Adv. Appl. Prob. **39,** 128–140 (2007)
Printed in Northern Ireland

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 a Markovian sequence. Using the Chen–Stein method, we provide a bound for the approximation error. This error converges to 0 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; clump; 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 application domains. With a huge number of biological sequences now available, genome analysis is an important consumer of probabilistic and statistical results on word occurrences (see [5, Chapter 6] or [9] for an overview). In particular, the number, N, of occurrences of a given word in a DNA sequence is a quantity of special interest to molecular biologists. Some words, called *motifs*, are recognized by proteins and occur 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 be treated as 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 be used to fit the composition of a DNA sequence in short words of length 1 up to length $m + 1$, 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 [7] or by using the distributions of both the waiting time till the first occurrence and the interarrival time between two occurrences [2], [10]. Several approximations have also been proposed for long sequences. The Gaussian distribution proposed in [6] appears to be a good approximation for words (and word families) having a sufficiently large expected count [11]. For an expectedly

Received 10 March 2006; revision received 28 August 2006.
[*] Postal address: INRA, Unité Mathématique, Informatique et Génome, Domaine de Vilvert, F-78352 Jouy-en-Josas, France.
[**] Email address: etienne.roquain@jouy.inra.fr
[***] Email address: sophie.schbath@jouy.inra.fr

rare word $\boldsymbol{w}$, i.e. one whose count, $N(\boldsymbol{w})$, satisfies the rare event condition $\mathrm{E}(N(\boldsymbol{w})) = O(1)$ as the length, n, of the sequence tends to ∞, Poisson approximations were 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 (with a geometric distribution). The compound Poisson distribution reduces to a Poisson distribution for nonoverlapping words. For an expectedly rare family of words $\mathcal{W}$, the authors of [8] proposed that the compound Poisson approximation of [12] be used for each count $N(\boldsymbol{w})$, $\boldsymbol{w} \in \mathcal{W}$, and that $N(\mathcal{W}) = \sum_{\boldsymbol{w}\in\mathcal{W}} N(\boldsymbol{w})$ be approximated by the sum of independent compound Poisson variables. Using the Chen–Stein method, a bound for the approximation error was given which explicitly depends on the degree of overlap between the words of the family $\mathcal{W}$. Unfortunately, this error bound does not converge to 0 given that there exists a couple of different words $(\boldsymbol{w}, \boldsymbol{w}') \in \mathcal{W}^2$ which overlap.

Also using the Chen–Stein method, we here propose a compound Poisson distribution more suitable to approximate the count, $N(\mathcal{W})$, of any expectedly rare word family $\mathcal{W}$. The main difference from [8] is that we will consider clumps composed of overlapping occurrences of $\mathcal{W}$, instead of separately considering clumps of $\boldsymbol{w}$ for each word $\boldsymbol{w} \in \mathcal{W}$. 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 was previously introduced in [3] to approximate the count of competing renewals, but the authors there focused only on the event that 'a mixed clump starts at a given position'. Here we will also have to take into account 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, which uses the Chen–Stein method for Poisson approximations. As a corollary, in Section 5 we propose 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, relative to the results of [3], is in the derivation of an explicit formula for the parameter of the limiting Poisson distribution. In Section 6 we present 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, $\boldsymbol{X} = (X_i)_{i\in\mathbb{Z}}$, generated by a homogeneous stationary Markov chain of order 1 on a finite alphabet $\mathcal{A}$. The generalization to higher-order Markov chains is discussed in the conclusion. The stationary distribution on $\mathcal{A}$ is denoted by μ, and $\Pi = [\pi(x, y)]_{x,y\in\mathcal{A}}$ denotes the transition matrix of the model.

Let $\mathcal{W}$ be a family of d different words, $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_d$, on the alphabet $\mathcal{A}$ with length at least 2. The length of any word $\boldsymbol{w}$ will be denoted by $|\boldsymbol{w}|$, and we define h to be the length of the longest word from the family $\mathcal{W}$:

$$h := \max\{|\boldsymbol{w}|\colon \boldsymbol{w} \in \mathcal{W}\}.$$

We make two assumptions on the word family $\mathcal{W}$: (i) it is *reduced*, meaning that, for all $\boldsymbol{w} \neq \boldsymbol{w}' \in \mathcal{W}$, $\boldsymbol{w}$ is not a substring of $\boldsymbol{w}'$ (this is a usual assumption when studying occurrences of word families, and is immediately satisfied if all the words of $\mathcal{W}$ have the same length); and (ii) each word $\boldsymbol{w} \in \mathcal{W}$ has a nonzero probability of occurring in $\boldsymbol{X}$ (this is a natural assumption). Owing to the Markov property, the occurrence probability of a $|\boldsymbol{w}|$-letter word

https://doi.org/10.1239/aap/1175266472 Published online by Cambridge University Press

$\boldsymbol{w} = w_1 w_2 \cdots w_{|\boldsymbol{w}|}$ in $\boldsymbol{X}$ is given by $\mu(w_1)\prod_{j=1}^{|\boldsymbol{w}|-1}\pi(w_j, w_{j+1})$ and will be simply denoted by $\mu(\boldsymbol{w})$ in what follows.

