Poisson approximation in biological context

Nicolas Vergne Miguel Abadi

October 27, 2005

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 generally modelled by a Markov chain. We obtain a bound of the error between the law of the number of occurrences of a word in a sequence (under a Markov model) and its Poisson approximation. A global bound is already given by a Chen-Stein method. Our method, the ψ-mixing method, gives local bounds. We search a number of occurrences from which we can regard the studied word as a rare word. If the word appears more often than this number in the biological sequence, we conclude that it is an overrepresented word and then we suppose a biological role. Our method always give a limit number, while it was impossible with the Chen-Stein method. Comparing the methods, we observe a better accuracy for the ψ-mixing method for the bound of the tails of distribution. We also present the software `PANOW`[1] dedicated to the computation of the error term and the limit number of occurrences for a studied word.

Keywords Poisson approximation, Chen-Stein method, mixing processes, Markov chains, rare words, DNA sequences

1 Introduction

Mathematics are widely used in the context of the analysis of biological sequences (such as DNA sequences or protein sequences). In particular, Markov chains are commonly used to model the dependancies between the letters (nucleotides, amino-acids) in these sequences (see Almagor [6], Blaisdell [10], Phillips et al. [23], Gelfand et al. [16]). We can then study the properties of a sequence through those of a Markov chain. In particular, we can compare the number of occurrences of a word in an observed sequence with its number of occurrences expected in a Markov model (see Schbath et al. [30] on DNA sequences). Hence, the number of occurences is a statistic to find significantly over or underrepresented words and then assess a biological relevance. Examples of these biologically relevant words are given in an abundant litterature. Nicodème et al. [20] discuss about this relevance of finding over or underrepresented words. See for example the paper about the Chi site of *Escherischia coli* (Smith et al. [31]). More information about this Chi site and its repartition along the genome appears in El Karoui et al. [14]. Uptake sequences are other examples of rare words (see a sample in Smith et al. [32]). Few examples of rare words are given in Nuel [21]. Other applications on the relevance of finding over or underrepresentated words are given in Stückle et al. [34], Robin [27] or Bodman and Ward [12].

The exact law of the number of a word occurences under the Markovian model is known (Robin and Daudin [28], Régnier [24]) but, because of numerical complexity, it is not used in practice so that we use approximations. The reader can see Nuel [22] to have an overview of the different approaches. In this paper we focus on the Poisson approximation (Godbole [17]). We approximate $\mathbb{P}(N(A) = k)$ by $\frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!}$ where $\mathbb{P}(N(A) = k)$ is the stationary probability under the Markov model that the number of occurrences $N(A)$ of a word A is equal to k, $\mathbb{P}(A)$ is the probability of a word A occurrence in a given position, and t is the length of the sequence. Our aim is to bound the error between the law of

[1] available at `http://stat.genopole.cnrs.fr/software/panowdir/`

the number of occurrences of the word A and its Poisson approximation. The common method in this purpose is the Chen-Stein method, developped by Chen on Poisson approximations (Chen [13]) after a work of Stein on normal approximations Stein [33]. Its principle is to bound the difference between the two laws in total variation distance for all subset of the definition domain. But this method gives us too large an error because we are only interested in this difference for the tails of the distributions. We give here a new method, based on the property of mixing processes. It offers an error term ϵ, for the number of occurrences k, of the word A:

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

$\epsilon(A, k)$ decays factorially fast in k. This result is complementary to the results of Abadi [3, 1] and also Abadi and Vergne [5, 4]. As Markov chains are mixing-processes, we can apply this result for biological applications. In order to use this new method for these biological applications, we determine all the constants of the results of the above papers which are necessary for our proof.

