COMPOUND POISSON APPROXIMATION OF WORD COUNTS IN DNA SEQUENCES

SOPHIE SCHBATH

ABSTRACT. Identifying words with unexpected frequencies is an important problem in the analysis of long DNA sequences. To solve it, we need an approximation of the distribution of the number of occurrences $N(W)$ of a word W. Modeling DNA sequences with m-order Markov chains, we use the Chen-Stein method to obtain Poisson approximations for two different counts. We approximate the "declumped" count of W by a Poisson variable and the number of occurrences $N(W)$ by a compound Poisson variable. Combinatorial results are used to solve the general case of overlapping words and to calculate the parameters of these distributions.

1. INTRODUCTION

Because of many important sequencing projects, biologists now have large sets of DNA sequences from many different organisms. They need quantitative tools and automatic methods to help them in analyzing sequences. Statistics, computer science and graphical representations have already provided a lot of useful ways to analyze sequences.

A simple representation of a sequence is a finite series $X_1X_2\cdots X_n$ of letters taken from the alphabet $\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, with the four letters corresponding to the four bases adenine, cytosine, guanine and thymine. Quite a number of sub-sequences, or words, have a known biological function. Each single occurrence may interact with proteins during biological processes as replication, translation, repairing. Moreover, the statistical repetition of a given word or of a group of words, may be related to a biological code (Trifonov, 1989).

Therefore, the question of identifying words W with an unexpected frequency with respect to a given model is of interest. Here we study the asymptotic distribution of $N(W)$, the number of occurrences of a given word W in a sequence. These occurrences can overlap if W has a periodic composition. In this paper, we model the sequence $X_1X_2\cdots X_n$ with an homogeneous Markov chain of order m on the state space $\mathcal{A}$. This simple model is useful to identify exceptional long words, given $(m+1)$-words frequencies.

Prum *et al.* (1995) and Schbath *et al.* (1995) study the normal approximation of $N(W)$ corresponding to the asymptotic frame where the expectation of $N(W)$ converges to infinity with n. If the expectation of $N(W)$ is bounded when n increases, we then say that W is rare, and Poisson approximations are more reasonable. We show here that the number of overlapping

Received by the journal February 1, 1995. Revised September 4, 1995. Accepted for publication October 26, 1995.

occurrences of a rare word can be approximated by a compound Poisson variable, which reduces to a Poisson variable if the word W cannot overlap itself. As we do not want the model to depend on n, we consider a series of words W_n, with length h_n converging slowly to infinity at a rate greater or equal to $\log n$, so that $\mathbb{E}N(W_n)$ is bounded.

In the literature, the case of i.i.d. variables X_i has been widely considered. The convergence of $N(W)$ for rare words to a Poisson variable is then proved either with generating functions or by using the Chen-Stein method (Chryssaphinou and Papastavridis (1988a, 1988b), Godbole (1991), Hirano and Aki (1993), Godbole and Schaffner (1993), Fu (1993)). When the sequence $(X_i)_{i=1\cdots n}$ is a first order Markov chain on $\{0,1\}$ and W is a run of ones, some of these authors show the convergence of $N(W)$ to a Poisson or compound Poisson variable when $h_n \to +\infty$. Others show the convergence when the transition probabilities $\pi(1,1)$ and $\pi(0,1)$ converge to zero. Geske *et al.* (1995) considered the case of a first order Markov chain with states in a general alphabet. They proved the compound Poisson convergence for rare words with a single *principal period*. More refinements are required in the general case when a word overlaps in more ways than those associated with one principal period.

In this paper, we consider this general case, using combinatory results to get the whole set of overlapped compound words based on W. We suppose that the Markov chain is of order one and stationary. We do not loose generality, because we may write an m-order chain on $\mathcal{A}$ as a first-order chain on $\mathcal{A}^m$. If W can appear in clumps, the probability of a second occurrence after a first occurrence is different from that of an isolated one, so that the number of occurrences is well approximated by a compound Poisson variable, while the number of clumps is approximated by a Poisson variable. This result is easily shown using the Chen-Stein method (Chen (1975), Arratia *et al.* (1989), Barbour *et al.* (1992b)). The method used here is very similar to the method in Karlin and Ost (1987), Arratia *et al.* (1990), Godbole and Schaffner (1993).

The Chen-Stein method gives a bound of the total variation distance between the distribution of a sum of non i.i.d. Bernoulli variables Y_i, $i \in I$, and the distribution of a Poisson variable with parameter $\lambda = \sum_{i\in I} \mathbb{E}Y_i$. In our case, Y_i is one if $W = w_1 w_2 \cdots w_h$ occurs at position i, zero otherwise:

$$Y_i = \mathbb{1}\{X_i = w_1,\, X_{i+1} = w_2, \ldots, X_{i+h-1} = w_h\},$$

and $N(W) = \sum_{i=1}^{n-h+1} Y_i$.

Chen-Stein theorem says that

$$d_{\mathrm{TV}}(\mathcal{L}(N(W)), \mathcal{P}\mathrm{o}(\lambda)) \leq 2(b_1 + b_2 + b_3)\,,$$

where

$$\begin{aligned} b_1 &= \sum_{i\in I}\sum_{j\in B_i} \mathbb{E}Y_i\,\mathbb{E}Y_j\,, \\ b_2 &= \sum_{i\in I}\sum_{j\in B_i\setminus\{i\}} \mathbb{E}(Y_iY_j)\,, \\ b_3 &= \sum_{i\in I} \mathbb{E}|\mathbb{E}(Y_i - \mathbb{E}Y_i|\sigma(Y_j, j\in B_i^c))|\,, \end{aligned}$$

$I = \{1, \ldots, n-h+1\}$ and $B_i \subset I$ is a neighborhood of i.

We choose B_i as a set of indices ℓ so that, for j not in B_i, there are no common X_k to Y_i and Y_j. Moreover, we want the X_k defining Y_i and those defining Y_j to be separated by at least r positions, with some $r > 0$. Taking $r/\log n$ greater than a certain constant is sufficient to have b_3 converging to zero, as shown in the Appendix. Therefore, we choose B_i as the indices ℓ such that

$$B_i = \{i - r - h + 2, \ldots, i + r + h - 2\} \cap I.$$

Therefore,

$$b_1 \le (n - h + 1)(2r + 2h - 3)\mu^2(W)$$

where $\mu(W) = \mathbb{E}Y_i$. If r and h are both $\mathrm{o}(n)$ and $h/\log n$ is greater than some fixed number C, then $n\mu(W) = \mathrm{O}(1)$ and b_1 is $\mathrm{o}(1)$. Let us consider this asymptotic framework.

