb{P}\left(\bigcap_{j=1}^{k-1}\mathcal{S}_j; N_{t_{k-1}+1}^{t_k-2n} = 0; T^{-t_k}(A); N_{t_k+1}^t = 0\right)\right.\\ &\quad\left. - \mathbb{P}\left(\bigcap_{j=1}^{k-1}\mathcal{S}_j; N_{t_{k-1}+1}^{t_k-2n} = 0\right)\mathbb{P}\left(A; N_1^{t-t_k} = 0\right)\right|\\ &\le \mathbb{P}\left(\bigcap_{j=1}^{k-1}; N_{t_{k-1}+1}^{t_k-2n} = 0\right)\mathbb{P}\left(A; N_1^{t-t_k} = 0\right)\psi(n)\\ &\le \left(\mathbb{P}(A)(1+\psi(n))\right)^k\psi(n)e^{-(t-(3k+1)n)\mathbb{P}(A)},\\ III &= \left|\mathbb{P}\left(\bigcap_{j=1}^{k-1}\mathcal{S}_j; N_{t_{k-1}+1}^{t_k-2n} = 0\right) - \mathbb{P}\left(\bigcap_{j=1}^{k-1}\mathcal{S}_j; N_{t_{k-1}+1}^{t_k-1} = 0\right)\right|\mathbb{P}\left(A; N_1^{t-t_k} = 0\right)\\ &\le \mathbb{P}\left(\bigcap_{j=1}^{k-1}\mathcal{S}_j; N_{t_{k-1}+1}^{t_k-2n} = 0; \bigcup_{t_k-2n+1}^{t_k-1} T^{-i}(A)\right)\mathbb{P}(A)\\ &\le 2\mathbb{P}(A)\left(\mathbb{P}(A)(1+\psi(n))\right)^k e^{-(t-(3k+1)n)\mathbb{P}(A)}.\end{aligned}$$

We use the inductive hypothesis for the term IV and the case with $k=1$ for the term V.

$$\begin{aligned} IV &= \left| \mathbb{P}\left(\bigcap_{j=1}^{k-1} \mathcal{S}_j; N_{t_{k-1}+1}^{t_k-1} = 0 \right) - \mathbb{P}(A)^{k-1} \prod_{j=1}^{k} \mathcal{P}_j \right| \mathbb{P}\left(A; N_1^{t-t_k} = 0\right) \\ &\leq C_1(k-1)(\mathbb{P}(A)(1+\psi(n)))^k e_\psi(A) e^{-(t-(3k+1)n)\mathbb{P}(A)}, \\ V &= \mathbb{P}(A)^{k-1} \prod_{j=1}^{k} \mathcal{P}_j \left| \mathbb{P}\left(A; N_1^{t-t_k} = 0\right) - \mathbb{P}(A)\mathcal{P}_{k+1} \right| \\ &\leq 2(\mathbb{P}(A)(1+\psi(n)))^k e_\psi(A) e^{-(t-(3k+1)n)\mathbb{P}(A)}. \end{aligned}$$

Finally, we obtain

$$I + II + III + IV + V \leq (3 + C_1(k-1) + 2)(\mathbb{P}(A) + \psi(n))^k e_\psi(A).$$

To conclude the proof, it is sufficient that $C_1 k = 3 + C_1(k-1) + 2$, therefore $C_1 = 5$. This ends the proof of the proposition. □

5. Proof of Theorem 3.3

In this section, we prove the main result of our work (see Section 3.2): an upper bound for the difference between the exact distribution of the number of occurrences of word A and the Poisson distribution of parameter $t\mathbb{P}(A)$. Throughout the proof, we will note in italic the terms computed by our software PANOW (see Section 6.1).

Proof. For $k=0$, the result comes from Proposition 4.8 ($\mathbb{P}(N^t=0) = \mathbb{P}(\tau_A > t)$). For $k > 2t/n$, since $A \notin \mathcal{B}_n$, we have $\mathbb{P}(N^t = k) = 0$. Hence,

$$\begin{aligned} \left| \mathbb{P}(N^t = k) - \frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!} \right| &= \frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!} \\ &\leq \frac{(t\mathbb{P}(A))^{k-1}}{(k-1)!} \frac{t\mathbb{P}(A)}{k} \\ &\leq \frac{1}{2} \frac{(t\mathbb{P}(A))^{k-1}}{(k-1)!} e_\psi(A). \end{aligned}$$

Indeed, since $\frac{t}{k} < \frac{n}{2}$ then $\frac{t\mathbb{P}(A)}{k} < \frac{n\mathbb{P}(A)}{2} \leq \frac{e_\psi(A)}{2}$.