Classically, the number of occurrences of a word family $\mathcal{W}$ in the finite sequence $X_1 \cdots X_n$ is defined as $N(\mathcal{W}) = \sum_{\boldsymbol{w}\in\mathcal{W}} \sum_{i=1}^{n-|\boldsymbol{w}|+1} Y_i(\boldsymbol{w})$, where $Y_i(\boldsymbol{w})$ is a Bernoulli variable which is equal to 1 if there is an occurrence of $\boldsymbol{w}$ starting at position i and is equal to 0 otherwise. Note that we will generalize $Y_i(\boldsymbol{w})$ to $Y_i(\mathcal{W})$, which will be equal to 1 if and only if there exists a word from $\mathcal{W}$ occurring at position i (i.e. if and only if there is an occurrence of $\mathcal{W}$ at position i). Here we will use another decomposition of the count, based on the occurrences of k-clumps. The notion of a clump makes no sense outside a sequence: a k-clump of $\mathcal{W}$ in a sequence is 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 $\mathcal{W} = \{\texttt{atta}, \texttt{ttat}\}$, the sequence gattagcattattac has a 1-clump of $\mathcal{W}$ at $i = 2$ and a 3-clump of $\mathcal{W}$ at $i = 8$ (shown underlined). We should be careful not to forget the occurrence of ttat in the 3-clump attatta. Therefore, we have

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

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

For convenience, we will work with the infinite sequence $\boldsymbol{X}$. We define $\tilde{Y}_{i,k}(\mathcal{W})$ to be a Bernoulli variable which is equal to 1 if a k-clump of $\mathcal{W}$ occurs at position i in $\boldsymbol{X}$ and is equal to 0 otherwise, and we let

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

Note that the count $N^{\infty}(\mathcal{W})$ can differ slightly 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 $\boldsymbol{X}$ 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 a $\boldsymbol{w} \in \mathcal{W}$ such that $|\boldsymbol{w}| \neq h$. However, the occurrence of the event $\{N(\mathcal{W}) \neq N^{\infty}(\mathcal{W})\}$ implies that there exists (at least) one occurrence of $\mathcal{W}$ starting at a position in $\{1, \ldots, h-1\}$ or $\{n - h + 2, \ldots, n\}$. This event occurs with probability less than $2(h-1)\mu(\mathcal{W})$, where $\mu(\mathcal{W}) := \mathrm{E}(Y_i(\mathcal{W})) = \sum_{\boldsymbol{w}\in\mathcal{W}} \mu(\boldsymbol{w})$ denotes the occurrence probability of $\mathcal{W}$ at a given position. Therefore, the total variation distance between the distributions of these two counts is bounded by $2h\mu(\mathcal{W})$, which tends to 0 as n tends to ∞ under both $h = o(n)$ and the rare event condition. (The total variation distance between two discrete distributions P and P' on $\mathbb{N}$ is defined by $\frac{1}{2}\sum_{x\in\mathbb{N}} |P(x) - P'(x)| \leq \min \mathrm{P}(N \neq N')$, where the minimum ranges over all couplings (N, N') of P and P'.) The two counts are then asymptotically equivalent – we will focus on $N^{\infty}(\mathcal{W})$.

We will now use the Chen–Stein theorem as stated in [1] to bound the total variation distance, d_{TV}, between the distribution of the vector $(\tilde{Y}_{i,k}(\mathcal{W}))_{i,k}$ and the joint distribution of independent Poisson variables $(Z_{i,k})_{i,k}$ such that $\mathrm{E}(Z_{i,k}) = \mathrm{E}(\tilde{Y}_{i,k}(\mathcal{W}))$, which expectations will be denoted by $\tilde{\mu}_k(\mathcal{W})$. With $Z_k := \sum_{i=1}^{n-h+1} Z_{i,k}$, the Chen–Stein theorem states that

$$d_{\mathrm{TV}}(\mathcal{D}((\tilde{N}_k^{\infty}(\mathcal{W}))_k), \mathcal{D}((Z_k)_k)) \leq b_1 + b_2 + b_3, \tag{2}$$

https://doi.org/10.1239/aap/1175266472 Published online by Cambridge University Press

where $\mathcal{D}(\cdot)$ denotes the distribution of its argument and

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

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

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

and where $B_{i,k} \subset \{1,\dots,n-h+1\}\times\mathbb{N}\setminus\{0\}$ 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 0 as n tends to ∞ under both $h = o(n)$ and the rare event condition $\mathrm{E}(N(\mathcal{W})) = O(1)$ (see Section 4). This means that the process $(\tilde{N}_k^{\infty}(\mathcal{W}))_k$ can be approximated by independent Poisson variables $(Z_k)_k$ with respective expectations $\tilde{\lambda}_k(\mathcal{W}) := \mathrm{E}(\tilde{N}_k^{\infty}(\mathcal{W})) = (n-h+1)\tilde{\mu}_k(\mathcal{W})$. From (1) and properties of the total variation distance, 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 by definition follows the compound Poisson distribution $\mathcal{CP}(\tilde{\lambda}_k(\mathcal{W})\colon k\geq 1)$. We can now 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 $(\tilde{\lambda}_k(\mathcal{W}))_{k\geq 1}$ such that $\tilde{\lambda}_k(\mathcal{W}) = (n-h+1)\tilde{\mu}_k(\mathcal{W})$, with $\tilde{\mu}_k(\mathcal{W})$ as given in (13), is bounded as follows:*

$$d_{\mathrm{TV}}(\mathcal{D}(N(\mathcal{W})), \mathcal{CP}(\tilde{\lambda}_k(\mathcal{W})\colon k\geq 1)) \leq Cnh\mu^2(\mathcal{W}) + C'n\mu(\mathcal{W})|\alpha|^h + 2h\mu(\mathcal{W}), \tag{6}$$

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

$$d_{\mathrm{TV}}(\mathcal{D}(N(\mathcal{W})), \mathcal{CP}(\tilde{\lambda}_k(\mathcal{W})\colon k\geq 1)) \to 0 \quad \text{as } n\to\infty.$$

The proof is presented in Section 4.