The second term b_2 is also written

$$b_2 \le 2 \sum_{i \in I} \sum_{\ell \in \{1,\ldots,r+h-2\}} \mathbb{E}(Y_i Y_{i+\ell})$$

using the symmetry of B_i. Considering separately the indices ℓ where $\ell < h$ corresponds to overlapping occurrences of W, and $\ell \ge h$ corresponds to non overlapping occurrences, we get two terms

$$\begin{aligned} b_2' &= 2 \sum_{i \in I} \sum_{1 \le \ell < h} \mathbb{E}(Y_i Y_{i+\ell}) \\ b_2'' &= 2 \sum_{i \in I} \sum_{h \le \ell \le r+h-2} \mathbb{E}(Y_i Y_{i+\ell}). \end{aligned}$$

For $\ell \ge h$,

$$\mathbb{E}(Y_i Y_{i+\ell}) = (\mu(w_1))^{-1} \Pi^{\ell-h+1}(w_1, w_h)\mu^2(W),$$

where Π is the transition matrix of the model and μ is the invariant probability. Therefore b_2'' is of order $\mathrm{O}(nr\mu^2(W))$ and converges to zero.
For $\ell < h$, there are non-zero terms if the word W overlaps, that is when there exists an integer p, $1 \le p \le h-1$, such that $w_i = w_{i+p}$ for $i = 1, \ldots, h-p$. Such an integer is called a *period* of W. We denote $\mathcal{P}(W)$ the set of periods of W. If ℓ is not in $\mathcal{P}(W)$, $\mathbb{E}(Y_i Y_{i+\ell}) = 0$. If ℓ is in $\mathcal{P}(W)$, $\mathbb{E}(Y_i Y_{i+\ell}) = \mu(W^{(p)}W)$ where $W^{(p)}W$ is the concatenated word $w_1 w_2 \cdots w_p w_1 \cdots w_h$. In general, b_2' is not $\mathrm{O}(nh\mu^2(W))$. For instance, if p is bounded, it is $\mathrm{O}(nh\mu(W))$, which does not converge to zero.

Therefore, it is necessary to declump the count of occurrences and consider $\widetilde{N}(W) = \sum_{i \in I} \widetilde{Y}_i$, where $\widetilde{Y}_i$ only counts occurrences which do not overlap a preceding occurrence :

$$\widetilde{Y}_i = Y_i(1 - Y_{i-1}) \cdots (1 - Y_{i-h+1}).$$

In the term $\widetilde{b}_2$, associated with $\widetilde{N}(W)$, we have now $\mathbb{E}(\widetilde{Y}_i \widetilde{Y}_{i+\ell}) = 0$ for $\ell < h$, solving the previous problem. For other cases, we use $\widetilde{Y}_i \le Y_i$. Therefore, if we take the neighborhood $\widetilde{B}_i = \{i - r - 2h + 3, \ldots, i + r + 2h - 3\} \cap I$, the second term $\widetilde{b}_2$ is of order $\mathrm{O}(n(r+h)\mu^2(W))$, while $\widetilde{b}_1$ is $\mathrm{O}(n(r+2h)\widetilde{\mu}^2(W))$. Moreover, we prove in the Appendix that the third term $\widetilde{b}_3$ is of order $\mathrm{O}(n\rho^r)$ for some $0 < \rho < 1$, where ρ only depends on the Markov chain. Since

$r > \log n / \log \rho^{-1}$, $\widetilde{b}_3$ converges to zero. We therefore have the following theorem.

THEOREM 1.

$$d_{TV}\Big(\mathcal{L}(\widetilde{N}(W)), \mathcal{P}o((n-h+1)\widetilde{\mu}(W))\Big) \leq$$
$$(n-h+1)\left(A_1(2h+r)\widetilde{\mu}^2(W) + A_2(h+r)\mu^2(W)\right) + A_3 n\rho^r ,$$

where A_1, A_2, A_3 are of order O(1).

REMARK 2. The definition of $\widetilde{N}(W) = \sum_{i\in I} \widetilde{Y}_i$ supposes that $X_0, X_{-1}, \dots, X_{-h+2}$ are known. In practical situation, we use the number of clumps $\widetilde{N}^*(W) = \sum_{i\in I} \widetilde{Y}_i^*$, where $\widetilde{Y}_1^* = Y_1$, $\widetilde{Y}_i^* = Y_i \prod_{j=1}^{i-1}(1-Y_j)$ if $1 < i < h$, and $\widetilde{Y}_i^* = \widetilde{Y}_i$ otherwise. As the total variation between $\widetilde{N}(W)$ and $\widetilde{N}^*(W)$ is bounded by $2\,\mathbb{P}\{\widetilde{N}(W) \neq \widetilde{N}^*(W)\}$, and

$$\mathbb{P}\{\widetilde{N}(W) \neq \widetilde{N}^*(W)\} \leq h\mu(W),$$

both counts have the same asymptotic distribution.

The next point is to calculate $\widetilde{\mu}(W) = \mathbb{E}\widetilde{Y}_i$. This is easy using the set $\mathcal{P}(W)$ of the periods of W, as shown in the next section. Finally, an estimation of $\widetilde{\mu}(W)$ is necessary for a statistical use of the theorem. We use the fact that the total variation distance between two Poisson variables of parameters λ and λ' is less than $|\lambda - \lambda'|$. In Markov chain model, using the LIL, the estimator of $\mu(W_n)$ defined with the MLE of the parameters (plug-in estimator) is such that $(n-h+1)\,(\widehat{\mu}(W_n) - \mu(W_n))$ =o(1) a.e. This solves the problem of the approximation of the declumped count.

To approximate the count $N(W)$, we write

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

where $\widetilde{N}^{(k)}(W)$ is the number of k-clumps. We say that a k-clump occurs at position i if there exists a concatenated word C composed of exactly k overlapping occurrences of W, and if there is an occurrence of C at position i which does not overlap any other occurrence of W in the sequence. We denote by $\widetilde{Y}_i^{(k)}$ the variable that is one if a k-clump occurs in i, and is zero otherwise. For example, for the word W =`AACAA`, the sequence `TGAACAAACAACAATAGAACAAAA` has a 3-clump at $i = 3$ and an isolated occurrence, or 1-clump, at $i = 18$. We use the process version of the Chen-Stein theorem as stated by Arratia *et al.* (1990) for the process $\mathbb{Y} = (\widetilde{Y}_i^{(k)})_{i,k}$. The process $(\widetilde{N}^{(k)}(W))_k$ is therefore approximated by a Poisson process $(Z^{(k)})_k$ with parameter $\Lambda = \Big((n-h+1)\widetilde{\mu}_k = (n-h+1)\mathbb{E}\widetilde{Y}_i^{(k)}\Big)_k$, and $Z = \sum_{k\geq 1} k\, Z^{(k)}$ is a compound Poisson variable with parameter Λ. Again, we need combinatory results to calculate $\widetilde{\mu}_k$ (Sections 2 and 3). The use of the process version of Chen-Stein theorem needs a definition of the neighborhood $B_{i,k}$ of the double index (i,k), and its application to our problem requires careful calculation of the bounds in order to deal with the case when we have words with more than one principal period. This is done in Section 4.

2. Combinatory results

In this section, we give a simple expression of the two variables, $\widetilde{Y}_i = Y_i(1-Y_{i-1})\cdots(1-Y_{i-h+1})$ and $\widetilde{Y}_i^{(k)}$, defined in the previous section, in order to calculate their expectations, denoted $\widetilde{\mu}$ and $\widetilde{\mu}_k$. If we develop the expression $Y_i(1-Y_{i-1})\cdots(1-Y_{i-h+1})$, we obtain 2^{h-1} terms that are products like $Y_iY_{i-j_1}\cdots Y_{i-j_\ell}$; a lot of them are equal to zero because they do not correspond to possible overlaps of W. If we know the periods of the word, we simply get the terms of this sum, which have a positive expectation. We show that they are necessarily written as Y_iY_{i-p} where p is a period of W. We then use the same ideas to describe the set of concatenated words C composed of exactly k overlapping occurrences of W. Conversely, every possible overlap corresponds to a period of the word.