Now, let us consider $1 \leq k \leq 2t/n$. We consider a sequence which contains exactly k occurrences of A. These occurrences can be isolated or can be in clumps. We define the following set:

$$\mathcal{T} = \mathcal{T}(t_1, t_2, ..., t_k) = \left\{ \bigcap_{j=1}^{k} (\tau_A^{(j)} = t_j); \tau_A^{(k+1)} > t \right\}.$$

We recall that we put $\mathcal{P}_j = \mathbb{P}(\tau_A > (t_j - t_{j-1}) - 2n)$, $\Delta_j = t_j - t_{j-1}$ for $j = 2, ..., k$, $\Delta_1 = t_1$ and $\Delta_{k+1} = t - t_k$. Define $I(\mathcal{T}) = \min_{2 \leq j \leq k} \{\Delta_j\}$. We say that the occurrences of A are isolated if $I(\mathcal{T}) \geq 2n$ and we say that there exists at least one clump if $I(\mathcal{T}) < 2n$. We also denote

$$B_k = \{\mathcal{T} | I(\mathcal{T}) < 2n\} \quad \text{and} \quad G_k = \{\mathcal{T} | I(\mathcal{T}) \geq 2n\}.$$

The set $\{N^t = k\}$ is the disjoint union between B_k and G_k, then

$$\mathbb{P}(N^t = k) = \mathbb{P}(B_k) + \mathbb{P}(G_k),$$

$$\left|\mathbb{P}(N^t=k)-\frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!}\right|\leq \mathbb{P}(B_k)+\left|\mathbb{P}(G_k)-\frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!}\right|.$$

We will prove an upper bound for the two quantities on the right hand side of the above inequality to conclude the proof of the theorem.

We prove an upper bound for $\mathbb{P}(\mathbf{B_k})$. Define $C(\mathcal{T})=\sum_{j=2}^{k}\mathbb{1}_{\{\Delta_j>2n\}}+1$. $C(\mathcal{T})$ computes how many clusters there are in a given $\mathcal{T}$. Suppose that $\mathcal{T}$ is such that $C(\mathcal{T})=1$ and fix the position t_1 of the first occurrence of A. Further, each occurrence inside the cluster (with the exception of the most left one which is fixed at t_1) can appear at distance d of the previous one, with $p_A\leq d\leq 2n$. Therefore, the ψ-mixing property leads to the bound

$$\begin{aligned}\mathbb{P}\left(\bigcup_{t_2,\ldots,t_k}\mathcal{T}(t_1,t_2,\ldots,t_k)\right) &\leq \mathbb{P}\left(\bigcap_{j=1}^{k}\bigcup_{\substack{n/2\leq t_{i+1}-t_i\leq 2n;\\ i=2,\ldots,k}}T^{-t_j}(A)\right) \qquad (5.1)\\ &\leq \mathbb{P}(A)e_\psi(A)^{k-1}e_\psi(A)e^{-(t-(3k+1)n)\mathbb{P}(A)}.\end{aligned}$$

Suppose now that $\mathcal{T}$ is such that $C(\mathcal{T})=i$. Assume also that the most left occurrence of the i clusters of $\mathcal{T}$ occurs at $t(1),\ldots,t(i)$, with $1\leq t(1)<\cdots<t(i)\leq t$ fixed. By the same argument used above, we have the inequalities

$$\begin{aligned}&\mathbb{P}\left(\bigcup_{\{t_1,\ldots,t_k\}\setminus\{t(1),\ldots,t(i)\}}\mathcal{T}(t_1,\ldots,t_k)\right)\\ \leq\ &\left(\mathbb{P}(A)(1+\psi(n))\right)^{i-1}e_\psi(A)^{k-i}e^{-(t-(3k+1)n)\mathbb{P}(A)}.\end{aligned}$$

