Journal of Computational Biology
Volume 5, Number 2, 1998
Mary Ann Liebert, Inc.
Pp. 223–253

Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains

GESINE REINERT[1] and SOPHIE SCHBATH[2]

Abstract

We derive a Poisson process approximation for the occurrences of clumps of multiple words, and a compound Poisson process approximation for the number of occurrences of multiple words in a sequence of letters generated by a stationary Markov chain. Using the Chen-Stein method we provide a bound on the error in the approximations. For rare words, these errors tend to zero as the length of the sequence increases to infinity. Modeling a DNA sequence as a stationary Markov chain, we show as an application that the compound Poisson approximation is efficient for the number of occurrences of rare stem-loop motifs.

Key words: Chen-Stein method, stem-loop motifs, compound Poisson approximation, Poisson process approximation, occurrences of multiple words.

1 Introduction

We consider a finite sequence S of letters in a finite alphabet $\mathcal{A}$. The letters of the sequence S can be chosen independently in the alphabet or, more generally, generated by a stationary Markov chain. Our main example will be a Markov chain on the four-letter DNA alphabet $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ but our results are valid for general finite alphabets such as $\{0, 1\}$ or the 20-letter amino acid alphabet. An abundant literature exists on the asymptotic distribution of the number of occurrences of a single word in such a sequence S. A normal approximation, valid for frequent words, is presented by Prum *et al.* (1995). Arratia *et al.* (1990), Geske *et al.* (1995) and Schbath (1995a) showed that the number of clumps of a single rare word can be approximated by a Poisson variable and that, in contrast, the number of occurrences of a single rare word is approximately compound Poisson distributed.

As soon as one is simultaneously interested in occurrences of different rare words in a sequence, the asymptotic joint distribution of the different counts is of interest. The novelty in this paper is to provide multidimensional results and to give sufficient conditions which ensure the asymptotic independence, which is very useful in practice. Our main result concerns the approximation of the m-tuple of counts of m different words by m independent compound Poisson variables. The parameters of the asymptotic distributions are identical to those obtained if the words were considered separately (Schbath, 1995a). This approximation is valid under some reasonable conditions on the overlapping structure between the m words, and it is good if the words are long related to the length of the sequence, so that the words are rare, (as a rule of thumb, a word is rare if its length is at least of order $\log n$, where n is the length of the sequence) and the words do not overlap too much between themselves. The main difficulty here is not given by the self-overlaps but by the possible overlaps between two different words.

[1]Department of Statistics, UCLA, Los Angeles, CA 90095, USA.
[2]Unité de Biométrie, INRA, 78352 Jouy-en-Josas, France.

Using the Chen-Stein method for Poisson process approximation as stated in Arratia *et al.* (1990) and refined in Barbour *et al.* (1992b), we give an explicit error bound for the approximation improving those obtained by Schbath (1995a) for a single word. A more direct approach could have been the Chen-Stein method for compound Poisson approximation, developed by Barbour *et al.* (1992a), Roos (1994) and Barbour and Utev (1997), which has been applied in this context to approximate the count of single words with the most simple self-overlapping structure (Roos and Stark, 1996), but it is not adapted for a multidimensional approximation to multiple words.

The multidimensional approximation for counts of multiple words, in particular the asymptotic statistical independence, is very useful and convenient to study statistical properties of any function of these counts. For instance, in the application section we consider the simple but frequent case where the quantity of interest is the count of a word family, which is the sum of the counts of all the words within the family. In the four-letter DNA alphabet, a word family could be a stem-loop motif or a set of words coding for the same amino-acid sequence. (Note that asymptotic statistical independence does not necessarily have a biological interpretation).

For non-rare words (short words related to the length of the sequence), a Gaussian approximation is more appropriate and corresponding results have been shown: Lundstrom (1990) was the first to derive a multidimensional Gaussian approximation (using the δ-method) for a m-tuple of counts; see Waterman (1995), Chapter 12, for an exposition. Prum *et al.* (1995) or Schbath *et al.* (1995) gave an explicit formula for the asymptotic covariance matrix. These results can be used to construct a Gaussian statistic based on the count of a word family. Recently, Tanushev (1996) presented the multidimensional Gaussian approximation for an m-tuple of renewal counts.

To approximate the counts of occurrences of rare words, we use the "declumping" approach: we first count the number of clumps of occurrences. To this purpose, we prove the multidimensional Poisson approximation for the joint distribution of counts of clumps (also called declumped counts) of m different words.

All the results given in this paper are not "naive" approximations, but instead give explicit bounds on the distance to a suitable Poisson or compound Poisson distribution. The bounds are in terms of word lengths, sequence length, letter probabilities (using a Markov model), and of the explicit overlapping structure between the words within the family. In particular, these bounds will tend to zero as the sequence length tends to infinity, if the words within the family do not overlap too much. In case of considerable overlaps, our bounds will not be informative. This is the scope of our results. It is intuitively obvious that, in case of considerable overlaps, an approximation with independent variables should not be valid.

In Section 2, we define an occurrence of a word and an occurrence of a clump in a sequence, as well as the number of occurrences of a word and the number of clumps of a word in a sequence. Moreover we introduce the decomposition of the count of a word with respect to the number of clumps; this decomposition is fundamental to prove the compound Poisson approximation via the Chen-Stein method given in Arratia *et al.* (1990). Next, in Section 3 we present the Poisson and the compound Poisson process approximations, related first to declumped counts, and then to counts of two words of same length. Sketches of the proofs are drawn in order to understand the different terms appearing in the error bounds. The complete proofs can be found in the appendices. Then these results are generalized in Section 4 to m words with not necessarily identical lengths. As an illustration (Section 5), we study the count of some stem-loop motifs, like `ATGGCNNNNNGCCAT` (`N` denotes any letter in the four-letter DNA alphabet), in a first order Markov chain on $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ modeling the λ-phage genome. We give the expected counts, the error bounds and the asymptotic distributions. As we will see, the error bounds are very small, and thus the method provides a good way to approximate a collection of counts. Finally, Appendices A and B contain the proofs. The proof of the results corresponding to two words of same length are fundamental and are detailed in Appendix A. The general case, for more than two words, is proved in Appendix B.

2 Preliminary notation and the Chen-Stein method

We consider a stationary Markov chain $\mathcal{M} = \{X_i\}_{i\in\mathbb{Z}}$ on a finite alphabet $\mathcal{A}$, with transition matrix Π such that $\Pi(x,y) > 0\ \forall x, y \in \mathcal{A}$. This implies that the Markov chain has a unique stationary distribution μ defined by

$$\mu(x) = \sum_{y\in\mathcal{A}} \mu(y)\Pi(y,x)\,, \quad \forall x \in \mathcal{A}.$$

Moreover, we have

$$\mu_{\min} := \min_{x\in\mathcal{A}} \mu(x) > 0\,.$$

The model where the letters $\{X_i\}_{i\in\mathbb{Z}}$ are chosen independently with probabilities $p_1, p_2, \ldots, p_{|\mathcal{A}|}$, where $|\mathcal{A}|$ denotes the size of the alphabet, corresponds to the transition matrix Π with identical rows $(p_1\ p_2\ \cdots\ p_{|\mathcal{A}|})$ and stationary distribution $\mu = (p_1, p_2, \ldots, p_{|\mathcal{A}|})$.

Let $\underline{u} = u_1u_2\cdots u_\ell$ be a word of length ℓ on $\mathcal{A}$. Say that an occurrence of $\underline{u}$ starts at position i in the infinite sequence $\mathcal{M}$ if $X_iX_{i+1}\cdots X_{i+\ell-1} = u_1u_2\cdots u_\ell$, and denote the indicator random variable of this event by $\mathbb{I}_i(\underline{u})$;

$$\mathbb{I}_i(\underline{u}) := \mathbb{I}\{\text{an occurrence of } \underline{u} \text{ starts at position } i \text{ in } \mathcal{M}\}\,.$$

The probability $\mu(\underline{u})$ that $\underline{u}$ starts at a given position in $\mathcal{M}$ is exactly the expectation of $\mathbb{I}_i(\underline{u})$ and is given by

$$\mu(\underline{u}) := \mathbb{E}\mathbb{I}_i(\underline{u}) = \mu(u_1)\Pi(u_1,u_2)\cdots\Pi(u_{\ell-1},u_\ell)\,.$$

In the finite sequence $S = X_1X_2\cdots X_n$ of length n, the number $N(\underline{u})$ of occurrences of $\underline{u}$ in S is defined by

$$N(\underline{u}) = \sum_{i=1}^{n-\ell+1} \mathbb{I}_i(\underline{u})\,, \tag{1}$$

and its expectation is

$$\mathbb{E}N(\underline{u}) = (n-\ell+1)\mu(\underline{u})\,. \tag{2}$$

Occurrences of a word may overlap in S or $\mathcal{M}$. Through this section, the example $S =$ **TAAGAAGAAGAAGAAGT** and $\underline{u} =$ **AAGAAGAA** is used. In this case, the word $\underline{u}$ occurs in S at positions 2, 5 and 8. The self-overlapping structure of a word can be described via the set of principal periods defined as follows. The lag between two overlapping occurrences of $\underline{u}$ is said to be a *period* of the word $\underline{u}$ (Guibas and Odlyzko (1981), Lothaire (1983)). A word may have several periods; for any word $\underline{u}$, the set $\mathcal{P}(\underline{u})$ of the periods of $\underline{u}$ is defined by

$$\mathcal{P}(\underline{u}) := \{p \in \{1,\ldots,\ell-1\} : u_i = u_{i+p}, \forall i = 1,\ldots,\ell-p\}\,.$$

The word $\underline{u}$ is a non-self-overlapping word if and only if $\mathcal{P}(\underline{u})$ is empty. The most relevant periods (see for instance (3)) are the ones which cannot be written as kp_0 with $k > 1$ and p_0 the minimal period. These periods are said to be *principal*; let $\mathcal{P}'(\underline{u})$ denote the set of the principal periods of $\underline{u}$. For example, $\mathcal{P}(\underline{u}) = \{3,6,7\}$ and $\mathcal{P}'(\underline{u}) = \{3,7\}$ for $\underline{u} =$ **AAGAAGAA**.

As announced in the introduction, we will also study the occurrences of clumps of a word $\underline{u}$. The concept of clump will be very important in the remainder of the paper. A *clump* of $\underline{u}$ in a sequence is a maximal set of overlapping occurrences of $\underline{u}$ in this sequence; no two clumps of $\underline{u}$ overlap in the sequence. Say that a clump of $\underline{u}$ starts at position i in the infinite sequence $\mathcal{M}$ if an occurrence of $\underline{u}$ starts at position i in $\mathcal{M}$ and if this occurrence is not overlapped by a preceding occurrence of $\underline{u}$. Denote the indicator random variable of this event by $\widetilde{\mathbb{I}}_i(\underline{u})$;

$$\begin{aligned}\widetilde{\mathbb{I}}_i(\underline{u}) &:= \mathbb{1}\{\text{a clump of } \underline{u} \text{ starts at position } i \text{ in } \mathcal{M}\}\\ &= \mathbb{I}_i(\underline{u}) \prod_{j=i-\ell+1}^{i-1} (1-\mathbb{I}_j(\underline{u})) .\end{aligned}$$

The probability $\widetilde{\mu}(\underline{u})$ that a clump of $\underline{u}$ starts at a given position in $\mathcal{M}$ is exactly the expectation of $\widetilde{\mathbb{I}}_i(\underline{u})$; Schbath (1995a) proved that

$$\widetilde{\mu}(\underline{u}) := \mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u}) = \mu(\underline{u}) - \sum_{p\in\mathcal{P}'(\underline{u})} \mu(\underline{u}^{(p)}\underline{u}) , \tag{3}$$

where $\underline{u}^{(p)} = u_1u_2\cdots u_p$ is the word composed of the first p letters of $\underline{u}$. Here is the sketch of the proof for equation (3). A clump of $\underline{u}$ starts at position i in the sequence if and only if there is an occurrence of $\underline{u}$ starting at position i and there are none of the $\underline{u}^{(p)}$ starting at $i-p$ where $p \in \mathcal{P}(\underline{u})$. In fact, it suffices to exclude the occurrences of all the $\underline{u}^{(p)}$ at $i-p$ where p is only a principal period of $\underline{u}$. Thus

$$\begin{aligned}\widetilde{\mathbb{I}}_i(\underline{u}) &= \mathbb{I}_i(\underline{u})\mathbb{1}\left\{\cap_{p\in\mathcal{P}'(\underline{u})}\{\text{no occurrence of } \underline{u}^{(p)} \text{ starts at } i-p\}\right\}\\ &= \mathbb{I}_i(\underline{u})\left(1-\mathbb{1}\left\{\cup_{p\in\mathcal{P}'(\underline{u})}\{\text{an occurrence of } \underline{u}^{(p)} \text{ starts at } i-p\}\right\}\right)\end{aligned}$$

One can then show that the events $\{$an occurrence of $\underline{u}^{(p)}$ starts at $i-p\}$ for $p \in \mathcal{P}'(\underline{u})$ are disjoint, meaning that any two of them cannot occur simultaneously (Schbath, 1995b). This leads to

$$\begin{aligned}\widetilde{\mathbb{I}}_i(\underline{u}) &= \mathbb{I}_i(\underline{u})\left(1-\sum_{p\in\mathcal{P}'(\underline{u})}\mathbb{I}_{i-p}(\underline{u}^{(p)})\right)\\ &= \mathbb{I}_i(\underline{u}) - \sum_{p\in\mathcal{P}'(\underline{u})}\mathbb{I}_{i-p}(\underline{u}^{(p)}\underline{u});\end{aligned}$$

equation (3) then easily follows.

Now, define $\widetilde{N}_{\text{inf}}(\underline{u})$ as the count

$$\widetilde{N}_{\text{inf}}(\underline{u}) := \sum_{i=1}^{n-\ell+1}\widetilde{\mathbb{I}}_i(\underline{u}),$$

so that $\widetilde{N}_{\text{inf}}(\underline{u})$ represents the number of clumps of $\underline{u}$ in the infinite sequence $\mathcal{M}$ but starting in S. Its expectation is

$$\mathbb{E}\widetilde{N}_{\text{inf}}(\underline{u}) = (n-\ell+1)\widetilde{\mu}(\underline{u}) . \tag{4}$$

For $1 \le i \le \ell-1$, the definition of $\widetilde{\mathbb{I}}_i(\underline{u})$ involves in particular the letters X_j with $i-\ell+1 \le j \le 0$. In practice, only the finite sequence $S = X_1X_2\dots X_n$ is observable and the observable number of clumps, denoted by $\widetilde{N}(\underline{u})$, may be different from $\widetilde{N}_{\rm inf}(\underline{u})$. In our example, we observe in S a unique clump of $\underline{u}$ starting at position 2 and ending at position 15, and $X_1 = \texttt{T}$ ensures that this observable clump is a real one in the infinite sequence. This might not be the case if the first letter X_1 was a `G`. Our quantity of interest is $\widetilde{N}(\underline{u})$, the observable number of clumps, but we will work with the easier count $\widetilde{N}_{\rm inf}(\underline{u})$, since we can control the boundary effect. Indeed, $\mathbb{P}(\widetilde{N}(\underline{u}) \neq \widetilde{N}_{\rm inf}(\underline{u}))$ is an upper bound on the total variation distance between $\widetilde{N}(\underline{u})$ and $\widetilde{N}_{\rm inf}(\underline{u})$ (see, e.g., Barbour *et al.* (1992b)), and we have

$$\begin{aligned}
\mathbb{P}(\widetilde{N}(\underline{u}) \neq \widetilde{N}_{\rm inf}(\underline{u})) &\leq \sum_{i=1}^{\ell-1} \mathbb{P}\Big(\mathbb{I}_i(\underline{u}) = 1, \mathbb{I}_j(\underline{u}) = 1 \text{ for some } j = i-\ell+1, \ldots, i-1\Big) \\
&= \sum_{i=1}^{\ell-1} \mathbb{P}\Big(\mathbb{I}_i(\underline{u}) = 1, \mathbb{I}_{i-p}(\underline{u}) = 1 \text{ for some } p \in \mathcal{P}(\underline{u})\Big) \\
&= \sum_{i=1}^{\ell-1} \mathbb{P}\Big(\mathbb{I}_i(\underline{u}) = 1, \mathbb{I}_{i-p}(\underline{u}) = 1 \text{ for some } p \in \mathcal{P}'(\underline{u})\Big) \\
&= \sum_{i=1}^{\ell-1} \mathbb{P}\Big(\mathbb{I}_{i-p}(\underline{u}^{(p)}\underline{u}) = 1 \text{ for some } p \in \mathcal{P}'(\underline{u})\Big) \\
&\leq (\ell-1) \sum_{p \in \mathcal{P}'(\underline{u})} \mu(\underline{u}^{(p)}\underline{u}) = (\ell-1)\big(\mu(\underline{u}) - \widetilde{\mu}(\underline{u})\big). \qquad (5)
\end{aligned}$$

As we think of $\mu(\underline{u})$ as being such that $n\mu(\underline{u})$ is bounded as n tends to infinity (the rare word condition), the above probability is very small for large n. Therefore, $\widetilde{N}(\underline{u})$ can be approximated by $\widetilde{N}_{\rm inf}(\underline{u})$; as soon as we approximate $\widetilde{N}_{\rm inf}(\underline{u})$, the approximation for $\widetilde{N}(\underline{u})$ follows.