We study now the set of periods of a given word.

Definition 3. Let $W = w_1\cdots w_h$ be a given word of length h and $p \in \{1,\ldots,h-1\}$. p is a *period* of W if $w_i = w_{i+p}$, $i \in \{1,\ldots,h-p\}$.
For each period p, the word $w_1\cdots w_p$, denoted $W^{(p)}$, is called a *root* of W.

Example 4. The periods of the word AAACAAAACAAA are 5, 9, 10 and 11.

Remark 5. The set $\mathcal{P}(W)$ of the periods of W is empty if and only if W cannot overlap itself.

If p is a period of W, we decompose the word as $W = (yz)^r y$ where yz is the root $W^{(p)}$, z is a non-empty word and $r \geq 1$ (Figure 1). We

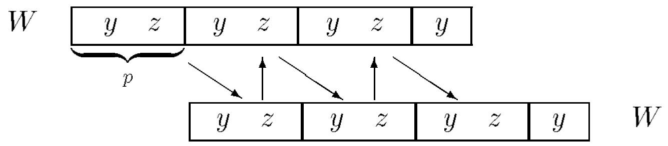

Figure 1. Periodic decomposition of a word W with period p.

also consider the *canonical decomposition* of W which is associated with the minimal period p_0. The root $W^{(p_0)}$ is then called the *minimal root.*

The question of finding the whole set of periods of a word W is usually solved in a recursive way (Guibas and Odlyzko, 1981). Suppose the minimal period p_0 is known and $yz = W^{(p_0)}$; then, the following theorem specifies how the other possible periods are obtained from the periods of the shorter word yzy.

Theorem 6. *If p_0 is the minimal period of a word W, and $(yz)^r y$ its canonical decomposition, the set $\mathcal{P}(W)$ of the periods of W is*

$$\mathcal{P}(W) = \{jp_0,\ 1 \leq j \leq r\} \cup \{(r-1)p_0 + q,\ q \in \mathcal{P}(yzy)\}.$$

The only difficult point is to prove (Section 3) that any period is either a multiple of the minimal period or corresponds to a second word W starting only in the last part yzy of the word.

Remark 7. Moreover, it will be shown in Section 3 that, if yz is the minimal root of W, then the periods of yzy are greater than the length of y, denoted $|y|$.

EXAMPLE (cont.) $p_0 = 5$, $r = 2$, periods 9 and 11 correspond to the periods 4 and 6 of AAACAAA$= yzy$.

When two occurrences of W overlap in the sequence, Figure 1 shows that the second occurrence is preceded by a root of W. Thus, an occurrence of W at position i does not overlap a preceding one if and only if there is no root of W just before this occurrence. Now, if an occurrence is not preceded by the minimal root $W^{(p_0)}$, it cannot be preceded by one of the roots $W^{(jp_0)}$. Therefore, we define the set $\mathcal{P}'(W)$ of the *principal periods* of W as the periods of W that are not a multiple of p_0. Consequently, an occurrence of W at position i does not overlap a preceding one if and only if there is no occurrence of any *principal root* $W^{(p)}$ at position $i-p$, for p in $\mathcal{P}'(W)$. Moreover, we show in Section 3 that it is not possible to observe two different principal roots, $W^{(p)}$ at position $i-p$ and $W^{(q)}$ at position $i-q$. Using these two results, we get an expression of $\widetilde{Y}_i$ as

$$\widetilde{Y}_i = Y_i(W) - \sum_{p\in\mathcal{P}'(W)} Y_{i-p}(W^{(p)}W), \tag{1}$$

where $Y_i(W')$ is the indicator of the occurrence of W' at position i in the infinite sequence $(X_i)_{i=-\infty}^{i=+\infty}$. Clearly, $Y_i(W) = Y_i$.

Taking the expectation in (1), we get

$$\widetilde{\mu} = \mu(W) - \sum_{p\in\mathcal{P}'(W)} \mu(W^{(p)}W). \tag{2}$$

This gives an easy way to calculate and then estimate the expected declumped count.

To extend this result to an occurrence of a k-clump at position i in the infinite sequence $(X_i)_{i=-\infty}^{i=+\infty}$, we write that $\widetilde{Y}_i^{(k)} = 1$ if and only if there is an occurrence at position i of a word C, composed of exactly k overlapping occurrences of W, which does not overlap any other occurrence of W in the sequence. We denote $\mathcal{C}_k$ the whole set of concatenated words composed of exactly k overlapping occurrences of W.
We noticed before that an occurrence at position i of C does not overlap a preceding occurrence of W if and only if there is no occurrence of any principal root $W^{(p)}$ at position $i-p$. We consider now suffixes of W, denoted $W_{(p)} = w_{h-p+1}\cdots w_h$. With the same methods, an occurrence of C does not overlap a following occurrence of W if no $W_{(p)}$ occurs just after C, with $p \in \mathcal{P}'(W)$. Since we have $\mathcal{C}_k = \{W^{(p_1)}W^{(p_2)}\cdots W^{(p_{k-1})}W,\ p_j \in \mathcal{P}'(W)\}$, we show in Section 3 that simultaneous occurrences of two different compound words of $\mathcal{C}_k$, C and C', at position i in the sequence is not possible. Therefore, we obtain the following expression of $\widetilde{Y}_i^{(k)}(W)$:

$$\begin{aligned}\widetilde{Y}_i^{(k)} &= \sum_{C\in\mathcal{C}_k}\Bigg(Y_i(C) - \sum_{p\in\mathcal{P}'(W)} Y_{i-p}\left(W^{(p)}C\right) - \sum_{q\in\mathcal{P}'(W)} Y_i\left(C\,W_{(q)}\right)\\ &\qquad + \sum_{p,q\in\mathcal{P}'(W)} Y_{i-p}\left(W^{(p)}\,C\,W_{(q)}\right)\Bigg).\end{aligned} \tag{3}$$

Using (3), we obtain expression of $\widetilde{\mu}_k = \mathbb{E}\widetilde{Y}_i^{(k)}$ as

$$\widetilde{\mu}_k = \sum_{C \in \mathcal{C}_k} \mu(C) \; - 2 \sum_{C' \in \mathcal{C}_{k+1}} \mu(C') \; + \sum_{C'' \in \mathcal{C}_{k+2}} \mu(C''). \tag{4}$$

A straightforward manipulation of formula (4) leads to the following expression which shows the geometric structure of $\widetilde{\mu}_k$ (Schbath, 1995): $\widetilde{\mu}_k = (1-A)^2 A^{k-1} \mu(W)$ with $A = \sum_{p \in \mathcal{P}'(W)} \prod_{j=1}^{p} \pi(w_j, w_{j+1})$.

REMARK 8. Since $\widetilde{Y}_i = \sum_{k \geq 1} \widetilde{Y}_i^{(k)}$, formula (4) has to and does verify

$$\widetilde{\mu} = \sum_{k \geq 1} \widetilde{\mu}_k. \tag{5}$$

We can also easily prove that

$$\sum_{k \geq 1} k \, \widetilde{\mu}_k = \mu(W) \; ; \tag{6}$$

this result has to be interpreted carefully taking care that $\check{N}(W)$ defined by $\check{N}(W) = \sum_{k \geq 1} k \sum_{i \in I} \widetilde{Y}_i^{(k)}$ is not equal to $N(W)$.

