Applied Probability Trust (28 August 2006)

IMPROVED COMPOUND POISSON APPROXIMATION FOR THE NUMBER OF OCCURRENCES OF ANY RARE WORD FAMILY IN A STATIONARY MARKOV CHAIN

ETIENNE ROQUAIN,* *and*

SOPHIE SCHBATH,* *Institut National de la Recherche Agronomique*

Abstract

We derive a new compound Poisson distribution with explicit parameters to approximate the number of overlapping occurrences of any set of words in Markovian sequences. Thanks to the Chen-Stein method, we provide a bound for the approximation error. This error converges to zero under the rare event condition, even for overlapping families which improves previous results. As a consequence, we also propose Poisson approximations for the declumped count and the number of competing renewals.

Keywords: Compound Poisson approximation; Chen-Stein method; multiple word count; clumps; period; Markov chain

2000 Mathematics Subject Classification: Primary 62E17

Secondary 60C05

1. Introduction

Word statistics in random sequences of letters have been popular for a long time because they arise in various applicative domains. With the huge amount of biological sequences now available, genome analysis is

* Postal address: INRA, Unité Mathématique, Informatique et Génome, France

Etienne.Roquain@jouy.inra.fr, Sophie.Schbath@jouy.inra.fr

an important consumer of probabilistic and statistical results on word occurrences (see chapter 6 of [5] or [7] for overviews). In particular the number N of occurrences of a given word in a DNA sequence is a special quantity of interest for molecular biologists; Some words, called *motifs*, are indeed recognized by proteins and take place in various biological processes. Over- and under-represented motifs are then looked for in many genomes. Moreover, biological motifs are often degenerated, i.e. some letters are ambiguous, and should rather be considered like families of fixed words.

The most popular random sequence models are the Markov chain models; They are widely used in genome analysis because they can fit the composition of a DNA sequence in short words of length 1 up to $(m+1)$ where m is the order of the Markov chain. Various results have been published on the word count distribution in Markov chains. The exact distribution can be obtained through its probability generating function ([11]) or thanks to the distributions of both the waiting time till the first occurrence and the inter-arrival time between two occurrences ([2], [9]). Several approximations have also been proposed for long sequences. The Gaussian distribution proposed by [6] appears to be a good approximation for words (and word families) having an expected count large enough ([10]). For an expectedly rare word $\mathbf{w}$, i.e. satisfying the rare event condition $\mathbb{E}N(\mathbf{w}) = \mathrm{O}(1)$ as the length n of the sequence tends to infinity, Poisson approximations have been first proposed ([4]) but compound Poisson approximations appear to be better ([12]). This result is based on the fact that (i) occurrences of a given word occur in clumps, (ii) clumps asymptotically form a Poisson process under the rare event condition, and

(iii) the numbers of occurrences per clump are asymptotically independent and identically distributed (geometric distribution). The compound Poisson distribution reduces to a Poisson distribution for non overlapping words. For an expectedly rare family of words $\mathcal{W}$, [8] proposed to use the compound Poisson approximation of [12] for each count $N(\mathbf{w}), \mathbf{w} \in \mathcal{W}$, and to approximate $N(\mathcal{W}) = \sum_{\mathbf{w}\in\mathcal{W}} N(\mathbf{w})$ by the sum of independent compound Poisson variables. Thanks to the Chen-Stein method, a bound for the approximation error was given; It explicitly depends on the degree of overlaps between the words from the family $\mathcal{W}$. Unfortunately, this error bound does not converge to zero as soon as there exists a couple of different words $(\mathbf{w}, \mathbf{w}') \in \mathcal{W}^2$ which overlap.

Still using the Chen-Stein method, we propose here a more suitable compound Poisson distribution to approximate the count $N(\mathcal{W})$ of any expectedly rare word family $\mathcal{W}$. The main difference with [8] is that we will consider clumps composed of overlapping occurrences of $\mathcal{W}$, instead of considering clumps of $\mathbf{w}$ for each word $\mathbf{w} \in \mathcal{W}$ separately. We will then directly adapt the method of [12] for a single word to a word family. The difficulty arises from the structure and the occurrence probabilities of such mixed clumps. The idea of studying mixed clumps has been previously introduced by [3] to approximate the count of competing renewals, but the authors were only focusing on the event "a mixed clump starts at a given position". Here, we will have to deal also with the exact size of the mixed clumps.

The paper is organized as follows. In Section 2, we state the approximation theorem for the count $N(\mathcal{W})$; The parameters of the limiting compound Poisson distribution will be explicitly derived in Section 3 which is the high point of the paper. Section 4 contains the proof of the approximation theorem: It uses the Chen-Stein method for Poisson approximations. As a corollary, Section 5 proposes a Poisson approximation for both the number of clumps of a word family $\mathcal{W}$ and the number of competing renewals of $\mathcal{W}$ in a Markov chain. Our contribution, compared to the result of [3], lies in the explicit formula for the parameter of the limiting Poisson distribution. Section 6 presents generalizations to high order Markov chains and to hidden Markov models.

2. Compound Poisson approximation for $N(\mathcal{W})$

In this paper, we consider a random sequence $\mathbf{X} = (X_i)_{i\in\mathbb{Z}}$ generated by an homogeneous stationary Markov chain of order 1 on a finite alphabet $\mathcal{A}$. Generalization to higher order Markov chains is discussed in the conclusion. The stationary distribution on $\mathcal{A}$ is denoted by μ and $\Pi = [\pi(x,y)]_{x,y\in\mathcal{A}}$ is the transition matrix of the model.

Let $\mathcal{W}$ be a family of d different words $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_d\}$ on the alphabet $\mathcal{A}$ with length at least 2. The length of any word $\mathbf{w}$ will be denoted by $|\mathbf{w}|$ and we define h as the length of the longest word from the family $\mathcal{W}$, $h := \max\{|\mathbf{w}|, \mathbf{w} \in \mathcal{W}\}$. We make two assumptions on the word family $\mathcal{W}$: (i) it is *reduced* meaning that, $\forall \mathbf{w} \neq \mathbf{w}' \in \mathcal{W}$, $\mathbf{w}$ is not a substring of $\mathbf{w}'$ (it is a usual assumption when studying occurrences of word families and it is immediately fullfilled if all the words from $\mathcal{W}$ have the same length), (ii) each word $\mathbf{w} \in \mathcal{W}$ has a non zero probability to occur in $\mathbf{X}$ (it is a natural assumption). Thanks to the Markov property, the occurrence probability of a $|\mathbf{w}|$-letter word $\mathbf{w} = w_1 w_2 \cdots w_{|\mathbf{w}|}$ in $\mathbf{X}$ is given by $\mu(w_1)\prod_{j=1}^{|\mathbf{w}|-1}\pi(w_j, w_{j+1})$ and will be simply denoted by $\mu(\mathbf{w})$ in the remainder.