This paper is organized in the following way. In section 2, we introduce the Chen-Stein method. In section 3, we present the new method on the Poisson approximation. In section 4, we state preliminary results. In section 5, we establish the proof of this result. 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 1 *For any two random variables X and Y with values in the same discrete space E, the total variation distance between their probability laws is defined by*

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

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 law of the number of occurrences of a word A in a sequence X and the Poisson law of parameter $t\mathbb{P}(A)$ where t is the length of the sequence and $\mathbb{P}(A)$ the stationary measure of A. There is an abundant literature about the Chen-Stein method (see Chen [13], Arratia et al. [7, 8], Barbour et al. [9]).

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 $\mathcal{F}$. Then, we work on a sequence $X = (X_1, \ldots, X_t)$ where each X_i belongs to $\mathcal{A}$. 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

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

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 $V(i) \subset \{1, 2, \cdots, t-n+1\}$ with $i \in V(i)$. $V(i)$ is a neighborood of i. The Chen-Stein method gives the following bound (see Arratia et al. [8]):

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

where

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

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

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

3 Preliminary notations and Poisson law

3.1 Preliminary notations

We focus in this study on Markov processes. The theorem given in the following subsection is established for more general mixing processes: the so called ψ-mixing processes.

Definition 2 *Let $\psi = (\psi(\ell))_{\ell \geq 0}$ be a sequence 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 we say that $A \in \mathcal{C}_n$ if its length is equal to n. 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 classic 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 $\{X_r = x_r, ..., X_s = x_s\}$. 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 define the periodicity of $A \in \mathcal{C}_n$ as the number p_A defined as follows:

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

p_A is called the principal period of the word A. Then, we denote by $\mathcal{R}_p = \mathcal{R}_p(n)$ the set of these $A \in \mathcal{C}_n$ of periodicity p and we also define $\mathcal{B}_n$ as the set of those $A \in \mathcal{C}_n$ with periodicity less than $\left[\frac{n}{2}\right]$:

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

$\mathcal{B}_n$ is the set of words which are self-overlaping before half their length (see Example 1). 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 note $r_A = \#\mathcal{R}(A)$. 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. Then, n_A is the smallest secondary period of A. Finally, we

introduce the following notation. For $0 \leq s < t$ integer, let $N_s^t = \sum_{i=s}^{t} \mathbb{1}_{T^{-i}A}$. So, N_s^t counts the number of occurrences of A between s and t. For the sake of simplicity, we also put $N_t = N_0^t$.

Example 1 *Consider the word $A = aaataaataaa$. As $p_A = 4$, $A \in \mathcal{B}_n$. See the following figure to note that $\mathcal{R}(A) = \{9; 10\}$, $r_A = 2$ and $n_A = 9$.*

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

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 of distribution of N_t.

Theorem 1 (ψ-mixing approximation) *Let $(X_m)_{m\in\mathbb{Z}}$ be a ψ-mixing process. There exists a constant $C_\psi > 0$, 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:*

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

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

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

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

4 Compute of constants¡

Our goal is to compute a bound as small as possible to control the error between the Poisson law and the law of the number of occurrences of a word. Thus, we determine the global constant C_ψ by means of intermediary bounds met in the proof. General bounds are interesting asymptotically in n, but for biological applications, n approx 10 or 20, it is not enough. Then along the proof, we will indicate the intermediary bounds that we compute. Before establishing the proof of this theorem, we point out here, for easy references, some results of Abadi [3], and some other useful results. In this paper, these results are given only in the ϕ-mixing context. Moreover the exact values of the constants are not given, while they are necessary for practical use of these methods. We give the values of all the constants met in the proofs of these results.