To obtain an upper bound for $\mathbb{P}(B_k)$ we must sum the above bound over all $\mathcal{T}$ such that $C(\mathcal{T})=i$ with i running from 1 to $k-1$. Fixed $C(\mathcal{T})=i$, the locations of the most left occurrences of A of each one of the i clusters can be chosen in at most C_t^i many ways. The cardinality of each one of the i clusters can be arranged in C_{k-1}^{i-1} many ways. (This corresponds to breaking the interval $(1/2,k+1/2)$ in i intervals at points chosen from $\{1+1/2,\ldots,k-1/2\}$.) Collecting these informations, we have that $\mathbb{P}(B_k)$ is bounded by

$$\begin{aligned}&\sum_{i=1}^{k-1}C_t^iC_{k-1}^{i-1}\left(\mathbb{P}(A)(1+\psi(n))\right)^ie_\psi(A)^{k-i}e^{-(t-(3k+1)n)\mathbb{P}(A)}\\ \leq\ &e^{-(t-(3k+1)n)\mathbb{P}(A)}e_\psi(A)^k\max_{1\leq i\leq k-1}\frac{\left(\lambda/e_\psi(A)\right)^i}{i!}\sum_{i=1}^{k-1}C_{k-1}^{i-1}\\ \leq\ &e^{-(t-(3k+1)n)\mathbb{P}(A)}e_\psi(A)\begin{cases}\dfrac{(2\lambda)^{k-1}}{(k-1)!} & k<\frac{\lambda}{e_\psi(A)}\\ \dfrac{(2\lambda)^{k-1}}{\left(\frac{\lambda}{e_\psi(A)}\right)!\left(\frac{\lambda}{e_\psi(A)}\right)^{k-1-\frac{\lambda}{e_\psi(A)}}} & k\geq\frac{\lambda}{e_\psi(A)}\end{cases}.\end{aligned}$$

This ends the proof of the bound for $\mathbb{P}(B_k)$.

We compute $\mathbb{P}(B_k)\leq\sum_{i=1}^{k-1}C_t^iC_{k-1}^{i-1}\left(\mathbb{P}(A)(1+\psi(n))\right)^ie_\psi(A)^{k-i}e^{-(t-(3k+1)n)\mathbb{P}(A)}$.

We prove an upper bound for $\left|\mathbb{P}(G_k) - \frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!}\right|$**.** It is bounded by four terms by the triangular inequality

$$\sum_{T\in G_k}\left|\mathbb{P}\left(\bigcap_{j=1}^{k}\left(\tau_A^{(j)}=t_j\right);\tau_A^{(k+1)}>t\right)-\mathbb{P}(A)^k\prod_{j=1}^{k+1}\mathcal{P}_j\right| \tag{5.2}$$

$$+\quad\sum_{T\in G_k}\mathbb{P}(A)^k\left|\prod_{j=1}^{k+1}\mathcal{P}_j-\prod_{j=1}^{k+1}e^{-(\Delta_j-2n)\mathbb{P}(A)}\right| \tag{5.3}$$

$$+\quad\sum_{T\in G_k}\mathbb{P}(A)^k\left|e^{-(t-2(k+1)n)\mathbb{P}(A)}-e^{-t\mathbb{P}(A)}\right| \tag{5.4}$$

$$+\quad\left|\frac{\#G_k k!}{t^k}\frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!}-\frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!}\right|. \tag{5.5}$$

We will bound these terms to obtain Theorem 3.3.
First, we bound the cardinal of G_k

$$\#G_k \le C_t^k \le \frac{t^k}{k!}.$$

Term (5.2) is bounded with Proposition 4.10

$$(5.2)\le C_1\frac{t^k}{(k-1)!}(\mathbb{P}(A)(1+\psi(n)))^k e_\psi(A)e^{-(t-(3k+1)n)\mathbb{P}(A)}.$$

Term (5.3) is bounded with Proposition 4.8

$$\begin{aligned}(5.3)\quad&\le\quad\frac{t^k}{k!}\mathbb{P}(A)^k\sum_{j=1}^{k+1}\prod_{i=1}^{j-1}\mathcal{P}_i\left|\mathcal{P}_j-e^{-(\Delta_j-2n)\mathbb{P}(A)}\right|\prod_{i=j+1}^{k+1}e^{-(\Delta_i-2n)\mathbb{P}(A)}\\&\le\quad\frac{t^k}{k!}\mathbb{P}(A)^k(k+1)C_p e_\psi(A)e^{-(\zeta_A-11e_\psi(A))t\mathbb{P}(A)}\\&\le\quad 2C_p\frac{(t\mathbb{P}(A))^k}{(k-1)!}e_\psi(A)e^{-(\zeta_A-11e_\psi(A))t\mathbb{P}(A)}\end{aligned}$$