Remark 1. The conditions $\mathrm{E}(N(\mathcal{W})) = O(1)$ and $h = o(n)$ imply that $n\mu(\mathcal{W}) = O(1)$, which is equivalent to the condition that $\log(n)/|\boldsymbol{w}| = O(1)$ for all $\boldsymbol{w}\in\mathcal{W}$, which in turn means that the compound Poisson approximation holds for families of sufficiently long 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 0 as n tends to ∞, for $h = o(n)$ and $\mathrm{E}(N(\mathcal{W})) = O(1)$.

An important task now is to calculate the parameters of the limiting compound Poisson distribution. We do this in the next section, and then provide an expression for $\tilde{\mu}_k(\mathcal{W})$ which is the occurrence probability of a k-clump of $\mathcal{W}$ occurring at a given position in the infinite sequence $\boldsymbol{X}$.

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. k successive overlapping occurrences of $\mathcal{W}$.

https://doi.org/10.1239/aap/1175266472 Published online by Cambridge University Press

3.1. Principal periods

For two words $\boldsymbol{w} = w_1 \cdots w_{|\boldsymbol{w}|}$ and $\boldsymbol{w}' = w'_1 \cdots w'_{|\boldsymbol{w}'|}$ of $\mathcal{W}$, an integer p, $1 \le p \le |\boldsymbol{w}| - 1$, such that $w'_i = w_{i+p}$ for $i = 1, \ldots, |\boldsymbol{w}| - p$ is called a *period* of $(\boldsymbol{w}, \boldsymbol{w}')$. We denote by $\mathcal{P}(\boldsymbol{w}, \boldsymbol{w}')$ the set of periods of $(\boldsymbol{w}, \boldsymbol{w}')$. For each couple of words $(\boldsymbol{w}, \boldsymbol{w}')$ and each period $p \in \mathcal{P}(\boldsymbol{w}, \boldsymbol{w}')$, the prefix $\boldsymbol{w}^{(p)} := w_1 \cdots w_p$ is called a *root* of $(\boldsymbol{w}, \boldsymbol{w}')$. The periods of $(\boldsymbol{w}, \boldsymbol{w}')$ are then the distances allowed between an occurrence of $\boldsymbol{w}$ and a further overlapping occurrence of $\boldsymbol{w}'$. For instance, $\mathcal{P}(\texttt{taca}, \texttt{acac}) = \{1, 3\}$.

If we now look at the possible distances between *successive* overlapping occurrences of $(\boldsymbol{w}, \boldsymbol{w}')$, it appears that some periods are not possible. For instance, the period $p = 3$ of $(\texttt{taca}, \texttt{acac})$ is not possible because an occurrence of taca at position i and an occurrence of acac at position $i + 3$ implies another occurrence of acac, in between (in fact at position $i + 1$). More generally, for two words $\boldsymbol{w}$ and $\boldsymbol{w}'$ of $\mathcal{W}$, a period $p \in \mathcal{P}(\boldsymbol{w}, \boldsymbol{w}')$ is said to be *principal* with respect to $\mathcal{W}$ if, for all $\boldsymbol{w}^\star \in \mathcal{W}$ and $j \in \mathcal{P}(\boldsymbol{w}, \boldsymbol{w}^\star)$, we have $p - j \notin \mathcal{P}(\boldsymbol{w}^\star, \boldsymbol{w}')$. This condition simply means that $\mathcal{W}$ cannot occur between an occurrence of $\boldsymbol{w}$ at a position i and an occurrence of $\boldsymbol{w}'$ at position $i + p$. We denote by $\mathcal{P}'_{\mathcal{W}}(\boldsymbol{w}, \boldsymbol{w}')$ the set of principal periods of $(\boldsymbol{w}, \boldsymbol{w}')$ 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 $\boldsymbol{w}$ then the set $\mathcal{P}'_{\{\boldsymbol{w}\}}(\boldsymbol{w}, \boldsymbol{w})$ coincides with the so-called *principal period set*, $\mathcal{P}'(\boldsymbol{w})$, of $\boldsymbol{w}$ introduced in [12].

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

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

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

Note that the same result holds for a later occurrence of $\mathcal{W}$ and a suffix

$$\boldsymbol{w}_{(p)} := w_{|\boldsymbol{w}|-p+1} \cdots w_{|\boldsymbol{w}|}$$

with $p \in \mathcal{P}'(\boldsymbol{w}, \boldsymbol{w}')$.

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

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

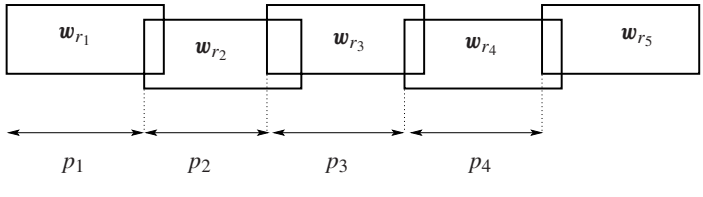

FIGURE 1: Structure of a word composed of exactly five occurrences of $\mathcal{W}$.

https://doi.org/10.1239/aap/1175266472 Published online by Cambridge University Press

occurrences of $\mathcal{W}$. Therefore, the word $\boldsymbol{c}$ has the form

$$\boldsymbol{c} = \boldsymbol{w}_{r_1}^{(p_1)} \cdots \boldsymbol{w}_{r_{k-1}}^{(p_{k-1})} \boldsymbol{w}_{r_k}. \tag{7}$$

To simplify the notation, the first word, $\boldsymbol{w}_{r_1}$, the second word, $\boldsymbol{w}_{r_2}$, and the last word, $\boldsymbol{w}_{r_k}$, of $\boldsymbol{c}$ are respectively denoted by $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$. We denote by $\mathcal{C}_k(\mathcal{W})$ the set of words of the form (7), by $\mathcal{C}_k^{(\boldsymbol{u};\boldsymbol{w})}(\mathcal{W})$ the subset of words of $\mathcal{C}_k(\mathcal{W})$ which begin with $\boldsymbol{u}$ and end with $\boldsymbol{w}$, and by $\mathcal{C}_k^{(\boldsymbol{u},\boldsymbol{v})}(\mathcal{W})$ the subset of words of $\mathcal{C}_k(\mathcal{W})$ which have $\boldsymbol{u}$ and $\boldsymbol{v}$ as the first two occurrences from $\mathcal{W}$. In the latter notation, when $\boldsymbol{v}$ is unknown, we replace it by a dot (e.g. we write $\mathcal{C}_k^{(\boldsymbol{u},\cdot)}(\mathcal{W})$).