Remark 1 *If $\underline{u}$ is not self-overlapping, meaning that $\underline{u}$ has no period, then $\widetilde{\mu}(\underline{u}) = \mu(\underline{u})$ and $\widetilde{N}(\underline{u}) = \widetilde{N}_{inf}(\underline{u}) = N(\underline{u})$.*

Now clumps of different sizes are distinguished. The size of a clump of $\underline{u}$ is the maximal number of overlapping occurrences of $\underline{u}$ contained in the clump. In our example, the unique clump is of size 3. The structure of a clump can be complex, depending on the self-overlapping structure of the underlying word. Let $\mathcal{C}_k(\underline{u})$ be the set of all the concatenated words composed of exactly k overlapping occurrences of $\underline{u}$. For $\underline{u} = \texttt{AAGAAGAA}$, we have for example

$$\begin{aligned}
\mathcal{C}_1(\underline{u}) &= \{\texttt{AAGAAGAA}\}, \\
\mathcal{C}_2(\underline{u}) &= \{\texttt{AAGAAGAAGAA}, \texttt{AAGAAGAAAGAAGAA}\} \text{ and} \\
\mathcal{C}_3(\underline{u}) &= \{\texttt{AAGAAGAAGAAGAA}, \texttt{AAGAAGAAGAAAGAAGAA}, \texttt{AAGAAGAAAGAAGAAGAA}, \texttt{AAGAAGAAAGAAGAAAGAAGAA}\}.
\end{aligned}$$

Note that the size of $\mathcal{C}_k(\underline{u})$ increases exponentially with k whenever $\underline{u}$ has more than one principal period, and the length of two k-clumps may differ a lot.
Say that a k-clump of $\underline{u}$ starts at position i in the infinite sequence $\mathcal{M}$ if and only if a clump of $\underline{u}$ starts at position i and this clump is composed of exactly k overlapping occurrences of $\underline{u}$. Denote the indicator random variable of this event by $\widetilde{\mathbb{I}}_{i,k}(\underline{u})$;

$$\widetilde{\mathbb{I}}_{i,k}(\underline{u}) := \mathbb{I}\{\text{a } k\text{-clump of } \underline{u} \text{ starts at position } i \text{ in } \mathcal{M}\}.$$

The probability $\widetilde{\mu}_k(\underline{u})$ that a k-clump of $\underline{u}$ starts at a given position in $\mathcal{M}$ is exactly the expectation of $\widetilde{\mathbb{I}}_{i,k}(\underline{u})$; Schbath (1995a) proved that

$$\widetilde{\mu}_k(\underline{u}) := \mathbb{E}\widetilde{\mathbb{I}}_{i,k}(\underline{u}) = (1 - A(\underline{u}))^2 A(\underline{u})^{k-1} \mu(\underline{u}), \qquad (6)$$

where

$$A(\underline{u}) = \sum_{p\in\mathcal{P}'(\underline{u})} \prod_{j=1}^{p} \Pi(u_j, u_{j+1}) = \sum_{p\in\mathcal{P}'(\underline{u})} \frac{\mu(\underline{u}^{(p+1)})}{\mu(u_1)}. \tag{7}$$

In order to give some insight into formula (6), we now prove it for a word $\underline{u}$ having a unique principal period p. Thus, for each $k \geq 1$, there exists a unique concatenated word C_k composed of exactly k overlapping occurrences of $\underline{u}$, and C_k is given by

$$C_k = \underbrace{\underline{u}^{(p)}\underline{u}^{(p)}\cdots\underline{u}^{(p)}}_{k-1}\underline{u}.$$

Now, a k-clump of $\underline{u}$ starts at position i if and only if an occurrence of C_k starts at position i and no occurrence of $\underline{u}^{(p)}$ starts at position $i-p$ and no occurrence of $\underline{u}_{(p)} := u_{\ell-p+1}\cdots u_{\ell-1}u_\ell$ (the p last letters of $\underline{u}$) starts at position $i+|C_k|$ where $|C_k|$ denotes the length of C_k. We can then write

$$\begin{aligned}\widetilde{\mathbb{I}}_{i,k}(\underline{u}) &= \left(1-\mathbb{I}_{i-p}(\underline{u}^{(p)})\right)\mathbb{I}_i(C_k)\left(1-\mathbb{I}_{i+|C_k|}(\underline{u}_{(p)})\right)\\ &= \mathbb{I}_i(C_k) - \mathbb{I}_{i-p}(\underline{u}^{(p)}C_k) - \mathbb{I}_i(C_k\underline{u}_{(p)}) + \mathbb{I}_{i-p}(\underline{u}^{(p)}C_k\underline{u}_{(p)}).\end{aligned}$$

Note that $\underline{u}^{(p)}C_k = C_{k+1} = C_k\underline{u}_{(p)}$ and that $\underline{u}^{(p)}C_k\underline{u}_{(p)} = C_{k+2}$. Therefore, we have

$$\widetilde{\mathbb{I}}_{i,k}(\underline{u}) = \mathbb{I}_i(C_k) - \mathbb{I}_{i-p}(C_{k+1}) - \mathbb{I}_i(C_{k+1}) + \mathbb{I}_{i-p}(C_{k+2}),$$

and taking the expectation leads to

$$\widetilde{\mu}_k(\underline{u}) = \mu(C_k) - 2\mu(C_{k+1}) + \mu(C_{k+2}).$$

Employing that $\mu(C_k) = \mu(\underline{u})\left(\pi(u_1,u_2)\cdots\pi(u_p,u_1)\right)^{k-1}$, we get

$$\widetilde{\mu}_k(\underline{u}) = \mu(\underline{u})\left(\pi(u_1,u_2)\cdots\pi(u_p,u_1)\right)^{k-1}\left(1-\pi(u_1,u_2)\cdots\pi(u_p,u_1)\right)^2.$$

Since, $u_1 = u_{p+1}$, using the definition (7) of $A(\underline{u})$ gives $\mu(\underline{u})A(\underline{u})^{k-1}(1-A(\underline{u}))^2$, resulting in exactly equation (6). When $\underline{u}$ has more than one principal period, for each k, C_k is not unique any more but the argument is similar.

We define $\widetilde{N}_k(\underline{u})$ as the count

$$\widetilde{N}_k(\underline{u}) := \sum_{i=1}^{n-\ell+1} \widetilde{\mathbb{I}}_{i,k}(\underline{u}), \tag{8}$$

so that $\widetilde{N}_k(\underline{u})$ represents the number of k-clumps of $\underline{u}$ in the infinite sequence $\mathcal{M}$ but starting in S. Again, because of the boundary effects, the count $N_{\text{inf}}(\underline{u})$ defined by

$$N_{\text{inf}}(\underline{u}) := \sum_{k\geq 1} k\widetilde{N}_k(\underline{u}) \tag{9}$$

is not equal to the count $N(\underline{u})$ of $\underline{u}$ in the finite sequence S, but their difference is negligible. Indeed, they can differ only if a clump in $\mathcal{M}$ overlaps positions 1 or n. Using the same techniques as for the number of clumps, the total variation distance between $N(\underline{u})$ and $N_{\text{inf}}(\underline{u})$ is bounded by

$$\mathbb{P}(N(\underline{u}) \neq N_{\text{inf}}(\underline{u})) \leq 2(\ell-1)\sum_{p\in\mathcal{P}'(\underline{u})} \mu(\underline{u}^{(p)}\underline{u}) = 2(\ell-1)\left(\mu(\underline{u}) - \widetilde{\mu}(\underline{u})\right). \tag{10}$$

(The counts $N(\underline{u})$ and $N_{\rm inf}(\underline{u})$ have the same expectation because equation (6) leads to $\sum_{k\geq 1} k\widetilde{\mu}_k(\underline{u}) = \mu(\underline{u})$.) Therefore, if $\underline{u}$ is a rare word, the count $N(\underline{u})$ can be approximated by $N_{\rm inf}(\underline{u})$; we will focus on $N_{\rm inf}(\underline{u})$ to apply the Chen-Stein method.

Remark 2 *If $\underline{u}$ is not self-overlapping, meaning that $\underline{u}$ has no period, then $\widetilde{\mu}_1(\underline{u}) = \mu(\underline{u})$, $\widetilde{\mu}_k(\underline{u}) = 0\ \forall k \geq 2$, and $N_{inf}(\underline{u}) = \widetilde{N}_1(\underline{u}) = N(\underline{u})$.*

The Chen-Stein method is a powerful tool for deriving Poisson approximations and compound Poisson approximations in terms of bounds on the total variation distance. For any two random processes $\underline{Y}$ and $\underline{Z}$ with values in the same space E, the total variation distance between their probability distributions is defined by

$$\begin{aligned} d_{\rm TV}(\mathcal{L}(\underline{Y}), \mathcal{L}(\underline{Z})) &= \sup_{B\subset E,\text{measurable}} |\mathbb{P}(\underline{Y}\in B) - \mathbb{P}(\underline{Z}\in B)| \\ &= \sup_{h:E\to[0,1],\text{measurable}} |\mathbb{E}h(\underline{Y}) - \mathbb{E}h(\underline{Z})|\,. \end{aligned}$$

The Chen-Stein method for Poisson approximation has been developed by Chen (1975); a friendly exposition is in Arratia *et al.* (1989) and a description with many examples can be found in Arratia *et al.* (1990) and Barbour *et al.* (1992b). We will use Theorem 1 in Arratia *et al.* (1990) with an improved bound by Barbour *et al.* (1992b) (Theorem 1.A and Theorem 10.A).

Theorem 1 (Arratia *et al.* (1990), Barbour *et al.* (1992b)) *Let I be an index set. For each $\alpha \in I$, let Y_α be a Bernoulli random variable with $p_\alpha = \mathbb{P}(Y_\alpha = 1) > 0$. Suppose that, for each $\alpha \in I$, we have chosen $B_\alpha \subset I$ with $\alpha \in B_\alpha$. Let Z_α, $\alpha \in I$, be independent Poisson variables with mean p_α. The total variation distance between the dependent Bernoulli process $\underline{Y} = (Y_\alpha, \alpha\in I)$ and the Poisson process $\underline{Z} = (Z_\alpha, \alpha \in I)$ satisfies*

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

where

$$b_1 = \sum_{\alpha\in I}\sum_{\beta\in B_\alpha} p_\alpha p_\beta \tag{11}$$

$$b_2 = \sum_{\alpha\in I}\sum_{\beta\in B_\alpha,\beta\neq\alpha} \mathbb{E}(Y_\alpha Y_\beta) \tag{12}$$

$$b_3 = \sum_{\alpha\in I} \mathbb{E}|\mathbb{E}\{Y_\alpha - p_\alpha|\sigma(Y_\beta, \beta\notin B_\alpha)\}|\,. \tag{13}$$

Moreover, if $W = \sum_{\alpha\in I} Y_\alpha$ and $\lambda = \sum_{\alpha\in I} p_\alpha < \infty$, then

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

Note that $b_3 = 0$ if Y_α is independent of $\sigma(Y_\beta, \beta \notin B_\alpha\}$. We think of B_α as a neighborhood of strong dependence of Y_α.

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

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

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

3 Occurrences of two words of same length

For simplicity, we first consider the case of only two words having the same length. The general case of multiple words with different lengths will then be a straightforward generalization and will be presented in Section 4.

Let $\underline{u} = u_1u_2\cdots u_\ell$ and $\underline{v} = v_1v_2\cdots v_\ell$ be two different words of length ℓ on $\mathcal{A}$. To describe the possible overlaps between $\underline{u}$ and $\underline{v}$, we define

$$\mathcal{P}(\underline{u},\underline{v}) \quad := \quad \{p \in \{1,\ldots,\ell-1\} : v_i = u_{i+p}, \forall i = 1,\ldots,\ell-p\}.$$

Thus $\mathcal{P}(\underline{u},\underline{v}) \neq \emptyset$ means that an occurrence of $\underline{v}$ can overlap an occurrence of $\underline{u}$ from the right, and $\mathcal{P}(\underline{v},\underline{u}) \neq \emptyset$ means that $\underline{v}$ can overlap $\underline{u}$ from the left. Note the lack of symmetry; for example, if $\underline{u} = \texttt{AAGAAGAA}$ and $\underline{v} = \texttt{AAGAATCA}$, we have $\mathcal{P}(\underline{u},\underline{v}) = \{3,6,7\}$ and $\mathcal{P}(\underline{v},\underline{u}) = \{7\}$.

This section is composed of two parts. First, we focus on the occurrences of clumps of $\underline{u}$ and on the occurrences of clumps of $\underline{v}$. We derive Poisson process approximations for the indicators of occurrences of clumps $(\widetilde{\mathbb{I}}_i(\underline{u}), \widetilde{\mathbb{I}}_i(\underline{v}))_{i=1,\ldots,(n-\ell+1)}$ (Theorem 2), and then for the pair of declumped counts $(\widetilde{N}_{\inf}(\underline{u}), \widetilde{N}_{\inf}(\underline{v}))$ (Corollary 1). A Poisson approximation is then obtained for $\widetilde{N}_{\inf}(\underline{u}) + \widetilde{N}_{\inf}(\underline{v})$ (Corollary 2). In the second part, we derive compound Poisson process approximations, first for the indicators of occurrences of k-clumps $(\widetilde{\mathbb{I}}_{i,k}(\underline{u}), \widetilde{\mathbb{I}}_{i,k}(\underline{v}))_{i=1,\ldots,(n-\ell+1),k\geq 1}$ (Theorem 3), next for the numbers of k-clumps $(\widetilde{N}_k(\underline{u}), \widetilde{N}_k(\underline{v}))_{k\geq 1}$ (Corollary 3), and then for the pair of counts $(N(\underline{u}), N(\underline{v}))$ (Corollary 4). A compound Poisson approximation is finally obtained for $N(\underline{u}) + N(\underline{v})$ (Corollary 5). As main tool we use the Stein-Chen method as stated in Theorem 1.

3.1 Poisson process approximation for the numbers of clumps

To prove a Poisson approximation for the numbers of clumps, we will apply Theorem 1. To this purpose we define the Bernoulli and the Poisson processes. We form the random vector $\underline{\widetilde{Y}} = (\widetilde{Y}_i)_{i\in I}$, with index set $I = \{1, 2, \ldots, 2(n-\ell+1)\}$, as

$$\widetilde{Y}_i = \begin{cases} \widetilde{\mathbb{I}}_i(\underline{u}) & \text{if } 1 \leq i \leq n-\ell+1\,, \\ \widetilde{\mathbb{I}}_{[i]}(\underline{v}) & \text{if } n-\ell+1 < i \leq 2(n-\ell+1)\,, \end{cases}$$

with the notation

$$[i] = \begin{cases} i & \text{if } 1 \leq i \leq n-\ell+1\,, \\ i-n+\ell-1 & \text{if } n-\ell+1 < i \leq 2(n-\ell+1)\,. \end{cases} \tag{14}$$

Thus $\underline{\widetilde{Y}}$ is the concatenated vector of the occurrence indicators for clumps of $\underline{u}$ and clumps of $\underline{v}$. Note that we do not consider "mixed" clumps of $\underline{u}$ and $\underline{v}$; each clump consists only in occurrences of $\underline{u}$, or only in occurrences of $\underline{v}$. However, a clump of $\underline{u}$ may overlap a clump of $\underline{v}$ in the sequence. For example, if $S = \texttt{TGAAGAAGAAAGAAGAATCAAGAATCACTC}$, $\underline{u} = \texttt{AAGAAGAA}$ and $\underline{v} = \texttt{AAGAATCA}$, we count in S one clump of $\underline{u}$ starting at position 3 and ending at position 17, and we count one clump of $\underline{v}$ starting at position 13 and ending at position 27.

Our goal is to approximate $\underline{\widetilde{Y}}$ by a vector $\underline{Z}$ with independent Poisson coordinates, and to bound the error in this approximation in terms of the total variation distance. Let $\underline{Z} = (Z_i)_{i\in I}$ be a random vector of independent Poisson variables such that

$$Z_i \sim \begin{cases} \mathrm{Po}(\widetilde{\mu}(\underline{u})) & \text{if } 1 \leq i \leq n-\ell+1\,, \\ \mathrm{Po}(\widetilde{\mu}(\underline{v})) & \text{if } n-\ell+1 < i \leq 2(n-\ell+1), \end{cases} \tag{15}$$

where $\widetilde{\mu}(\cdot)$ is defined in equation (3).

To apply Theorem 1, we choose the following neighborhood of $i \in I$:

$$B_i := \{j \in I : |[j]-[i]| \leq 3\ell - 3\},$$

where the notation $[j]$ is defined in (14). Note that, when i is just a little smaller than $n-\ell+1$, and when j is just a little larger than $n-\ell+1$, so that $j-i \leq 3\ell-3$, the $[\cdot]$ notation used in the definition of B_i ensures that j is not in the neighborhood of i.

Deriving a bound for the total variation distance between $\underline{\widetilde{Y}}$ and $\underline{Z}$ consists now in bounding the quantities b_1, b_2 and b_3 given in (11), (12) and (13). To formulate the result, we define the following quantities; for any two words $\underline{w}_1 = w_{1,1}\cdots w_{1,\ell}$ and $\underline{w}_2 = w_{2,1}\cdots w_{2,\ell}$ of length ℓ, put

$$\Omega(x,y) = \sum_{s=1}^{2\ell-2} \Pi^s(x,y)\,,\ x,y \in \mathcal{A}, \tag{16}$$

$$M(\underline{w}_1,\underline{w}_2) = \begin{cases} \displaystyle\sum_{p \in \mathcal{P}(\underline{w}_1,\underline{w}_2)} \frac{1}{\mu(\underline{w}_2^{(\ell-p)})} & \text{if } \underline{w}_1 \neq \underline{w}_2, \\ 0 & \text{if } \underline{w}_1 = \underline{w}_2 \end{cases} \tag{17}$$

$$T_1(\underline{w}_1,\underline{w}_2) = 2(n-\ell+1)\mu(\underline{w}_1)\widetilde{\mu}(\underline{w}_2)\left(\frac{\Omega(w_{2,\ell},w_{1,1})}{\mu(w_{1,1})} + M(\underline{w}_2,\underline{w}_1)\right). \tag{18}$$

The quantities $M(\underline{u},\underline{v})$ and $M(\underline{v},\underline{u})$ can be seen as measures of the overlapping structure between $\underline{u}$ and $\underline{v}$. If $\underline{u}$ and $\underline{v}$ cannot overlap, these quantities are equal to zero; otherwise, the more they can overlap, the larger are $M(\underline{u},\underline{v})$ and $M(\underline{v},\underline{u})$.

Moreover, we need some quantities related to the transition matrix. Let $(\alpha_t)_{t=1,\ldots,|\mathcal{A}|}$ be the eigenvalues of Π such that $|\alpha_1| \geq |\alpha_2| \geq \cdots \geq |\alpha_{|\mathcal{A}|}|$. The Perron-Frobenius Theorem (see, e.g., Karlin and Taylor (1975)) ensures that $\alpha_1 = 1$ and $|\alpha_2| < 1$; we abbreviate α_2 by α. As in Schbath (1995b), we diagonalize the transition matrix $\Pi = PDP^{-1}$ with $D = \mathrm{Diag}(1,\alpha,\alpha_3,\cdots,\alpha_{|\mathcal{A}|})$. For all $t \in \{1,\ldots,|\mathcal{A}|\}$, I_t denotes the $|\mathcal{A}|\times|\mathcal{A}|$ matrix such that all its entries are equal to 0 except $I_t(t,t)=1$, and we define $Q_t := PI_tP^{-1}$. Thus,

$$\Pi = \sum_{t=1}^{|\mathcal{A}|} \alpha_t Q_t. \tag{19}$$

Now, we define $\gamma_1(\ell,a)$ for any $a \in \mathcal{A}$ by

$$\gamma_1(\ell,a) = \sum_{x,y\in\mathcal{A}} \mu(x) \max_{b\in\mathcal{A}} \left| \frac{1}{\mu(b)} \sum_{(t,t')\neq(1,1)} \frac{\alpha_t^\ell \alpha_{t'}^\ell}{\alpha^\ell} Q_t(x,b)Q_{t'}(a,y) - \sum_{t=2}^{|\mathcal{A}|} \frac{\alpha_t^{4\ell-2}}{\alpha^\ell} Q_t(x,y) \right|. \tag{20}$$

Note that $\gamma_1(\ell,a)$ depends on ℓ and is bounded by a constant.