where C_p is defined in Proposition 4.8.
We compute

$$\begin{aligned}(5.3)\quad&\le\quad\frac{(t\mathbb{P}(A))^k}{(k-1)!}\frac{k+1}{k}\\&\quad\left[(8+C_a t\mathbb{P}(A)+C_a+2C_b)\varepsilon(A)+11t\mathbb{P}(A)e_\psi(A)\right]e^{-(\zeta_A-11e_\psi(A))t\mathbb{P}(A)}.\end{aligned}$$

Term (5.4) is bounded by

$$(5.4)\le\frac{t^k}{k!}\mathbb{P}(A)^k(k+1)2n\mathbb{P}(A)e^{-t\mathbb{P}(A)}e^{2(k+1)n\mathbb{P}(A)}.$$

To bound term (5.5), we bound the following difference

$$\left|\frac{\#G_k k!}{t^k}-1\right|\le\left|\frac{(t-k(4n))^k}{t^k}-1\right|\le\frac{k\,(k+4n)}{t}.$$

Then, we have

$$(5.5) \leq \frac{k\,(k+4n)}{t}\,\frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!}.$$

Now, we just have to add the five bounds to obtain the theorem with the constant $C_\psi = 1 + C_1 + 2C_p + 8 + 8$. Proposition 4.10 shows that $C_1 = 5$ and Proposition 4.8 with Theorem 4.3 that $C_p = 116$. Then, we prove the theorem with $C_\psi = 254$. □

6. Biological applications

With the explicit value of the constant C_ψ of Theorem 3.3, and more particularly thanks to all the intermediary bounds given in the proof of this theorem, we can develop an algorithm to apply this formula to the study of rare words in biological sequences. In order to compare different methods, we also compute the bounds corresponding to a ϕ-mixing, process for which a proof of Poisson approximation is given in (Abadi and Vergne, in preparation). Let us recall the definition of such a mixing process.

Definition 6.1. Let $\phi = (\phi(\ell))_{\ell \geq 0}$ be a sequence decreasing to zero. We say that $(X_m)_{m\in\mathbb{Z}}$ is a ϕ-mixing process if for all integers $\ell \geq 0$, the following holds

$$\sup_{n\in\mathbb{N}, B\in\mathcal{F}_{\{0,.,n\}}, C\in\mathcal{F}_{\{n\geq 0\}}} \frac{|\mathbb{P}(B\cap T^{-(n+\ell+1)}(C)) - \mathbb{P}(B)\mathbb{P}(C)|}{\mathbb{P}(B)} = \phi(\ell),$$

where the supremum is taken over the sets B and C, such that $\mathbb{P}(B) > 0$.

Note that obviously, ψ-mixing implies ϕ-mixing. Then, we obtain two new methods for the detection of over- or under-represented words in biological sequences and we compare them to the Chen-Stein method.

We recall that Markov models are ψ-mixing processes and then also ϕ-mixing processes. Then, we first need to know the functions ψ and ϕ for a Markov model. It turns out that we can use

$$\psi(\ell) = \phi(\ell) = K\nu^\ell \text{ with } K > 0 \text{ and } 0 < \nu < 1,$$

where K and ν have to be estimated (see Meyn and Tweedie (1993)). There are several estimations of K and ν. We choose ν equal to the second eigenvalue of the transition matrix of the model and $K = \left(\inf_{j\in\{1,\ldots,|\mathcal{A}|^k\}} \mu_j\right)^{-1}$ where $|\mathcal{A}|$ is the alphabet size, k the order of the Markov model and μ the stationary distribution of the Markov model.