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

$$\begin{aligned}\tilde{Y}_{i,k}(\mathcal{W}) = \sum_{\boldsymbol{u}\in\mathcal{W}} \sum_{\boldsymbol{w}\in\mathcal{W}} \sum_{\boldsymbol{c}\in\mathcal{C}_k^{(\boldsymbol{u};\boldsymbol{w})}(\mathcal{W})} \Bigg(& Y_i(\boldsymbol{c}) - \sum_{\boldsymbol{u}'\in\mathcal{W}} \sum_{p\in\mathcal{P}'(\boldsymbol{u}',\boldsymbol{u})} Y_{i-p}(\boldsymbol{u}'^{(p)}\boldsymbol{c}) \\ & - \sum_{\boldsymbol{w}'\in\mathcal{W}} \sum_{q\in\mathcal{P}'(\boldsymbol{w},\boldsymbol{w}')} Y_i(\boldsymbol{c}\boldsymbol{w}'_{(q)}) \\ & + \sum_{\boldsymbol{u}'\in\mathcal{W}} \sum_{\boldsymbol{w}'\in\mathcal{W}} \sum_{p\in\mathcal{P}'(\boldsymbol{u}',\boldsymbol{u})} \sum_{q\in\mathcal{P}'(\boldsymbol{w},\boldsymbol{w}')} Y_{i-p}(\boldsymbol{u}'^{(p)}\boldsymbol{c}\boldsymbol{w}'_{(q)}) \Bigg).\end{aligned} \tag{8}$$

Thus, by taking the expectation in (8), we obtain the equality

$$\begin{aligned}\tilde{\mu}_k(\mathcal{W}) &= \sum_{\boldsymbol{c}\in\mathcal{C}_k(\mathcal{W})} \mu(\boldsymbol{c}) - 2 \sum_{\boldsymbol{c}'\in\mathcal{C}_{k+1}(\mathcal{W})} \mu(\boldsymbol{c}') + \sum_{\boldsymbol{c}''\in\mathcal{C}_{k+2}(\mathcal{W})} \mu(\boldsymbol{c}'') \\ &= p_k(\mathcal{W}) - 2p_{k+1}(\mathcal{W}) + p_{k+2}(\mathcal{W}),\end{aligned} \tag{9}$$

where $p_k(\mathcal{W})$ and $p_k^{(\boldsymbol{u},\cdot)}(\mathcal{W})$ respectively denote the occurrence probabilities of a word of $\mathcal{C}_k(\mathcal{W})$ and a word of $\mathcal{C}_k^{(\boldsymbol{u},\cdot)}(\mathcal{W})$ occurring at a given position. The expression for $\tilde{\mu}_k(\mathcal{W})$ can thus be deduced from the one for the $p_k(\mathcal{W})$. The computation of $p_k(\mathcal{W})$ is done recursively. For all $k \geq 1$ and $\boldsymbol{u} = u_1 \cdots u_{|\boldsymbol{u}|} \in \mathcal{W}$,

$$\begin{aligned}p_1^{(\boldsymbol{u},\cdot)}(\mathcal{W}) &= \mu(\boldsymbol{u}), \\ p_{k+1}^{(\boldsymbol{u},\cdot)}(\mathcal{W}) &= \sum_{\boldsymbol{v}\in\mathcal{W}} \sum_{\boldsymbol{c}\in\mathcal{C}_{k+1}^{(\boldsymbol{u},\boldsymbol{v})}(\mathcal{W})} \mu(\boldsymbol{c}) \\ &= \sum_{\boldsymbol{v}\in\mathcal{W}} \sum_{p\in\mathcal{P}'(\boldsymbol{u},\boldsymbol{v})} \sum_{\boldsymbol{c}'\in\mathcal{C}_k^{(\boldsymbol{v},\cdot)}(\mathcal{W})} \mu(\boldsymbol{u}^{(p)}\boldsymbol{c}') \\ &= \sum_{\boldsymbol{v}\in\mathcal{W}} \frac{1}{\mu(v_1)} \sum_{p\in\mathcal{P}'(\boldsymbol{u},\boldsymbol{v})} \mu(\boldsymbol{u}^{(p+1)}) \sum_{\boldsymbol{c}'\in\mathcal{C}_k^{(\boldsymbol{v},\cdot)}(\mathcal{W})} \mu(\boldsymbol{c}') \\ &= \sum_{\boldsymbol{v}\in\mathcal{W}} A_{\boldsymbol{u},\boldsymbol{v}} p_k^{(\boldsymbol{v},\cdot)}(\mathcal{W}),\end{aligned} \tag{10}$$

https://doi.org/10.1239/aap/1175266472 Published online by Cambridge University Press

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

$$A_{\boldsymbol{u},\boldsymbol{v}} = \frac{\mu(u_1)}{\mu(v_1)} \sum_{p \in \mathcal{P}'(\boldsymbol{u},\boldsymbol{v})} \prod_{t=1}^{p} \pi(u_t, u_{t+1}). \tag{11}$$