Classically the number of occurrences $N(\mathcal{W})$ of a word family $\mathcal{W}$ in the finite sequence $X_1 \cdots X_n$ is defined as $N(\mathcal{W}) = \sum_{\mathbf{w}\in\mathcal{W}} \sum_{i=1}^{n-|\mathbf{w}|+1} Y_i(\mathbf{w})$ where $Y_i(\mathbf{w})$ is the Bernoulli variable which is equal to 1 if there is an occurrence of $\mathbf{w}$ starting at position i and 0 otherwise. Note that we will generalize this Bernoulli variable to $Y_i(\mathcal{W})$ which will be equal to 1 if and only if it exists a word from $\mathcal{W}$ occurring at position i (e.g. iff an occurrence of $\mathcal{W}$ occurs at position i). Here we will use another decomposition of the count based on the occurrences of k-clumps. The notion of clump has no sense out of a sequence. A k-clump of $\mathcal{W}$ in a sequence is indeed a maximal set of k overlapping occurrences of $\mathcal{W}$ in this sequence. Therefore, a k-clump of $\mathcal{W}$ occurs at position i in a sequence if and only if a word composed of exactly k overlapping occurrences of the family $\mathcal{W}$ occurs at position i without overlapping any other occurrence of the family $\mathcal{W}$ in this sequence. For example, for the family $\mathcal{W} = \{$`atta,ttat`$\}$, the sequence g<u>atta</u>gc<u>attatta</u>c has a 1-clump of $\mathcal{W}$ at $i = 2$ and a 3-clump of $\mathcal{W}$ at $i = 8$. One should be careful not forgetting the occurrence of `ttat` in the 3-clump `attatta`. Therefore, we have

$$N(\mathcal{W}) = \sum_{k\geq 1} k\widetilde{N}_k(\mathcal{W}),$$

where $\widetilde{N}_k(\mathcal{W})$ is the number of k-clumps of $\mathcal{W}$ in $X_1 \cdots X_n$.

For convenience, we will work in the infinite sequence $\mathbf{X}$: We define $\widetilde{Y}_{i,k}(\mathcal{W})$ like the Bernoulli variable which is one if a k-clump of $\mathcal{W}$ occurs at position i in $\mathbf{X}$, and 0 otherwise, and we put:

$$N^{\infty}(\mathcal{W}) := \sum_{k\geq 1} k\widetilde{N}_k^{\infty}(\mathcal{W}) \qquad \text{with} \qquad \widetilde{N}_k^{\infty}(\mathcal{W}) := \sum_{i=1}^{n-h+1} \widetilde{Y}_{i,k}(\mathcal{W}). \quad (1)$$

Note that the count $N^\infty(\mathcal{W})$ can slightly differ from the real observed count $N(\mathcal{W})$ of $\mathcal{W}$ in the finite sequence $X_1 \cdots X_n$ because clumps of $\mathcal{W}$ in $\mathbf{X}$ may start before position 1 and/or end after position n, and occurrences of $\mathcal{W}$ in $X_1 \cdots X_n$ may start after position $n-h+1$ if there exists $\mathbf{w} \in \mathcal{W}$ such that $|\mathbf{w}| \neq h$. However, the event $\{N(\mathcal{W}) \neq N^\infty(\mathcal{W})\}$ implies that it exists (at least) an occurrence of $\mathcal{W}$ starting in $\{1, \dots, h-1\}$ or in $\{n-h+2, \dots, n\}$. This event occurs with a probability less than $2(h-1)\mu(\mathcal{W})$, where $\mu(\mathcal{W}) = \mathbb{E}Y_i(\mathcal{W}) = \sum_{\mathbf{w}\in\mathcal{W}} \mu(\mathbf{w})$ denotes the occurrence probability of $\mathcal{W}$ at a given position. Therefore, the total variation distance* between the distribution of these two counts is bounded by $2h\mu(\mathcal{W})$ which tends to zero as $n \to \infty$ under both $h = \mathrm{o}(n)$ and the rare event condition. Both counts are then asymptotically equivalent and we will focus on $N^\infty(\mathcal{W})$.

We will now use the Chen-Stein theorem as stated by [1] to bound the total variation distance between the distribution of the vector $(\widetilde{Y}_{i,k}(\mathcal{W}))_{i,k}$ and the joint distribution of independent Poisson variables $(Z_{i,k})_{i,k}$ such that $\mathbb{E}Z_{i,k} = \mathbb{E}\widetilde{Y}_{i,k}(\mathcal{W})$; These expectations will be denoted by $\widetilde{\mu}_k(\mathcal{W})$. By defining $Z_k := \sum_{i=1}^{n-h+1} Z_{i,k}$, the Chen-Stein theorem says that

$$d_{\mathrm{tv}}\Big(\mathcal{L}\big((\widetilde{N}_k^\infty(\mathcal{W}))_k\big), \mathcal{L}\big((Z_k)_k\big)\Big) \leq b_1 + b_2 + b_3, \qquad (2)$$

* The total variation distance between two discrete distributions P and P' on $\mathbb{N}$ is defined like $\frac{1}{2}\sum_{x\in\mathbb{N}} |P(x) - P'(x)| \leq \min \mathbb{P}(N \neq N')$ where the minimum ranges over all coupling (N, N') of P and P'.

where

$$b_1 = \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{(j,\ell)\in B_{i,k}} \mathbb{E}(\widetilde{Y}_{i,k}(\mathcal{W}))\mathbb{E}(\widetilde{Y}_{j,\ell}(\mathcal{W})) \tag{3}$$

$$b_2 = \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{(j,\ell)\in B_{i,k}\setminus\{(i,k)\}} \mathbb{E}(\widetilde{Y}_{i,k}(\mathcal{W})\widetilde{Y}_{j,\ell}(\mathcal{W})) \tag{4}$$

$$b_3 = \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \mathbb{E}\Big|\mathbb{E}\big(\widetilde{Y}_{i,k}(\mathcal{W}) - \widetilde{\mu}_k(\mathcal{W})|\sigma(\widetilde{Y}_{j,\ell}(\mathcal{W}), (j,\ell)\notin B_{i,k})\big)\Big|, \tag{5}$$

and where $B_{i,k} \subset \{1,\dots,n-h+1\}\times\mathbb{N}^\star$ is a neighborhood of (i,k). As we will see, for a particular choice of the neighborhood $B_{i,k}$, the quantities b_1, b_2 and b_3 will tend to zero as $n\to\infty$, $h = \mathrm{o}(n)$ and under the rare event condition $\mathbb{E}N(\mathcal{W}) = \mathrm{O}(1)$ (cf. Section 4). It means that the process $(\widetilde{N}_k^\infty(\mathcal{W}))_k$ can be approximated by independent Poisson variables $(Z_k)_k$ with expectation $\widetilde{\lambda}_k(\mathcal{W}) := \mathbb{E}\widetilde{N}_k^\infty(\mathcal{W}) = (n-h+1)\widetilde{\mu}_k(\mathcal{W})$. Thanks to equation (1) and total variation distance properties, it also means that, under the same asymptotic conditions, the count $N^\infty(\mathcal{W})$ can be approximated by $\sum_{k\geq 1} kZ_k$ which follows by definition the compound Poisson distribution $\mathcal{CP}(\widetilde{\lambda}_k(\mathcal{W}), k\geq 1)$. We then state the following approximation theorem.