Theorem 2 *With the previous notation, we have*

$$\begin{aligned} d_{TV}\left(\mathcal{L}(\underline{\widetilde{Y}}),\mathcal{L}(\underline{Z})\right) \leq\ & (n-\ell+1)(6\ell-5)\left(\widetilde{\mu}(\underline{u})+\widetilde{\mu}(\underline{v})\right)^2 \\ & + T_1(\underline{u},\underline{u}) + T_1(\underline{u},\underline{v}) + T_1(\underline{v},\underline{u}) + T_1(\underline{v},\underline{v}) \\ & + (n-\ell+1)|\alpha|^\ell\Big(\gamma_1(\ell,u_\ell)\widetilde{\mu}(\underline{u}) + \gamma_1(\ell,v_\ell)\widetilde{\mu}(\underline{v})\Big), \end{aligned}$$

where T_1 and γ_1 are given by (18) and (20). If the X_i's are independent, then the bound simplifies to

$$\begin{aligned}
d_{TV}\left(\mathcal{L}(\underline{\widetilde{Y}}), \mathcal{L}(\underline{Z})\right) \quad &\leq \quad (n-\ell+1)(6\ell-5)\left(\widetilde{\mu}(\underline{u})+\widetilde{\mu}(\underline{v})\right)^2 + 4(n-\ell+1)(\ell-1)(\mu(\underline{u})+\mu(\underline{v}))(\widetilde{\mu}(\underline{u})+\widetilde{\mu}(\underline{v})) \\
&\quad +2(n-\ell+1)\left(\mu(\underline{u})\widetilde{\mu}(\underline{v})M(\underline{v},\underline{u})+\mu(\underline{v})\widetilde{\mu}(\underline{u})M(\underline{u},\underline{v})\right),
\end{aligned}$$

where M is given by (17).

Poisson approximations should be good for rare events. Here, rare words mean that $\mathbb{E}N(\underline{u})$ and $\mathbb{E}N(\underline{v})$ are bounded away from 0 and ∞; we use the notation $\mathbb{E}N(\underline{u}) \asymp 1$ and $\mathbb{E}N(\underline{v}) \asymp 1$. For a fixed alphabet, this asymptotic framework is equivalent to $\ell \asymp \log n$ because $\mu(u_1)(\min_{x,y\in\mathcal{A}}\Pi(x,y))^{\ell-1} \leq \mu(\underline{u}) \leq \mu(u_1)(\max_{x,y\in\mathcal{A}}\Pi(x,y))^{\ell-1}$. If $\mathbb{E}N(\underline{u}) \asymp 1$ and $\mathbb{E}N(\underline{v}) \asymp 1$, then $T_1(\underline{u},\underline{v}) \asymp n^{-1}\log n + M(\underline{u},\underline{v})n^{-1}$. Since γ_1 is bounded by a constant, we obtain

$$d_{TV}\left(\mathcal{L}(\underline{\widetilde{Y}}), \mathcal{L}(\underline{Z})\right) = \mathrm{O}\left(n^{-1}\log n\right) + \mathrm{O}\left((M(\underline{u},\underline{v})+M(\underline{v},\underline{u}))n^{-1}\right) + \mathrm{O}\left(|\alpha|^{\log n}\right).$$

As noticed above, $M(\underline{u},\underline{v})$ and $M(\underline{v},\underline{u})$ are quantities that measure the importance of the overlapping structure between $\underline{u}$ and $\underline{v}$. If $\underline{u}$ and $\underline{v}$ cannot overlap too much, meaning that $M(\underline{u},\underline{v})$ and $M(\underline{v},\underline{u})$ are small, then the approximation will be good. In the extreme case where $\underline{u} = \mathtt{AAA}\cdots\mathtt{AAA}$ and $\underline{v} = \mathtt{TAA}\cdots\mathtt{AAA}$, if the letters are independent, we have $M(\underline{v},\underline{u}) = \sum_{s=1}^{\ell-1}\mu(\mathtt{A})^{-s} \asymp \mu(\mathtt{A})^{-\log n}$, so the error bound may fail to converge to zero as n tends to infinity. This confirms the intuition that, because of considerable overlaps, the occurrences of $\underline{u}$ and $\underline{v}$ will not be independent even asymptotically.

Proof: We will successively bound the quantities b_1, b_2 and b_3 given in (11), (12) and (13).

Bounding $\mathbf{b_1}$: We can decompose $b_1 = \sum_{i\in I}\sum_{j\in B_i}\mathbb{E}\widetilde{Y}_i\mathbb{E}\widetilde{Y}_j$ as $b_1 = b_1'(\underline{u}) + b_1'(\underline{v}) + b_1'(\underline{u},\underline{v}) + b_1'(\underline{v},\underline{u})$ with

$$\begin{aligned}
b_1'(\underline{u}) \quad &= \quad \sum_{i=1}^{n-\ell+1}\sum_{j\in B_i\cap\{1,\dots,n-\ell+1\}}\mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u})\mathbb{E}\widetilde{\mathbb{I}}_j(\underline{u}) \\
b_1'(\underline{u},\underline{v}) \quad &= \quad \sum_{i=1}^{n-\ell+1}\sum_{j\in B_i\cap\{n-\ell+2,\dots,2(n-\ell+1)\}}\mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u})\mathbb{E}\widetilde{\mathbb{I}}_{[j]}(\underline{v}).
\end{aligned}$$

As $\mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u}) = \widetilde{\mu}(\underline{u})$ from (3), we have

$$\begin{aligned}
b_1'(\underline{u}) \quad &= \quad \sum_{i=1}^{n-\ell+1}\sum_{j\in B_i\cap\{1,\dots,n-\ell+1\}}\widetilde{\mu}(\underline{u})^2 \\
&\leq \quad (n-\ell+1)(6\ell-5)\widetilde{\mu}(\underline{u})^2,
\end{aligned}$$

where the inequality results from the neighborhood B_i having less or equal to $6\ell-5$ elements. Similarly,

$$b_1'(\underline{u},\underline{v}) \quad \leq \quad (n-\ell+1)(6\ell-5)\widetilde{\mu}(\underline{u})\widetilde{\mu}(\underline{v}),$$

and thus

$$b_1 \leq (n-\ell+1)(6\ell-5)\left(\widetilde{\mu}(\underline{u})+\widetilde{\mu}(\underline{v})\right)^2. \tag{21}$$

Bounding $\mathbf{b_2}$: We write b_2 as

$$b_2 = \sum_{i \in I} \sum_{j \in B_i \setminus \{i\}} \mathbb{E}\widetilde{Y}_i \widetilde{Y}_j = b_2'(\underline{u}) + b_2'(\underline{v}) + b_2'(\underline{u}, \underline{v}) + b_2'(\underline{v}, \underline{u})$$

with

$$\begin{aligned} b_2'(\underline{u}) &= \sum_{i=1}^{n-\ell+1} \sum_{j \in B_i \cap \{1, \dots, n-\ell+1\}, j \neq i} \mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u})\widetilde{\mathbb{I}}_j(\underline{u}), \\ b_2'(\underline{u}, \underline{v}) &= \sum_{i=1}^{n-\ell+1} \sum_{j \in B_i \cap \{n-\ell+2, \dots, 2(n-\ell+1)\}} \mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u})\widetilde{\mathbb{I}}_{[j]}(\underline{v}). \end{aligned}$$

For $b_2'(\underline{u})$, note that, due to the declumping, we have $\mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u})\widetilde{\mathbb{I}}_j(\underline{u}) = 0$ if $i \neq j$ and $|i - j| < \ell$. Moreover, for all $s \geq \ell$ we have

$$\begin{aligned} \mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u})\widetilde{\mathbb{I}}_{i+s}(\underline{u}) &\leq \mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u})\mathbb{I}_{i+s}(\underline{u}) \\ &= \widetilde{\mu}(\underline{u})\Pi^{i+s-(i+\ell-1)}(u_\ell, u_1)\Pi(u_1, u_2)\dots\Pi(u_{\ell-1}, u_\ell) \\ &= \widetilde{\mu}(\underline{u})\Pi^{s-\ell+1}(u_\ell, u_1)\frac{\mu(\underline{u})}{\mu(u_1)}, \qquad (22) \\ \mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u})\widetilde{\mathbb{I}}_{i-s}(\underline{u}) &\leq \widetilde{\mu}(\underline{u})\Pi^{s-\ell+1}(u_\ell, u_1)\frac{\mu(\underline{u})}{\mu(u_1)}. \end{aligned}$$

Thus,

$$\begin{aligned} b_2'(\underline{u}) &= \sum_{i=1}^{n-\ell+1} \sum_{\substack{j \in \{1, \dots, n-\ell+1\} \\ \ell \leq |j-i| \leq 3\ell-3}} \mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u})\widetilde{\mathbb{I}}_j(\underline{u}) \\ &\leq \sum_{i=1}^{n-\ell+1} \sum_{s=\ell}^{3\ell-3} \mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u})\widetilde{\mathbb{I}}_{i+s}(\underline{u}) + \sum_{i=1}^{n-\ell+1} \sum_{s=\ell}^{3\ell-3} \mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u})\widetilde{\mathbb{I}}_{i-s}(\underline{u}) \\ &\leq 2(n-\ell+1)\widetilde{\mu}(\underline{u})\frac{\mu(\underline{u})}{\mu(u_1)} \sum_{s=1}^{2\ell-2} \Pi^s(u_\ell, u_1) \\ &= 2(n-\ell+1)\frac{\mu(\underline{u})\widetilde{\mu}(\underline{u})}{\mu(u_1)}\Omega(u_\ell, u_1). \qquad (23) \end{aligned}$$

For $b_2'(\underline{u}, \underline{v})$ we use similar arguments. First note that, if $p \notin \mathcal{P}(\underline{u}, \underline{v})$ and $1 \leq p \leq \ell - 1$ then $\mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u})\widetilde{\mathbb{I}}_{i+p}(\underline{v}) = 0$. Furthermore, if $p \in \mathcal{P}(\underline{u}, \underline{v})$, then an occurrence of $\underline{v}$ at position $i + p$ overlaps an occurrence of $\underline{u}$ at position i and we have

$$\begin{aligned} \mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u})\widetilde{\mathbb{I}}_{i+p}(\underline{v}) &\leq \mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u})\mathbb{I}_{i+p}(\underline{v}) \\ &= \widetilde{\mu}(\underline{u})\Pi(u_\ell, v_{\ell-p+1})\Pi(v_{\ell-p+1}, v_{\ell-p+2})\dots\Pi(v_{\ell-1}, v_\ell) \\ &= \widetilde{\mu}(\underline{u})\mu(\underline{v})\frac{1}{\mu(\underline{v}^{(\ell-p)})}. \end{aligned}$$

Lastly, if $p \geq \ell$ then we can argue as for (22), obtaining

$$\mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u})\widetilde{\mathbb{I}}_{i+p}(\underline{v}) \leq \widetilde{\mu}(\underline{u})\Pi^{p-\ell+1}(u_\ell, v_1)\frac{\mu(\underline{v})}{\mu(v_1)}.$$

Summing separately over $i < [j] < i+\ell$ and $[j] \geq i+\ell$, and using symmetric arguments when $[j] < i$, yields

$$\begin{aligned} b'_2(\underline{u},\underline{v}) &\leq (n-\ell+1)\left(\mu(\underline{u})\widetilde{\mu}(\underline{v})\left(\frac{\Omega(v_\ell,u_1)}{\mu(u_1)} + \sum_{p\in\mathcal{P}(\underline{v},\underline{u})}\frac{1}{\mu(\underline{u}^{(\ell-p)})}\right)\right. \\ &\qquad \left. +\mu(\underline{v})\widetilde{\mu}(\underline{u})\left(\frac{\Omega(u_\ell,v_1)}{\mu(v_1)} + \sum_{p\in\mathcal{P}(\underline{u},\underline{v})}\frac{1}{\mu(\underline{v}^{(\ell-p)})}\right)\right) \\ &= (n-\ell+1)\left(\mu(\underline{u})\widetilde{\mu}(\underline{v})\left(\frac{\Omega(v_\ell,u_1)}{\mu(u_1)} + M(\underline{v},\underline{u})\right) + \mu(\underline{v})\widetilde{\mu}(\underline{u})\left(\frac{\Omega(u_\ell,v_1)}{\mu(v_1)} + M(\underline{u},\underline{v})\right)\right), \end{aligned}$$

where we use notation (16) and (17). In view of (18), we proved that $b'_2(\underline{u}) \leq T_1(\underline{u},\underline{u})$ and $b'_2(\underline{u},\underline{v}) + b'_2(\underline{v},\underline{u}) \leq T_1(\underline{u},\underline{v}) + T_1(\underline{v},\underline{u})$ establishing that

$$b_2 \leq T_1(\underline{u},\underline{u}) + T_1(\underline{u},\underline{v}) + T_1(\underline{v},\underline{u}) + T_1(\underline{v},\underline{v}). \tag{24}$$

Bounding $\mathbf{b_3}$: We write

$$b_3 = \sum_{i\in I}\mathbb{E}|\mathbb{E}(\widetilde{Y}_i - \mathbb{E}\widetilde{Y}_i \mid \widetilde{\mathcal{Y}}_i)| = b'_3(\underline{u}) + b'_3(\underline{v}),$$

where $\widetilde{\mathcal{Y}}_i = \sigma(\widetilde{Y}_j, j \notin B_i)$ and

$$b'_3(\underline{u}) = \sum_{i=1}^{n-\ell+1}\mathbb{E}|\mathbb{E}(\widetilde{\mathbb{I}}_i(\underline{u}) - \mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u}) \mid \widetilde{\mathcal{Y}}_i)|.$$

Since $\widetilde{\mathcal{Y}}_i \subset \mathcal{X}_{[i]} = \sigma(X_1,\ldots,X_{[i]-2\ell+1},X_{[i]+2\ell-1},\ldots,X_n)$, properties on conditional expectation and the Markov property give

$$b'_3(\underline{u}) \leq \sum_{i=1}^{n-\ell+1}\mathbb{E}|\mathbb{E}(\widetilde{\mathbb{I}}_i(\underline{u}) - \mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u}) \mid X_{i-2\ell+1}, X_{i+2\ell-1})|.$$

To evaluate the right-hand term, we introduce the set of possible words of length $\ell-1$ preceding a clump of $\underline{u}$:

$$\mathcal{G}(\underline{u}) = \{\underline{g} = g_1\cdots g_{\ell-1} : \text{ for all } p\in\mathcal{P}(\underline{u}), g_{\ell-p}\cdots g_{\ell-1} \neq \underline{u}^{(p)}\}. \tag{25}$$

Thus a clump of $\underline{u}$ starts at position i in $\mathcal{M}$ if and only if one of the words $\underline{gu}$, $\underline{g}\in\mathcal{G}(\underline{u})$, starts at position $i-\ell+1$. Therefore, we can write

$$\widetilde{\mathbb{I}}_i(\underline{u}) = \sum_{\underline{g}\in\mathcal{G}(\underline{u})}\mathbb{I}_{i-\ell+1}(\underline{gu}). \tag{26}$$

This gives us

$$
\begin{aligned}
&\mathbb{E}|\mathbb{E}(\widetilde{\mathbb{I}}_i(\underline{u}) - \widetilde{\mathbb{EI}}_i(\underline{u}) \mid X_{i-2\ell+1}, X_{i+2\ell-1})| \\
&\quad = \mathbb{E}| \sum_{\underline{g}\in\mathcal{G}(\underline{u})} \mathbb{E}(\mathbb{I}_{i-\ell+1}(\underline{gu}) - \mathbb{E}(\mathbb{I}_{i-\ell+1}(\underline{gu})) \mid X_{i-2\ell+1}, X_{i+2\ell-1})| \\
&\quad = \sum_{x,y\in\mathcal{A}} \left| \sum_{\underline{g}\in\mathcal{G}(\underline{u})} \mathbb{E}(\mathbb{I}_{i-\ell+1}(\underline{gu}) - \mathbb{E}(\mathbb{I}_{i-\ell+1}(\underline{gu})) \mid X_{i-2\ell+1} = x, X_{i+2\ell-1} = y) \right| \mathbb{P}(X_{i-2\ell+1} = x, X_{i+2\ell-1} = y) \\
&\quad = \sum_{x,y\in\mathcal{A}} \left| \sum_{\underline{g}\in\mathcal{G}(\underline{u})} \mathbb{P}(\mathbb{I}_{i-\ell+1}(\underline{gu}) = 1,\, X_{i-2\ell+1} = x, X_{i+2\ell-1} = y) - \mu(\underline{gu})\mathbb{P}(X_{i-2\ell+1} = x, X_{i+2\ell-1} = y) \right| \\
&\quad = \sum_{x,y\in\mathcal{A}} \left| \sum_{\underline{g}\in\mathcal{G}(\underline{u})} \left\{ \mu(x)\Pi^{\ell}(x, g_1)\frac{\mu(\underline{gu})}{\mu(g_1)}\Pi^{\ell}(u_\ell, y) - \mu(\underline{gu})\mu(x)\Pi^{4\ell-2}(x, y) \right\} \right| .
\end{aligned}
$$

Now we use the decomposition (19), $\Pi = \sum_{t=1}^{|\mathcal{A}|} \alpha_t Q_t$, where $(\alpha_t)_{t=1,\dots,|\mathcal{A}|}$ are the eigenvalues of Π such that $|\alpha_1| \geq |\alpha_2| \geq \cdots \geq |\alpha_{|\mathcal{A}|}|$. Recall that $\alpha_1 = 1$ and $|\alpha_2| < 1$; we abbreviate α_2 by α. Since $(1, 1, \dots, 1)^T$ is a right-eigenvector of Π for the eigenvalue 1, we diagonalize $\Pi = PDP^{-1}$ with $D = \mathrm{Diag}(1, \alpha, \alpha_3, \cdots, \alpha_{|\mathcal{A}|})$ such that the first column of P is $(1, 1, \dots, 1)^T$; then the first row of P^{-1} is the vector of the stationary distribution $\mu = (\mu(a), a \in \mathcal{A})$. As $Q_t := PI_tP^{-1}$, we thus have $Q_1(a, b) = \mu(b), \forall a, b \in \mathcal{A}$. We will use that $\Pi^\ell = PD^\ell P^{-1}$ and $D^\ell = \sum_{t=1}^{|\mathcal{A}|} \alpha_t^\ell I_t$ so that $\Pi^\ell = \sum_{t=1}^{|\mathcal{A}|} \alpha_t^\ell Q_t$. Thus,

$$
\begin{aligned}
&\mathbb{E}|\mathbb{E}(\widetilde{\mathbb{I}}_i(\underline{u}) - \widetilde{\mathbb{EI}}_i(\underline{u}) \mid X_{i-2\ell+1}, X_{i+2\ell-1})| \\
&\quad \leq |\alpha|^\ell \sum_{x,y\in\mathcal{A}} \mu(x) \left| \sum_{\underline{g}\in\mathcal{G}(\underline{u})} \mu(\underline{gu}) \left\{ \frac{1}{\mu(g_1)} \sum_{(t,t')} \frac{\alpha_t^\ell \alpha_{t'}^\ell}{\alpha^\ell} Q_t(x, g_1)Q_{t'}(u_\ell, y) - \sum_{t=1}^{|\mathcal{A}|} \frac{\alpha_t^{4\ell-2}}{\alpha^\ell} Q_t(x, y) \right\} \right| \\
&\quad = |\alpha|^\ell \sum_{x,y\in\mathcal{A}} \mu(x) \left| \sum_{\underline{g}\in\mathcal{G}(\underline{u})} \mu(\underline{gu}) \left\{ \frac{1}{\mu(g_1)} \sum_{(t,t')\neq(1,1)} \frac{\alpha_t^\ell \alpha_{t'}^\ell}{\alpha^\ell} Q_t(x, g_1)Q_{t'}(u_\ell, y) - \sum_{t=2}^{|\mathcal{A}|} \frac{\alpha_t^{4\ell-2}}{\alpha^\ell} Q_t(x, y) \right\} \right| \\
&\quad \leq |\alpha|^\ell \gamma_1(\ell, u_\ell) \sum_{\underline{g}\in\mathcal{G}(\underline{u})} \mu(\underline{gu}) ,
\end{aligned}
$$

where $\gamma_1(\ell, u_\ell)$ is defined in (20). From (26) we have

$$
\sum_{\underline{g}\in\mathcal{G}(\underline{u})} \mu(\underline{gu}) = \widetilde{\mu}(\underline{u})
$$

and it follows that

$$
b_3'(\underline{u}) \leq \gamma_1(\ell, u_\ell)(n - \ell + 1)\widetilde{\mu}(\underline{u})|\alpha|^\ell
$$

and

$$
b_3 \leq (n - \ell + 1)|\alpha|^\ell \Big(\gamma_1(\ell, u_\ell)\widetilde{\mu}(\underline{u}) + \gamma_1(\ell, v_\ell)\widetilde{\mu}(\underline{v}) \Big) . \tag{27}
$$

Combining the bounds (21), (24), (27) for b_1, b_2 and b_3 establishes the first part of the theorem.