We recall that we aim at guessing a relevant biological role of a word in a sequence using its number of occurrences. Thus we compare the number of occurrences expected in the Markov chain that models the sequence and the observed number of occurrences. It is recommended to choose a degree of significance s to quantify this relevance. We fix arbitrarily a degree of significance and we want to calculate the smallest number of occurrences u necessary for $\mathbb{P}(N > u) < s$, where N is the number of occurrences of the studied word. If the number of occurrences counted in the sequence is larger than this u, we can consider the word to be relevant with a degree of significance s. We have

$$\mathbb{P}(N > u) \leq \sum_{k=u}^{+\infty} \left(\mathbb{P}_{\mathcal{P}}(N = k) + Error(k)\right)$$

where $\mathbb{P}_{\mathcal{P}}(N = k)$ is the probability under the Poisson model that N is equal to k and $Error(k)$ is the error between the exact distribution and its Poisson approximation,

bounded using Theorem 3.3. Then, we search the smallest threshold u such that

$$\sum_{k=u}^{+\infty} (\mathbb{P}_{\mathcal{P}}(N = k) + Error(k)) < s. \tag{6.1}$$

Then, we have $\mathbb{P}(N > u) < s$ and we consider the word relevant with a degree of significance s if it appears more than u times in the sequence.

In order to compare the different methods, we compare the thresholds that they give. Obviously, the smaller the degree of significance, the more relevant the studied word is. But for a fixed degree of significance, the best method is the one which gives the smallest threshold u. Indeed, to give the smallest u is equivalent to give the smallest error in the tail of the distribution between the exact distribution of the number of occurrences of word A and the Poisson distribution with parameter $t\mathbb{P}(A)$.

6.1. *Software availability.* We developed `PANOW`, dedicated to the determination of threshold u for given words. This software is written in ANSI `C++` and developed on x86 GNU/Linux systems with GCC 3.4, and successfully tested with GCC latest versions on Sun and Apple Mac OSX systems. It relies on `seq++` library (Miele et al. (2005)).

Compilation and installation are compliant with the GNU standard procedure. It is available at `http://stat.genopole.cnrs.fr/sg/software/panow/`. Online documentation is also available. `PANOW` is licensed under the GNU General Public License (`http://www.gnu.org`).

6.2. *Comparisons between the three different methods.*

6.2.1. *Comparisons using synthetic data.* We can compare the mixing methods and the Chen-Stein method through the values of threshold u obtained with `PANOW` using (Abadi and Vergne, in preparation) in the first case and Reinert and Schbath (1998) in the second one. We recall that the method which gives the smallest threshold u is the best method for a fixed degree of significance. Table 6.2 offers a good outline of the possibilities and limits of each method. It displays some results on different words randomly selected (no biological meaning for any of these words). Table 6.2 has been obtained with an order one

TABLE 6.2. Table of thresholds u obtained by the three methods (sequence length t equal to 10^6). For each one of the three methods and for each word, we compute the threshold which permits to consider the word as an over-represented word or not, for degree of significance s equal to 0.1 or 0.01. IMP means that the method can not return a result.

Words	$t = 10^6$					
	$s = 0.1$			$s = 0.01$		
	CS	ϕ	ψ	CS	ϕ	ψ
cccg	IMP	IMP	IMP	IMP	IMP	IMP
aagcgc	IMP	1301	378	IMP	1304	392
cgagcttc	18	38	18	IMP	40	22
ttgggctg	14	27	14	18	29	17
gtgcggag	16	32	16	22	34	20
agcaaata	19	39	19	IMP	41	23

Markov model using a random transition matrix and for a degree of significance of 0.1 and 0.01. IMP means that the method can not return a result. There are several reasons for that and we explain them in the following paragraph. Analysing many results, we notice some differences between the methods.

Firstly, none of the methods gives us a result in all the cases. We recall that the Chen Stein method gives a bound (CS) using the total variation distance. If the degree of significance s that we choose is smaller than the bound of Chen-Stein, we never find a threshold u such that

$$CS + \sum_{k=u}^{+\infty} \mathbb{P}_{\mathcal{P}}(N = k) < s.$$

Then, each time that the given bound is higher than the significance degree, use of the Chen Stein method is impossible. Therefore there are many examples that we can not study with this method. Obviously, it is interesting to have a small degree of significance s and that may be impossible by this restriction of the Chen-Stein method. For example, this problem appears for the words `aagcgc` and `cgagcttc` in Table 6.2. For this second word, the Chen-Stein bound is equal to 0.0107954. Hence, we can use this method for a significance degree s equal to 0.1 but not for a significance degree of 0.01. The same phenomena appears for the word `agcaaata` (the Chen-Stein bound is equal to 0.0120193).