Theorem 1. *For every word family $\mathcal{W}$, the total variation distance between the distribution of $N(\mathcal{W})$ and the compound Poisson distribution with parameters $(\widetilde{\lambda}_k(\mathcal{W}))_{k\geq 1}$ such that $\widetilde{\lambda}_k(\mathcal{W}) = (n-h+1)\widetilde{\mu}_k(\mathcal{W})$ given by (15) is bounded as follows:*

$$d_{tv}\Big(\mathcal{L}\big(N(\mathcal{W})\big), \mathcal{CP}\big(\widetilde{\lambda}_k(\mathcal{W}), k\geq 1\big)\Big) \leq Cnh\mu^2(\mathcal{W}) + C'n\mu(\mathcal{W})|\alpha|^h + 2h\mu(\mathcal{W}), \tag{6}$$

where $C>0$ and $C'>0$ are two explicit constants that only depend on the transition matrix Π, and α is the second largest eigenvalue in modulus *of* Π *(*$|\alpha|<1$*). Therefore, if* $n\mu(\mathcal{W})=O(1)$ *and* $h=o(n)$*, we have*

$$d_{tv}\Big(\mathcal{L}\big(N(\mathcal{W})\big),\mathcal{CP}\big(\widetilde{\lambda}_k(\mathcal{W}),k\geq 1\big)\Big)\xrightarrow[n\to\infty]{} 0. \tag{7}$$

The proof is done in Section 4.

Remark 2.1. The condition $n\mu(\mathcal{W})=\mathrm{O}(1)$ is equivalent to $\log(n)/|\mathbf{w}|=\mathrm{O}(1)$, $\forall\mathbf{w}\in\mathcal{W}$, which means that the compound Poisson approximation holds for families of long enough words.

The Chen-Stein method usually does not provide an optimal bound. Our concern here is just to show that the bound given by (6) converges to zero when $n\to\infty$, $h=\mathrm{o}(n)$ and $n\mu(\mathcal{W})=\mathrm{O}(1)$.

An important task is now to calculate the parameters of the limiting compound Poisson distribution. In the next section, we will then provide the expression of $\widetilde{\mu}_k(\mathcal{W})$ which is the occurrence probability of a k-clump of $\mathcal{W}$ at a given position in the infinite sequence $\mathbf{X}$.

3. Occurrence probability of a k-clump of $\mathcal{W}$

We first have to look at the typical distances allowed between successive occurrences of $\mathcal{W}$ in a k-clump i.e. successive overlapping occurrences of $\mathcal{W}$.

3.1. Principal periods

For two words $\mathbf{w}=w_1\cdots w_{|\mathbf{w}|}$ and $\mathbf{w}'=w'_1\cdots w'_{|\mathbf{w}'|}$ of $\mathcal{W}$, an integer p, $1\leq p\leq|\mathbf{w}|-1$, such that $w'_i=w_{i+p}$ for $i=1,\ldots,|\mathbf{w}|-p$ is called a **period** of $(\mathbf{w},\mathbf{w}')$. We denote by $\mathcal{P}(\mathbf{w},\mathbf{w}')$ the set of periods of $(\mathbf{w},\mathbf{w}')$.

For each words $\mathbf{w}, \mathbf{w}'$ and each period $p \in \mathcal{P}(\mathbf{w}, \mathbf{w}')$, the prefix $\mathbf{w}^{(p)} := w_1 \dots w_p$ is called a **root** of $(\mathbf{w}, \mathbf{w}')$. The periods of $(\mathbf{w}, \mathbf{w}')$ are then all the distances allowed between an occurrence of $\mathbf{w}$ and a further overlapping occurrence of $\mathbf{w}'$. For instance $\mathcal{P}(\mathtt{taca}, \mathtt{acac}) = \{1, 3\}$.

If we now look at the possible distance between **successive** overlapping occurrences of $(\mathbf{w}, \mathbf{w}')$, it appears that some periods are not possible. For instance, the period 3 of $(\mathtt{taca}, \mathtt{acac})$ is not possible because an occurrence of `taca` at position i and an occurrence of `acac` at position $i + 3$ implies an other occurrence of `acac` in between (position $i + 1$). More generally, for two words $\mathbf{w}$ and $\mathbf{w}'$ of $\mathcal{W}$, a period $p \in \mathcal{P}(\mathbf{w}, \mathbf{w}')$ is said **principal** with respect to $\mathcal{W}$ if for all $\mathbf{w}^\star \in \mathcal{W}$ and $j \in \mathcal{P}(\mathbf{w}, \mathbf{w}^\star)$, we have $p - j \notin \mathcal{P}(\mathbf{w}^\star, \mathbf{w}')$. This condition simply means that $\mathcal{W}$ cannot occur between an occurrence of $\mathbf{w}$ at a position i and an occurrence of $\mathbf{w}'$ at $i+p$. We denote by $\mathcal{P}'_{\mathcal{W}}(\mathbf{w}, \mathbf{w}')$ the set of principal periods of $(\mathbf{w}, \mathbf{w}')$ with respect to $\mathcal{W}$. When there will be no ambiguity, we will omit the subscript $\mathcal{W}$. If $\mathcal{W}$ is composed of a unique word $\mathbf{w}$, possible distances between two successive overlapping occurrences of $\mathbf{w}$ coincide with the so-called principal period set $\mathcal{P}'(\mathbf{w})$ of $\mathbf{w}$ introduced in [12].

A direct consequence of the definition of a principal period is the following lemma.

Lemma 1. *(i) An occurrence of $\mathbf{w}' \in \mathcal{W}$ at position i does overlap a preceding occurrence of $\mathcal{W}$ in the sequence if and only if it exists a word $\mathbf{w} \in \mathcal{W}$ and a principal period $p \in \mathcal{P}'(\mathbf{w}, \mathbf{w}')$ such that there is an occurrence of the principal root $\mathbf{w}^{(p)}$ at position $i - p$ in the sequence.*

(ii) In the previous assertion, the word $\mathbf{w}$ and the period p are unique.

Note that the same result holds with a further occurrence and a suffix $\mathbf{w}_{(p)} := w_{|\mathbf{w}|-p+1} \dots w_{|\mathbf{w}|}$, $p \in \mathcal{P}'(\mathbf{w}, \mathbf{w}')$.

3.2. Computation of $\widetilde{\mu}_k(\mathcal{W})$

We can now describe more explicitly what is a k-clump of $\mathcal{W}$ in a sequence. Let a word $\mathbf{c}$ composed of exactly k successive overlapping occurrences $(\mathbf{w}_{r_1}, \mathbf{w}_{r_2}, \dots, \mathbf{w}_{r_k})$ of the family $\mathcal{W}$ with $r_1, \dots, r_k \in \{1, \dots, d\}$. Then, for $j \in \{1, \dots, k-1\}$, each occurrence $\mathbf{w}_{r_j}$ overlaps the occurrence $\mathbf{w}_{r_{j+1}}$ with the corresponding period $p_j \in \mathcal{P}(\mathbf{w}_{r_j}, \mathbf{w}_{r_{j+1}})$ (cf. Figure 1). Moreover, these periods p_j are necessarily principal because $\mathbf{c}$ has to contain exactly k overlapping occurrences of $\mathcal{W}$. Therefore, the word $\mathbf{c}$ has the form

$$\mathbf{c} = \mathbf{w}_{r_1}^{(p_1)} \dots \mathbf{w}_{r_{k-1}}^{(p_{k-1})} \mathbf{w}_{r_k}. \tag{8}$$

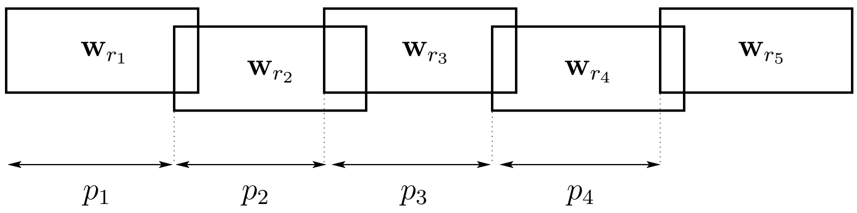

FIGURE 1: Structure of a word composed of exactly 5 occurrences of $\mathcal{W}$.