3. Combinatory proofs

All the proofs below are based on the property that the minimal root yz cannot be simultaneously written as xx' and $x'x$. This property follows from the Theorem of Lothaire (1983).

THEOREM 9 (LOTHAIRE (1983)). *Two nonempty words x and x' commute if and only if they are powers of the same word.*

The next proposition is a simple application of this theorem.

PROPOSITION 10. *If yz is the minimal root of W, then the periods of yzy are greater than $|y|$.*

Proof. If yzy has a period less or equal to $|y|$, the minimal root yz has a non trivial decomposition as $yz = xx' = x'x$ (in Figure 2, the word x is in black). Using Theorem 9, we get the result that yz is a power of another word, and therefore yz is not the minimal root of W. □

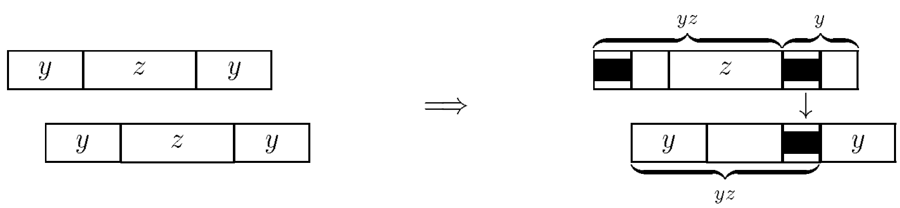

FIGURE 2. yzy has a period less or equal to $|y|$.

PROOF OF THEOREM 6. We show that if $(yz)^r y$ is the canonical decomposition of W, then there can be two occurrences of W at position i and $i+\ell$ only if ℓ is a multiple of the minimal period p_0 or is of the form $\ell = (r-1)p_0 + q$ where q is a period of yzy. Suppose that two words W occur with a shift of ℓ positions. Obviously, if $\ell > (r-1)p_0$ then $q = \ell - (r-1)p_0$ is a period of yzy, and ℓ is in $\mathcal{P}(W)$. If $\ell < (r-1)p_0$, we show that ℓ is a multiple of p_0. If it is not, we consider separately the two cases corresponding to a second occurrence of W starting in a word y (Figure 3.a), or in a word z (Figure 3.b). In the two cases, the minimal root yz can be decomposed as $yz = xx' = x'x$ where the words x, in black in the figure, and x' are not empty. By Theorem 9, this is a contradiction. Such an overlap is not possible. □

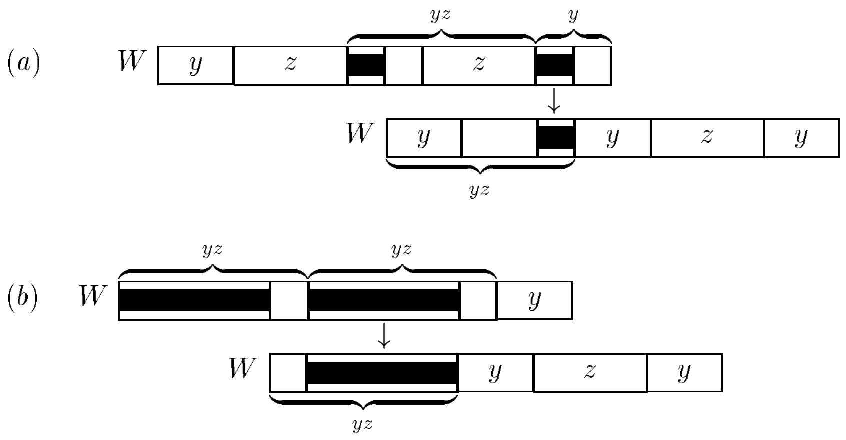

FIGURE 3. Simultaneous occurrences of W at position i and $i+\ell$ in two cases: (a) ℓ is written $r'p_0 + \ell'$ with $0 < \ell' \leq |y|$, (b) ℓ is written $r''p_0 + \ell''$ with $|y| < \ell'' < |yz|$.

PROPOSITION 11. *If W occurs at position i in the sequence, there cannot be occurrences of two different principal roots $W^{(p)}$ at position $i-p$ and $W^{(s)}$ at position $i-s$.*

Proof. By Theorem 6, we classify the principal periods into three classes: the minimal period p_0 (class I), the periods of the form $(r-1)p_0 + q$ such that $|y| < q < |yz|$ (class II), and the periods of the form $(r-1)p_0 + q$ such that $|yz| < q < |yzy|$ (class III). To prove the proposition, we have to consider the five cases corresponding to (p, s) in classes (I-II), (I-III), (II-II), (II-III) and (III-III). We only study the two cases (I-III) and (III-III); for the other cases, the same method is used.

- Case (I-III): We suppose that $W^{(p_0)} = yz$ occurs at position $i - p_0$ and $W^{(s)} = (yz)^r y'$ occurs at position $i-s$ where y' is a root of y. This overlap, represented in Figure 4, implies that $yz = xy' = y'x$ for a nonempty word x, because y' is a root of y. Theorem 9 completes the proof; this overlap is not possible.
- Case (III-III): We suppose that $W^{(p)} = (yz)^r y'$ occurs at position $i-p$ and $W^{(s)} = (yz)^r y''$ occurs at position $i-s$ where y' and y'' are two different roots of y. This overlap, represented in Figure 5, leads to $yz = xx' = x'x$. Theorem 9 completes the proof.

□

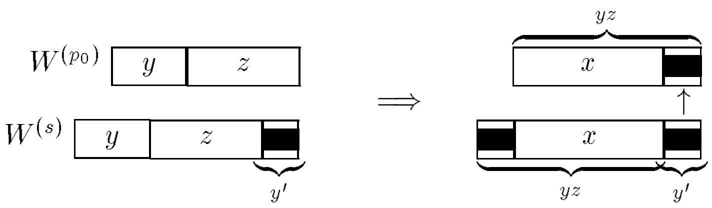

FIGURE 4. Simultaneous occurrences of $W^{(p_0)}$ at position $i - p_0$ and $W^{(s)}$ at position $i - s$ when s is in class (III).

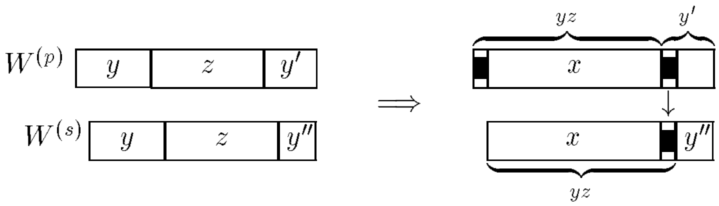

FIGURE 5. Simultaneous occurrences of $W^{(p)}$ at position $i-p$ and $W^{(s)}$ at position $i - s$ when p and s are in class (III).