For the second part of the theorem, if the X_i's are independent, then we have $b_3 = 0$, $\Pi^s = \Pi$, $\Omega(w_{2,\ell}, w_{1,1}) = 2(\ell-1)\mu(w_{1,1})$ and

$$T_1(\underline{w}_1, \underline{w}_2) = 2(n-\ell+1)\mu(\underline{w}_1)\widetilde{\mu}(\underline{w}_2)\left(2(\ell-1) + M(\underline{w}_2, \underline{w}_1)\right) .$$

■

Remark 3 *The geometric convergence of b_3 can be shown more elegantly using the β-mixing property of a Markov chain, as in Schbath (1995a), but that method does not give an explicit bound.*

From Theorem 2, equation (4) and inequality (5), we immediately obtain a multidimensional Poisson approximation for the joint distribution for the number of clumps of $\underline{u}$ and the number of clumps of $\underline{v}$.

Corollary 1 *Let $Z^{(1)}$ and $Z^{(2)}$ be independent Poisson variables with $\mathbb{E}Z^{(1)} = (n-\ell+1)\widetilde{\mu}(\underline{u})$ and $\mathbb{E}Z^{(2)} = (n-\ell+1)\widetilde{\mu}(\underline{v})$. With the previous notation we have*

$$\begin{aligned}
(i) \quad d_{TV}\left(\mathcal{L}\big(\widetilde{N}_{inf}(\underline{u}), \widetilde{N}_{inf}(\underline{v})\big), \mathcal{L}(Z^{(1)}, Z^{(2)})\right) \quad &\leq \quad (n-\ell+1)(6\ell-5)\big(\widetilde{\mu}(\underline{u}) + \widetilde{\mu}(\underline{v})\big)^2 \\
&\quad + T_1(\underline{u}, \underline{u}) + T_1(\underline{u}, \underline{v}) + T_1(\underline{v}, \underline{u}) + T_1(\underline{v}, \underline{v}) \\
&\quad + (n-\ell+1)|\alpha|^\ell \Big(\gamma_1(\ell, u_\ell)\widetilde{\mu}(\underline{u}) + \gamma_1(\ell, v_\ell)\widetilde{\mu}(\underline{v})\Big), \\
(ii) \quad d_{TV}\left(\mathcal{L}\big(\widetilde{N}(\underline{u}), \widetilde{N}(\underline{v})\big), \mathcal{L}(Z^{(1)}, Z^{(2)})\right) \quad &\leq \quad d_{TV}\left(\mathcal{L}\big(\widetilde{N}_{inf}(\underline{u}), \widetilde{N}_{inf}(\underline{v})\big), \mathcal{L}(Z^{(1)}, Z^{(2)})\right) \\
&\quad + (\ell-1)\big(\mu(\underline{u}) - \widetilde{\mu}(\underline{u}) + \widetilde{\mu}(\underline{v}) - \mu(\underline{v})\big) .
\end{aligned}$$

PROOF: Let f be the function such that $f(\underline{\widetilde{Y}}) = \big(\widetilde{N}_{\rm inf}(\underline{u}), \widetilde{N}_{\rm inf}(\underline{v})\big) = \Big(\sum_{i=1}^{n-\ell+1} \widetilde{Y}_i(\underline{u}), \sum_{i=n-\ell+2}^{2(n-\ell+1)} \widetilde{Y}_i(\underline{v})\Big)$. Then, with $\underline{Z}$ given by (15), $f(\underline{Z}) = \Big(\sum_{i=1}^{n-\ell+1} Z_i, \sum_{i=n-\ell+2}^{2(n-\ell+1)} Z_i\Big)$ has the same distribution as $\big(Z^{(1)}, Z^{(2)}\big)$ where $Z^{(1)}$ and $Z^{(2)}$ are two independent Poisson variables with expectation $(n-\ell+1)\widetilde{\mu}(\underline{u})$ and $(n-\ell+1)\widetilde{\mu}(\underline{v})$. By definition,

$$\begin{aligned}
d_{\rm TV}(\mathcal{L}(f(\underline{\widetilde{Y}})), \mathcal{L}(f(\underline{Z}))) \quad &= \quad \sup_{h:\mathbb{N}^2 \to [0,1],\ \text{measurable}} |\mathbb{E}h(f(\underline{\widetilde{Y}})) - \mathbb{E}h(f(\underline{Z}))| \\
&\leq \quad d_{\rm TV}(\mathcal{L}(\underline{\widetilde{Y}}), \mathcal{L}(\underline{Z})) .
\end{aligned}$$

Statement (i) then holds from Theorem 2. The triangle inequality and inequality (5) lead to statement (ii). ■

Theorem 1 also allows us to derive a Poisson approximation for $\widetilde{N}_{\rm inf}(\underline{u}, \underline{v}) := \widetilde{N}_{\rm inf}(\underline{u}) + \widetilde{N}_{\rm inf}(\underline{v})$ the total number of clumps of $\underline{u}$ and $\underline{v}$. Let λ be its expectation, $\lambda := \mathbb{E}\widetilde{N}_{\rm inf}(\underline{u}, \underline{v}) = (n-\ell+1)(\widetilde{\mu}(\underline{u}) + \widetilde{\mu}(\underline{v}))$. Combining the bounds (21), (24), (27) for b_1, b_2 and b_3 and using inequality (5) immediately gives the following corollary.

Corollary 2 *With the previous notation,*

$$(i)\quad d_{TV}\left(\mathcal{L}(\widetilde{N}_{inf}(\underline{u},\underline{v})),\mathrm{Po}(\lambda)\right) \le \frac{1-e^{-\lambda}}{\lambda}\Big\{(n-\ell+1)(6\ell-5)\left(\widetilde{\mu}(\underline{u})+\widetilde{\mu}(\underline{v})\right)^2 + T_1(\underline{u},\underline{u})+T_1(\underline{u},\underline{v})+T_1(\underline{v},\underline{u})+T_1(\underline{v},\underline{v})\Big\} + \min\left(1,\sqrt{\frac{2}{\lambda e}}\right)(n-\ell+1)|\alpha|^\ell\Big(\gamma_1(\ell,u_\ell)\widetilde{\mu}(\underline{u})+\gamma_1(\ell,v_\ell)\widetilde{\mu}(\underline{v})\Big).$$

If the X_i's are independent, then the bound simplifies to

$$d_{TV}\left(\mathcal{L}(\widetilde{N}_{inf}(\underline{u},\underline{v})),\mathrm{Po}(\lambda)\right) \le \frac{1-e^{-\lambda}}{\lambda}(n-\ell+1)\Big\{(6\ell-5)\left(\widetilde{\mu}(\underline{u})+\widetilde{\mu}(\underline{v})\right)^2 +4(\ell-1)(\mu(\underline{u})+\mu(\underline{v}))(\widetilde{\mu}(\underline{u})+\widetilde{\mu}(\underline{v})) +2\left(\mu(\underline{u})\widetilde{\mu}(\underline{v})M(\underline{v},\underline{u})+\mu(\underline{v})\widetilde{\mu}(\underline{u})M(\underline{u},\underline{v})\right)\Big\}.$$

$$(ii)\quad d_{TV}\left(\mathcal{L}(\widetilde{N}(\underline{u})+\widetilde{N}(\underline{v})),\mathrm{Po}(\lambda)\right) \le d_{TV}\left(\mathcal{L}(\widetilde{N}_{inf}(\underline{u},\underline{v})),\mathrm{Po}(\lambda)\right)+(\ell-1)\left(\mu(\underline{u})-\widetilde{\mu}(\underline{u})+\mu(\underline{v})-\widetilde{\mu}(\underline{v})\right).$$

Remark 4 *If both $\underline{u}$ and $\underline{v}$ are non self-overlapping words, the previous corollaries lead to Poisson approximations of the counts of $\underline{u}$ and $\underline{v}$ because of Remark 1.*

3.2 Compound Poisson approximation for the numbers of occurrences

The process $(\mathbb{I}_i(\underline{u}))_{i\ge 1}$ of occurrences of a word $\underline{u}$ cannot be approximated by a Poisson process using Theorem 1 due to the self-overlaps of $\underline{u}$, which may cause clumps to appear; indeed, b_2 will be large. So we cannot obtain a Poisson approximation for the count if we apply Theorem 1 directly to decomposition (1). However, using the decomposition (9) leads to a compound Poisson approximation. Again we start by defining the Bernoulli and the Poisson processes of interest. We define the random vector $\underline{\widetilde{\mathbb{Y}}}=(\widetilde{Y}_{i,k})_{(i,k)\in I}$, with index set $I=\{1,2,\dots,2(n-\ell+1)\}\times\{1,2,\dots\}$, by

$$\widetilde{Y}_{i,k}=\begin{cases}\widetilde{\mathbb{I}}_{i,k}(\underline{u}) & \text{if } 1\le i\le n-\ell+1\,,\\ \widetilde{\mathbb{I}}_{[i],k}(\underline{v}) & \text{if } n-\ell+1<i\le 2(n-\ell+1),\end{cases}$$

where $[i]$ is given by (14). Recall that $\widetilde{\mathbb{I}}_{i,k}(\underline{u})$ is the indicator that a $k-$clump of $\underline{u}$ starts at position i in the infinite sequence $\mathcal{M}$.

Our goal is to approximate $\underline{\widetilde{\mathbb{Y}}}$ by a vector $\underline{\mathbb{Z}}$ with independent Poisson coordinates and to bound the total variation distance between the two. Since the number of k-clumps $\widetilde{N}_k(\underline{u})$ is defined as the sum of indicators $\widetilde{\mathbb{I}}_{i,k}(\underline{u})$ (see formula (8)), the approximation of the vector $(\widetilde{N}_k(\underline{u}),\widetilde{N}_k(\underline{v}))_{k\ge 1}$ by a vector of independent Poisson variables will follow immediately. Moreover, since $N_{\mathrm{inf}}(\underline{u})=\sum_{k\ge 1}k\widetilde{N}_k(\underline{u})$, the approximation of $(N_{\mathrm{inf}}(\underline{u}),N_{\mathrm{inf}}(\underline{v}))$ by a pair of independent compound Poisson variables will hold. Finally, $(N(\underline{u}),N(\underline{v}))$ can be approximated by $(N_{\mathrm{inf}}(\underline{u}),N_{\mathrm{inf}}(\underline{v}))$ using (10).

Let $\underline{\mathbb{Z}}=(Z_{i,k})_{(i,k)\in I}$ be a random vector of independent Poisson variables such that

$$Z_{i,k}\sim\begin{cases}\mathrm{Po}(\widetilde{\mu}_k(\underline{u})) & \text{if } 1\le i\le n-\ell+1\,,\\ \mathrm{Po}(\widetilde{\mu}_k(\underline{v})) & \text{if } n-\ell+1<i\le 2(n-\ell+1)\,.\end{cases}$$

To apply Theorem 1, we choose as neighborhood of (i,k)

$$B_{i,k}:=\{(j,k')\in I:-(k'+3)(\ell-1)\le[j]-[i]\le(k+3)(\ell-1)\}\,. \tag{28}$$

The choice of such a neighborhood is explained in Schbath (1995a). Theorem 1 gives $d_{\mathrm{TV}}\left(\mathcal{L}(\underline{\widetilde{\mathbb{Y}}}),\mathcal{L}(\underline{\mathbb{Z}})\right)\leq b_1+b_2+b_3$, where

$$b_1 = \sum_{(i,k)\in I}\sum_{(j,k')\in B_{i,k}} \mathbb{E}\widetilde{Y}_{i,k}\mathbb{E}\widetilde{Y}_{j,k'} \tag{29}$$

$$b_2 = \sum_{(i,k)\in I}\sum_{(j,k')\in B_{i,k}\setminus\{(i,k)\}} \mathbb{E}\widetilde{Y}_{i,k}\widetilde{Y}_{j,k'} \tag{30}$$

$$b_3 = \sum_{(i,k)\in I} \mathbb{E}|\mathbb{E}(\widetilde{Y}_{i,k}-\mathbb{E}\widetilde{Y}_{i,k}\mid \sigma(\widetilde{Y}_{j,k'},(j,k')\notin B_{i,k}))|\,. \tag{31}$$

Our task consists now in bounding b_1, b_2, b_3 given in (29), (30) and (31). To bound b_2 we make the following assumption:

(A1) $\underline{u}$ is not a substring of any $C_v\in\mathcal{C}_2(\underline{v})$ and $\underline{v}$ is not a substring of any $C_u\in\mathcal{C}_2(\underline{u})$.

As we noticed in Section 2, clumps of $\underline{u}$ and clumps of $\underline{v}$ may overlap in the sequence. This assumption guarantees that a clump of $\underline{u}$ and a clump of $\underline{v}$ can overlap on at most $\ell-1$ letters. Heuristically, if $\underline{u}$ and $\underline{v}$ do not satisfy Assumption (A1), the approximation of $\underline{\widetilde{\mathbb{Y}}}$ with independent Poisson variables should not be valid. Note that this assumption is not required to approximate the declumped counts because, in that case, the clump overlaps are not taken into account. In the present framework, however, they must be considered, which complicates the way to bound b_1, b_2 and b_3. This will be carried out in Appendix A.

To formulate the results we need some more notation. For any words $\underline{w}_1$ and $\underline{w}_2$ of length ℓ, we define

$$\begin{aligned} T_3(\underline{w}_1,\underline{w}_2) &= (n-\ell+1)\mu(\underline{w}_1)\mu(\underline{w}_2)\left(\frac{\Omega(w_{1,\ell},w_{2,1})}{\mu(w_{2,1})}+\frac{\Omega(w_{2,\ell},w_{1,1})}{\mu(w_{1,1})}\right)\\ &\quad+\frac{(n-\ell+1)}{\mu_{\min}}\Big((\ell-2)\big(\mu(\underline{w}_1)\widetilde{\mu}(\underline{w}_2)+\widetilde{\mu}(\underline{w}_1)\mu(\underline{w}_2)\big)+2\widetilde{\mu}(\underline{w}_1)\widetilde{\mu}(\underline{w}_2)\Big)\\ &\quad+(n-\ell+1)\mu(\underline{w}_1)\mu(\underline{w}_2)\left(M(\underline{w}_1,\underline{w}_2)+M(\underline{w}_2,\underline{w}_1)\right)\,, \end{aligned} \tag{32}$$

where Ω and M are given by (16) and (17). Moreover, we define

$$\gamma_2(\ell)=\sum_{x,y\in\mathcal{A}}\mu(x)\max_{a,b\in\mathcal{A}}\left(\frac{1}{\mu(b)}\sum_{(t,t')\neq(1,1)}\left|\frac{\alpha_t^\ell\alpha_{t'}^\ell}{\alpha^\ell}Q_t(x,b)Q_{t'}(a,y)\right|+\sum_{t=2}^{|\mathcal{A}|}\left|\frac{\alpha_t^{5\ell-3}}{\alpha^\ell}Q_t(x,y)\right|\right)\,. \tag{33}$$

Theorem 3 *If $\underline{u}$ is not a substring of any $C_v\in\mathcal{C}_2(\underline{v})$ and if $\underline{v}$ is not a substring of any $C_u\in\mathcal{C}_2(\underline{u})$, then*

$$\begin{aligned} d_{TV}\left(\mathcal{L}(\underline{\widetilde{\mathbb{Y}}}),\mathcal{L}(\underline{\mathbb{Z}})\right) &\leq (n-\ell+1)\big(\widetilde{\mu}(\underline{u})+\widetilde{\mu}(\underline{v})\big)\Big((6\ell-5)(\widetilde{\mu}(\underline{u})+\widetilde{\mu}(\underline{v}))+2(\ell-1)(\mu(\underline{u})+\mu(\underline{v}))\Big)\\ &\quad+T_3(\underline{u},\underline{u})+T_3(\underline{u},\underline{v})+T_3(\underline{v},\underline{u})+T_3(\underline{v},\underline{v})+\gamma_2(\ell)(n-\ell+1)\big(\widetilde{\mu}(\underline{u})+\widetilde{\mu}(\underline{v})\big)|\alpha|^\ell, \end{aligned}$$

where T_3 and γ_2 are given by (32) and (33).