To simplify the notations, the first word $\mathbf{w}_{r_1}$ (resp. the second word $\mathbf{w}_{r_2}$, the last word $\mathbf{w}_{r_k}$) of $\mathbf{c}$ is denoted by $\mathbf{u}$ (resp. $\mathbf{v}$, $\mathbf{w}$). We denote by $\mathcal{C}_k(\mathcal{W})$ the set of the words of the form (8) and by $\mathcal{C}_k^{(\mathbf{u};\mathbf{w})}(\mathcal{W})$ (resp. $\mathcal{C}_k^{(\mathbf{u},\mathbf{v})}(\mathcal{W})$) the subset of words $\mathbf{c}$ of $\mathcal{C}_k(\mathcal{W})$ which begin with the word $\mathbf{u}$ and end with $\mathbf{w}$ (resp. which have $\mathbf{u}$ and $\mathbf{v}$ as first two occurrences from $\mathcal{W}$). In the latter notation, when $\mathbf{v}$ is unknown, we replace it by a dot (e.g. $\mathcal{C}_k^{(\mathbf{u},\cdot)}(\mathcal{W})$).

A k-clump of $\mathcal{W}$ in $\mathbf{X}$ which begins with $\mathbf{u}$ and ends with $\mathbf{w}$ is then a word $\mathbf{c} \in \mathcal{C}_k^{(\mathbf{u};\mathbf{w})}(\mathcal{W})$ not preceded by any root $\mathbf{u}'^{(p)}$, $\mathbf{u}' \in \mathcal{W}$, $p \in \mathcal{P}'(\mathbf{u}', \mathbf{u})$ and not followed by any suffix $\mathbf{w}'_{(q)}$, $\mathbf{w}' \in \mathcal{W}$, $q \in \mathcal{P}'(\mathbf{w}, \mathbf{w}')$ in $\mathbf{X}$. Since simultaneous occurrences of two different elements of $\mathcal{C}_k(\mathcal{W})$ at position i are impossible in the sequence and using Lemma 1, we obtain the following expression of $\widetilde{Y}_{i,k}(\mathcal{W})$:

$$
\begin{aligned}
\widetilde{Y}_{i,k}(\mathcal{W}) \;=\; & \sum_{\mathbf{u}\in\mathcal{W}} \sum_{\mathbf{w}\in\mathcal{W}} \sum_{\mathbf{c}\in\mathcal{C}_k^{(\mathbf{u};\mathbf{w})}(\mathcal{W})} \Bigg(Y_i(\mathbf{c}) - \sum_{\mathbf{u}'\in\mathcal{W}} \sum_{p\in\mathcal{P}'(\mathbf{u}',\mathbf{u})} Y_{i-p}(\mathbf{u}'^{(p)}\mathbf{c}) \\
& - \sum_{\mathbf{w}'\in\mathcal{W}} \sum_{q\in\mathcal{P}'(\mathbf{w},\mathbf{w}')} Y_i(\mathbf{c}\mathbf{w}'_{(q)}) \\
& + \sum_{\mathbf{u}'\in\mathcal{W}} \sum_{\mathbf{w}'\in\mathcal{W}} \sum_{p\in\mathcal{P}'(\mathbf{u}',\mathbf{u})} \sum_{q\in\mathcal{P}'(\mathbf{w},\mathbf{w}')} Y_{i-p}(\mathbf{u}'^{(p)}\mathbf{c}\mathbf{w}'_{(q)}) \Bigg). \qquad (9)
\end{aligned}
$$

Thus by taking the expectation in expression (9), we obtain the equality:

$$
\begin{aligned}
\widetilde{\mu}_k(\mathcal{W}) &= \sum_{\mathbf{c}\in\mathcal{C}_k(\mathcal{W})} \mu(\mathbf{c}) - 2 \sum_{\mathbf{c}'\in\mathcal{C}_{k+1}(\mathcal{W})} \mu(\mathbf{c}') + \sum_{\mathbf{c}''\in\mathcal{C}_{k+2}(\mathcal{W})} \mu(\mathbf{c}'') \\
&= p_k(\mathcal{W}) - 2p_{k+1}(\mathcal{W}) + p_{k+2}(\mathcal{W}), \qquad (10)
\end{aligned}
$$

where $p_k(\mathcal{W})$ (resp. $p_k^{(\mathbf{u},.)}(\mathcal{W})$) denotes the occurrence probability of a word of $\mathcal{C}_k(\mathcal{W})$ (resp. $\mathcal{C}_k^{(\mathbf{u},.)}(\mathcal{W})$) at a given position. Then, the expression of $\widetilde{\mu}_k(\mathcal{W})$ can be deduced from the one of the $p_k(\mathcal{W})$'s. The computation of $p_k(\mathcal{W})$ is done recursively. For all $k \geq 1$ and $\mathbf{u} = u_1 \dots u_{|\mathbf{u}|} \in \mathcal{W}$,

$$\begin{aligned}
p_1^{(\mathbf{u},.)}(\mathcal{W}) &= \mu(\mathbf{u}) \\
p_{k+1}^{(\mathbf{u},.)}(\mathcal{W}) &= \sum_{\mathbf{v}\in\mathcal{W}} \sum_{\mathbf{c}\in\mathcal{C}_{k+1}^{(\mathbf{u},\mathbf{v})}(\mathcal{W})} \mu(\mathbf{c}) \\
&= \sum_{\mathbf{v}\in\mathcal{W}} \sum_{p\in\mathcal{P}'(\mathbf{u},\mathbf{v})} \sum_{\mathbf{c}'\in\mathcal{C}_k^{(\mathbf{v},.)}(\mathcal{W})} \mu(\mathbf{u}^{(p)}\mathbf{c}') \\
&= \sum_{\mathbf{v}\in\mathcal{W}} \frac{1}{\mu(v_1)} \sum_{p\in\mathcal{P}'(\mathbf{u},\mathbf{v})} \mu(\mathbf{u}^{(p+1)}) \sum_{\mathbf{c}'\in\mathcal{C}_k^{(\mathbf{v},.)}(\mathcal{W})} \mu(\mathbf{c}') \\
&= \sum_{\mathbf{v}\in\mathcal{W}} A_{\mathbf{u},\mathbf{v}} p_k^{(\mathbf{v},.)}(\mathcal{W}), \qquad (11)
\end{aligned}$$

where $A_{\mathbf{u},\mathbf{v}}$ is the probability that an occurrence of $\mathbf{v} = v_1 \cdots v_{|\mathbf{v}|}$ overlaps a previous occurrence of $\mathbf{u}$ in the sequence and that there are no other occurrence of $\mathcal{W}$ in between:

$$A_{\mathbf{u},\mathbf{v}} = \frac{\mu(u_1)}{\mu(v_1)} \sum_{p\in\mathcal{P}'(\mathbf{u},\mathbf{v})} \prod_{t=1}^{p} \pi(u_t, u_{t+1}). \qquad (12)$$