PROPOSITION 12. *Simultaneous occurrences at position i of two different composed words in $\mathcal{C}_k$ are not possible.*

Proof. Let C and C' be two different composed words in $\mathcal{C}_k$. We can decompose these words into principal roots of W in the following way:
$C = W^{(p_1)} W^{(p_2)} \cdots W^{(p_{k-1})} W$ and $C' = W^{(q_1)} W^{(q_2)} \cdots W^{(q_{k-1})} W$ with $p_j, q_j \in \mathcal{P}'(W)$, $j = 1 \cdots k - 1$. Since C and C' are different, there exists at least one pair (p_j, q_j) such that $p_j \neq q_j$; we denote j the first index such that p_j is not equal to q_j. If C and C' both occur at position i in the sequence, the two different principal roots $W^{(p_j)}$ and $W^{(q_j)}$ occur also at the same position. Moreover the two composed words $W^{(p_j)} \cdots W^{(p_{k-1})} W$ and $W^{(q_j)} \cdots W^{(q_{k-1})} W$ appear at the same position. Some tedious manipulation, using the three classes of principal roots, as in the proof of Proposition 11, yields to a decomposition of the minimal root yz in all the cases, which is impossible by Theorem 9. □

4. COMPOUND POISSON APPROXIMATION

We noticed in Section 1 that to approximate the count $N(W)$, we need to study the occurrences of all k-clumps of this word in the sequence. Therefore, let us suppose that the infinite sequence $\{X_i\}_{i=-\infty}^{i=+\infty}$ is observed; we define the following count

$$\check{N}(W) = \sum_{k \geq 1} k \sum_{i \in I} \widetilde{Y}_i^{(k)} ,$$

where $I = \{1, \ldots, n - h + 1\}$ and $\widetilde{Y}_i^{(k)}$ is the indicator of the occurrence of a k-clump at position i in the infinite sequence, calculated in (3).

COMPARISON OF $\check{N}(W)$ AND $N(W)$. The counts $\check{N}(W)$ and $N(W)$ are not equal but their difference is negligible in probability. Actually, they can only differ in two cases: a clump of W starts before position $n-h+2$ and stops beyond position n in the infinite sequence $\{X_i\}_{i=-\infty}^{i=+\infty}$; such a clump, if it exists, is unique. Consequently, all the occurrences of W contained in this clump are taken into account in $\check{N}(W)$, but only those we can observe in the finite sequence $X_1\cdots X_n$ are counted in $N(W)$. In this case, there is necessarily an occurrence of W at one of the positions $n-2h+3,\ldots,n-h+1$. The second case is when a clump starts before position 1 and stops beyond position $h-1$ in the infinite sequence. Here, no occurrence of W contained in this clump is counted in $\check{N}(W)$, but those observed in $X_1X_2\cdots X_n$ are taken into account in $N(W)$. Necessarily, W occurs at one of the $(h-1)$ first positions of the finite sequence. Therefore, we have

$$\mathrm{P}\{N(W)\neq\check{N}(W)\}\leq 2h\,\mu(W).$$

Moreover, these bias complement each other as shown by (6) which is $\mathbb{E}\check{N}(W)=\mathbb{E}N(W)$. Our problem is now just to approximate the variable $\check{N}(W)$. □

We now consider the Poisson process $\mathbb{Z}=(Z_i^{(k)})_{i\in I,k\geq 1}$ such that $Z_i^{(k)}$ are independent Poisson variables with mean $\widetilde{\mu}_k=\mathbb{E}(\widetilde{Y}_i^{(k)})$. A process version of the Chen-Stein theorem, as formulated in Arratia *et al.* (1990), allows us to approximate the process $\widetilde{\mathbb{Y}}=(\widetilde{Y}_i^{(k)})_{i\in I,k\geq 1}$ by the Poisson process $\mathbb{Z}$. The total variation is bounded by

$$d_{\mathrm{TV}}(\mathcal{L}(\widetilde{\mathbb{Y}}),\,\mathcal{L}(\mathbb{Z}))\leq 2(2b_1^*+2b_2^*+b_3^*)\,,$$

where the expressions of b_1^*,b_2^* and b_3^* are similar to those of b_1,b_2 and b_3 given in Section 1, for the variables $\widetilde{Y}_i^{(k)}$ with the double index (i,k).

Therefore, with the same error bound, we approximate the count $\check{N}(W)$ by the variable $\sum_{k\geq 1}k\sum_{i\in I}Z_i^{(k)}$, which is distributed according to the compound Poisson distribution $\mathrm{CP}(\Lambda)$, where the discrete measure $\Lambda=\sum_{k\geq 1}\Lambda_k\delta_k$ is defined by $\Lambda_k=(n-h+1)\widetilde{\mu}_k$, calculated in (4). Using the triangular inequality, we then have

$$d_{\mathrm{TV}}(\mathcal{L}(N(W)),\,\mathrm{CP}(\Lambda))\leq 2(2b_1^*+2b_2^*+b_3^*)\,+4h\,\mu(W). \tag{7}$$

Our task is now to show that b_1^*,b_2^* and b_3^* tend to zero. First, we need to choose a suitable neighborhood of the double index $(i,k)\in I^*$, where $I^*=I\times\mathbb{N}\backslash\{0\}$.

Choice of the neighborhood $B_{i,k}$. We define $Z(i,k)=\{j\,:\,i-h\leq j\leq i+(k+1)h\}$ (Figure 6). Therefore, $Z(i,k)$ contains the positions j of the letters X_j that define the variable $\widetilde{Y}_i^{(k)}$. Actually, the length of a compound word C in $\mathcal{C}_k$ is less than kh, and we need to know $(h-1)$ letters at the ends of C to be sure that an occurrence of C at position i is in fact an occurrence of a k-clump of composition C.
Moreover, in the term b_3^*, we need to have at least r positions $(r>0)$ between the sets $Z(i,k)$ and $Z(j,\ell)$, so we define (i,k) and (j,ℓ) as not neighbors if $Z(i,k)$ and $Z(j,\ell)$ are separated with at least r positions, that is if one of

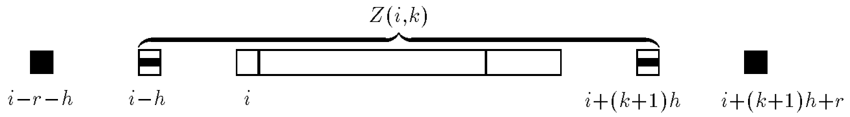

FIGURE 6. Construction of the neighborhood of the index (i, k).

the two following inequalities holds (Figure 6):

$$\begin{aligned} i + (k+1)h + r &< j - h, \\ j + (\ell+1)h + r &< i - h. \end{aligned}$$

Therefore, the neighborhood $B_{i,k}$ of (i, k) is the set of indices (j, ℓ) that are neighbors on (i, k):

$$B_{i,k} = \{(j, \ell) \,:\, -(\ell+2)h - r \leq j - i \leq (k+2)h + r\} \,\cap\, I^*.$$

A BOUND FOR THE TERM b_1^*. By definition, the first term b_1^* is written

$$b_1^* = \sum_{(i,k)\in I^*} \sum_{(j,\ell)\in B_{i,k}} \mathbb{E}(\widetilde{Y}_i^{(k)})\mathbb{E}(\widetilde{Y}_j^{(\ell)}).$$

$\mathbb{E}(\widetilde{Y}_i^{(k)}) = \widetilde{\mu}_k$ and $\mathbb{E}(\widetilde{Y}_j^{(\ell)}) = \widetilde{\mu}_\ell$ does not depend on j; we then separate the indices $j \in B_{i,k}$ such that $j < i$ and $j \geq i$, and we have

$$b_1^* \leq \sum_{i\in I}\sum_{k\geq 1}\sum_{\ell\geq 1} [1 + (k+2)h + r]\widetilde{\mu}_k\widetilde{\mu}_\ell + \sum_{i\in I}\sum_{k\geq 1}\sum_{\ell\geq 1} [(\ell+2)h + r]\widetilde{\mu}_k\widetilde{\mu}_\ell.$$

Using the symmetry between k and ℓ, the above expression reduces to

$$b_1^* \leq 2\sum_{i\in I}\left(\sum_{\ell\geq 1}\widetilde{\mu}_\ell\right)\left(\sum_{k\geq 1}[(k+2)h + r + 1]\widetilde{\mu}_k\right).$$

Therefore, it follows from (5) and (6) that

$$b_1^* \leq 2(n - h + 1)\,(h\mu + (2h + r + 1)\widetilde{\mu})\,\widetilde{\mu}, \tag{8}$$

where μ denotes $\mu(W)$.