Therefore, if we introduce the vectorial notation $\vec{p}_k(\mathcal{W})$ for the vector $[p_k^{(\boldsymbol{u},\cdot)}(\mathcal{W})]_{\boldsymbol{u}\in\mathcal{W}}$ and A for the matrix $[A_{\boldsymbol{u},\boldsymbol{v}}]_{\boldsymbol{u},\boldsymbol{v}\in\mathcal{W}}$, (10) can be written as follows: $\vec{p}_{k+1}(\mathcal{W}) = A\vec{p}_k(\mathcal{W})$ for all $k \geq 1$. Similarly, we have $\vec{p}_1(\mathcal{W}) = \vec{\mu}(\mathcal{W}) := [\mu(\boldsymbol{w})]_{\boldsymbol{w}\in\mathcal{W}}$, leading to

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

Denoting by $\|\cdot\|_1$ the 1-norm of $\mathbb{R}^d$ defined by $\|\vec{z}\|_1 = \sum_{r=1}^{d} |z_r|$ for all $\vec{z} = (z_1, \ldots, z_d) \in \mathbb{R}^d$, we can conclude that

$$p_k(\mathcal{W}) = \|\vec{p}_k(\mathcal{W})\|_1 = \|A^{k-1}\vec{\mu}(\mathcal{W})\|_1. \tag{12}$$

Combining relations (9) and (12) yields our final expression for $\tilde{\mu}_k(\mathcal{W})$:

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

This establishes the following proposition.

Proposition 1. *For all families $\mathcal{W}$, the occurrence probability of a k-clump of $\mathcal{W}$ is given by*

$$\tilde{\mu}_k(\mathcal{W}) = \|A^{k-1}(I - A)^2\vec{\mu}(\mathcal{W})\|_1, \tag{13}$$

where I is the identity matrix of $\mathbb{R}^d$, A is the matrix of coefficients $[A_{\boldsymbol{u},\boldsymbol{v}}]_{\boldsymbol{u},\boldsymbol{v}\in\mathcal{W}}$ defined in (11), $\vec{\mu}(\mathcal{W})$ is the vector $[\mu(\boldsymbol{w})]_{\boldsymbol{w}\in\mathcal{W}}$, and $\|\cdot\|_1$ is the 1-norm of $\mathbb{R}^d$.

Remarks 2. 1. Theorem 1 and Proposition 1 generalize [12, Theorem 13]: indeed, for a single-word family $\mathcal{W} = \{\boldsymbol{w}\}$, (13) reduces to $\tilde{\mu}_k(\boldsymbol{w}) = a_{\boldsymbol{w}}^{k-1}(1 - a_{\boldsymbol{w}})^2\mu(\boldsymbol{w})$, where $a_{\boldsymbol{w}}$ is the probability of there being two successive overlapping occurrences of $\boldsymbol{w}$ and is given by $a(\boldsymbol{w}) = \sum_{p\in\mathcal{P}'(\boldsymbol{w})} \prod_{t=1}^{p} \pi(w_t, w_{t+1})$ with $\mathcal{P}'(\boldsymbol{w}) := \mathcal{P}'_{\{\boldsymbol{w}\}}(\boldsymbol{w}, \boldsymbol{w})$.

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

3. From (9), we can moreover show that

$$\sum_{k\geq 1} k\tilde{\mu}_k(\mathcal{W}) = \mu(\mathcal{W}), \tag{14}$$

$$\sum_{k\geq 1} \tilde{\mu}_k(\mathcal{W}) = \|(I - A)\vec{\mu}(\mathcal{W})\|_1. \tag{15}$$

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 $I := \{1, \ldots, n - h + 1\} \times \mathbb{N} \setminus \{0\}$, and then bound the three quantities b_1, b_2, and b_3 defined respectively by (3), (4), and (5). To do so, we will adapt the setup presented in [12] for a single word.

https://doi.org/10.1239/aap/1175266472 Published online by Cambridge University Press

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

For each $(i,k) \in I$, we define a set $Z(i,k) \subset \mathbb{Z}$ which contains all the indices j of the letters X_j used in the definition of $\tilde{Y}_{i,k}(\mathcal{W})$. We can take $Z(i,k) = \{s \in \mathbb{Z} \text{ such that } i-h \le s \le i+(k+1)h\}$, 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 ensure 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:

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

This implies that if $\tilde{Y}_{i,k}(\mathcal{W}) = \tilde{Y}_{j,\ell}(\mathcal{W}) = 1$ 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

$$\begin{aligned} b_1 &= \sum_{i=1}^{n-h+1} \sum_{k\ge1} \sum_{(j,\ell)\in B_{i,k}} \mathrm{E}(\tilde{Y}_{i,k}(\mathcal{W}))\,\mathrm{E}(\tilde{Y}_{j,\ell}(\mathcal{W})) \\ &\le \sum_{i=1}^{n-h+1} \sum_{k\ge1} \sum_{\ell\ge1} \sum_{j=i-(\ell+3)h}^{i+(k+3)h} \tilde{\mu}_k(\mathcal{W})\tilde{\mu}_\ell(\mathcal{W}). \end{aligned}$$

Let $\tilde{\mu}(\mathcal{W})$ be the probability of a clump of $\mathcal{W}$ occurring at a given position; it satisfies $\tilde{\mu}(\mathcal{W}) = \sum_{k\ge1} \tilde{\mu}_k(\mathcal{W}) \le \mu(\mathcal{W})$. Using the symmetry between i and j and between k and l, and (14), we can write

$$\begin{aligned} b_1 &\le 2\tilde{\mu}(\mathcal{W}) \sum_{i=1}^{n-h+1} \sum_{k\ge1} ((k+3)h+1)\tilde{\mu}_k(\mathcal{W}), \\ &\le 2(n-h+1)\tilde{\mu}(\mathcal{W})([\mu(\mathcal{W})+3\tilde{\mu}(\mathcal{W})]h+\tilde{\mu}(\mathcal{W})) \\ &\le 10nh\mu^2(\mathcal{W}). \end{aligned} \tag{16}$$