Therefore, if we introduce the vectorial notations $\vec{p}_k(\mathcal{W})$ for the vector $[p_k^{(\mathbf{u},.)}(\mathcal{W})]_{\mathbf{u}\in\mathcal{W}}$ and A for the matrix $[A_{\mathbf{u},\mathbf{v}}]_{\mathbf{u},\mathbf{v}\in\mathcal{W}}$, expression (11) becomes $\forall k \geq 1$, $\vec{p}_{k+1}(\mathcal{W}) = A\vec{p}_k(\mathcal{W})$. Similarly we have $\vec{p}_1(\mathcal{W}) = \vec{\mu}(\mathcal{W}) := [\mu(\mathbf{w})]_{\mathbf{w}\in\mathcal{W}}$, leading to

$$\vec{p}_k(\mathcal{W}) = A^{k-1}\vec{\mu}(\mathcal{W}), \qquad (13)$$

If we denote by $||.||_1$ the 1-norm of $\mathbb{R}^d$ defined by $\forall z = (z_1, \dots, z_d) \in \mathbb{R}^d$, $||z||_1 = \sum_{r=1}^d |z_r|$, we can conclude that

$$\begin{aligned}
p_k(\mathcal{W}) &= ||\vec{p}_k(\mathcal{W})||_1 \\
&= ||A^{k-1}\vec{\mu}(\mathcal{W})||_1. \qquad (14)
\end{aligned}$$

Combining relations (10) and (14) yields to the final expression of $\widetilde{\mu}_k(\mathcal{W})$:

$$\widetilde{\mu}_k(\mathcal{W}) = ||A^{k-1}(I-A)^2\vec{\mu}(\mathcal{W})||_1.$$

This establishes the following proposition.

Proposition 3.1. *For all family $\mathcal{W}$, the occurrence probability of a k-clump of $\mathcal{W}$ is given by*

$$\widetilde{\mu}_k(\mathcal{W}) = ||A^{k-1}(I-A)^2\vec{\mu}(\mathcal{W})||_1, \tag{15}$$

where A is the matrix of coefficients $[A_{\mathbf{u},\mathbf{v}}]_{\mathbf{u},\mathbf{v}\in\mathcal{W}}$ defined in (12), $\vec{\mu}(\mathcal{W})$ is the vector $[\mu(\mathbf{w})]_{\mathbf{w}\in\mathcal{W}}$, and $||.||_1$ is the 1-norm of $\mathbb{R}^d$.

Remarks 3.1.

1. Proposition 3.1 generalizes the result of [12]: indeed, for a single word $\mathcal{W} = \{\mathbf{w}\}$, equation (15) reduces to $\widetilde{\mu}_k(\mathbf{w}) = a_{\mathbf{w}}^{k-1}(1-a_{\mathbf{w}})^2\mu(\mathbf{w})$, where $a_{\mathbf{w}}$ is the occurrence probability of two successive overlapping occurrences of $\mathbf{w}$ and is given by $a(\mathbf{w}) = \sum_{p\in\mathcal{P}'(\mathbf{w})}\prod_{t=1}^{p}\pi(w_t, w_{t+1})$ with $\mathcal{P}'(\mathbf{w}) := \mathcal{P}'_{\{\mathbf{w}\}}(\mathbf{w},\mathbf{w})$.

2. For a family $\mathcal{W}$ such that, $\forall \mathbf{w} \neq \mathbf{w}' \in \mathcal{W}$, $\mathbf{w}$ does not overlap $\mathbf{w}'$ (i.e. $\mathcal{P}(\mathbf{w},\mathbf{w}') = \emptyset$), A is a diagonal matrix, and we find $\widetilde{\mu}_k(\mathcal{W}) = \sum_{\mathbf{w}\in\mathcal{W}} a_{\mathbf{w}}^{k-1}(1-a_{\mathbf{w}})^2\mu(\mathbf{w})$ like in [8].

3. From (10), we can moreover show that

$$\sum_{k\geq 1} k\widetilde{\mu}_k(\mathcal{W}) = \mu(\mathcal{W}) \tag{16}$$

$$\sum_{k\geq 1} \widetilde{\mu}_k(\mathcal{W}) = ||(I-A)\vec{\mu}(\mathcal{W})||_1. \tag{17}$$

4. Proof of the approximation theorem

To prove Theorem 1, we first have to choose the neighborhoods $B_{i,k}$ for all $(i,k) \in I$, where $I := \{1,\dots,n-h+1\} \times \mathbb{N}^\star$, and then to bound the three quantities b_1, b_2 and b_3 defined by (3),(4) and (5). For this, we will adapt the setup presented in [12] for a single word.

4.1. Choice of the neighborhood $B_{i,k}$

We define for each $(i,k) \in I$, a set $Z(i,k) \subset \mathbb{Z}$ which contains all the indexes j of the letters X_j used in the definition of $\widetilde{Y}_{i,k}(\mathcal{W})$. We can take $Z(i,k) = \{s \in \mathbb{Z}$ such that $i-h \le s \le i+(k+1)h\}$ because the length of a k-clump is less than kh and we have to know the $h-1$ letters before and after the clump to control that it does not overlap other occurrences. We now define the neighborhood of (i,k) as the set of $(j,\ell) \in I$ such that $Z(i,k)$ and $Z(j,\ell)$ are separated by at most h positions:

$$B_{i,k} = \{(j,\ell) \in I \text{ such that } -(\ell+3)h \le j-i \le (k+3)h\}.$$

It implies that if $\widetilde{Y}_{i,k}(\mathcal{W}) = \widetilde{Y}_{j,\ell}(\mathcal{W}) = 1$ with $(j,\ell) \notin B_{i,k}$, then the two clumps will be separated by more than $3h$ letters.

4.2. Bounding b_1

From definition (3) we have

$$\begin{aligned} b_1 &= \sum_{i=1}^{n-h+1} \sum_{k\ge 1} \sum_{(j,\ell)\in B_{i,k}} \mathbb{E}(\widetilde{Y}_{i,k}(\mathcal{W}))\mathbb{E}(\widetilde{Y}_{j,\ell}(\mathcal{W})) \\ &\le \sum_{i=1}^{n-h+1} \sum_{k\ge 1} \sum_{\ell\ge 1} \sum_{j=i-(\ell+3)h}^{i+(k+3)h} \widetilde{\mu}_k(\mathcal{W})\widetilde{\mu}_\ell(\mathcal{W}). \end{aligned}$$

Let $\widetilde{\mu}(\mathcal{W})$ be the occurrence probability of a clump of $\mathcal{W}$ at a given position; It satisfies $\widetilde{\mu}(\mathcal{W}) = \sum_{k\ge 1} \widetilde{\mu}_k(\mathcal{W}) \le \mu(\mathcal{W})$. By using the symmetry between i and j and equation (16), we can write

$$\begin{aligned} b_1 &\leq 2\widetilde{\mu}(\mathcal{W}) \sum_{i=1}^{n-h+1} \sum_{k\geq 1}((k+3)h+1)\widetilde{\mu}_k(\mathcal{W}), \\ &\leq 2(n-h+1)\widetilde{\mu}(\mathcal{W})\Big(\big[\mu(\mathcal{W}) + 3\widetilde{\mu}(\mathcal{W})\big]h + \widetilde{\mu}(\mathcal{W})\Big) \\ &\leq 10nh\mu^2(\mathcal{W}). \end{aligned} \tag{18}$$

The last inequality is just obtained by bounding $\widetilde{\mu}(\mathcal{W})$ by $\mu(\mathcal{W})$.