Proposition 1 (Proposition 11 in Abadi [3]) *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, $A \in \mathcal{C}_n$, and any $f \in \left[4n, \frac{1}{2\mathbb{P}(A)}\right]$ such that*

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

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

$$\left|\mathbb{P}\left(\tau_A > kf\right) - \mathbb{P}\left(\tau_A > f - 2\Delta\right)^k\right| \leq C_a \varepsilon\left(A\right) k \mathbb{P}\left(\tau_A > f - 2\Delta\right)^k, \quad (1)$$

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

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

Remark 1 *Both inequalities provide an approximation between the hitting time law and a geometric law for $t = kf$. The difference between them is that in the first one, the geometric inside the modulus is the same as in the upper bound, while in the second one, the geometric inside the modulus is larger than the one in the upper bound. That is, the second one gives a larger error. We will use both in the proof of Theorem 2.*

We denote $\mathcal{N}_j^i = \left\{\tau_A \circ T^{if+j\Delta} > f - j\Delta\right\}$ and $\mathcal{N} = \left\{\tau_A > f - 2\Delta\right\}$ for the sake of simplicity.

Proof. For the details of the proof, we refer to Proposition 11 in Abadi [3]. We only determine the values of constants C_a and C_b.

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

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

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

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

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

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

We apply this inequality in the inequalities (14) and (15) of [3] to get

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

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

We also have $(c) \leq \mathbb{P}\left(\mathcal{N}\right)^{k-1} \mathbb{P}\left(\mathcal{N}; \tau_A \circ T^{f-2\Delta} \leq 2\Delta\right) \leq \varepsilon\left(A\right) \mathbb{P}\left(\mathcal{N}\right)^{k-1}$.

We obtain (1): $\left|\mathbb{P}\left(\tau_A > kf\right) - \mathbb{P}\left(\mathcal{N}\right)^k\right| \leq 24k\varepsilon\left(A\right) \mathbb{P}\left(\mathcal{N}\right)^k$.

We deduce (2): $\left|\mathbb{P}\left(\tau_A > kf\right) - \mathbb{P}\left(\tau_A > f\right)^k\right| \leq 25k\varepsilon\left(A\right) \mathbb{P}\left(\mathcal{N}\right)^k$.

Then, $C_a = 24$ and $C_b = 25$. □

Theorem 2 (Theorem 1 in Abadi [3]) *Let $(X_m)_{m\in\mathbb{Z}}$ be a ψ-mixing process. Then, there exist constants $C_h > 0$, $0 < \Xi_1 < 1 \leq \Xi_2 < \infty$, such that for all $n \in \mathbb{N}$ and $A \in \mathcal{C}_n$, there exists $\xi_A \in [\Xi_1, \Xi_2]$, for which the following inequality holds for all $t > 0$:*

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

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

We prove an upper bound between the rescaled hitting time and the mean one exponential law. 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.

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

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

There are three steps in the proof of the theorem. First, we consider t of the form $t = kf$ with k a positive integer. Secondly, we prove the theorem for t of the form $t = (k + p/q)f$ with k, p positive integers and $1 \leq p \leq q$ with $q = \frac{1}{2\varepsilon(A)}$. We also put $r = (p/q)f$. Finally, we consider the remaining t. Here, we do not detail the two first steps (for that, see Abadi [3]), but only the last one. Let t be any positive real. We write $t = kf + r$, with k a positive integer and r such that $0 \leq r < f$. We can choose a $\bar{t}$ such that $\bar{t} < t$ and $\bar{t} = (k + p/q)f$ with p, q as before. We have

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

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

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

The second term is bounded as in the two first steps of the proof. We apply inequalities (1) and (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, the third term is bounded using the Mean Value Theorem:

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

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

Lemma 1 *$(X_m)_{m\in\mathbb{Z}}$ 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 $\mathbb{P}(B \cap C) \leq \mathbb{P}(B)(\mathbb{P}(C) + \psi(g))$. Dividing the above inequality by $\mathbb{P}(A)$ and the lemma follows. □

Proposition 2 *Let $(X_m)_{m\in\mathbb{Z}}$ be a ψ-mixing process. Let $A \in \mathcal{R}_p(n)$. Then the following holds:*

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

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

and similarly

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

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

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

Remark 2 *The above proposition establishes a relation between hitting and return times with an error uniform in t. In particular, (b) says that they coincide if and only if $\zeta_A = 1$, namely, the string A is non-self-repeating.*