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

4.3. Bounding b_2

From definition (4), we have

$$b_2 = \sum_{i=1}^{n-h+1} \sum_{k\ge1} \sum_{(j,\ell)\in B_{i,k}\setminus\{(i,k)\}} \mathrm{E}(\tilde{Y}_{i,k}(\mathcal{W})\tilde{Y}_{j,\ell}(\mathcal{W})).$$

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

$$b_2 \le 2 \sum_{i=1}^{n-h+1} \sum_{k\ge1} \sum_{\ell\ge1} \sum_{j=i+1}^{i+(k+3)h} \mathrm{E}(\tilde{Y}_{i,k}(\mathcal{W})\tilde{Y}_{j,\ell}(\mathcal{W})).$$

Let $\tilde{Y}_j(\mathcal{W}) = \sum_{\ell\ge1} \tilde{Y}_{j,\ell}(\mathcal{W})$ denote a Bernoulli variable that is equal to 1 if a clump of $\mathcal{W}$ occurs at position j and is equal to 0 otherwise. Since $\tilde{Y}_{i,k}(\mathcal{W}) = \sum_{\boldsymbol{c}\in\mathcal{C}_k(\mathcal{W})} \tilde{Y}_{i,k}(\mathcal{W})Y_i(\boldsymbol{c})$,

https://doi.org/10.1239/aap/1175266472 Published online by Cambridge University Press

we have

$$b_2 \le 2 \sum_{i=1}^{n-h+1} \sum_{k\ge 1} \sum_{j=i+1}^{i+(k+3)h} \mathrm{E}(\tilde{Y}_{i,k}(\mathcal{W})\tilde{Y}_j(\mathcal{W})),$$
$$\le 2 \sum_{i=1}^{n-h+1} \sum_{k\ge 1} \sum_{\boldsymbol{c}\in\mathcal{C}_k(\mathcal{W})} \sum_{j=i+1}^{i+(k+3)h} \mathrm{E}(\tilde{Y}_{i,k}(\mathcal{W})Y_i(\boldsymbol{c})\tilde{Y}_j(\mathcal{W})).$$

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

$$b_2 \le 2 \sum_{i=1}^{n-h+1} \sum_{k\ge 1} \sum_{\boldsymbol{c}\in\mathcal{C}_k(\mathcal{W})} \sum_{j=i+|\boldsymbol{c}|}^{i+(k+3)h} \mathrm{E}(\tilde{Y}_{i,k}(\mathcal{W})Y_i(\boldsymbol{c})\tilde{Y}_j(\mathcal{W}))$$
$$\le 2 \sum_{i=1}^{n-h+1} \sum_{k\ge 1} \sum_{\boldsymbol{c}\in\mathcal{C}_k(\mathcal{W})} \sum_{j=i+|\boldsymbol{c}|+h}^{i+(k+3)h} \mathrm{E}(\tilde{Y}_{i,k}(\mathcal{W})Y_i(\boldsymbol{c})Y_j(\mathcal{W}))$$
$$+ 2 \sum_{i=1}^{n-h+1} \sum_{k\ge 1} \sum_{\boldsymbol{c}\in\mathcal{C}_k(\mathcal{W})} \sum_{j=i+|\boldsymbol{c}|}^{i+|\boldsymbol{c}|+h-1} \mathrm{E}(\tilde{Y}_{i,k}(\mathcal{W})Y_i(\boldsymbol{c})Y_j(\mathcal{W})).$$

The first and second terms on the right-hand side will respectively be denoted by b_{21} and b_{22}. Let us bound b_{21}. Note that the random variable $\tilde{Y}_{i,k}(\mathcal{W})Y_i(\boldsymbol{c})$ only involves the letters $X_{i-h+1}, \dots, X_{i+|\boldsymbol{c}|+h-1}$, whereas $Y_j(\mathcal{W})$ involves the letters $X_j, \dots, X_{j+h-1}$. Therefore, for every position j which satisfies $j \ge i+|\boldsymbol{c}|+h$, the Markov property yields

$$\mathrm{E}(\tilde{Y}_{i,k}(\mathcal{W})Y_i(\boldsymbol{c})Y_j(\mathcal{W})) \le \frac{\mu(\mathcal{W})}{\mu_{\min}} \mathrm{E}(\tilde{Y}_{i,k}(\mathcal{W})Y_i(\boldsymbol{c})),$$

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

$$b_{21} \le 2(n-h+1)\frac{\mu(\mathcal{W})}{\mu_{\min}} \sum_{k\ge 1} (k+2)h\tilde{\mu}_k(\mathcal{W})$$
$$\le 2(n-h+1)\frac{\mu(\mathcal{W})}{\mu_{\min}} (\mu(\mathcal{W}) + 2\tilde{\mu}(\mathcal{W}))h$$
$$\le \frac{6nh}{\mu_{\min}} \mu^2(\mathcal{W}). \tag{17}$$

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

$$\mathrm{E}(\tilde{Y}_i(\mathcal{W})Y_i(\boldsymbol{c})Y_j(\mathcal{W})) \le \frac{\mu(\mathcal{W})}{\mu_{\min}} \mathrm{E}(\tilde{Y}_i(\mathcal{W})Y_i(\boldsymbol{c})).$$