Comments on the order of the error bound in Theorem 2 holds also for Theorem 3. If $\mathbb{E}N(\underline{u})\asymp 1$ and $\mathbb{E}N(\underline{v})\asymp 1$, meaning that $\underline{u}$ and $\underline{v}$ are considered as rare events in the sequence S, we have $\ell\asymp\log n$, $T_3(\underline{u},\underline{v})\asymp n^{-1}\log n+(M(\underline{u},\underline{v})+M(\underline{v},\underline{u}))n^{-1}$, and

$$d_{\mathrm{TV}}\left(\mathcal{L}(\underline{\widetilde{\mathbb{Y}}}),\mathcal{L}(\underline{\mathbb{Z}})\right)\leq \mathrm{O}\left(n^{-1}\log n\right)+\mathrm{O}\left((M(\underline{u},\underline{v})+M(\underline{v},\underline{u}))n^{-1}\right)+\mathrm{O}\left(|\alpha|^{\log n}\right).$$

Therefore, if $\underline{u}$ and $\underline{v}$ cannot overlap too much (this is measured by $M(\underline{u},\underline{v})+M(\underline{v},\underline{u})$), the error bound is very small for large n.

Remark 5 *From Theorem 3 we would also obtain a Poisson process approximation of* $\underline{\widetilde{Y}}$ *like in Theorem 2. However, for Theorem 2, condition (A1) is not needed and the bounds are better.*

From Theorem 3 and equation (8) we have the following corollary. The proof is very similar to the one of Corollary 1 with the function f such that $f(\widetilde{\mathbb{Y}}) = (\widetilde{N}_k(\underline{u}), \widetilde{N}_k(\underline{v}))_{k\geq 1}$. Therefore, we omit the proof.

Corollary 3 *For each* $k \geq 1$, *let* $Z_k^{(1)}$ *and* $Z_k^{(2)}$ *be independent Poisson variables with expectation* $\mathbb{E}Z_k^{(1)} = (n-\ell+1)\widetilde{\mu}_k(\underline{u})$ *and* $\mathbb{E}Z_k^{(2)} = (n-\ell+1)\widetilde{\mu}_k(\underline{v})$. *With same notation as Theorem 3 and under the assumption (A1), we have*

$$\begin{aligned}
&d_{TV}\Big(\mathcal{L}\big((\widetilde{N}_k(\underline{u}), \widetilde{N}_k(\underline{v}))_{k\geq 1}\big), \mathcal{L}\big((Z_k^{(1)}, Z_k^{(2)})_{k\geq 1}\big)\Big) \leq \\
&\quad (n-\ell+1)\big(\widetilde{\mu}(\underline{u}) + \widetilde{\mu}(\underline{v})\big)\Big((6\ell-5)(\widetilde{\mu}(\underline{u}) + \widetilde{\mu}(\underline{v})) + 2(\ell-1)(\mu(\underline{u}) + \mu(\underline{v}))\Big) \\
&\quad + T_3(\underline{u}, \underline{u}) + T_3(\underline{u}, \underline{v}) + T_3(\underline{v}, \underline{u}) + T_3(\underline{v}, \underline{v}) + \gamma_2(\ell)(n-\ell+1)\big(\widetilde{\mu}(\underline{u}) + \widetilde{\mu}(\underline{v})\big)|\alpha|^\ell.
\end{aligned}$$

From Theorem 3, equation (9) and inequality (10), we have the following two corollaries. Again the proofs are omitted; they are very similar to the one of Corollary 1 (use $f(\widetilde{\mathbb{Y}}) = (N_{\mathrm{inf}}(\underline{u}), N_{\mathrm{inf}}(\underline{v}))$ for Corollary 4, and $f(N(\underline{u}), N(\underline{v})) = N(\underline{u}) + N(\underline{v})$ for Corollary 5).

Corollary 4 *With same notation as Corollary 3,* $\mathrm{CP}^{(1)}$ *and* $\mathrm{CP}^{(2)}$ *denote respectively the compound Poisson distribution of* $\sum_{k\geq 1} kZ_k^{(1)}$ *and* $\sum_{k\geq 1} kZ_k^{(2)}$. *Under the assumption (A1), we have*

$$\begin{aligned}
(i)\quad &d_{TV}\Big(\mathcal{L}\big(N_{inf}(\underline{u}), N_{inf}(\underline{v})\big), \mathrm{CP}^{(1)} \otimes \mathrm{CP}^{(2)}\Big) \leq \\
&\quad (n-\ell+1)\big(\widetilde{\mu}(\underline{u}) + \widetilde{\mu}(\underline{v})\big)\Big((6\ell-5)(\widetilde{\mu}(\underline{u}) + \widetilde{\mu}(\underline{v})) + 2(\ell-1)(\mu(\underline{u}) + \mu(\underline{v}))\Big) \\
&\quad + T_3(\underline{u}, \underline{u}) + T_3(\underline{u}, \underline{v}) + T_3(\underline{v}, \underline{u}) + T_3(\underline{v}, \underline{v}) + \gamma_2(\ell)(n-\ell+1)\big(\widetilde{\mu}(\underline{u}) + \widetilde{\mu}(\underline{v})\big)|\alpha|^\ell.
\end{aligned}$$

$$\begin{aligned}
(ii)\quad d_{TV}\Big(\mathcal{L}\big(N(\underline{u}), N(\underline{v})\big), \mathrm{CP}^{(1)} \otimes \mathrm{CP}^{(2)}\Big) \quad &\leq \quad d_{TV}\Big(\mathcal{L}\big(N_{inf}(\underline{u}), N_{inf}(\underline{v})\big), \mathrm{CP}^{(1)} \otimes \mathrm{CP}^{(2)}\Big) \\
&\quad + 2(\ell-1)\big(\mu(\underline{u}) - \widetilde{\mu}(\underline{u}) + \widetilde{\mu}(\underline{v}) - \mu(\underline{v})\big).
\end{aligned}$$

Corollary 5 *Let* $(Z_k)_{k\geq 1}$ *be independent Poisson variables with expectation* $\mathbb{E}Z_k = (n-\ell+1)(\widetilde{\mu}_k(\underline{u}) + \widetilde{\mu}_k(\underline{v}))$; CP *denotes the compound Poisson distribution of* $\sum_{k\geq 1} kZ_k$. *With same notation as Theorem 3 and under the assumption (A1), we have*

$$\begin{aligned}
&d_{TV}(\mathcal{L}(N(\underline{u}) + N(\underline{v})), \mathrm{CP}) \leq \\
&\quad (n-\ell+1)(\widetilde{\mu}(\underline{u}) + \widetilde{\mu}(\underline{v}))\Big((6\ell-5)\big(\widetilde{\mu}(\underline{u}) + \widetilde{\mu}(\underline{v})\big) + 2(\ell-1)(\mu(\underline{u}) + \mu(\underline{v}))\Big) \\
&\quad + T_3(\underline{u}, \underline{u}) + T_3(\underline{u}, \underline{v}) + T_3(\underline{v}, \underline{u}) + T_3(\underline{v}, \underline{v}) + \gamma_2(\ell)(n-\ell+1)\big(\widetilde{\mu}(\underline{u}) + \widetilde{\mu}(\underline{v})\big)|\alpha|^\ell \\
&\quad + 2(\ell-1)\big(\mu(\underline{u}) - \widetilde{\mu}(\underline{u}) + \mu(\underline{v}) - \widetilde{\mu}(\underline{v})\big).
\end{aligned}$$

The compound Poisson distributions defined in Corollaries 4 and 5 reduce to simple Poisson distributions when both $\underline{u}$ and $\underline{v}$ are not self-overlapping words, because of Remark 2.

4 Occurrences of m words of different lengths

We now generalize the results of the previous section to m words of possibly different lengths. We consider m different words $\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_m$ of length $\ell_1, \ell_2, \ldots, \ell_m$, respectively; ℓ will denote the maximal length, and $\ell_{\min}$ will denote the minimal length. We use the notation

$$\underline{u}_r = u_{r,1} u_{r,2} \cdots u_{r,\ell_r}, \quad \forall\, r \in \{1, \ldots, m\}.$$

Moreover we assume that

(A2) $\qquad \forall r \neq r'\,,\ \underline{u}_r$ is not a substring of $\underline{u}_{r'}$.

Since $\underline{u}_r$ and $\underline{u}_{r'}$ do not have necessarily the same length, the definition of the set $\mathcal{P}(\underline{u}_r, \underline{u}_{r'})$ is generalized as

$$\mathcal{P}(\underline{u}_r, \underline{u}_{r'}) := \{p \in \{1, \ldots, \ell_r - 1\} : u_{r',i} = u_{r,i+p},\ \text{for all } i = 1, \ldots, \ell_r - p\}.$$

Under Assumption (A2), if $p \in \mathcal{P}(\underline{u}_r, \underline{u}_{r'})$ then $p > \ell_r - \ell_{r'}$ and the last $(\ell_r - p)$ letters of $\underline{u}_r$ are equal to the first $(\ell_r - p)$ letters of $\underline{u}_{r'}$ ($\underline{u}_{r'}$ can overlap $\underline{u}_r$ from the right).

In view of (16), (17), (18), (20), (41), (32) and (33) we define the corresponding quantities for all r and r' in $\{1, \ldots, m\}$, and for all $a \in \mathcal{A}$;

$$\begin{aligned}
\Omega_r &= \sum_{s=1}^{3\ell-\ell_r-2} \Pi^s, \\
\Omega_{r,r'} &= \sum_{s=1}^{\ell_r+\ell_{r'}-2} \Pi^s, \\
M(\underline{u}_r, \underline{u}_{r'}) &= \begin{cases} \displaystyle\sum_{p \in \mathcal{P}(\underline{u}_r, \underline{u}_{r'})} \frac{1}{\mu(\underline{u}_{r'}^{(\ell_r - p)})} & \text{if } r \neq r'\,, \\ 0 & \text{if } r = r'\,, \end{cases} \\
T_1(\underline{u}_r, \underline{u}_{r'}) &= (2n - \ell_r - \ell_{r'} + 2)\mu(\underline{u}_r)\widetilde{\mu}(\underline{u}_{r'}) \left(\frac{\Omega_{r'}(u_{r',\ell_{r'}}, u_{r,1})}{\mu(u_{r,1})} + M(\underline{u}_{r'}, \underline{u}_r) \right), \\
T_2(\underline{u}_r, \underline{u}_{r'}) &= (n - \ell_r + 1)\Big((\ell - 1)(\widetilde{\mu}(\underline{u}_r)\mu(\underline{u}_{r'}) + \mu(\underline{u}_r)\widetilde{\mu}(\underline{u}_{r'})) + (6\ell - 5)\widetilde{\mu}(\underline{u}_r)\widetilde{\mu}(\underline{u}_{r'}) \Big), \\
T_3(\underline{u}_r, \underline{u}_{r'}) &= (n - \ell_r + 1)\mu(\underline{u}_r)\mu(\underline{u}_{r'}) \left(\frac{\Omega_{r,r'}(u_{r,\ell_r}, u_{r',1})}{\mu(u_{r',1})} + \frac{\Omega_{r,r'}(u_{r',\ell_{r'}}, u_{r,1})}{\mu(u_{r,1})} \right) \\
&\quad + \frac{(n - \ell_r + 1)(6\ell - 3\ell_r - 3\ell_{r'} + 2)}{\mu_{\min}} \widetilde{\mu}(\underline{u}_r)\widetilde{\mu}(\underline{u}_{r'}) \\
&\quad + \frac{(n - \ell_r + 1)(\ell - 2)}{\mu_{\min}} \Big(\mu(\underline{u}_r)\widetilde{\mu}(\underline{u}_{r'}) + \mu(\underline{u}_{r'})\widetilde{\mu}(\underline{u}_r) \Big) \\
&\quad + (n - \ell_r + 1)\mu(\underline{u}_r)\mu(\underline{u}_{r'}) \left(M(\underline{u}_r, \underline{u}_{r'}) + M(\underline{u}_{r'}, \underline{u}_r) \right), \\
\gamma_1(\ell_r, \ell, a) &= \sum_{x,y \in \mathcal{A}} \mu(x) \max_{b \in \mathcal{A}} \left| \frac{1}{\mu(b)} \sum_{(t,t') \neq (1,1)} \frac{\alpha_t^{2\ell-\ell_r} \alpha_{t'}^{2\ell-\ell_r}}{\alpha^\ell} Q_t(x,b) Q_{t'}(a,y) - \sum_{t=2}^{|\mathcal{A}|} \frac{\alpha_t^{4\ell-2}}{\alpha^\ell} Q_t(x,y) \right|, \\
\gamma_2(\ell_r, \ell) &= \sum_{x,y \in \mathcal{A}} \mu(x) \max_{a,b \in \mathcal{A}} \left(\frac{1}{\mu(b)} \sum_{(t,t') \neq (1,1)} \left| \frac{\alpha_t^{2\ell-\ell_r} \alpha_{t'}^{2\ell-\ell_r}}{\alpha^\ell} Q_t(x,b) Q_{t'}(a,y) \right| + \sum_{t=2}^{|\mathcal{A}|} \left| \frac{\alpha_t^{5\ell-3}}{\alpha^\ell} Q_t(x,y) \right| \right).
\end{aligned} \tag{34}$$

As the derivation of the bounds is very similar to the derivation presented in Section 3 when only two words of same length were considered, we give only the results here; the proof is carried out in Appendix B.

4.1 Poisson process approximation for the numbers of clumps

Firstly, we choose as index set $I = \{1, 2, \ldots, m(n+1) - \sum_{s=1}^{m} \ell_s\}$; it can be written as the disjoint union $I = \bigcup_{r=1}^{m} I_r$ with

$$I_r = \left\{ (r-1)(n+1) - \sum_{s=1}^{r-1} \ell_s + 1, \ldots, r(n+1) - \sum_{s=1}^{r} \ell_s \right\}. \tag{35}$$

We define $[i]$ by

$$[i] := i - (r-1)(n+1) + \sum_{s=1}^{r-1} \ell_s \quad \text{with } r = r(i) \text{ such that } i \in I_r\,. \tag{36}$$

To apply Theorem 1, the Bernoulli process $\underline{\widetilde{Y}} = (\widetilde{Y}_i)_{i\in I}$ and the Poisson process $\underline{Z} = (Z_i)_{i\in I}$ are given by

$$\begin{aligned} \widetilde{Y}_i &= \widetilde{\mathbb{I}}_{[i]}(\underline{u}_r), \\ Z_i &\sim \mathrm{Po}(\widetilde{\mu}(\underline{u}_r)), \end{aligned}$$

where r is such that $i \in I_r$. For $i \in I$, we choose the neighborhood $B_i := \{j \in I : |[j] - [i]| \leq 3\ell - 3\}$.

Theorem 4 *Under assumption (A2) and with the notation (34), we have*

$$\begin{aligned} d_{TV}\left(\mathcal{L}(\underline{\widetilde{Y}}), \mathcal{L}(\underline{Z})\right) \;&\leq\; (n - \ell_{\min} + 1)(6\ell - 5)\left(\sum_{r=1}^{m} \widetilde{\mu}(\underline{u}_r)\right)^2 + \sum_{1\leq r,r'\leq m} T_1(\underline{u}_r, \underline{u}_{r'}) \\ &\quad + |\alpha|^{\ell} \sum_{r=1}^{m} \gamma_1(\ell_r, \ell, u_{r,\ell_r})(n - \ell_r + 1)\widetilde{\mu}(\underline{u}_r)\,. \end{aligned}$$

Corollary 6 *Let $(Z^{(r)})_{r=1,\ldots,m}$ be independent Poisson variables with $\mathbb{E}Z^{(r)} = (n - \ell_r + 1)\widetilde{\mu}(\underline{u}_r)$. With the previous notation and under assumption (A2), we have*

(*i*) $$d_{TV}\left(\mathcal{L}\big((\widetilde{N}_{inf}(\underline{u}_r))_{r=1,\ldots,m}\big), \mathcal{L}\big((Z^{(r)})_{r=1,\ldots,m}\big)\right) \leq$$
$$(n - \ell_{\min} + 1)(6\ell - 5)\left(\sum_{r=1}^{m} \widetilde{\mu}(\underline{u}_r)\right)^2 + \sum_{1\leq r,r'\leq m} T_1(\underline{u}_r, \underline{u}_{r'}) + |\alpha|^{\ell} \sum_{r=1}^{m} \gamma_1(\ell_r, \ell, u_{r,\ell_r})(n - \ell_r + 1)\widetilde{\mu}(\underline{u}_r)\,.$$

(*ii*) $$\begin{aligned} d_{TV}\left(\mathcal{L}\big((\widetilde{N}(\underline{u}_r))_{r=1,\ldots,m}\big), \mathcal{L}\big((Z^{(r)})_{r=1,\ldots,m}\big)\right) \;&\leq\; d_{TV}\left(\mathcal{L}\big((\widetilde{N}_{inf}(\underline{u}_r))_{r=1,\ldots,m}\big), \mathcal{L}\big((Z^{(r)})_{r=1,\ldots,m}\big)\right) \\ &\quad + \sum_{r=1}^{m} (\ell_r - 1)\big(\mu(\underline{u}_r) - \widetilde{\mu}(\underline{u}_r)\big)\,. \end{aligned}$$

4.2 Compound Poisson process approximation for the numbers of occurrences

Firstly, we choose as index set $I = \{1, 2, \dots, m(n+1) - \sum_{s=1}^{m} \ell_s\} \times \{1, 2, \dots\}$. To apply Theorem 1, the Bernoulli process $\underline{\widetilde{\mathbb{Y}}} = (\widetilde{Y}_{i,k})_{(i,k)\in I}$ and the Poisson process $\underline{\mathbb{Z}} = (Z_{i,k})_{(i,k)\in I}$ are now defined as

$$\begin{aligned}\widetilde{Y}_{i,k} &= \widetilde{\mathbb{I}}_{[i],k}(\underline{u}_r),\\ Z_{i,k} &\sim \mathrm{Po}(\widetilde{\mu}_k(\underline{u}_r)),\end{aligned}$$

where $r = r(i)$ is such that $i \in I_r$; I_r and $[i]$ are given by (35) and (36). For $(i,k) \in I$, the neighborhood is still $B_{i,k} := \{(j,k') \in I : -(k'+3)(\ell-1) \le [j] - [i] \le (k+3)(\ell-1)\}$.
The assumption (A1) for two words changes now to

(A3) $\quad \forall r \neq r'$, $\underline{u}_r$ is not a substring of any composed word in $\mathcal{C}_2(\underline{u}_{r'})$.

We will show in Appendix B that the following theorem holds.

Theorem 5 *Under assumptions (A2), (A3) and with the notation (34), we have*

$$d_{TV}\Big(\mathcal{L}(\underline{\widetilde{\mathbb{Y}}}), \mathcal{L}(\underline{\mathbb{Z}})\Big) \le \sum_{1\le r,r'\le m} T_2(\underline{u}_r, \underline{u}_{r'}) + \sum_{1\le r,r'\le m} T_3(\underline{u}_r, \underline{u}_{r'}) + |\alpha|^{\ell} \sum_{r=1}^{m} \gamma_2(\ell_r, \ell)(n - \ell_r + 1)\widetilde{\mu}(\underline{u}_r) .$$

From Theorem 5 and equation (8), the following corollary is easily obtained.