A BOUND FOR THE TERM b_2^*. The second term b_2^* is sum of $\mathbb{E}\widetilde{Y}_i^{(k)}\widetilde{Y}_j^{(\ell)}$ with (j, ℓ) in $B_{i,k}$ but not equal to (i, k). Since $\widetilde{Y}_i^{(k)}\widetilde{Y}_i^{(\ell)} = 0$ if $k \neq \ell$, and using the symmetry of $B_{i,k}$, b_2^* is also written

$$b_2^* \leq 2 \sum_{(i,k)\in I^*} \sum_{\ell\geq 1} \sum_{j=i+1}^{i+(k+2)h+r} \mathbb{E}\left(\widetilde{Y}_i^{(k)}\widetilde{Y}_j^{(\ell)}\right).$$

A straightforward majoration is not enough. Bounding $\sum_{\ell\geq 1}\widetilde{Y}_j^{(\ell)} = \widetilde{Y}_j$ by Y_j gives either $b_2^* \leq \mathrm{O}(n(3h + r)\mu)$ if we only use that $Y_j \leq 1$, or $b_2^* \leq \mathrm{O}\left(n\mu^2\sum_k[(k+2)h + r]A^{k-1}.\mathrm{card}(\mathcal{C}_k)\right)$ if we bound $\mathbb{E}\left(\widetilde{Y}_i^{(k)}Y_j\right)$ by $\mu\sum_{C\in\mathcal{C}_k}\mu(C)$ and use that $\mu(C) \leq \mu\, A^{k-1}$ for a certain $|A| < 1$, as in Geske *et al.* (1995). In the second case, the upper bound can diverge because $\mathrm{card}(\mathcal{C}_k) = \mathrm{card}(\mathcal{P}'(W))^{k-1}$; we can only conclude if $\mathrm{card}(\mathcal{P}'(W)) = 1$ which is exactly the framework of Geske *et al.* (1995).

We have to take into account that other products $\widetilde{Y}_i^{(k)}\widetilde{Y}_j^{(\ell)}$ are equal to zero, because the ℓ-clump at position j cannot overlap the k-clump at position i, and to identify some of them in order to prove that b_2^* tends to zero.

To describe these products, we need to know more about the compound word C that appears at i; therefore, we write

$$\widetilde{Y}_i^{(k)} = \sum_{U\in\mathcal{U},\, C\in\mathcal{C}_k,\, V\in\mathcal{V}} Y_{i-h}(UCV), \tag{9}$$

where $\mathcal{U}$ is the set of h-words that do not end by a principal root of W

$$\mathcal{U} = \{u_1\cdots u_h \mid \forall p \in \mathcal{P}'(W),\ u_{h-p+1}\cdots u_h \neq w_1\cdots w_p\}, \tag{10}$$

and $\mathcal{V}$ is the set of h-words that do not start by the p last letters of W, p in $\mathcal{P}'(W)$

$$\mathcal{V} = \{v_1\cdots v_h \mid \forall p \in \mathcal{P}'(W),\ v_1\cdots v_p \neq w_{h-p+1}\cdots w_h\}. \tag{11}$$

Using the same decomposition for $\widetilde{Y}_j^{(\ell)}$ gives

$$b_2^* \le 2 \sum_{(i,k)\in I^*} \sum_{U,C,V} \sum_{\ell\ge 1} \sum_{U'\in\mathcal{U},\, C'\in\mathcal{C}_\ell,\, V'\in\mathcal{V}} \sum_{j=i+|C|}^{i+(k+2)h+r} \mathbb{E}\left(Y_{i-h}(UCV)Y_{j-h}(U'C'V')\right);$$

in this formula we dropped the indexes j smaller than $i+|C|$ because $i+1 \le j < i+|C|$ would mean that the compound word C' overlaps C, which is not possible. Considering separately the indices $i+|C| \le j < i+|C|+2h$ that correspond to an overlap between UCV and $U'C'V'$, and the indexes $j \ge i+|C|+2h$ that correspond to non overlapping occurrences of UCV and $U'C'V'$, we get two terms b_{21}^* and b_{22}^*.

• For $j \ge i+|C|+2h$, we bound $\mathbb{E}(Y_{i-h}(UCV)Y_{j-h}(U'C'V'))$ by $\mu(u_1')^{-1}\mu(UCV)\mu(U'C'V')$, where u_1' is the first letter of U'. As it does not depend on j, we obtain

$$b_{22}^* \le 2\sum_{i\in I}\sum_{k\ge 1}\sum_{U,C,V}\sum_{\ell\ge 1}\sum_{U',C',V'} ((k+2)h+r-|C|-2h+1)\frac{\mu(UCV)\mu(U'C'V')}{\mu(u_1')},$$

Since $|C| \ge h$ and $\mu(u_1')$ is greater or equal to $\inf_{x\in\mathcal{A}}\mu(x)$, denoted $\mu_{\inf}$, we have

$$b_{22}^* \le \frac{2}{\mu_{\inf}}\sum_{i\in I}\left(\sum_{k\ge 1}((k-1)h+r+1)\sum_{U,C,V}\mu(UCV)\right)\left(\sum_{\ell\ge 1}\sum_{U',C',V'}\mu(U'C'V')\right).$$

Finally, gathering (5), (6) and (9) gives

$$\sum_{\ell\ge 1}\sum_{U',C',V'}\mu(U'C'V') = \sum_{\ell\ge 1}\widetilde{\mu}_\ell = \widetilde{\mu}$$

and

$$\sum_{k\ge 1} k \sum_{U,C,V}\mu(UCV) = \sum_{k\ge 1} k\widetilde{\mu}_k = \mu,$$

and therefore

$$b_{22}^* \le \frac{2}{\mu_{\inf}}(n-h+1)(h\mu+(r-h+1)\widetilde{\mu})\widetilde{\mu}. \tag{12}$$

• For $i+|C| \le j < i+|C|+2h$, the term b_{21}^* is written

$$b_{21}^* = 2 \sum_{(i,k)\in I^*} \sum_{U,C,V} \sum_{\ell\ge 1} \sum_{U',C',V'} \sum_{j=i+|C|}^{i+|C|+2h-1} \mathbb{E}\left(Y_{i-h}(UCV)Y_{j-h}(U'C'V')\right).$$

Summing first on ℓ, U' ,C' and V', we obtain

$$b_{21}^* = 2 \sum_{(i,k)\in I^*} \sum_{U,C,V} \sum_{j=i+|C|}^{i+|C|+2h-1} \mathbb{E}\left(Y_{i-h}(UCV)\widetilde{Y}_j(W)\right).$$