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

$$b_{22} \le \frac{2(n-h+1)h}{\mu_{\min}} \mu(\mathcal{W}) \sum_{k\ge 1} \sum_{\boldsymbol{c}\in\mathcal{C}_k(\mathcal{W})} \mathrm{E}(\tilde{Y}_i(\mathcal{W})Y_i(\boldsymbol{c})) \le \frac{2nh}{\mu_{\min}} \mu^2(\mathcal{W}). \tag{18}$$

https://doi.org/10.1239/aap/1175266472 Published online by Cambridge University Press

Indeed,

$$\sum_{k\ge1}\sum_{\boldsymbol{c}\in\mathcal{C}_k(\mathcal{W})}\mathrm{E}(\tilde{Y}_i(\mathcal{W})Y_i(\boldsymbol{c})) = \sum_{k\ge1}\mathrm{P}(\text{a } K\text{-clump of } \mathcal{W} \text{ with } K\ge k \text{ starts at position } i)$$
$$= \sum_{k\ge1}\sum_{K\ge k}\tilde{\mu}_K(\mathcal{W}) = \sum_{K\ge1}K\tilde{\mu}_K(\mathcal{W}) = \mu(\mathcal{W}).$$

Finally, combining (17) and (18) leads to

$$b_2 \le \frac{8nh}{\mu_{\min}}\mu^2(\mathcal{W}). \tag{19}$$

4.4. Bounding b_3

From definition (5), we have

$$b_3 = \sum_{i=1}^{n-h+1}\sum_{k\ge1}\mathrm{E}(|\mathrm{E}(\tilde{Y}_{i,k}(\mathcal{W}) - \tilde{\mu}_k(\mathcal{W}) \mid \sigma(\tilde{Y}_{j,\ell}(\mathcal{W})\colon (j,\ell)\notin B_{i,k}))|).$$

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

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

Therefore, owing to the Markov property, we have

$$b_3 \le \sum_{i=1}^{n-h+1}\sum_{k\ge1}\sum_{\boldsymbol{rcs}\in\mathcal{C}'_k}\mathrm{E}(|\mathrm{E}(Y_{i-h}(\boldsymbol{rcs}) - \mu(\boldsymbol{rcs}) \mid \sigma(\dots, X_{i-2h}, X_{i+|\boldsymbol{c}|+2h}, \dots))|)$$
$$\le \sum_{i=1}^{n-h+1}\sum_{k\ge1}\sum_{\boldsymbol{rcs}\in\mathcal{C}'_k}\mathrm{E}(|\mathrm{E}(Y_{i-h}(\boldsymbol{rcs}) - \mu(\boldsymbol{rcs}) \mid X_{(i-h)-h}, X_{(i-h)+|\boldsymbol{rcs}|+h})|).$$

Now we use the following result, proved in [13]: for all words $\boldsymbol{w}$ and all integers j and t,

$$\mathrm{E}(|\mathrm{E}(Y_j(\boldsymbol{w}) - \mu(\boldsymbol{w}) \mid X_{j-t}, X_{j+|\boldsymbol{w}|+t})|) \le C'\mu(\boldsymbol{w})|\alpha|^t,$$

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

$$b_3 \le \sum_{i=1}^{n-h+1}\sum_{k\ge1}\sum_{\boldsymbol{rcs}\in\mathcal{C}'_k} C'\mu(\boldsymbol{rcs})|\alpha|^h.$$

Finally, the equality $\sum_{k\ge1}\sum_{\boldsymbol{rcs}\in\mathcal{C}'_k}\mu(\boldsymbol{rcs}) = \tilde{\mu}(\mathcal{W})$ yields

$$b_3 \le C'(n-h+1)|\alpha|^h\tilde{\mu}(\mathcal{W}) \le C'n\mu(\mathcal{W})|\alpha|^h. \tag{20}$$

Inequalities (16), (19), and (20) establish Theorem 1.

https://doi.org/10.1239/aap/1175266472 Published online by Cambridge University Press

5. Clumps and competing renewals

When counting the occurrences of a word or word family in a finite sequence $X_1 \cdots X_n$, one may be interested in counting only nonoverlapping occurrences, for instance clumps or renewals. A renewal can be recursively defined as follows: an occurrence is a renewal if and only if either it is the first occurrence 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 follow from Theorem 1.

First, inequalities (2), (16), (19), and (20) lead to

$$d_{\mathrm{TV}}(\mathcal{D}(\tilde{N}^{\infty}(\mathcal{W})), \mathcal{P}(\tilde{\lambda})) \le Cnh\mu^2(\mathcal{W}) + C'n\mu(\mathcal{W})|\alpha|^h,$$

where $\tilde{N}^{\infty}(\mathcal{W}) := \sum_{k\ge 1} \tilde{N}_k^{\infty}(\mathcal{W})$, $\mathcal{P}(\cdot)$ denotes the Poisson distribution, and, using (15),

$$\tilde{\lambda} := \mathrm{E}(\tilde{N}^{\infty}(\mathcal{W})) = (n-h+1)\|(I-A)\vec{\mu}(\mathcal{W})\|_1.$$

Moreover, $\tilde{N}^{\infty}(\mathcal{W})$ asymptotically has the same distribution as the number, $\tilde{N}(\mathcal{W})$, of clumps of $\mathcal{W}$ in $X_1 \cdots X_n$: $\mathrm{P}(\tilde{N}^{\infty}(\mathcal{W}) \ne \tilde{N}(\mathcal{W})) \le h\mu(\mathcal{W})$ (by same argument as for N^{∞}). Therefore, under both $h = o(n)$ and the rare event condition $\mathrm{E}(N(\mathcal{W})) = O(1)$, the total variation distance between the distribution of $\tilde{N}(\mathcal{W})$ and the Poisson distribution $\mathcal{P}(\tilde{\lambda})$ tends to 0 as n tends to ∞.