The ϕ- and ψ-mixing methods are not based on the total variation distance. Then, whatever the degree of significance s and if the studied word satisfies the three following weak properties, we always give a threshold u, contrary to the Chen Stein method. In spite of these three conditions, our methods enable us to study a much broader panel of words than the Chen-Stein method. Indeed, for these two methods, the only problematic cases arise either when function e_ψ (see Theorem 3.3) is larger than 1 or for a "high" parameter of the Poisson distribution ("high" means larger than 500) or when the word periodicity is smaller than half its length (see assumptions in Theorem 3.3: $A \notin \mathcal{B}_n$). In fact, the first case does not occur very frequently (in any case in Table 6.2). The reason why the function e_ψ (or a similar function in the ϕ-mixing case) has to be smaller than 1 is that, for numerical reasons, the error term has to be decreasing with the number of occurrences k and without this condition on e_ψ we can not ensure this decrease. We have to compute error terms for a finite number of values of k but in order to reduce the computation time, when error term becomes smaller than a certain value (we choose 10^{-300}), we suppose all the following error terms equals to this value. That is why error term has to be decreasing. The second problem, a "high" parameter of the Poisson distribution, is just a computational difficulty and once again it does not occur very frequently (only for the word `cccg` in Table 6.2 for instance). We would like to insist on the main advantage of our methods: we can fix any significance degree s and, except in the very rare cases mentioned above, we will find a threshold u, contrary to the Chen-Stein method.

Also, we can use our methods for any Markov chain order. Indeed, `PANOW` runs fast enough contrary to the R program used to compute the Chen-Stein bound of Reinert and Schbath (1998). Note that, in program `PANOW`, we give another method to compute the Chen-Stein bound (see Abadi (2001b)) and this method gives approximately the same Chen-Stein bound.

The second main observation we can make is that, when it works, the Chen-Stein method gives either a similar threshold u than the ψ-mixing method, or a smaller one. This means that the ψ-mixing method out-performs the Chen-Stein method.

Thirdly we notice that the ψ-mixing method is always better than the ϕ-mixing one. Obviously, this result was expected by the definitions of these mixing processes and also by

the theorems because of the extra factor $e^{-(t-(3k+1)n)\mathbb{P}(A)}$ (see Theorem 3.3 and Theorem 2 in (Abadi and Vergne, in preparation). We are interested by the real impact of this factor on the threshold u: it is significantly better in the case of a ψ-mixing process.

Finally, let us remember you that Chen-Stein method give any result in a great number of cases where our method works. And it is more the case when our model of interest is a Markov model of order greater than 2. Indeed, Chen-Stein bounds for Markov model of order greater than 2 are very high and then cannot give any result whereas our local method works easily.

6.2.2. *Biological comparisons.* Now, we present a few results obtained on real biological examples with order one Markov models. There are many categories of words which have relevant biological functions (promoters, terminators, repeat sequences, chi sites, uptake sequences, bend sites, signal peptides, binding sites, restriction sites, ...). Some of them are highly present in the sequence, some others are almost absent. Then, it turns out to be interesting to consider the over or the under-representation of words to find words biologically relevant.

In this section, we test our methods on words already known to be relevant. We focus our study on Chi sites or uptake sequences. Chi sites of bacterias protect the genome by stopping its degradation performed by a particular enzyme. The function of this enzyme is to destroy viruses which could appear into the bacteria. Viruses do not contain Chi sites and then are exterminated. It turns out that Chi sites are highly present in the bacterial genome. Uptake sequences are abundant sequence motifs, often located downstream of ORFs, that are used to facilitate the within-species horizontal transfer of DNA.

Example 1

First, we consider the Chi of *Escherichia coli*, `gctggtgg`, (see Table 6.3), for different degrees of significance. We use complete sequence of *Escherichia coli K12* (Blattner et al. (1997)). Sequence length is equal to 4639221. We recall that for a fixed significance

TABLE 6.3. Table of thresholds u obtained by the three methods for the Chi of *Escherichia coli*: `gctggtgg` (sequence length t equal to 4639221). For each one of the three methods we compute the threshold which permits to consider the word as an over-represented word or not, for degree of significance s. IMP means that the method can not return a result. "counts" correspond to the number of occurrences observed in the sequence.