Now, we use that $\widetilde{Y}_j(W) \le Y_j(W)$ and $\sum_V Y_{i-h}(UCV) \le Y_{i-h}(UC)$ to get

$$b_{21}^* \le 2 \sum_{(i,k)\in I^*} \sum_{U,C} \sum_{j=i+|C|}^{i+|C|+2h-1} \mathbb{E}\left(Y_{i-h}(UC)\, Y_j(W)\right).$$

As $j \ge i+|C|$, the occurrence of W at position j does not overlap the occurrence of UC at position $i-h$, so we bound $\mathbb{E}(Y_{i-h}(UC)\, Y_j(W))$ by $\mu(w_1)^{-1}\mu(UC)\mu(W)$, which does not depend on j. Therefore, we have

$$b_{21}^* \le \frac{4}{\mu(w_1)}\, h\, \mu(W) \sum_{i\in I} \sum_{k\ge 1} \sum_{U\in\mathcal{U}, C\in\mathcal{C}_k} \mu(UC).$$

Finally, we show that $\sum_k \sum_{U,C} \mu(UC) = \mu(W)$; indeed, $\sum_{U\in\mathcal{U},C\in\mathcal{C}_k} \mu(UC)$ is the probability that the sequence presents, at position i, a clump composed of at least k occurrences of W. Therefore

$$\sum_{k\ge 1} \sum_{U\in\mathcal{U},C\in\mathcal{C}_k} \mu(UC) = \sum_{k\ge 1} \sum_{K\ge k} \widetilde{\mu}_K = \sum_{K\ge 1} K\, \widetilde{\mu}_K = \mu(W).$$

Consequently, we get

$$b_{21}^* \le \frac{4}{\mu(w_1)}(n-h+1)h\mu^2(W). \tag{13}$$

A BOUND FOR THE TERM b_3^*. We prove in the Appendix, there exists some constant $0<\rho<1$ depending on the Markov chain such that

$$b_3^* \le \mathrm{O}(n^2\rho^r)\,. \tag{14}$$

Since $r > 2\log n/\log\rho^{-1}$, b_3^* converges to zero. The power r of ρ proceed from the choice of the neighborhood we made before.

Using (7), (8), (12), (13) and (14), we thus proved the following result on the compound Poisson approximation of $\mathcal{L}(N(W))$.

THEOREM 13.

$$d_{TV}(\mathcal{L}(N(W)),\, CP(\Lambda)) \le (n-h+1)\left(A_1 h\,\mu^2 + A_2 h\,\mu\widetilde{\mu} + A_3 h\,\widetilde{\mu}^2 + A_4 r\widetilde{\mu}^2 + A_5\widetilde{\mu}^2\right) + A_6 n^2\rho^r + 4h\,\mu$$

where A_1, A_2, A_3, A_4, A_5 and A_6 are of order O(1) and $CP(\Lambda)$ represents the compound Poisson distribution of $\sum_{k\ge 1} k\, Z^{(k)}$ where $Z^{(k)}$ are Poisson variables independent with means $(n-h+1)\widetilde{\mu}_k$; the expression of $\widetilde{\mu}_k$ is given by (4).

The corollary below easily follows.

COROLLARY 14. *Let W_n be a sequence of words with length $h_n = o(n)$ such that $n\mu(W_n) = O(1)$. We then have*

$$\lim_{n\to+\infty} d_{TV}(\mathcal{L}(N(W_n)),\ CP(\Lambda_n)) = 0\,,$$

where $CP(\Lambda_n)$ represents the compound Poisson distribution defined in Theorem 13, associated with the word W_n.

An alternative method could be Stein's method for compound Poisson approximation (Barbour *et al.* (1992a), Roos (1994)); it would be interesting to see if this direct method could potentially improve our results. For simple words, it gives the same bound on the total variation but a different compound Poisson distribution. This compound Poisson distribution has a finite number of components that are not related to the k-clumps.

APPENDIX

We first calculate the term b_3, defined in section 1, which appears in the error bound of the Poisson approximation for the sum $\sum_{i\in I} Y_i$. The method presented below uses the property of β-mixing of the Markov chain (X_i), and can be easily adapted to the calculation of the terms $\widetilde{b}_3$ and b_3^* associated with $\widetilde{Y}_i$ and $\widetilde{Y}_i^{(k)}$. A direct method, using the transition matrix diagonalization, also leads to the geometric convergence of these terms.

CALCULATION OF THE TERM b_3. Our aim is to bound b_3 defined by

$$b_3 = \sum_{i\in I} \mathbb{E}\,|\ \mathbb{E}\,(Y_i - \mathbb{E}Y_i \mid \sigma(Y_j,\ j\in B_i^c))\ |\,,$$

where $B_i = \{i-h-r+2, \ldots, i+h+r-2\}\cap I$. It is clear that $\sigma(Y_j,\ j\in B_i^c)$ is included in the larger sigma-algebra $\sigma(X_1,\ldots,X_{i-r},X_{i+h+r-1},\ldots,X_n)$, so we have

$$b_3 \le \sum_{i\in I} \mathbb{E}|\ \mathbb{E}(Y_i - \mathbb{E}Y_i \mid \sigma(X_1,\ldots,X_{i-r},X_{i+h+r-1},\ldots,X_n))\ |.$$

Let $Y_i' = Y_i - \mathbb{E}Y_i$ be the recentered variable, and $\mathcal{G}$, $\mathcal{Y}$, $\mathcal{H}$, $\mathcal{B}$ be the following σ-algebras: $\mathcal{G} = \sigma(X_1,\ldots,X_{i-r})$, $\mathcal{Y} = \sigma(X_i,\ldots,X_{i+h-1})$, $\mathcal{H} = \sigma(X_{i+r+h-1},\ldots,X_n)$ and $\mathcal{B} = \sigma(\mathcal{G}\cup\mathcal{H})$. Thus, Y_i' is $\mathcal{Y}$-measurable. First note that $||\mathbb{E}(Y_i' \mid \mathcal{B})||_1 = \sup\{|\mathbb{E}Y_i'\phi|,\ \phi\ \mathcal{B}\text{-mes},\ |\phi|\le 1\}$. Since ϕ is $\mathcal{B}$-measurable, it can be interpreted as a function of two variables G $\mathcal{G}$-measurable and H $\mathcal{H}$-measurable. Then, denoting the tensor product $\mathrm{P}_G\otimes\mathrm{P}_Y\otimes\mathrm{P}_H$ of the marginals P_G, P_Y and P_H, we have

$$\begin{aligned}\mathbb{E}Y_i'\phi &= \int Y_i'\phi(G,H)\,\mathrm{dP}\\ &= \int Y_i'\phi(G,H)\,\mathrm{d}(\mathrm{P}-\mathrm{P}_G\otimes\mathrm{P}_Y\otimes\mathrm{P}_H)\end{aligned}$$

because $\int Y_i'\,\mathrm{dP}_Y = 0$. Moreover, $|Y_i'\phi|\le 1$ leads to

$$\begin{aligned}|\mathbb{E}Y_i'\phi| &\le d_{\mathrm{TV}}(\mathrm{P},\ \mathrm{P}_G\otimes\mathrm{P}_Y\otimes\mathrm{P}_H)\\ &\le d_{\mathrm{TV}}(\mathrm{P},\ \mathrm{P}_{(G,Y)}\otimes\mathrm{P}_H) + d_{\mathrm{TV}}(\mathrm{P}_{(G,Y)}\otimes\mathrm{P}_H,\ \mathrm{P}_G\otimes\mathrm{P}_Y\otimes\mathrm{P}_H)\\ &\le d_{\mathrm{TV}}(\mathrm{P},\ \mathrm{P}_{(G,Y)}\otimes\mathrm{P}_H) + d_{\mathrm{TV}}(\mathrm{P}_{(G,Y)},\ \mathrm{P}_G\otimes\mathrm{P}_Y)\,.\end{aligned}$$