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

$$d_{\mathrm{TV}}(\mathcal{D}(R(\mathcal{W})), \mathcal{D}(\tilde{N}(\mathcal{W}))) \le \mathrm{P}(R(\mathcal{W}) \ne \tilde{N}(\mathcal{W})) \le \frac{1}{\mu_{\min}} nh\mu^2(\mathcal{W}), \tag{21}$$

where, recall, $\mu_{\min} = \min_{\boldsymbol{w}\in\mathcal{W}} \mu(w_1) > 0$. Indeed, we note 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 $R(\mathcal{W}) = \tilde{N}(\mathcal{W})$. Thus, if $R(\mathcal{W}) \ne \tilde{N}(\mathcal{W})$ then there 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 let $\boldsymbol{u}$ be the occurrence from $\mathcal{W}$ it starts with. Then an occurrence of $\boldsymbol{u}$ starts at position i and an occurrence of $\mathcal{W}$ starts between positions $i+|\boldsymbol{u}|$ and $i+|\boldsymbol{u}|+h-1$; this occurs with probability $h\mu(\boldsymbol{u})\mu(\mathcal{W})/\mu_{\min}$. Summing over $i \in \{1, \ldots, n-h+1\}$ and $\boldsymbol{u} \in \mathcal{W}$ leads to inequality (21). Owing to the triangular inequality, we then obtain the following Poisson approximation for the number of competing renewals:

$$d_{\mathrm{TV}}(\mathcal{D}(R(\mathcal{W})), \mathcal{P}(\tilde{\lambda})) = \mathrm{O}(nh\mu^2(\mathcal{W}) + n\mu(\mathcal{W})|\alpha|^h + h\mu(\mathcal{W})).$$

If $\mathrm{E}(N(\mathcal{W})) = O(1)$ and $h = o(n)$, then the total variation distance between the distribution of $R(\mathcal{W})$ and the Poisson distribution $\mathcal{P}(\tilde{\lambda})$ tends to 0 as n tends to ∞. This Poisson distribution is in fact very close to the natural limiting Poisson distribution with parameter $\mathrm{E}(R(\mathcal{W}))$ proposed in [3], because their respective parameters are asymptotically equivalent under the rare condition and $h = o(n)$. However, in practice calculating $\mathrm{E}(R(\mathcal{W}))$ requires solving a system of equations, whereas the expression for $\tilde{\lambda}$ 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

https://doi.org/10.1239/aap/1175266472 Published online by Cambridge University Press

length n. The error of approximation converges to 0 given that the word family $\mathcal{W}$ is expectedly rare ($\mathrm{E}(N(\mathcal{W})) = O(1)$) and the maximal word length is of order less than n.

Our results can easily be extended to the case of a Markov chain of order m, $2 \le m \le \min\{|\boldsymbol{w}|\colon \boldsymbol{w} \in \mathcal{W}\} - 1$. It suffices to consider the sequence $\boldsymbol{X}^*$ obtained by letting $X_i^* := X_i X_{i+1} \cdots X_{i+m-1}$, which is a Markov chain of order 1 on the alphabet $\mathcal{A}^* := \mathcal{A}^m$. Moreover, an occurrence of $\mathcal{W}$ in $\boldsymbol{X}$ corresponds to an occurrence of $\mathcal{W}^*$ in $\boldsymbol{X}^*$, 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 $(\boldsymbol{u}, \boldsymbol{v})$-indexed coefficient is given by

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

and $\pi(\cdot,\cdot)$ and $\mu(\cdot)$ respectively denote the transition probabilities and the stationary distribution of the model. This compound Poisson distribution has been included in the R'MES software (see http://genome.jouy.inra.fr/ssb/rmes/), used 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 as a hidden Markov chain. Indeed, a hidden Markov chain $(\boldsymbol{X}, \boldsymbol{S})$ on the alphabet $\mathcal{A}$ with state space $\{1, \ldots, s\}$ can be written as an order-1 Markov chain $\overline{\boldsymbol{X}}$ on the alphabet $\mathcal{A} \times \{1, \ldots, s\}$, and an occurrence of a given word $\boldsymbol{w}$ in $\boldsymbol{X}$ corresponds to an occurrence of a word family $\overline{\mathcal{W}}$ in $\overline{\boldsymbol{X}}$. For instance, if there are two states, 1 and 2, the word family $\overline{\mathcal{W}}$ associated with $\boldsymbol{w} = \mathtt{aca}$ 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 $\mathtt{a}_j$ and $\mathtt{c}_j$ respectively stand for the letters a and c in state j.

Acknowledgements

The authors thank an anonymous reviewer for his/her helpful comments. This work was 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. *Statist. Sci*. **5,** 403–434.

[2] Chryssaphinou, O. and Papastavridis, S. (1990). The occurrence of sequence patterns in repeated dependent experiments. *Theory Prob. Appl*. **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. *Combin. Prob. Comput*. **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 de Turckheim, É. (1995). Finding words with unexpected frequencies in DNA sequences. *J. R. Statist. Soc. B* **57,** 205–220.

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

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

[9] Reinert, G., Schbath, S. and Waterman, M. (2000). Probabilistic and statistical properties of words. *J. Comput. Biol.* **7,** 1–46.

[10] ROBIN, S. AND DAUDIN, J.-J. (1999). Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Prob.* **36,** 179–193.
[11] ROBIN, S. AND SCHBATH, S. (2001). Numerical comparison of several approximations of the word count distribution in random sequences. *J. Comput. Biol.* **8,** 349–359.
[12] SCHBATH, S. (1995). Compound Poisson approximation of word counts in DNA sequences. *ESAIM Prob. Statist*. **1,** 1–16.
[13] 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. Doctoral Thesis, Université René Descartes, Paris V.