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

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

Term (3) is bounded with Lemma 1 by

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

Term (3) is bounded using the ψ-mixing property by $\mathbb{P}_A(\tau_A > M)\,\psi(g)$. The modulus in (3) is bounded using stationarity by $\mathbb{P}(\tau_A \leq g) \leq g\mathbb{P}(A)$. This ends the proof of both inequalities of item (a).
Item (b) for $t \geq 2n$ is proven similarly to item (a) with $t = M + M'$, $M = p$, and $g = w$ with $1 \leq w \leq n_A$. Consider now $p \leq t < 2n$.

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

First and second equalities follow by definition of τ_A and $\mathcal{R}(A)$. The inequality follows by Lemma 1. □

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

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

Hence, we have

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

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

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

The above modulus is bounded by

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

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 2 (b) and the fact that $\mathbb{P}(\tau_A > h) \geq 1/2$ we have

$$|\rho_i - \zeta_A| \leq \frac{2e_1(A)}{\mathbb{P}(\tau_A > h)} \leq 4e_\psi(A).$$

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 3 *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 2 and the second with Lemma 2 :

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

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

Definition 3 *Given $A \in \mathcal{C}_n$, we define for $j \in \mathbb{N}$, the j-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 *Let $(X_m)_{m\in\mathbb{Z}}$ be a ψ-mixing. Then, for all $A \notin \mathcal{B}_n$, all $k \in \mathbb{N}$, and all $0 \leq t_1 < t_2 < \ldots < t_k \leq t$ for which $\min\limits_{2\leq j\leq k}\{t_j - t_{j-1}\} > 2n$, there exists a positive constant C_1 independent of A, n, t and k such that*

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

Proof. We will show this proposition by induction on k. We put $\mathcal{P}_j = \mathbb{P}(\tau_A > (t_j - t_{j-1}) - 2n)$. We also 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

$$\begin{aligned}
& \left| \mathbb{P}\left(\tau_A = t_1\,;\, \tau_A^{(2)} > t\right) - \mathbb{P}(A) \prod_{j=1}^{2} \mathcal{P}_j \right| \\
\leq\ & \left| \mathbb{P}\left(\tau_A = t_1\,;\, \tau_A^{(2)} > t\right) - \mathbb{P}\left(\tau_A = t_1\,;\, N_{t_1+2n}^{t} = 0\right) \right| && (3) \\
+\ & \left| \mathbb{P}\left(\tau_A = t_1\,;\, N_{t_1+2n}^{t} = 0\right) - \mathbb{P}\left(\tau_A = t_1\right) \mathcal{P}_2 \right| && (4) \\
+\ & \left| \mathbb{P}(A\,;\, \tau > t_1 - 1) - \mathbb{P}\left(A\,;\, N_{2n}^{t_1-1} = 0\right) \right| \mathcal{P}_2 && (5) \\
+\ & \left| \mathbb{P}\left(A\,;\, N_{2n}^{t_1-1} = 0\right) \mathcal{P}_2 - \mathbb{P}(A) \prod_{j=1}^{2} \mathcal{P}_j \right|. && (6)
\end{aligned}$$

Term (3) is equal to $\mathbb{P}\left(\tau_A = t_1; \bigcup_{i=t_1+1}^{t_1+2n} T^{-i}A; N_{t_1+2n}^{t} = 0\right)$. For $1 \leq i \leq p_A$, the leading term of the above sum is zero. Thus, using mixing property

$$\begin{aligned}
(3) &= \mathbb{P}\left(\tau_A = t_1; \bigcup_{i \in \mathcal{R}(A) \cup i = t_1 + p_A}^{t_1+2n} T^{-i}A; N_{t_1+2n}^{t} = 0\right) \\
&\leq 2\mathbb{P}(A)\mathbb{P}(A)(r_A + n)(1 + \psi(n))\mathbb{P}\left(N_{t_1+2n}^{t} = 0\right) \\
&\leq 2\mathbb{P}(A) e_\psi(A) e^{-(t-(3k+1)n)\mathbb{P}(A)}
\end{aligned}$$

Term (4) is bounded using ψ-mixing property

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

Analogous computations are used to bound terms (5) and (6).