4.3. Bounding b_2

From definition (4) we have

$$b_2 = \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{(j,\ell)\in B_{i,k}\setminus\{(i,k)\}} \mathbb{E}(\widetilde{Y}_{i,k}(\mathcal{W})\widetilde{Y}_{j,\ell}(\mathcal{W}))$$

Since two clumps of different size cannot occur at the same position, the term corresponding to $i = j$ disappears in the sum, and again by symmetry we get

$$b_2 \leq 2 \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{\ell\geq 1} \sum_{j=i+1}^{i+(k+3)h} \mathbb{E}(\widetilde{Y}_{i,k}(\mathcal{W})\widetilde{Y}_{j,\ell}(\mathcal{W})).$$

Let denote by $\widetilde{Y}_j(\mathcal{W}) = \sum_{\ell\geq 1} \widetilde{Y}_{j,\ell}(\mathcal{W})$ the Bernoulli variable that is one if a clump of $\mathcal{W}$ occurs at position j and zero otherwise. Since $\widetilde{Y}_{i,k}(\mathcal{W}) = \sum_{\mathbf{c}\in\mathcal{C}_k(\mathcal{W})} \widetilde{Y}_{i,k}(\mathcal{W})Y_i(\mathbf{c})$, we have

$$\begin{aligned} b_2 &\leq 2 \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{j=i+1}^{i+(k+3)h} \mathbb{E}(\widetilde{Y}_{i,k}(\mathcal{W})\widetilde{Y}_j(\mathcal{W})), \\ &\leq 2 \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{\mathbf{c}\in\mathcal{C}_k(\mathcal{W})} \sum_{j=i+1}^{i+(k+3)h} \mathbb{E}\big(\widetilde{Y}_{i,k}(\mathcal{W})Y_i(\mathbf{c})\widetilde{Y}_j(\mathcal{W})\big). \end{aligned}$$

Since a clump of length $|\mathbf{c}|$ which begins at position i cannot overlap a clump starting at position $i+1 \leq j < i + |\mathbf{c}|$, and since $\widetilde{Y}_j(\mathcal{W}) \leq Y_j(\mathcal{W})$, it follow that

$$\begin{aligned}
b_2 &\leq 2\sum_{i=1}^{n-h+1}\sum_{k\geq 1}\sum_{\mathbf{c}\in\mathcal{C}_k(\mathcal{W})}\sum_{j=i+|\mathbf{c}|}^{i+(k+3)h}\mathbb{E}\big(\widetilde{Y}_{i,k}(\mathcal{W})Y_i(\mathbf{c})\widetilde{Y}_j(\mathcal{W})\big)\\
&\leq 2\sum_{i=1}^{n-h+1}\sum_{k\geq 1}\sum_{\mathbf{c}\in\mathcal{C}_k(\mathcal{W})}\sum_{j=i+|\mathbf{c}|+h}^{i+(k+3)h}\mathbb{E}\big(\widetilde{Y}_{i,k}(\mathcal{W})Y_i(\mathbf{c})Y_j(\mathcal{W})\big)\\
&\quad +2\sum_{i=1}^{n-h+1}\sum_{k\geq 1}\sum_{\mathbf{c}\in\mathcal{C}_k(\mathcal{W})}\sum_{j=i+|\mathbf{c}|}^{i+|\mathbf{c}|+h-1}\mathbb{E}\big(\widetilde{Y}_{i,k}(\mathcal{W})Y_i(\mathbf{c})Y_j(\mathcal{W})\big).
\end{aligned}$$

The first term (resp. the second term) in the right-hand side is denoted by b_{21} (resp. b_{22}). Let bound b_{21}. Note that the random variable $\widetilde{Y}_{i,k}(\mathcal{W})Y_i(\mathbf{c})$ only involves the letters $X_{i-h+1}\dots X_{i+|\mathbf{c}|+h-1}$ whereas $Y_j(\mathcal{W})$ involves $X_j\dots X_{j+h-1}$. Therefore, for every position j which satisfies $j\geq i+|\mathbf{c}|+h$, the property of Markov gives

$$\mathbb{E}\big(\widetilde{Y}_{i,k}(\mathcal{W})Y_i(\mathbf{c})Y_j(\mathcal{W})\big)\leq\frac{\mu(\mathcal{W})}{\mu_{\min}}\mathbb{E}\big(\widetilde{Y}_{i,k}(\mathcal{W})Y_i(\mathbf{c})\big),$$

where $\mu_{\min}=\min_{\mathbf{w}\in\mathcal{W}}\mu(w_1)>0$. Since the sum over j contains less than $(k+2)h$ terms, we get

$$\begin{aligned}
b_{21} &\leq 2(n-h+1)\frac{\mu(\mathcal{W})}{\mu_{\min}}\sum_{k\geq 1}(k+2)h\widetilde{\mu}_k(\mathcal{W})\\
&\leq 2(n-h+1)\frac{\mu(\mathcal{W})}{\mu_{\min}}\big(\mu(\mathcal{W})+2\widetilde{\mu}(\mathcal{W})\big)h\\
&\leq \frac{6nh}{\mu_{\min}}\mu^2(\mathcal{W}). \qquad (19)
\end{aligned}$$

To bound b_{22}, we write $\mathbb{E}\big(\widetilde{Y}_{i,k}(\mathcal{W})Y_i(\mathbf{c})Y_j(\mathcal{W})\big)\leq\mathbb{E}\big(\widetilde{Y}_i(\mathcal{W})Y_i(\mathbf{c})Y_j(\mathcal{W})\big)$ and we note that the random variable $\widetilde{Y}_i(\mathcal{W})Y_i(\mathbf{c})$ involves the letters $X_{i-h+1}\dots X_{i+|\mathbf{c}|-1}$ whereas $Y_j(\mathcal{W})$ involves $X_j\dots X_{j+h-1}$. Therefore, for every position j which satisfies $j\geq i+|\mathbf{c}|$, we have

$$\mathbb{E}\big(\widetilde{Y}_i(\mathcal{W})Y_i(\mathbf{c})Y_j(\mathcal{W})\big)\leq\frac{\mu(\mathcal{W})}{\mu_{\min}}\mathbb{E}\big(\widetilde{Y}_i(\mathcal{W})Y_i(\mathbf{c})\big).$$

Thus, we derive the following bound for b_{22}:

$$
\begin{aligned}
b_{22} &\leq \frac{2(n-h+1)h}{\mu_{\min}}\mu(\mathcal{W})\sum_{k\geq 1}\sum_{\mathbf{c}\in\mathcal{C}_k(\mathcal{W})}\mathbb{E}\big(\widetilde{Y}_i(\mathcal{W})Y_i(\mathbf{c})\big) \\
&\leq \frac{2nh}{\mu_{\min}}\mu^2(\mathcal{W}). \qquad (20)
\end{aligned}
$$

Indeed,

$$
\begin{aligned}
&\sum_{k\geq 1}\sum_{\mathbf{c}\in\mathcal{C}_k(\mathcal{W})}\mathbb{E}\big(\widetilde{Y}_i(\mathcal{W})Y_i(\mathbf{c})\big) = \qquad (21) \\
&\quad = \sum_{k\geq 1}\mathbb{P}(\text{a } K\text{-clump of } \mathcal{W} \text{ starts at position } i \text{ with } K \geq k) \\
&\quad = \sum_{k\geq 1}\sum_{K\geq k}\widetilde{\mu}_K(\mathcal{W}) = \sum_{K\geq 1}K\widetilde{\mu}_K(\mathcal{W}) = \mu(\mathcal{W}).
\end{aligned}
$$