s	Chen-Stein	ϕ-mixing	ψ-mixing	counts
0.1	87	193	83	499
0.01	IMP	195	92	499
0.0001	IMP	197	99	499
10^{-239}	IMP	549	498	499

degree, the smaller the threshold u, the best the method is. Then, we can conclude that the ψ-mixing method gives the most interesting results. Chi of *E. coli* could be considered as an over-represented one from 99 occurrences for a significance degree s of 0.0001. Because Chen-Stein bound is equal to 0.067726, Chen-Stein method does not permit to conclude for significance degrees of 0.01 and 0.001. Moreover, it is well known that Chi of *E. coli* is a very relevant word in this bacteria. Then, we expect a very small

significance degree for this word. Unfortunately, the minimal significance degree which could be obtained by Chen-Stein method is, in fact, the Chen-Stein bound: 0.067726. Our method allows to obtain very small significance degree and the minimal significance degree for which Chi of *E. coli* is considered as an over-represented word by the ψ-mixing method, is given at the last line of Table 6.3: it is equal to 10^{-239}. Note also that the thresholds u increase with the significance degrees s. To understand this fact, it is sufficient to look at inequality (6.1). But they increase slowly while significance degrees s decreases. It could be surprising but it is due to the error term which decreases very fast from a certain number of occurrences.

Example 2

Second, we consider the Chi of *Haemophilus influenzae* and its uptake sequence (see Table 6.4), for a significance degree s equal to 0.01. We use complete sequence of *Haemophilus influenzae* (Fleishmann et al. (1995)). Sequence length is equal to 1830138. We observe

TABLE 6.4. Table of thresholds u obtained by the three methods for the Chi and the uptake sequence of *Haemophilus influenzae* (sequence length t equal to 1830138). For each one of the three methods and for each word, we compute the threshold which permits to consider the word as an over-represented word or not, for degree of significance equal to 0.01. IMP means that the method can not return a result. "counts" correspond to the number of occurrences observed in the sequence.

Words	Chen-Stein	ϕ-mixing	ψ-mixing	counts
gatggtgg (chi)	23	36	22	20
gctggtgg (chi)	21	32	20	44
ggtggtgg (chi)	16	IMP	IMP	57
gttggtgg (chi)	30	45	26	37
aagtgcggt (uptake)	13	17	13	737

that in all the cases the ψ-mixing method is the best one because it gives the smallest u, except for the word `ggtggtgg` which has a periodicity less than $\left[\frac{n}{2}\right]$ (and then we can not study it: see assumptions in Theorem 3.3). We can not assume the good significance of the first Chi (`gatggtgg`) because we count only 20 occurrences in the sequence, whereas 23 occurrences are necessary to consider this word as exceptional. On the other hand, the uptake sequence is very significant (and then very relevant). Indeed, we could fix a significance degree equal to 10^{-224} and consider it as an over-represented word from 736 occurrences with the ψ-mixing method. As `aagtgcggt` is counted 737 times in the sequence, we obtain the well-known fact that this word is biologically relevant.

7. Conclusions and perspectives

To conclude this paper, we recall the advantages of our new methods. We give an error valid for all the values k of the random variable N^t corresponding to the number of occurrences of word A in a sequence of length t. Then, we can find a minimal number of occurrences to consider a word as biologically relevant for a very large number of words and for all degrees of significance. That is the main advantage of our methods on the Chen-Stein one which is based on the total variation distance and for which small degrees of significance can not be obtained. Results of our ψ-mixing method and the Chen-Stein

method remain similar but our method has less limitations. Note that our methods provide performing results for general modelling processes such as Markov chains as well as every ϕ- and ψ-mixing processes.

In terms of perspectives, as we expect more significant results, we hope to improve these methods adapting them directly to Markov chains instead of ψ- or ϕ-mixing. Moreover, it is well-known that a compound Poisson approximation is better for self-overlapping words (see Reinert et al. (2000) and Reinert and Schbath (1998)). An error term for the compound Poisson approximation for self-overlapping words can be easily derived from our results.

Acknowledgments

The authors would like to thank Bernard Prum for his support and his useful comments. The authors would like to thank Sophie Schbath for her program, Vincent Miele for his very relevant help in the conception of the software and Catherine Matias for her invaluable advices.

References

M. Abadi. Exponential approximation for hitting times in mixing processes. *Mathematical Physics Electronic Journal* **7** (2001a).

M. Abadi. *Instantes de ocorrência de eventos raros em processos misturadores.* Ph.D. thesis, Universidade de São Paulo (2001b).