Corollary 7 *Let $(Z_k^{(r)})_{k\ge 1, r\in\{1,\dots,m\}}$ be independent Poisson variables with expectation $\mathbb{E}Z_k^{(r)} = (n - \ell_r + 1)\widetilde{\mu}_k(\underline{u}_r)$. Under assumptions (A2), (A3) and with the notation (34), we have*

$$d_{TV}\Big(\mathcal{L}\big((\widetilde{N}_k(\underline{u}_r))_{k\ge 1, r\in\{1,\dots,m\}}\big), \mathcal{L}\big((Z_k^{(r)})_{k\ge 1, r\in\{1,\dots,m\}}\big)\Big) \le$$
$$\sum_{1\le r,r'\le m} T_2(\underline{u}_r, \underline{u}_{r'}) + \sum_{1\le r,r'\le m} T_3(\underline{u}_r, \underline{u}_{r'}) + |\alpha|^{\ell} \sum_{r=1}^{m} \gamma_2(\ell_r, \ell)(n - \ell_r + 1)\widetilde{\mu}(\underline{u}_r) .$$

From Theorem 5, equation (9) and inequality (10), we have the following two corollaries.

Corollary 8 *With same notation as Corollary 7,* $\mathrm{CP}^{(r)}$ *denotes the compound Poisson distribution of $\sum_{k\ge 1} kZ_k^{(r)}$. Under assumption (A2) and (A3), we have*

$$(i) \qquad d_{TV}\Big(\mathcal{L}\big(N_{\mathit{inf}}(\underline{u}_1), \dots, N_{\mathit{inf}}(\underline{u}_m)\big), \mathrm{CP}^{(1)} \otimes \cdots \otimes \mathrm{CP}^{(m)}\Big) \le$$
$$\sum_{1\le r,r'\le m} T_2(\underline{u}_r, \underline{u}_{r'}) + \sum_{1\le r,r'\le m} T_3(\underline{u}_r, \underline{u}_{r'}) + |\alpha|^{\ell} \sum_{r=1}^{m} \gamma_2(\ell_r, \ell)(n - \ell_r + 1)\widetilde{\mu}(\underline{u}_r) .$$

$$(ii) \quad d_{TV}\Big(\mathcal{L}\big(N(\underline{u}_1), \dots, N(\underline{u}_m)\big), \mathrm{CP}^{(1)} \otimes \cdots \otimes \mathrm{CP}^{(m)}\Big) \le$$
$$d_{TV}\Big(\mathcal{L}\big(N_{\mathit{inf}}(\underline{u}_1), \dots, N_{\mathit{inf}}(\underline{u}_m)\big), \mathrm{CP}^{(1)} \otimes \cdots \otimes \mathrm{CP}^{(m)}\Big) + 2\sum_{r=1}^{m} (\ell_r - 1)\big(\mu(\underline{u}_r) - \widetilde{\mu}(\underline{u}_r)\big) .$$

Corollary 9 *Let* $(Z_k)_{k\geq 1}$ *be independent Poisson variables with expectation* $\mathbb{E}Z_k = \sum_{r=1}^{m}(n-\ell_r+1)(\widetilde{\mu}_k(\underline{u}_r))$*;* CP *denotes the compound Poisson distribution of* $\sum_{k\geq 1} kZ_k$*. Under assumptions (A2), (A3) and with the notation (34), we have*

$$d_{TV}\left(\mathcal{L}\left(\sum_{r=1}^{m} N(\underline{u}_r)\right), \mathrm{CP}\right) \leq \sum_{1\leq r,r'\leq m} T_2(\underline{u}_r,\underline{u}_{r'}) + \sum_{1\leq r,r'\leq m} T_3(\underline{u}_r,\underline{u}_{r'})$$
$$+|\alpha|^{\ell}\sum_{r=1}^{m}\gamma_2(\ell_r,\ell)(n-\ell_r+1)\widetilde{\mu}(\underline{u}_r) + 2\sum_{r=1}^{m}(\ell_r-1)\left(\widetilde{\mu}(\underline{u}_r)-\mu(\underline{u}_r)\right).$$

Because of Remark 2, the compound Poisson distributions defined in Corollaries 8 and 9 reduce to simple Poisson distributions if every $\underline{u}_r$, $r=1,\dots,m$, is a non self-overlapping word.

5 Application

To illustrate the goodness of the approximation for collections of words in DNA sequences, we consider motifs on the four-letter DNA alphabet $\mathcal{A}=\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$ having the structure

$$a_1a_2\cdots a_r(\mathtt{N})_s\overline{a_r}\cdots\overline{a_2}\,\overline{a_1}, \tag{37}$$

where the integers r and s are fixed, and $a_i\in\mathcal{A}$ for $i=1,\dots,r$. Moreover, `N` represents any letter in the alphabet $\mathcal{A}$, $(\mathtt{N})_s$ denotes s consecutive, possibly different letters `N`, and for each $i=1,\dots,r$, $\overline{a_i}$ is the complementary of a_i. Here, we use that `A` and `T` are complementary, as are `C` and `G`. For instance, the motif `AGGCNNGCCT` is of type (37), involving the collection of the sixteen words $(\mathtt{AGGC}ab\mathtt{GCCT})_{a,b\in\mathcal{A}}$ of length 10.

These motifs are particularly interesting because of their possible stem-loop structure; the prefix $a_1a_2\cdots a_r$ could form a stem with the suffix $\overline{a_r}\cdots\overline{a_2}\,\overline{a_1}$, leading to a loop of length s as shown by Figure 1. Such a structure may lead to errors when the polymerase replicates the genome. This phenomenon can also occur with RNA sequences.

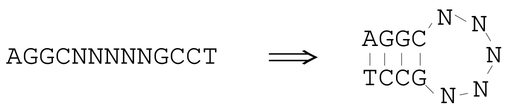

Figure 1: Stem-loop structure

Since clumps of these kind of motifs are not particularly biologically relevant, we will focus in this section on their number of occurrences in a DNA sequence. The number of occurrences of the motif `AGGCNNGCCT`, for instance, is denoted by $N(\mathtt{AGGCNNGCCT})$ and can be easily obtained by summing the numbers of occurrences of each word $\mathtt{AGGC}ab\mathtt{GCCT}$, $a,b\in\mathcal{A}$:

$$N(\mathtt{AGGCNNGCCT}) = \sum_{a,b\in\mathcal{A}} N(\mathtt{AGGC}ab\mathtt{GCCT}).$$

The expected count of `AGGCNNGCCT` in a sequence is then

$$\mathbb{E}N(\mathtt{AGGCNNGCCT}) = \sum_{a,b\in\mathcal{A}} \mathbb{E}N(\mathtt{AGGC}ab\mathtt{GCCT}),$$

where the right-hand terms are given in (2). Using Corollary 9, we can calculate the error bound while approximating the count $N(\texttt{AGGCNNGCCT})$ by a compound Poisson variable $\sum_{k\geq 1} kZ_k$ such that Z_k's are independent Poisson variables with expectation

$$\mathbb{E}Z_k = \sum_{a,b\in\mathcal{A}} (1 - A(\texttt{AGGCNNGCCT}))^2 A(\texttt{AGGCNNGCCT})^{k-1} \mathbb{E}N(\texttt{AGGCNNGCCT}),$$

where A is given in (7).

In the application below, we consider a sequence of length $n = 48,502$ generated by a first order Markov chain on the alphabet $\mathcal{A}$, with the following transition matrix

$$\Pi = \begin{pmatrix} 0.2994 & 0.2086 & 0.2215 & 0.2705 \\ 0.2830 & 0.2198 & 0.2740 & 0.2232 \\ 0.2540 & 0.2820 & 0.2480 & 0.2160 \\ 0.1813 & 0.2232 & 0.3164 & 0.2791 \end{pmatrix}. \qquad (38)$$

These values correspond approximately to the genome of the bacteriophage *Lambda*, in the sense that the length of the *Lambda* genome is 48,502bp, and $\Pi(a, b)$ is precisely the ratio between the count of the 2-letter word ab and the count of the letter a in the genome of *Lambda*.

Table 1 gives the expected counts in the sequence of the motifs $\texttt{AGGC(N)}_s\texttt{GCCT}$, $\texttt{ATGC(N)}_s\texttt{GCAT}$, $\texttt{ATGGC(N)}_s\texttt{GCCAT}$, $\texttt{ATTGGC(N)}_s\texttt{GCCAAT}$ and $\texttt{ACTGGC(N)}_s\texttt{GCCAGT}$, where s takes the values 1, 2, 3, 4 and 5. All these motifs have a very small expected count, which is in agreement with the rare word condition. Increasing the loop size s seems to only have a small effect on the expected count of the stem-loop motifs. This might be explained by the trade-off between each simple word in the motif being less likely to occur with increasing s, but the number of words in the motif being increasing with increasing s.

	$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$
$\texttt{AGGC(N)}_s\texttt{GCCT}$	0.691	0.698	0.699	0.699	0.699
$\texttt{ATGC(N)}_s\texttt{GCAT}$	1.681	1.698	1.698	1.698	1.698
$\texttt{ATGGC(N)}_s\texttt{GCCAT}$	0.092	0.093	0.093	0.093	0.093
$\texttt{ATTGGC(N)}_s\texttt{GCCAAT}$	7.65e-03	7.73e-03	7.74e-03	7.74e-03	7.73e-03
$\texttt{ACTGGC(N)}_s\texttt{GCCAGT}$	2.79e-03	2.82e-03	2.82e-03	2.82e-03	2.82e-03

Table 1: Expected counts of some stem-loop motifs in a sequence of length 48,502 generated by the transition matrix given in (38)

The motif $\texttt{ATGC(N)}_s\texttt{GCAT}$ has more than twice the expected count as the motif $\texttt{AGGC(N)}_s\texttt{GCCT}$ because of its more involved self-overlapping structure (see Figures 2 and 3). $\texttt{ATGGC(N)}_s\texttt{GCCAT}$ admits as many self-overlaps as $\texttt{ATGC(N)}_s\texttt{GCAT}$ but is about 16 times less frequent, which is not surprising since it has an additional two letters. Similar comments hold for $\texttt{ATTGGC(N)}_s\texttt{GCCAAT}$ and $\texttt{ACTGGC(N)}_s\texttt{GCCAGT}$.

Table 2 presents the global error bound in the compound Poisson approximation given by Corollary 9, for these stem-loop motifs. For a stem of fixed length, increasing the size s of the loop does not change the order of the bound substantially; it only gently increases. For computational reasons we restrict our study to loop sizes less or equal than 5, even though relevant biological loop sizes are slightly larger. This result is particularly interesting because it reveals a balance effect: increasing the loop size s enormously increases the number of words within the family and thus, the number of summands in the error bound, but on the other hand, increasing s makes the words rarer so the individual summands will be smaller.

However, there is a penalty of a sort for ATGC(N)$_s$GCAT as soon as $s \geq 4$, because these stem-loop motifs do not satisfy assumption (A3) any more. Indeed, for $s = 4$, (ATGCATGCGCAT, ATGCGCATGCAT) is the only pair that violates (A3): ATGCGCATGCAT is a substring of ATGCATGCGCATGCATGCGCAT composed of two overlapping occurrences of ATGCATGCGCAT. Therefore, we cannot apply Corollary 9 for these motifs and the associated error bounds are missing in Table 2.

	$s=1$	$s=2$	$s=3$	$s=4$	$s=5$
AGGC(N)$_s$GCCT	1.63e-03	2.23e-03	2.36e-03	2.57e-03	2.80e-03
ATGC(N)$_s$GCAT	1.09e-02	1.27e-02	1.56e-02	—	—
ATGGC(N)$_s$GCCAT	3.94e-05	4.46e-05	5.50e-05	5.69e-05	6.08e-05
ATTGGC(N)$_s$GCCAAT	3.25e-07	3.63e-07	3.91e-07	4.60e-07	4.76e-07
ACTGGC(N)$_s$GCCAGT	3.96e-08	4.38e-08	4.72e-08	5.64e-08	5.84e-08

Table 2: Global error bound in the compound Poisson approximation of some stem-loop motif counts, in a sequence of length 48,502 generated by the transition matrix given in (38)

Comparing the results for the motifs AGGC(N)$_s$GCCT and ATGC(N)$_s$GCAT, the bounds are roughly six times larger for ATGC(N)$_s$GCAT because of its more complicated overlapping structure. Consider, for instance, the motif ATGCNNNGCAT; Figure 2 describes the different possible overlaps that can occur between two words belonging to this motif. Let u and v be two words belonging to the ATGCNNNGCAT family; the set of periods $\mathcal{P}(u,v)$ is then necessarily equal to either $\{5,9\}$ (16 pairs) or $\{9\}$ ($4^6 - 16$ pairs). Moreover, the set $\mathcal{P}(u)$ is equal to $\{9\}$ for all words.

```
A T G C a b c G C A T
                  A T G C a' b' c' G C A T
                  ↑
                  9
A T G C a A T G C A T
          A T G C A T b G C A T
          ↑
          5
```

Figure 2: Self-overlaps of the family ATGCNNNGCAT

In contrast, $\mathcal{P}(u) = \emptyset$ for all words u in the AGGCNNNGCCT family. Figure 3 describes the possible overlaps between two words of the AGGCNNNGCCT family. Among all the pairs (u,v) in the AGGCNNNGCCT family, only 16 pairs have a nonempty period set $\mathcal{P}(u,v)$, which in this case equals $\{5\}$.

In fact, a motif composed of words that can overlap a lot between themselves will produce large quantities M and hence large quantities T_3 (given by (34)). The overlapping structure of a motif has an important influence on the error bound through the terms T_3. This explains the large bounds of order 10^{-2} for ATGC(N)$_s$GCAT.

However, Table 2 shows that, adding just one letter to the stem of the motifs ATGC(N)$_s$GCAT in such a way that the high overlapping structure of the motif is preserved, for instance ATGGC(N)$_s$GCCAT, is sufficient to reduce the global error bound considerably to the order of 10^{-5}. Adding another letter leads to an error bound of order 10^{-7} for ATTGGC(N)$_s$GCCAAT. The motifs ACTGGC(N)$_s$GCCAGT have a low overlapping structure and their bounds are clearly better (or order 10^{-8}). Undeniably, the approximation is better for long and thus for rare words (see Table 1). Moreover, the effect of increasing the overlapping structure of the motif is balanced with the effect of increasing the length of the stem motif, leading to a small expected count. The rule seems to be that the approximation is better for smaller expected motif counts.

```
A G G C a A G G C C T
          A G G C C T b G C C T
          ↑
          5
```

Figure 3: Self-overlaps of the family AGGCNNNGCCT

For the bounds to be informative, we need to compare them to the probabilities that one might want to approximate. As example we observe that the number of occurrences in the bacteriophage *Lambda* is zero for the motifs ATGGC(N)$_s$GCCAT, ATTGGC(N)$_s$GCCAAT and ACTGGC(N)$_s$GCCAGT, for $s = 1, 2, 3$. One might want to assess whether these counts are exceptionally low. To this purpose, we calculate the approximated probabilities that the counts of these motifs are (at most) zero, in a random sequence of length 48,502 generated by the transition matrix given in (38). The results are presented in Table 3; the error bounds given in Table 2 are of considerably smaller magnitude.

	$s=1$	$s=2$	$s=3$
ATGGC(N)$_s$GCCAT	0.91247	0.91160	0.91158
ATTGGC(N)$_s$GCCAAT	0.99237	0.99230	0.99229
ACTGGC(N)$_s$GCCAGT	0.99721	0.99719	0.99719

Table 3: Approximated probability to observe zero occurrences of some stem-loop motifs, in a sequence of length 48,502 generated by the transition matrix given in (38)

We now focus on the weight of each of the four terms appearing in the global error bound given by Corollary 9. The four terms correspond, respectively, to the bounds of b_1, b_2, b_3 and the boundary effect calculated in Section B.2. Table 4 gives the bound for each term and for the motifs ATGGC(N)$_s$GCCAT, ATTGGC(N)$_s$GCCAAT and ACTGGC(N)$_s$GCCAGT, s varying from 1 to 5. It is obvious that b_3 is negligible and decreases rapidly with increasing s. The boundary effect is also negligible and decreases smoothly as s increases. On the other hand, b_1 and b_2 are the main terms and their bound are about of the same order. These bounds increase very slightly while increasing the loop size s, leading to a small increase of the global error bound. Note that there is no boundary effect for ACTGGC(N)$_s$GCCAGT, $s = 1, \ldots, 5$; the explanation is that these motifs are composed of non-self-overlapping words (see (10)). The compound Poisson approximation of $N($ACTGGC(N)$_s$GCCAGT$)$ in this case is a simple Poisson approximation, even if these stem-loop motifs involve words that can overlap between them.

	b_1	b_2	b_3	boundary effect	global bound
ATGGCNGCCAT	1.40e-05	2.55e-05	7.09e-13	2.60e-10	3.95e-05
ATGGCNNGCCAT	1.57e-05	2.92e-05	5.98e-14	7.60e-11	4.49e-05
ATGGCNNNGCCAT	1.71e-05	3.80e-05	4.99e-15	2.12e-11	5.52e-05
ATGGCNNNNGCCAT	1.86e-05	3.83e-05	4.17e-16	9.82e-11	5.69e-05
ATGGCNNNNNGCCAT	2.00e-05	4.08e-05	3.48e-17	1.61e-12	6.08e-05
ATTGGCNGCCAAT	1.17e-07	2.07e-07	4.13e-16	2.18e-12	3.25e-07
ATTGGCNNGCCAAT	1.29e-07	2.34e-07	3.49e-17	6.27e-13	3.63e-07
ATTGGCNNNGCCAAT	1.39e-07	2.51e-07	2.91e-18	1.73e-13	3.91e-07
ATTGGCNNNNGCCAAT	1.49e-07	3.11e-07	2.43e-19	4.73e-14	4.60e-07
ATTGGCNNNNNGCCAAT	1.59e-07	3.17e-07	2.03e-20	1.29e-14	4.76e-07
ACTGGCNGCCAGT	1.56e-08	2.40e-08	1.50e-16	0	3.96e-08
ACTGGCNNGCCAGT	1.72e-08	2.66e-08	1.27e-17	0	4.38e-08
ACTGGCNNNGCCAGT	1.85e-08	2.86e-08	1.06e-18	0	4.72e-08
ACTGGCNNNNGCCAGT	1.98e-08	3.66e-08	8.85e-20	0	5.64e-08
ACTGGCNNNNNGCCAGT	2.11e-08	3.73e-08	7.39e-21	0	5.84e-08

Table 4: Weight of the different terms involved in the global error bound for the compound Poisson approximation of some stem-loop motif counts, in a sequence of length 48,502 generated by the transition matrix given in (38)

Rather than fixing the loop size s, one might be interested in considering a range for the loop size. For instance, let us consider the set of motifs {ATTGGCNGCCAAT, ATTGGCNNGCCAAT}, meaning that $s \in \{1,2\}$. Of course Corollary 9 is still valid since the words may have different lengths, and Table 5 gives the global error bound corresponding to the families ATTGGC(N)$_s$GCCAAT when $s \in \{1,2\}$, $s \in \{2,3\}$ and $s \in \{1,2,3\}$. The error bounds are still very good. We can notice, for instance, that the bound associated to $s \in \{1,2\}$ is larger than the sum of the two error bounds for $s = 1$ and $s = 2$ given in Table 4. If we denote $b_1^{\{1\}}$, $b_1^{\{2\}}$, $b_1^{\{1,2\}}$ the bounds of b_1 for the cases $s = 1$, $s = 2$ and $s \in \{1,2\}$, respectively, and similarly for the terms b_2 and b_3, we have from Section B.2 that $b_1^{\{1,2\}} \gg b_1^{\{1\}} + b_1^{\{2\}}$ and $b_2^{\{1,2\}} \gg b_2^{\{1\}} + b_2^{\{2\}}$ because of the quadratic structure of b_1 and b_2. In contrast, $b_3^{\{1,2\}} \leq b_3^{\{1\}} + b_3^{\{2\}}$ and the boundary effect is additive.