Finally, combining equations (19) and (20) leads to

$$
b_2 \leq \frac{8nh}{\mu_{\min}}\mu^2(\mathcal{W}). \qquad (22)
$$

4.4. Bounding b_3

From definition (5), we have

$$
b_3 = \sum_{i=1}^{n-h+1}\sum_{k\geq 1}\mathbb{E}\Big|\mathbb{E}\big(\widetilde{Y}_{i,k}(\mathcal{W}) - \widetilde{\mu}_k(\mathcal{W})|\sigma(\widetilde{Y}_{j,\ell}(\mathcal{W}),(j,\ell)\notin B_{i,k})\big)\Big|.
$$

We denote by $\mathcal{C}'_k$ the set of the words $\mathbf{rcs}$ such that $\mathbf{c}\in\mathcal{C}_k$, $|\mathbf{r}| = |\mathbf{s}| = h$ and $\mathbf{c}$ is a k-clump of $\mathcal{W}$ in the sequence $\mathbf{rcs}$. An occurrence of a word of $\mathcal{C}'_k$ is then equivalent to an occurrence of a k-clump of $\mathcal{W}$: $\widetilde{Y}_{i,k}(\mathcal{W}) = \sum_{\mathbf{rcs}\in\mathcal{C}'_k} Y_{i-h}(\mathbf{rcs})$. Moreover, for all $\mathbf{c}\in\mathcal{C}_k$, we deduce from the definition of the neighborhood $B_{i,k}$ that

$$
\sigma(\widetilde{Y}_{j,\ell}(\mathcal{W}),(j,\ell)\notin B_{i,k}) \subset \sigma(\ldots, X_{i-2h-1}, X_{i-2h}, X_{i+|\mathbf{c}|+2h}, X_{i+|\mathbf{c}|+2h+1},\ldots).
$$

Therefore, thanks to the Markov property, we have

$$\begin{aligned} b_3 &\leq \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{\mathbf{rcs}\in\mathcal{C}'_k} \mathbb{E}\Big|\mathbb{E}\big(Y_{i-h}(\mathbf{rcs}) - \mu(\mathbf{rcs})|\sigma(\ldots, X_{i-2h}, X_{i+|\mathbf{c}|+2h}, \ldots)\big)\Big| \\ &\leq \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{\mathbf{rcs}\in\mathcal{C}'_k} \mathbb{E}\Big|\mathbb{E}\big(Y_{i-h}(\mathbf{rcs}) - \mu(\mathbf{rcs})|X_{(i-h)-h}, X_{(i-h)+|\mathbf{rcs}|+h}\big)\Big| \end{aligned}$$

Now we use the following result proved by [13]: for all word $\mathbf{w}$ and all integers j and t,

$$\mathbb{E}\big|\mathbb{E}(Y_j(\mathbf{w}) - \mu(\mathbf{w})|X_{j-t}, X_{j+|\mathbf{w}|+t})\big| \leq C'\mu(\mathbf{w})|\alpha|^t, \tag{23}$$

where C' is a positive constant that only depend on the matrix Π, and α is the second largest eigenvalue in modulus of the matrix Π (we have $|\alpha| < 1$). It leads to

$$b_3 \leq \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{\mathbf{rcs}\in\mathcal{C}'_k} C'\mu(\mathbf{rcs})|\alpha|^h.$$

Finally, the equality $\sum_{k\geq 1}\sum_{\mathbf{rcs}\in\mathcal{C}'_k}\mu(\mathbf{rcs}) = \widetilde{\mu}(\mathcal{W})$ gives

$$\begin{aligned} b_3 &\leq C'(n-h+1)|\alpha|^h\widetilde{\mu}(\mathcal{W}) \\ &\leq C'n\mu(\mathcal{W})|\alpha|^h. \end{aligned} \tag{24}$$

Inequalities (18), (22) and (24) establish Theorem 1.

5. Clumps and competing renewals

When counting the occurrences of a word or a word family in a finite sequence $X_1\cdots X_n$, one may be interested in counting only non overlapping occurrences, for instance clumps or renewals. A renewal can be defined as follows: an occurrence is a renewal if and only if either it is the first one or it does not overlap a previous renewal. For a word family, they are called *competing* renewals. Various results have been obtained for the distribution of the number of clumps and the number of competing renewals (see [5], Chapter 6 and references therein). New Poisson approximations directly follows from Theorem 1.

First of all, inequalities (2), (18), (22) and (24) lead to

$$d_{\text{tv}}\big(\mathcal{L}(\widetilde{N}^{\infty}(\mathcal{W})), \mathcal{P}(\widetilde{\lambda})\big) \leq Cnh\mu^2(\mathcal{W}) + C'n\mu(\mathcal{W})|\alpha|^h, \tag{25}$$

where $\widetilde{N}^{\infty}(\mathcal{W}) := \sum_{k\geq 1} \widetilde{N}^{\infty}_k(\mathcal{W})$ and, by using (17), $\widetilde{\lambda} = \mathbb{E}(\widetilde{N}^{\infty}(\mathcal{W})) = (n-h+1)||(I-A)\vec{\mu}(\mathcal{W})||_1$. Moreover, $\widetilde{N}^{\infty}(\mathcal{W})$ has asymptotically the same distribution than the number $\widetilde{N}(\mathcal{W})$ of clumps of $\mathcal{W}$ in $X_1 \cdots X_n$: $\mathbb{P}(\widetilde{N}^{\infty}(\mathcal{W}) \neq \widetilde{N}(\mathcal{W})) \leq h\mu(\mathcal{W})$ (same argument than for N^{∞}). Therefore, under both $h = \mathrm{o}(n)$ and the rare condition $n\mu(\mathcal{W}) = \mathrm{O}(1)$, the total variation distance between the distribution of $\widetilde{N}(\mathcal{W})$ and the Poisson distribution $\mathcal{P}(\widetilde{\lambda})$ tends to zero.

Second, one can show that the distribution of the number $R(\mathcal{W})$ of competing renewals of $\mathcal{W}$ is asymptotically identical to the one of the number of clumps:

$$d_{\text{tv}}\big(\mathcal{L}(R(\mathcal{W})), \mathcal{L}(\widetilde{N}(\mathcal{W}))\big) \leq \mathbb{P}(R(\mathcal{W}) \neq \widetilde{N}(\mathcal{W})) \leq \frac{1}{\mu_{\min}} nh\mu^2(\mathcal{W}), \tag{26}$$

where $\mu_{\min} = \min_{\mathbf{w}\in\mathcal{W}} \mu(w_1) > 0$. Indeed, we can notice that if all the clumps are such that the occurrence of $\mathcal{W}$ they start with overlaps the occurrence of $\mathcal{W}$ they end with, then $R(\mathcal{W}) = \widetilde{N}(\mathcal{W})$. Thus, if $R(\mathcal{W}) \neq \widetilde{N}(\mathcal{W})$, then it exists (at least) one clump whose first and last occurrences from $\mathcal{W}$ do not overlap. Let i be the position of such a clump and $\mathbf{u}$ be the occurrence from $\mathcal{W}$ it starts with. It implies that both an occurrence of $\mathbf{u}$ starts at position i and an occurrence of $\mathcal{W}$ starts between positions