M. Abadi. Sharp error terms and necessary conditions for exponential hitting times in mixing processes. *Annals of Probability* **32**, 243–264 (2004).

H. Almagor. A Markov analysis of DNA sequences. *J.Theor. Biol.* **104**, 633–645 (1983).

R. Arratia, L. Goldstein and L. Gordon. Two moments suffice for Poisson approximations: the Chen-Stein method. *Ann. Prob.* **17**, 9–25 (1989).

R. Arratia, L. Goldstein and L. Gordon. Poisson approximation and the Chen-Stein method. *Statist. Sci.* **5**, 403–434 (1990).

A. D. Bardour, L. H. Y. Chen and W. L. Loh. Compound Poisson approximation for nonnegative random variables via Stein's method. *Ann. Prob.* **20**, 1843–1866 (1992).

B. E. Blaisdell. Markov chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eucaryotic nuclear DNA sequences both protein-coding and noncoding. *J. Mol. Evol.* **21**, 278–288 (1985).

F. R. Blattner, G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau and Y. Shao. The complete genome sequence of escherichia coli k-12. *Science* **277**, 1453–1474 (1997).

L. H. Y. Chen. Poisson approximation for dependant trials. *Ann. Prob.* **3**, 534–545 (1975).

S. A. Douglass. *Introduction to Mathematical Analysis*. Addison-Wesley, Boston (1996). Chapter 8.

M. El Karoui, V. Biaudet, S. Schbath and A. Gruss. Characteristics of Chi distribution on different bacterial genomes. *Res. Microbiol.* **150**, 579–587 (1999).

R. D. Fleishmann, M. D. Adams, O. White and R. A. Clayton. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science* **269**, 496–512 (1995).

M. S. Gelfand, C. G. Kozhukhin and P. A. Pevzner. Extendable words in nucleotide sequences. *Bioinformatics* **8**, 129–135 (1992).

A. P. Godbole. Poisson approximations for runs and patterns of rare events. *Adv. Appl. Prob.* **23**, 851–865 (1991).

S. Karlin, C. Burge and A. M. Campbell. Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucl. Acids Res.* **20**, 1363–1370 (1992).

S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, Heidelberg (1993).

V. Miele, P. Y. Bourguignon, D. Robelin, G. Nuel and H. Richard. seq++ : analyzing biological sequences with a range of Markov-related models. *Bioinformatics* **21**, 2783–2784 (2005).

P. Nicodème, T. Doerks and M. Vingron. Proteome analysis based on motif statistics. *Bioinformatics* **18 (Suppl. 2)**, 5161–5171 (2002).

G. Nuel. LD-SPatt: Large Deviations Statistics for Patterns on Markov Chains. *Comp. Biol.* **11**, 1023–1033 (2004).

G. J. Philips, J. Arnold and R. Ivarie. The effect of codon usage on the oligonucleotide composition of the e. coli genome and identification of over- and underrepresented sequences by Markov chain analysis. *Nucl. Acids Res.* **15**, 2627–2638 (1987).

B. Prum, F. Rodolphe and E. de Turckheim. Finding words with unexpected frequencies in DNA sequences. *J. R. Statis. Soc. B* **11**, 190–192 (1995).

M. Régnier. A unified approach to word occurrence probabilities. *Discr. Appl. Math.* **104**, 259–280 (2000).

G. Reinert and S. Schbath. Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains. *J. Comput. Biol.* **5**, 223–253 (1998).

G. Reinert, S. Schbath and M. S. Waterman. Probabilistic and Statistical Properties of Words: An Overview. *J. Comput. Biol.* **7** (2000).

S. Robin and J. J. Daudin. Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Prob.* **36** (1999).

G. R. Smith, S. M. Kunes, D. W. Schultz, A. Taylor and K. L. Triman. Structure of chi hotspots of generalized recombination. *Cell* **24**, 429–436 (1981).

H. O. Smith, M. L. Gwinn and S. L. Salzberg. DNA uptake signal sequences in naturally transformable bacteria. *Res. Microbiol.* **150**, 603–616 (1999).

C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. Sixth Berkeley Symp. Math. Statist. Probab.*, volume 2, pages 583–602. University of California Press (1972).

J. van Helden, B. André and J. Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* **281**, 872–842 (1998).

J. van Helden, M. del Olmo and J. E. Pérez-Ortín. Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucl. Acids Res.* **28**, 1000–1010 (2000).