	ATGGC(N)$_s$GCCAT		ATTGGC(N)$_s$GCCAAT	
	global bound	expected count	global bound	expected count
$s \in \{1,2\}$	1.84e-04	0.184	1.49e-06	0.015
$s \in \{2,3\}$	2.09e-04	0.185	1.61e-06	0.015
$s \in \{1,2,3\}$	4.73e-04	0.277	3.72e-06	0.023

Table 5: Global error bound for families of words of different lengths and expected count, in a sequence of length 48,502 generated by the transition matrix given in (38)

6 Conclusion

The two most important approximation results we have shown in this paper are related to numbers of clumps and numbers of occurrences of multiple rare words in a finite sequence generated by a Markov chain. First, one can approximate the vector $(\widetilde{N}(\underline{u}_1), \ldots, \widetilde{N}(\underline{u}_m))$ of numbers of clumps of several words $\underline{u}_1, \ldots, \underline{u}_m$, by a vector $(Z^{(1)}, \ldots, Z^{(m)})$ of independent Poisson variables with expectation $(\widetilde{\mu}(\underline{u}_1), \ldots, \widetilde{\mu}\underline{u}_m)$ given by (3). Second, one can approximate the vector of numbers of occurrences $(N(\underline{u}_1), \ldots, N(\underline{u}_m))$ by a suitable vector $(CP^{(1)}, \ldots, CP^{(m)})$ of independent compound Poisson variables; more precisely, the compound Poisson variables $CP^{(r)}$ are defined by $\sum_{k=1}^{\infty} kZ_k^{(r)}$ where $Z_k^{(r)}$ are independent Poisson variables with expectation $\widetilde{\mu}_k(\underline{u_r})$ given by (6).

The goodness of the approximation is measured via the distance in total variation, which is bounded using the Chen-Stein method. The error bound depends on the overlapping structure between the different words $\underline{u}_1, \ldots, \underline{u}_m$ and increases with the number of words; however, this can be balanced if the words have a very small expected count. The numerical examples (Section 5) show that the approximation are the better the rarer the words are, even if they have an involved overlapping structure.

These multidimensional results are particularly useful as soon as we do not want to study a unique word but a word family or a motif. Multidimensional Gaussian approximations for counts of multiple words and counts of renewals of multiple words have been previously presented by Lundstrom (1990) and Tanushev (1996), but Gaussian approximations are not efficient for rare words and hence for long words. Our approaches are thus complementary.

All our results are valid if the sequence is generated by a Markov chain of order $d > 1$; it suffices to write the sequence in the larger alphabet $\mathcal{A}^d$ that produces a one-order Markov chain. Moreover, the limiting behavior of the total variation distance between the processes we have considered here is similar when the parameters of the Markov chain have to be estimated. Indeed, using the triangular inequality, the estimation of the parameters will result in an additional error term which converges to zero as the length of the sequence tends to infinity (Schbath, 1995b). However, we do not have an explicit error bound for this term.

A Proof of Theorem 3

Our aim is to bound b_1, b_2 and b_3 given by (29), (30), (31); applying Theorem 1 will then lead to Theorem 3.

A.1 Bounding $\mathbf{b_1}$

We decompose $b_1 = b'_1(\underline{u}) + b'_1(\underline{v}) + b'_1(\underline{u}, \underline{v}) + b'_1(\underline{v}, \underline{u})$ with

$$\begin{aligned} b'_1(\underline{u}) &= \sum_{i=1}^{n-\ell+1} \sum_{k\geq 1} \sum_{\substack{(j,k')\in B_{i,k} \\ j\in\{1,\ldots,n-\ell+1\}}} \widetilde{\mathbb{EI}}_{i,k}(\underline{u}) \widetilde{\mathbb{EI}}_{j,k'}(\underline{u}) \\ b'_1(\underline{u}, \underline{v}) &= \sum_{i=1}^{n-\ell+1} \sum_{k\geq 1} \sum_{\substack{(j,k')\in B_{i,k} \\ j\in\{n-\ell+2,\ldots,2(n-\ell+1)\}}} \widetilde{\mathbb{EI}}_{i,k}(\underline{u}) \widetilde{\mathbb{EI}}_{[j],k'}(\underline{v}). \end{aligned}$$

From (6) and (28), we have

$$\begin{aligned} b_1'(\underline{u}) &\le \sum_{i=1}^{n-\ell+1} \sum_{k\ge 1} \sum_{k'\ge 1} \sum_{j=i-(k'+3)(\ell-1)}^{i+(k+3)(\ell-1)} \widetilde{\mu}_k(\underline{u})\widetilde{\mu}_{k'}(\underline{u}) \\ &\le (n-\ell+1)\sum_{k\ge 1}\sum_{k'\ge 1}\big((k+k'+6)(\ell-1)+1\big)\widetilde{\mu}_k(\underline{u})\widetilde{\mu}_{k'}(\underline{u}). \end{aligned}$$

Now, from (6) and (3) we use that

$$\sum_{k\ge 1} \widetilde{\mu}_k(\underline{u}) = \widetilde{\mu}(\underline{u}), \tag{39}$$

$$\sum_{k\ge 1} k\widetilde{\mu}_k(\underline{u}) = \mu(\underline{u}), \tag{40}$$

and we obtain

$$b_1'(\underline{u}) \le (n-\ell+1)\bigg(2(\ell-1)\widetilde{\mu}(\underline{u})\mu(\underline{u}) + (6\ell-5)\widetilde{\mu}(\underline{u})^2\bigg). \tag{41}$$

Using (39) and (40), for $b_1'(\underline{u},\underline{v})$ we have

$$\begin{aligned} b_1'(\underline{u},\underline{v}) &\le \sum_{i=1}^{n-\ell+1} \sum_{k\ge 1} \sum_{k'\ge 1} \sum_{[j]=i-(k'+3)(\ell-1)}^{i+(k+3)(\ell-1)} \widetilde{\mu}_k(\underline{u})\widetilde{\mu}_{k'}(\underline{v}) \\ &\le (n-\ell+1)\sum_{k\ge 1}\sum_{k'\ge 1}((k+k'+6)(\ell-1)+1)\widetilde{\mu}_k(\underline{u})\widetilde{\mu}_{k'}(\underline{v}) \\ &\le (n-\ell+1)\bigg((\ell-1)(\widetilde{\mu}(\underline{u})\mu(\underline{v}) + \mu(\underline{u})\widetilde{\mu}(\underline{v})) + (6\ell-5)\widetilde{\mu}(\underline{u})\widetilde{\mu}(\underline{v})\bigg). \end{aligned}$$

Combining these results gives

$$b_1 \le (n-\ell+1)\big(\widetilde{\mu}(\underline{u})+\widetilde{\mu}(\underline{v})\big)\bigg((6\ell-5)(\widetilde{\mu}(\underline{u})+\widetilde{\mu}(\underline{v})) + 2(\ell-1)(\mu(\underline{u})+\mu(\underline{v}))\bigg).$$

A.2 Bounding $\mathbf{b_2}$

In essence, we follow the set up we used to obtain (24), taking care of the possible overlaps between clumps of $\underline{u}$ and clumps of $\underline{v}$. We write

$$b_2 = \sum_{(i,k)\in I}\ \sum_{(j,k')\in B_{i,k}\setminus\{(i,k)\}} \mathbb{E}\widetilde{Y}_{i,k}\widetilde{Y}_{j,k'} = b_2'(\underline{u}) + b_2'(\underline{v}) + b_2'(\underline{u},\underline{v}) + b_2'(\underline{v},\underline{u}),$$

with

$$\begin{aligned} b_2'(\underline{u}) &= \sum_{i=1}^{n-\ell+1}\sum_{k\ge 1} \sum_{\substack{(j,k')\in B_{i,k}\setminus\{(i,k)\} \\ j\in\{1,\ldots,n-\ell+1\}}} \mathbb{E}\widetilde{\mathbb{I}}_{i,k}(\underline{u})\widetilde{\mathbb{I}}_{j,k'}(\underline{u}) \\ b_2'(\underline{u},\underline{v}) &= \sum_{i=1}^{n-\ell+1}\sum_{k\ge 1} \sum_{\substack{(j,k')\in B_{i,k} \\ j\in\{n-\ell+2,\ldots,2(n-\ell+1)\}}} \mathbb{E}\widetilde{\mathbb{I}}_{i,k}(\underline{u})\widetilde{\mathbb{I}}_{[j],k'}(\underline{v})\,. \end{aligned}$$

We will consider three cases: firstly when the k-clump starting at i and the k'-clump starting at $[j]$ overlap in $\mathcal{M}$, secondly when the clumps do not overlap but the "enlarged clumps" – including the $(\ell-1)$ preceding letters and the $(\ell-1)$ following letters of the clumps – overlap, and thirdly when the enlarged clumps do not overlap in $\mathcal{M}$.
For this purpose, we need to consider the set of words of length $\ell-1$ that can precede a clump of $\underline{u}$, and the set of words of length $\ell-1$ that can follow a clump of $\underline{u}$. As we defined $\mathcal{G}(\underline{u})$ the set of words allowed before a clump of $\underline{u}$ in (25), $\mathcal{D}(\underline{u})$ is the set of words allowed after a clump of $\underline{u}$:

$$\mathcal{D}(\underline{u}) = \{\underline{d} = d_1 \cdots d_{\ell-1} : \forall p \in \mathcal{P}(\underline{u}), d_1 \cdots d_p \neq u_{\ell-p+1} \cdots u_\ell\}.$$

Recall that $\mathcal{C}_k(\underline{u})$ is the set of all the concatenated words composed of exactly k overlapping occurrences of $\underline{u}$. Thus, a k-clump of $\underline{u}$ starts at position i in the infinite sequence $\mathcal{M}$ if and only if one of the words $\underline{gCd}$, where $\underline{g} \in \mathcal{G}(\underline{u})$, $\underline{C} \in \mathcal{C}_k(\underline{u})$ and $\underline{d} \in \mathcal{D}(\underline{u})$, occurs at position $i-\ell+1$. From Schbath (1995a), no two different C and C' in $\mathcal{C}_k(\underline{u})$ can occur simultaneously at position i. Therefore, we can write:

$$\widetilde{\mathbb{I}}_{i,k}(\underline{u}) = \sum_{\underline{g}\in\mathcal{G}(\underline{u}),\underline{C}\in\mathcal{C}_k(\underline{u}),\underline{d}\in\mathcal{D}(\underline{u})} \mathbb{I}_{i-\ell+1}(\underline{gCd}). \tag{42}$$

Bounding $\mathbf{b_2'}(\underline{u})$: Replacing $\widetilde{\mathbb{I}}_{i,k}(\underline{u})$ in $b_2'(\underline{u})$ by (42), and $\widetilde{\mathbb{I}}_{j,k'}(\underline{u})$ by its analogue decomposition $\widetilde{\mathbb{I}}_{j,k'}(\underline{u}) = \sum \mathbb{I}_{j-\ell+1}(\underline{g'C'd'})$ where $\underline{g'} \in \mathcal{G}(\underline{u})$, $\underline{C'} \in \mathcal{C}_{k'}(\underline{u})$ and $\underline{d'} \in \mathcal{D}(\underline{u})$, gives

$$b_2'(\underline{u}) = \sum_{i=1}^{n-\ell+1} \sum_{k\geq 1} \sum_{\substack{(j,k')\in B_{i,k}\setminus\{(i,k)\} \\ j\in\{1,\ldots,n-\ell+1\}}} \sum_{\substack{\underline{g}\in\mathcal{G}(\underline{u}),\,\underline{d}\in\mathcal{D}(\underline{u}) \\ \underline{C}\in\mathcal{C}_k(\underline{u})}} \sum_{\substack{\underline{g'}\in\mathcal{G}(\underline{u}),\,\underline{d'}\in\mathcal{D}(\underline{u}) \\ \underline{C'}\in\mathcal{C}_{k'}(\underline{u})}} \mathbb{E}\mathbb{I}_{i-\ell+1}(\underline{gCd})\mathbb{I}_{j-\ell+1}(\underline{g'C'd'}).$$

Since two clumps of $\underline{u}$ cannot overlap in the sequence, the composed words $\underline{C}$ and $\underline{C'}$ starting at i and j cannot overlap. However, we decompose $b_2'(\underline{u})$ as $b_{21}'(\underline{u}) + b_{22}'(\underline{u})$ corresponding to the following two situations:

1. $\underline{gCd}$ and $\underline{g'C'd'}$ overlap in the sequence (at most on $2(\ell-1)$ letters), that is,

$$j \in \{i-|\underline{C'}|-2\ell+3, \ldots, i-|\underline{C'}|\} \cup \{i+|\underline{C}|, \ldots, i+|\underline{C}|+2\ell-3\}.$$

By symmetry, we have

$$b_{21}'(\underline{u}) \leq 2 \sum_{i=1}^{n-\ell+1} \sum_{k\geq 1}\sum_{k'\geq 1} \sum_{\substack{\underline{g}\in\mathcal{G}(\underline{u}),\,\underline{d}\in\mathcal{D}(\underline{u}) \\ \underline{C}\in\mathcal{C}_k(\underline{u})}} \sum_{\substack{\underline{g'}\in\mathcal{G}(\underline{u}),\,\underline{d'}\in\mathcal{D}(\underline{u}) \\ \underline{C'}\in\mathcal{C}_{k'}(\underline{u})}} \sum_{j=i+|\underline{C}|}^{i+|\underline{C}|+2\ell-3} \mathbb{E}\mathbb{I}_{i-\ell+1}(\underline{gCd})\mathbb{I}_{j-\ell+1}(\underline{g'C'd'}).$$

Summing first over k', $\underline{g'}$, $\underline{C'}$, $\underline{d'}$, then over $\underline{d}$, and using that $\widetilde{\mathbb{I}}_j(\underline{u}) \leq \mathbb{I}_j(\underline{u})$ gives

$$b_{21}'(\underline{u}) \leq 2 \sum_{i=1}^{n-\ell+1} \sum_{k\geq 1} \sum_{\underline{g}\in\mathcal{G}(\underline{u}),\,\underline{C}\in\mathcal{C}_k(\underline{u})} \sum_{j=i+|\underline{C}|}^{i+|\underline{C}|+2\ell-3} \mathbb{E}\mathbb{I}_{i-\ell+1}(\underline{gC})\mathbb{I}_j(\underline{u}).$$

Now, $\underline{gC}$ at position $i-\ell+1$ and $\underline{u}$ at position j do not overlap; it follows that

$$\mathbb{E}\mathbb{I}_{i-\ell+1}(\underline{gC})\mathbb{I}_j(\underline{u}) = \mu(\underline{gC})\Pi^{j-i-|\underline{C}|+1}(u_\ell, u_1)\frac{\mu(\underline{u})}{\mu(u_1)}.$$

Therefore,

$$b'_{21}(\underline{u}) \leq 2(n-\ell+1)\frac{\mu(\underline{u})}{\mu(u_1)}\Omega(u_\ell, u_1)\sum_{k\geq 1}\sum_{\underline{g}\in\mathcal{G}(\underline{u}),\,\underline{C}\in\mathcal{C}_k(\underline{u})}\mu(\underline{gC})$$

where Ω is given in (16). Finally, note that

$$\sum_{k\geq 1}\sum_{\underline{g}\in\mathcal{G}(\underline{u}),\,\underline{C}\in\mathcal{C}_k(\underline{u})}\mu(\underline{gC}) = \sum_{k\geq 1}\sum_{k^*\geq k}\widetilde{\mu}_{k^*}(\underline{u}) = \sum_{k^*\geq 1}k^*\widetilde{\mu}_{k^*}(\underline{u}) = \mu(\underline{u})\,, \tag{43}$$

which leads to

$$b'_{21}(\underline{u}) \leq 2(n-\ell+1)\frac{\mu^2(\underline{u})}{\mu(u_1)}\Omega(u_\ell, u_1)\,. \tag{44}$$

2. $\underline{gCd}$ and $\underline{g'C'd'}$ do not overlap in the sequence, that is,

$$j \in \{i-(k'+3)(\ell-1), \ldots, i-|\underline{C'}|-2\ell+2\} \cup \{i+|\underline{C}|+2\ell-2, \ldots, i+(k+3)(\ell-1)\}\,.$$

By symmetry, we have

$$b'_{22}(\underline{u}) \leq 2\sum_{i=1}^{n-\ell+1}\sum_{k\geq 1}\sum_{k'\geq 1}\sum_{\substack{\underline{g}\in\mathcal{G}(\underline{u}),\,\underline{d}\in\mathcal{D}(\underline{u})\\ \underline{C}\in\mathcal{C}_k(\underline{u})}}\sum_{\substack{\underline{g'}\in\mathcal{G}(\underline{u}),\,\underline{d'}\in\mathcal{D}(\underline{u})\\ \underline{C'}\in\mathcal{C}_{k'}(\underline{u})}}\sum_{j=i+|\underline{C}|+2\ell-2}^{i+(k+3)(\ell-1)}\mathbb{E}\mathbb{I}_{i-\ell+1}(\underline{gCd})\mathbb{I}_{j-\ell+1}(\underline{g'C'd'})\,.$$

We first sum over k', $\underline{C'}$, $\underline{d'}$:

$$b'_{22}(\underline{u}) \leq 2\sum_{i=1}^{n-\ell+1}\sum_{k\geq 1}\sum_{\substack{\underline{g}\in\mathcal{G}(\underline{u}),\,\underline{d}\in\mathcal{D}(\underline{u})\\ \underline{C}\in\mathcal{C}_k(\underline{u})}}\sum_{\underline{g'}\in\mathcal{G}(\underline{u})}\sum_{j=i+|\underline{C}|+2\ell-2}^{i+(k+3)(\ell-1)}\mathbb{E}\mathbb{I}_{i-\ell+1}(\underline{gCd})\mathbb{I}_{j-\ell+1}(\underline{g'u})\,.$$

Now, $\underline{gCd}$ and $\underline{g'u}$ do not overlap, leading to

$$\sum_{\underline{g'}\in\mathcal{G}(\underline{u})}\mathbb{E}\mathbb{I}_{i-\ell+1}(\underline{gCd})\mathbb{I}_{j-\ell+1}(\underline{g'u}) \leq \sum_{\underline{g'}\in\mathcal{G}(\underline{u})}\mu(\underline{gCd})\frac{\mu(\underline{g'u})}{\mu(g'_1)} \leq \mu(\underline{gCd})\frac{\widetilde{\mu}(\underline{u})}{\mu_{\min}}\,.$$

Moreover, note that the sum over j has less than $k(\ell-2)+1$ terms because $|\underline{C}| \geq \ell+k-1$. Therefore,

$$b'_{22}(\underline{u}) \leq 2(n-\ell+1)\frac{\widetilde{\mu}(\underline{u})}{\mu_{\min}}\sum_{k\geq 1}\big(k(\ell-2)+1\big)\widetilde{\mu}_k(\underline{u})\,.$$

Finally, using (39) and (40), we obtain

$$b'_{22}(\underline{u}) \leq 2(n-\ell+1)\frac{\widetilde{\mu}(\underline{u})}{\mu_{\min}}\left((\ell-2)\mu(\underline{u})+\widetilde{\mu}(\underline{u})\right)\,. \tag{45}$$

Combining (44) and (45) gives $b'_2(\underline{u}) \leq T_3(\underline{u},\underline{u})$ where T_3 is given in (32).