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:

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

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

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

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

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

Finally, we obtain

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

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 1

For $k = 0$, the result comes from Proposition 3 ($\mathbb{P}(N_t = 0) = \mathbb{P}(\tau_A > t)$).
For $k > 2t/n$, as $A \notin \mathcal{B}_n$, 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 $I(\mathcal{T}) = \min_{2 \le j \le k} \{\Delta_j\}$. We say that the occurrences of A are isolated if $I(\mathcal{T}) \ge 2n$ and we say that there exists at least one clump if $I(\mathcal{T}) < 2n$. We also denote

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

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

$$\mathbb{P}(N_t = k) = \mathbb{P}(B_k) + \mathbb{P}(G_k),$$

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

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

We will 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 $p_A \le d \le 2n$. Therefore, the ψ-mixing property leads to the bound

$$\begin{aligned} \mathbb{P}\left(\bigcup_{t_2, \ldots, t_k} \mathcal{T}(t_1, t_2, \ldots, t_k) \right) &\le \mathbb{P}\left(\bigcap_{j=1}^{k} \bigcup_{\substack{n/2 \le t_{i+1} - t_i \le 2n; \\ i=2,\ldots,k}} T^{-t_j} A \right) \\ &\le \mathbb{P}(A) e_\psi(A)^{k-1} e_\psi(A) e^{-(t-(3k+1)n)\mathbb{P}(A)}. \end{aligned} \tag{7}$$

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 $1 \le t(1) < \ldots < t(i) \le t$ fixed. By the same argument used above, we have the inequalities

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

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 that runs 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 at most in $\binom{t}{i}$ many ways. The cardinality of each one of the i clusters can be arranged in $\binom{k-1}{i-1}$ many ways. (This corresponds to break 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

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

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

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

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

$$\sum_{T\in G_k} \left| \mathbb{P}\left(\bigcap_{j=1}^{k} \left(\tau_A^{(j)} = t_j\right) \,;\, \tau_A^{(k+1)} > t \right) - \mathbb{P}(A)^k \prod_{j=1}^{k+1} \mathcal{P}_j \right| \tag{8}$$

$$+ \sum_{T\in G_k} \mathbb{P}(A)^k \left| \prod_{j=1}^{k+1} \mathcal{P}_j - \prod_{j=1}^{k+1} e^{-(\Delta_j - 2n)\mathbb{P}(A)} \right| \tag{9}$$

$$+ \sum_{T\in G_k} \mathbb{P}(A)^k \left| e^{-(t-2(k+1)n)\mathbb{P}(A)} - e^{-t\mathbb{P}(A)} \right| \tag{10}$$

$$+ \left| \frac{\#G_k\, k!}{t^k} \frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!} - \frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!} \right|. \tag{11}$$

We will bound these terms to obtain Theorem 1.
First, we bound the cardinal of G_k

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

Term (8) is bounded with Proposition 4

$$(8) \leq C_1 \frac{t^k}{(k-1)!} (\mathbb{P}(A)(1+\psi(n))^k e_\psi(A) e^{-(t-(3k+1)n)\mathbb{P}(A)}.$$

Term (9) is bounded with Proposition 3

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

where C_p is defined in Proposition 3.
We compute

$$(9) \quad \leq \quad \frac{(t\mathbb{P}(A))^k}{(k-1)!}\frac{k+1}{k}\left[(8+C_a t\mathbb{P}(A)+C_a+2C_b)\varepsilon(A)+11t\mathbb{P}(A)e_*(A)\right]e^{-(\zeta_A-11e_*(A))t\mathbb{P}(A)}.$$