$i+|\mathbf{u}|$ and $i+|\mathbf{u}|+h-1$; this occurs with probability $h\mu(\mathbf{u})\mu(\mathcal{W})/\mu_{\min}$. Summing over $i \in \{1,\dots,n-h+1\}$ and $\mathbf{u} \in \mathcal{W}$ leads to inequality (26). Thanks to the triangular inequality, we then get the following Poisson approximation for the number of competing renewals

$$d_{\mathrm{tv}}\big(\mathcal{L}(R(\mathcal{W})), \mathcal{P}(\widetilde{\lambda})\big) = \mathrm{O}(nh\mu^2(\mathcal{W}) + n\mu(\mathcal{W})|\alpha|^h + h\mu(\mathcal{W})). \qquad (27)$$

If $n\mu(\mathcal{W}) = \mathrm{O}(1)$ and $h = \mathrm{o}(n)$ then the total variation distance between the distribution of $R(\mathcal{W})$ and the Poisson distribution $\mathcal{P}(\widetilde{\lambda})$ tends to zero. This Poisson distribution is in fact very close to the natural limiting Poisson distribution with parameter $\mathbb{E}R(\mathcal{W})$ proposed by [3] because both parameters are asymptotically equivalent under the rare condition and $h = \mathrm{o}(n)$. However, calculating $\mathbb{E}R(\mathcal{W})$ requires in practice to solve a system of equations whereas the expression of $\widetilde{\lambda}$ is explicit.

6. Generalizations and Conclusion

We have provided a new compound Poisson distribution with explicit parameters to approximate the count of overlapping occurrences of a word family in a stationary Markov chain of length n. The error of approximation converges to zero as soon as the word family $\mathcal{W}$ is expectedly rare ($\mathbb{E}N(\mathcal{W}) = \mathrm{O}(1)$) and the maximal word length is of smaller order than n.

Our results can be easily extended to the case of a Markov chain of order m, with $2 \le m \le \min\{|\mathbf{w}|, \mathbf{w} \in \mathcal{W}\} - 1$. It suffices to consider the sequence $\mathbf{X}^*$ obtained by putting $X_i^* := X_i X_{i+1} \cdots X_{i+m-1}$: $\mathbf{X}^*$ is a Markov chain of order 1 on the $\mathcal{A}^* := \mathcal{A}^m$ alphabet. Moreover, an occurrence of $\mathcal{W}$ in $\mathbf{X}$ corresponds to an occurrence of $\mathcal{W}^*$ in $\mathbf{X}^*$ and vice versa, where $\mathcal{W}^*$ is the word family $\mathcal{W}$ written on the new alphabet $\mathcal{A}^*$. The parameters of the limiting compound Poisson distribution will then be $||A_{(m)}^{k-1}(I - A_{(m)})^2\vec{\mu}(\mathcal{W})||_1$, where $A_{(m)}$ is the matrix whose coefficient $[\mathbf{u}, \mathbf{v}]$ is given by

$$\frac{\mu(u_1 \cdots u_m)}{\mu(v_1 \cdots v_m)} \sum_{\substack{p \in \mathcal{P}'(\mathbf{u},\mathbf{v}) \\ p \leq |\mathbf{u}|-m}} \prod_{t=1}^{p} \pi(u_t \ldots u_{t+m-1}, u_{t+m}),$$

with $\pi(\cdot,\cdot)$ and $\mu(\cdot)$ standing for the transition probabilities and the stationary distribution of the model. This compound Poisson distribution has been included into the *R'MES* software* to find exceptional motifs in DNA sequences.

Our compound Poisson approximation for the count of any rare word family in a Markov chain, together with a Gaussian approximation or the exact distribution, is extremely useful when one models the sequence by a hidden Markov chain. Indeed, a hidden Markov chain $(\mathbf{X}, \mathbf{S})$ on the alphabet $\mathcal{A}$ with state space $\{1, \ldots, s\}$ can be written like a one-order Markov chain $\overline{\mathbf{X}}$ on the alphabet $\mathcal{A} \times \{1, \ldots, s\}$ and an occurrence of a given word $\mathbf{w}$ in $\mathbf{X}$ would correspond to an occurrence of a word family $\overline{\mathcal{W}}$ in $\overline{\mathbf{X}}$. For instance, if there are two states 1 and 2, the word family $\overline{\mathcal{W}}$ associated with $\mathbf{w} = \mathtt{aca}$ is $\{\mathtt{a}_1\mathtt{c}_1\mathtt{a}_1, \mathtt{a}_1\mathtt{c}_1\mathtt{a}_2, \mathtt{a}_1\mathtt{c}_2\mathtt{a}_1, \mathtt{a}_2\mathtt{c}_1\mathtt{a}_1, \mathtt{a}_1\mathtt{c}_2\mathtt{a}_2, \mathtt{a}_2\mathtt{c}_1\mathtt{a}_2,$ $\mathtt{a}_2\mathtt{c}_2\mathtt{a}_1, \mathtt{a}_2\mathtt{c}_2\mathtt{a}_2\}$ where $\mathtt{a}_j$ (resp. $\mathtt{c}_j$) stands for the letter $\mathtt{a}$ (resp. $\mathtt{c}$) in state j.

* `http://genome.jouy.inra.fr/ssb/rmes/`

Acknowledgement

The authors thank an anonymous reviewer for his/her helpful comments. This work has been supported by the French Action Concertée Incitative IMPBio.

References

[1] ARRATIA, R., GOLDSTEIN, L. and GORDON, L. (1990). Poisson approximation and the Chen-Stein method. *Statistical Science.* **5** 403–434.<

[2] CHRYSSAPHINOU, O. and PAPASTAVRIDIS, S. (1990). The occurrence of sequence patterns in repeated dependent experiments. *Theory of Probability and its Applications.* **35** 145–152.

[3] CHRYSSAPHINOU, O., PAPASTAVRIDIS, S. and VAGGELATOU, E. (2001). Poisson approximation for the non-overlapping appearances of several words in Markov chains. *Combinatorics, Probability and Computing.* **10** 293–308.

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

[5] LOTHAIRE, M. (2005). *Applied combinatorics on words.* Cambridge University Press.

[6] PRUM, B., RODOLPHE, F. and TURCKHEIM, É. DE (1995). Finding words with unexpected frequencies in DNA sequences. *J. R. Statist. Soc. B.* **57** 205–220.

[7] REINERT, G., SCHBATH, S. and WATERMAN, M. (2000). Probabilistic and statistical properties of words. *J. Comp. Biol.* **7** 1–46.

[8] REINERT, G. and SCHBATH, S. (1998). Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains. *J. Comp. Biol.* **5** 223–253.

[9] Robin, S. and Daudin, J.-J. (1999). Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Prob.* **36** 179–193.

[10] Robin, S. and Schbath, S. (2001). Numerical comparison of several approximations of the word count distribution in random sequences. *J. Comp. Biol.* **8** 349–359.

[11] Régnier, M. (2000). A unified approach to word occurrence probabilities. *Discrete Applied Mathematics.* **104** 259–280.

[12] Schbath, S. (1995a). Compound Poisson approximation of word counts in DNA sequences. *ESAIM: Probability and Statistics.* **1** 1–16.

[13] Schbath, S. (1995b). *Étude asymptotique du nombre d'occurrences d'un mot dans une chaîne de Markov et application à la recherche de mots de fréquence exceptionnelle dans les séquences d'ADN.* PhD thesis, Université René Descartes, Paris V.