From Doukhan (1994, p3), we have $d_{\mathrm{TV}}(\mathrm{P}, \mathrm{P}_{(G,Y)} \otimes \mathrm{P}_H) = \beta(\mathcal{G} \otimes \mathcal{Y}, \mathcal{H})$ and $d_{\mathrm{TV}}(\mathrm{P}_{(G,Y)}, \mathrm{P}_G \otimes \mathrm{P}_Y) = \beta(\mathcal{G}, \mathcal{Y})$ where the coefficients β are the β-mixing coefficients associated with the sequence (X_i), and $\mathcal{G} \otimes \mathcal{Y}$ is the tensor product of the σ-algebras $\mathcal{G}$ and $\mathcal{Y}$. Since $\mathcal{G} \otimes \mathcal{Y} \subset \sigma(X_1, \ldots, X_{i+h-1})$ and the β-mixing coefficients are less than the φ-mixing coefficients (Doukhan, 1994, p4), we use the following result (Doukhan, 1994, p88): there exist constants C and $0 < \rho < 1$ only depending on the Markov chain (X_i), such that

$$\varphi(\mathcal{G} \otimes \mathcal{Y}, \mathcal{H}) \leq C\,\rho^r \quad \text{and} \quad \varphi(\mathcal{G}, \mathcal{Y}) \leq C\,\rho^r.$$

Finally, we obtain

$$b_3 \leq \mathrm{O}(n\rho^r),$$

and b_3 converges to zero provided $r > \log n / \log \rho^{-1}$.

Calculation of the term $\widetilde{b}_3$. The calculation of $\widetilde{b}_3$ defined by

$$\widetilde{b}_3 = \sum_{i \in I} \mathbb{E} \Big| \ \mathbb{E}\left(\widetilde{Y}_i - \mathbb{E}\widetilde{Y}_i \mid \sigma(\widetilde{Y}_j,\ j \in \widetilde{B}_i^c) \right) \Big| ,$$

where $\widetilde{B}_i = \{i - 2h - r + 3, \ldots, i + 2h + r - 3\} \cap I$, easily follows noting that $\sigma(\widetilde{Y}_j, \widetilde{Y}_j \in \widetilde{B}_i^c) \subset \sigma(X_1, \ldots, X_{i-h-r+1}, X_{i+h+r-1}, \ldots, X_n)$. Indeed, $|\widetilde{Y}_i - \mathbb{E}\widetilde{Y}_i| \leq 1$ and $(\widetilde{Y}_i - \mathbb{E}\widetilde{Y}_i)$ is $\mathcal{Y}'$-measurable with $\mathcal{Y}' = \sigma(X_{i-h+1}, \ldots, X_{i+h-1})$. Therefore, since we choose the σ-algebras $\mathcal{G} = \sigma(X_1, \ldots, X_{i-h-r+1})$ and $\mathcal{H} = \sigma(X_{i+r+h-1}, \ldots, X_n)$, we obtain

$$||\mathbb{E}(\widetilde{Y}_i - \mathbb{E}\widetilde{Y}_i \mid \sigma(\mathcal{G} \cup \mathcal{H}))||_1 \leq \mathrm{O}(\rho^r)$$

for some $0 < \rho < 1$. Thus, we have

$$\widetilde{b}_3 \leq \mathrm{O}(n\rho^r),$$

and $\widetilde{b}_3$ converges to zero provided $r > \log n / \log \rho^{-1}$.

Calculation of the term b_3^*. Using the same arguments for the indicator $\widetilde{Y}_i^{(k)}$, we obtain the calculation of b_3^* defined by

$$b_3^* = \sum_{(i,k) \in I^*} \mathbb{E} \Big| \ \mathbb{E}\left(\widetilde{Y}_i^{(k)} - \mathbb{E}\widetilde{Y}_i^{(k)} \mid \sigma(\widetilde{Y}_j^{(\ell)},\ (j,\ell) \in B_{i,k}^c) \right) \Big| .$$

The construction of the neighborhood $B_{i,k}$, in section 4, yields

$$||\mathbb{E}(\widetilde{Y}_i^{(k)} - \mathbb{E}\widetilde{Y}_i^{(k)} \mid \sigma(\widetilde{Y}_j^{(\ell)},\ (j,\ell) \in B_{i,k}^c))||_1 \leq \mathrm{O}(\rho^r)$$

for some $0 < \rho < 1$. Obviously $k \leq n$, so we have

$$b_3^* \leq \mathrm{O}(n^2\rho^r),$$

and b_3^* converges to zero provided $r > 2 \log n / \log \rho^{-1}$.

Acknowledgements

I am grateful to Élisabeth de Turckheim and Bernard Prum for their valuable comments during this work and to the referees for their helpful and encouraging comments on an earlier version of this paper.

References

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

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

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

Barbour, A. D., Holst, L. and Janson, S. (1992b). *Poisson approximation.* Oxford-University Press.

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

Chryssaphinou, O. and Papastavridis, S. (1988a). A limit theorem for the number of non-overlapping occurrences of a pattern in a sequence of independent trials. *J. Appl. Prob.* **25** 428–431.

Chryssaphinou, O. and Papastavridis, S. (1988b). A limit theorem on the number of overlapping appearances of a pattern in a sequence of independent trials. *Prob. Theory Rel. Fields.* **79** 129–143.

Doukhan, P. (1994). *Mixing: Properties and Examples.* L.N.S. 85, Springer-Verlag.

Fu, J. C. (1993). Poisson convergence in reliability of a large linearly connected system as related to coin tossing. *Statistica Sinica.* **3** 261–275.

Geske, M. X., Godbole, A. P., Schaffner, A. A., Skolnick, A. M. and Wallstrom, G. L. (1995). Compound Poisson approximations for word patterns under Markovian hypotheses. *J. Appl. Prob.* To appear.

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

Godbole, A. P. and Schaffner, A. A. (1993). Improved poisson approximations for word patterns. *Adv. Appl. Prob.* **25** 334–347.

Guibas, L. J. and Odlyzko, A. M. (1981). Periods in strings. *J. Combinatorial Theory A.* **30** 19–42.

Hirano, K. and Aki, S. (1993). On number of occurrences of sucess runs of specified length in a two-state Markov chain. *Statistica Sinica.* **3** 313–320.

Karlin, S. and Ost, F. (1987). Counts of long aligned word matches among random letter sequences. *Ann. Prob.* **19** 293–351.

Lothaire, M. (1983). *Combinatorics on words.* Addison-Wesley.

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.

Roos, M. (1994). Stein's method for compound Poisson approximation : the local approach. *Ann. Appl. Prob.* **4** 1177–1187.

Schbath, S. (1995). *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.

Schbath, S., Prum, B. and Turckheim, É. de (1995). Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *J. Comp. Biol.* **2** 417–437.

Trifonov, E. N. (1989). The multiple codes of nucleotide sequences. *Bull. Math. Biol.* **51** 417–432.

Sophie Schbath, INRA, Laboratoire de Biométrie, F78352 Jouy-en-Josas Cedex.