Term (10) is bounded by

$$(10) \leq \frac{t^k}{k!}\mathbb{P}(A)^k(k+1)2n\mathbb{P}(A)e^{-t\mathbb{P}(A)}e^{2(k+1)n\mathbb{P}(A)}.$$

To bound term (11), we bound the following difference

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

Then, we have

$$(11) \leq \frac{k\,(k+4n)}{t}\frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!}.$$

Now, we just have to add the five bounds to obtain the theorems with the constante $C_\psi = 1 + C_1 + 2C_p + 8 + 8$. Proposition 4 shows that $C_1 = 5$ and Theorem 2 that $C_p = 116$. Then, we obtain $C_\psi = 254$. □

6 Biological applications

With the explicit value of the constant C_ψ of Theorem 1, and more particulary thanks to all the intermediary bounds given in the proof of this theorem, we can develop the program to apply this formula to the study of rare words in biological sequences. In order to compare different methods, we also compute the bound for a ϕ-mixing, for which a proof of Poisson approximation is given in Abadi and Vergne [4]. Let us recall the definition of such a mixing.

Definition 4 *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 underrepresented words in biological sequences and we compare them to the Chen-Stein metod.

In order to apply our formulas, we note that DNA sequences are very often modelled by Markov chains (Almagor [6], Blaisdell [10]), particulary for the identification of over or underrepresented words (Phillips et al. [23], Nuel [22], Régnier [24], Reinert et al. [26], Reinert and Schbath [25], Schbath et al. [30]). Many other models, based or not on Markov chains, are used to find rare words (Karlin et al. [18], Stückle et al. [34], Bodman and Ward [12], Roth et al. [29]).

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. 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. There are several estimations of K and ν. We choose ν equal to the second eigenvalue of the transition matrix of the model and $K = \frac{1}{\inf\limits_{j\in\{1,\dots,|\mathcal{A}|^k\}} \mu_j}$ 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 to guess an 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 have to calculate the smallest number of occurrences u necessary to $\mathbb{P}(N > u) < s$, where N is the number of occurrences of the word to study. 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

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

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 law and its Poisson approximation, bounded using Theorem 1. Then, we search the smallest u such that

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

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.

6.1 Software availability

We developed `PANOW`, dedicated to the determination of 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. [19]).

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

6.2 Comparaisons between the three different methods

6.2.1 Theorical tests

We can compare the mixing methods and the Chen-Stein method through the values of u obtained with `PANOW` using Abadi and Vergne [5] in the first case and Reinert and Schbath [25] in the second one. In order to study the different possiblities of results of all the methods, Table 1 offers a good outline of the possibilities and limits of each method to us. It displays some results on different words randomly selected (no biological sense for any of these words). Table 1 has been obtained with an order one Markov model with a random transition matrix (once again, there is no biological sense with this model. It does not model a real biological sequence) 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 give us a result in all the cases. For the Chen-Stein method, the bound we obtain can be higher than the significance degree we fix and so we can not find u. Therefore there are many examples that we can not study with this method. Moreover, it is interesting to have

Table 1: Table of the u obtained by the three methods (sequence length equal to 10^6). For each one of the three methods and for each word, we compute the limit number of occurrences so that we can consider the word as a rare word. 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

a small s and because of this restriction of the Chen-Stein method, we can not have it. For example, this problem appears for the words `aagcgc` and `cgagcttc` in Table 1. For this second word, we notice that we have a Chen-Stein bound for a s equal to 0.1, but not for a s equal to 0.01. Indeed, for this word, the Chen-Stein bound is equal to 0.0107954. The same thing appears for the word `agcaaata` (the Chen-Stein bound is equal to 0.0120193). For the ϕ- or ψ-mixing method, the two only difficult cases arise for an error function (e_ψ in the ψ-mixing case) greater than 1 or for a "high" parameter of the Poisson law ("high" means greater than 500). Note that the first case does not appear very frequently (in any case in Table 1). The reason why the error function has to be greater than 1 is that the error term has to be decreasing with the number of occurrences k and without this condition we are not sure about this fact. The second problem is just a computational difficulty and once again it does not appear very frequently (only for the word `cccg` in Table 1 for instance). We would like to insist on the main advantage of our methods: we can fix any s and we will find a u, contrary to the Chen-Stein method. Also, we can use our methods for any Markov chains order. Indeed, `PANOW` runs fast enough contrary to the R program used to compute the Chen-Stein bound of Reinert and Schbath [25]. Note that we compute another way to calculate the Chen-Stein bound (see Abadi [2]) and this way gives approximately the same Chen-Stein bound.

Secondly we notice that the ψ-mixing method is always better than the ϕ-mixing one. Obviously, this result was expected by the theorems because of the extra factor $e^{-(t-(3k+1)n)\mathbb{P}(A)}$ (see Theorem 1), but we were interested in the real impact of this factor on the limit number of occurrences u.

The third main observation we can make is that, when it works, the Chen-Stein method and the ψ-mixing method gives very similar u.

6.2.2 Biological tests

Now, we present a few results obtained on real biological examples. 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 consider the over or the underrepresentation 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 to Chi sites or uptake sequences. Chi sites of bacterias protect their genome by stopping its degradation performed by a particular enzyme. The function of this enzyme is to cancel virus which could appear into the bacteria. Virus 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.

Exemple 1

Firstly, we consider the Chi of *Escherichia coli* (see Table 2), for different degrees of significance.

We use complete sequence of *Escherichia coli K12* (Blattner et al. [11]). We can conclude that the

Table 2: Chi of *Escherichia coli*: `gctggtgg`

s	Chen-Stein	ϕ-mixing	ψ-mixing	counts
0.1	88	194	83	499
0.01	IMP	196	92	499
0.0001	IMP	198	99	499

ψ-mixing method gives the most interesting results. The word could be considered as a rare word from 99 occurrences for a s of 0.0001. The Chi of *E. coli* is biologically very relevant and that explains the significance of the results. In another way, we note that u increases slowly while s decreases. It could be surprising but it is due to the error term which decreases very fast from a certain number of occurrences.

Exemple 2

Secondly, we consider the Chi of *Haemophilus influenzae* and its uptake sequence (see Table 3), for a s of 0.01. We use complete sequence of *Haemophilus influenzae* (Fleischmann et al. [15]). We observe

Table 3: The Chi and the uptake sequence of *Haemophilus influenzae*

Words	Chen-Stein	ϕ-mixing	ψ-mixing	counts
gatggtgg (chi)	23	39	22	20
gctggtgg (chi)	21	34	20	44
ggtggtgg (chi)	16	IMP	IMP	57
gttggtgg (chi)	30	48	27	37
aagtgcggt (uptake)	13	19	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 on Theorem 1). 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) because it is counted 737 times in the sequence. We could give a very small degree of significance s.

7 Conclusions and perspectives

To conclude this paper, we recall the advantages of our new methods. Thanks to an error holding for all the values of k, contrary to the Chen-Stein error which is based on the total variation distance, we can find a number of occurencies minimal to consider a word as biologically relevant for almost all words and for all degrees of significance. Results of our method and the Chen-Stein method remain similar but our method has less limitations. Note that our method provides performing results for general modelling processes such as Markov chains as well as every ϕ and ψ mixing.

In term of perspectives, we plan to adapt `PANOW` to the study of words which are underrepresented (there are in the sequence significantly fewer of these words than expected by the model). Secondly, as we expected 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. [26] and Reinert and Schbath [25]). An error term for the compound Poisson approximation for self-overlapping words can be easily derived from our results.