Bounding $\mathbf{b'_2}(\underline{u},\underline{v})$: Our purpose is now to prove that $b'_2(\underline{u},\underline{v}) \leq T_3(\underline{u},\underline{v})$. In view of (42), we write

$$b'_2(\underline{u},\underline{v}) = \sum_{i=1}^{n-\ell+1}\sum_{k\geq 1}\sum_{\substack{(j,k')\in B_{i,k}\\ j\in\{n-\ell+2,\ldots,2(n-\ell+1)\}}}\sum_{\substack{\underline{g}\in\mathcal{G}(\underline{u}),\,\underline{d}\in\mathcal{D}(\underline{u})\\ \underline{C}\in\mathcal{C}_k(\underline{u})}}\sum_{\substack{\underline{g'}\in\mathcal{G}(\underline{v}),\,\underline{d'}\in\mathcal{D}(\underline{v})\\ \underline{C'}\in\mathcal{C}_{k'}(\underline{v})}}\mathbb{E}\mathbb{I}_{i-\ell+1}(\underline{gCd})\mathbb{I}_{[j]-\ell+1}(\underline{g'C'd'})\,.$$

Again, we decompose $b'_2(\underline{u},\underline{v})$ as $b'_{21}(\underline{u},\underline{v}) + b'_{22}(\underline{u},\underline{v}) + b'_{23}(\underline{u},\underline{v})$, with an additional term corresponding to the case when $\underline{C}$ and $\underline{C'}$ overlap in the sequence (at most on $(\ell-1)$ letters, from (A1)).

1. When $\underline{gCd}$ and $\underline{g'C'd'}$ overlap in the sequence but $\underline{C}$ and $\underline{C'}$ do not overlap (which happens for at most $2(\ell-1)$ letters), that is, for

$$[j] \in \{i - |\underline{C'}| - 2\ell + 3, \dots, i - |\underline{C'}|\} \cup \{i + |\underline{C}|, \dots, i + |\underline{C}| + 2\ell - 3\},$$

the term $b'_{21}(\underline{u}, \underline{v})$ can be bounded using the same techniques as for $b'_{21}(\underline{u})$. We obtain

$$b'_{21}(\underline{u}, \underline{v}) \leq (n - \ell + 1)\mu(\underline{u})\mu(\underline{v}) \left(\frac{\Omega(u_\ell, v_1)}{\mu(v_1)} + \frac{\Omega(v_\ell, u_1)}{\mu(u_1)} \right). \tag{46}$$

2. When $\underline{gCd}$ and $\underline{g'C'd'}$ do not overlap in the sequence, that is, for

$$j \in \{i - (k'+3)(\ell-1), \dots, i - |\underline{C'}| - 2\ell + 2\} \cup \{i + |\underline{C}| + 2\ell - 2, \dots, i + (k+3)(\ell-1)\},$$

similar arguments as used to bound $b'_{22}(\underline{u})$ give

$$b'_{22}(\underline{u}, \underline{v}) \leq (n - \ell + 1)\frac{\widetilde{\mu}(\underline{u})}{\mu_{\min}}\big((\ell-2)\mu(\underline{v}) + \widetilde{\mu}(\underline{v})\big) + (n - \ell + 1)\frac{\widetilde{\mu}(\underline{v})}{\mu_{\min}}\big((\ell-2)\mu(\underline{u}) + \widetilde{\mu}(\underline{u})\big). \tag{47}$$

3. We now consider the case when $\underline{C}$ and $\underline{C'}$ overlap in the sequence, that is,

$$[j] \in \{i + |\underline{C}| - \ell, \dots, i + |\underline{C}| - 1\} \cup \{i - |\underline{C'}| + 1, \dots, i - |\underline{C'}| + \ell\}.$$

We denote by $b'_{23}(\underline{u}, \underline{v})$ the quantity corresponding to this case. By symmetry, we focus on $j > i$. If $\underline{C}$ and $\underline{C'}$ overlap, Assumption (A2) ensures that only the last occurrence of $\underline{u}$ in $\underline{C}$ overlaps with the first occurrence of $\underline{v}$ in $\underline{C'}$. The last occurrence of $\underline{u}$ in $\underline{C}$ starts at position $i + |\underline{C}| - \ell$. We know that an occurrence of $\underline{v}$ starting at position $[j]$ may overlap $\underline{u}$ at $i + |\underline{C}| - \ell$ only if $[j] = i + |\underline{C}| - \ell + p$ with $p \in \mathcal{P}(\underline{u}, \underline{v})$. Therefore,

$$\begin{aligned}
&\sum_{i=1}^{n-\ell+1} \sum_{k\geq 1}\sum_{k'\geq 1} \sum_{\substack{\underline{g}\in\mathcal{G}(\underline{u}),\, \underline{d}\in\mathcal{D}(\underline{u}) \\ \underline{C}\in\mathcal{C}_k(\underline{u})}} \sum_{\substack{\underline{g'}\in\mathcal{G}(\underline{u}),\, \underline{d'}\in\mathcal{D}(\underline{u}) \\ \underline{C'}\in\mathcal{C}_{k'}(\underline{u})}} \sum_{[j]=i+|\underline{C}|-\ell}^{i+|\underline{C}|-1} \mathbb{E}\mathbb{I}_{i-\ell+1}(\underline{gCd})\mathbb{I}_{[j]-\ell+1}(\underline{g'C'd'}) \\
= &\sum_{i=1}^{n-\ell+1} \sum_{k\geq 1}\sum_{k'\geq 1} \sum_{\substack{\underline{g}\in\mathcal{G}(\underline{u}),\, \underline{d}\in\mathcal{D}(\underline{u}) \\ \underline{C}\in\mathcal{C}_k(\underline{u})}} \sum_{\substack{\underline{g'}\in\mathcal{G}(\underline{u}),\, \underline{d'}\in\mathcal{D}(\underline{u}) \\ \underline{C'}\in\mathcal{C}_{k'}(\underline{u})}} \sum_{p\in\mathcal{P}(\underline{u},\underline{v})} \mathbb{E}\mathbb{I}_{i-\ell+1}(\underline{gCd})\mathbb{I}_{i+|\underline{C}|-2\ell+p+1}(\underline{g'C'd'}) \\
\leq &\sum_{i=1}^{n-\ell+1} \sum_{k\geq 1} \sum_{\underline{g}\in\mathcal{G}(\underline{u}),\, \underline{C}\in\mathcal{C}_k(\underline{u})} \sum_{p\in\mathcal{P}(\underline{u},\underline{v})} \mathbb{E}\mathbb{I}_{i-\ell+1}(\underline{gC})\mathbb{I}_{i+|\underline{C}|-\ell+p}(\underline{v});
\end{aligned}$$

the last inequality comes from summing over $\underline{d}$, k', $\underline{g'}$, $\underline{C'}$, $\underline{d'}$, and then using that $\widetilde{\mathbb{I}}_{i+|\underline{C}|-\ell+p}(\underline{v}) \leq \mathbb{I}_{i+|\underline{C}|-\ell+p}(\underline{v})$. Now, $\underline{gC}$ starting at $i-\ell+1$ and $\underline{v}$ starting at $i + |\underline{C}| - \ell + p$ overlap at most on $(\ell-1)$ letters; we have

$$\mathbb{E}\mathbb{I}_{i-\ell+1}(\underline{gC})\mathbb{I}_{i+|\underline{C}|-\ell+p}(\underline{v}) \leq \mu(\underline{gC})\frac{\mu(\underline{v})}{\mu(\underline{v}^{(\ell-p)})}.$$

Finally, using (43) we obtain

$$b'_{23}(\underline{u}, \underline{v}) \leq (n - \ell + 1)\mu(\underline{u})\mu(\underline{v})(M(\underline{u}, \underline{v}) + M(\underline{v}, \underline{u})), \tag{48}$$

where M is given by (17)

Combining (46), (47) and (48) gives $b'_2(\underline{u}, \underline{v}) \leq T_3(\underline{u}, \underline{v})$. Thus

$$b_2 \leq T_3(\underline{u}, \underline{u}) + T_3(\underline{u}, \underline{v}) + T_3(\underline{v}, \underline{u}) + T_3(\underline{v}, \underline{v}).$$

A.3 Bounding $\mathbf{b_3}$

It remains to bound b_3. We write

$$b_3 = \sum_{(i,k)\in I} \mathbb{E}|\mathbb{E}(\widetilde{Y}_{i,k} - \mathbb{E}\widetilde{Y}_{i,k} \mid \widetilde{\mathcal{Y}}_{i,k})| = b_3'(\underline{u}) + b_3'(\underline{v}),$$

where $\widetilde{\mathcal{Y}}_{i,k} = \sigma(\widetilde{Y}_{j,k'}, (j,k') \notin B_{i,k})$ and

$$b_3'(\underline{u}) = \sum_{i=1}^{n-\ell+1} \sum_{k\geq 1} \mathbb{E}|\mathbb{E}(\widetilde{\mathbb{I}}_{i,k}(\underline{u}) - \mathbb{E}\widetilde{\mathbb{I}}_{i,k}(\underline{u}) \mid \widetilde{\mathcal{Y}}_{i,k})|.$$

Since $\widetilde{\mathcal{Y}}_{i,k} \subset \mathcal{X}_{[i],k} = \sigma(X_1, \ldots, X_{[i]-2\ell+1}, X_{[i]+(k+2)(\ell-1)+1}, \ldots, X_n)$, using the Markov property gives

$$b_3'(\underline{u}) \leq \sum_{i=1}^{n-\ell+1} \sum_{k\geq 1} \mathbb{E}|\mathbb{E}(\widetilde{\mathbb{I}}_{i,k}(\underline{u}) - \mathbb{E}\widetilde{\mathbb{I}}_{i,k}(\underline{u}) \mid X_{i-2\ell+1}, X_{i+(k+2)(\ell-1)+1})|.$$

To bound the right-hand term, we follow the set-up used to prove Theorem 2, but we employ the decomposition (42). This leads to

$$\begin{aligned} b_3'(\underline{u}) \quad &\leq \quad |\alpha|^\ell (n-\ell+1) \sum_{k\geq 1} \sum_{x,y\in\mathcal{A}} \mu(x) \sum_{\substack{\underline{g}\in\mathcal{G}(\underline{u}),\, \underline{d}\in\mathcal{D}(\underline{u}) \\ \underline{C}\in\mathcal{C}_k(\underline{u})}} \mu(\underline{gCd}) \\ &\left| \frac{1}{\mu(g_1)} \sum_{(t,t')\neq(1,1)} \frac{\alpha_t^\ell \alpha_{t'}^{(k+2)(\ell-1)+1-|\underline{C}|-\ell+2}}{\alpha^\ell} Q_t(x,g_1) Q_{t'}(d_{\ell-1},y) - \sum_{t=2}^{|\mathcal{A}|} \frac{\alpha_t^{(k+4)(\ell-1)+2}}{\alpha^\ell} Q_t(x,y) \right|. \end{aligned}$$

Since $|\underline{C}| \leq k(\ell-1)+1$ we have $(k+2)(\ell-1)+1-|\underline{C}|-\ell+2 \geq \ell$; moreover $(k+4)(\ell-1)+2 \geq 5\ell-3$. It follows that

$$\begin{aligned} b_3'(\underline{u}) \quad &\leq \quad |\alpha|^\ell (n-\ell+1) \sum_{k\geq 1} \sum_{\substack{\underline{g}\in\mathcal{G}(\underline{u}),\, \underline{d}\in\mathcal{D}(\underline{u}) \\ \underline{C}\in\mathcal{C}_k(\underline{u})}} \mu(\underline{gCd}) \sum_{x,y\in\mathcal{A}} \mu(x) \\ &\quad \left\{ \frac{1}{\mu(g_1)} \sum_{(t,t')\neq(1,1)} \left| \frac{\alpha_t^\ell \alpha_{t'}^\ell}{\alpha^\ell} Q_t(x,g_1) Q_{t'}(d_{\ell-1},y) \right| + \sum_{t=2}^{|\mathcal{A}|} \left| \frac{\alpha_t^{5\ell-3}}{\alpha^\ell} Q_t(x,y) \right| \right\} \\ &\leq \quad |\alpha|^\ell (n-\ell+1)\gamma_2(\ell) \sum_{k\geq 1} \sum_{\substack{\underline{g}\in\mathcal{G}(\underline{u}),\, \underline{d}\in\mathcal{D}(\underline{u}) \\ \underline{C}\in\mathcal{C}_k(\underline{u})}} \mu(\underline{gCd}) \\ &\leq \quad |\alpha|^\ell \gamma_2(\ell)(n-\ell+1)\widetilde{\mu}(\underline{u}), \end{aligned} \tag{49}$$

where $\gamma_2(\ell)$ is defined in (33) and the last equality comes from (42) and (39). Therefore,

$$b_3 \leq \gamma_2(\ell)(n-\ell+1)\big(\widetilde{\mu}(\underline{u}) + \widetilde{\mu}(\underline{v})\big)|\alpha|^\ell;$$

this establishes the theorem.

B Proofs of results in Section 4

Theorems 4 and 5 are generalizations of Theorem 2 and 3 for m words $\underline{u}_1, \ldots, \underline{u}_m$ of possibly different lengths. The proofs of Theorems 4 and 5 follow easily from the set up for two words of same length and will be briefly derived. The key argument is the decomposition of the index set I (see (35)). The only novelty comes from the different lengths; the maximal and the minimal lengths will then appear in the bounds and the assumption (A2) is now needed to prove Theorem 4.

Corollaries 7 to 9 can be proved using exactly the same technique as for Corollaries 1 to 5; thusly their proof will be omitted.

B.1 Proof of Theorem 4

We bound b_1, b_2 and b_3 given by Theorem 1 using the steps carried out in detail in Section 3.1 to prove Theorem 2.

Bounding $\mathbf{b_1}$:

$$\begin{aligned} b_1 &= \sum_{i\in I}\sum_{j\in B_i} \mathbb{E}\widetilde{Y}_i \mathbb{E}\widetilde{Y}_j = \sum_{r=1}^{m}\sum_{i=1}^{n-\ell_r+1}\sum_{r'=1}^{m}\sum_{|j-i|\le 3\ell-3} \widetilde{\mu}(\underline{u}_r)\widetilde{\mu}(\underline{u}_{r'}) \\ &\le (n-\ell_{\min}+1)(6\ell-5)\left(\sum_{r=1}^{m}\widetilde{\mu}(\underline{u}_r)\right)^2; \end{aligned} \tag{50}$$

this is the contribution from b_1 to the bound on the error in the approximation.

Bounding $\mathbf{b_2}$: For b_2 we write

$$b_2 = \sum_{i\in I}\sum_{j\in B_i\setminus\{i\}} \mathbb{E}\widetilde{Y}_i\widetilde{Y}_j = \sum_{r=1}^{m} b_2'(\underline{u}_r) + \sum_{r=1}^{m}\sum_{r'\neq r} b_2'(\underline{u}_r, \underline{u}_{r'}),$$

where

$$\begin{aligned} b_2'(\underline{u}_r) &= \sum_{i=1}^{n-\ell_r+1}\sum_{j\in B_i\cap I_r,\,[j]\neq i} \mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u}_r)\widetilde{\mathbb{I}}_{[j]}(\underline{u}_r) \\ b_2'(\underline{u}_r, \underline{u}_{r'}) &= \sum_{i=1}^{n-\ell_r+1}\sum_{j\in B_i\cap I_{r'}} \mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u}_r)\widetilde{\mathbb{I}}_{[j]}(\underline{u}_{r'}). \end{aligned}$$

For $b_2'(\underline{u}_r)$ we obtain as in (23) that

$$\begin{aligned} b_2'(\underline{u}_r) &\le \sum_{i=1}^{n-\ell_r+1}\sum_{s=\ell_r}^{3\ell-3} \mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u}_r)\widetilde{\mathbb{I}}_{i+s}(\underline{u}_r) + \sum_{i=1}^{n-\ell_r+1}\sum_{s=\ell_r}^{3\ell-3} \mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u}_r)\widetilde{\mathbb{I}}_{i-s}(\underline{u}_r) \\ &\le 2(n-\ell_r+1)\widetilde{\mu}(\underline{u}_r)\frac{\mu(\underline{u}_r)}{\mu(u_{r,1})}\sum_{s=1}^{3\ell-\ell_r-2}\Pi^s(u_{r,\ell_r}, u_{r,1}) \\ &= 2(n-\ell_r+1)\frac{\mu(\underline{u}_r)\widetilde{\mu}(\underline{u}_r)}{\mu(u_{r,1})}\Omega_r(u_{r,\ell_r}, u_{r,1}). \end{aligned}$$

To bound $b_2'(\underline{u}_r, \underline{u}_{r'})$ we need to consider the overlaps between $\underline{u}_r$ and $\underline{u}_{r'}$. If $p \notin \mathcal{P}(\underline{u}_r, \underline{u}_{r'})$ and $1 \le p \le \ell_r - 1$ then $\mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u}_r)\widetilde{\mathbb{I}}_{i+p}(\underline{u}_{r'}) = 0$. If $p \ge \ell_r$ then

$$\mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u}_r)\widetilde{\mathbb{I}}_{i+p}(\underline{u}_{r'}) \le \widetilde{\mu}(\underline{u}_r)\Pi^{p-\ell_r+1}(u_{r,\ell_r}, u_{r',1})\frac{\mu(\underline{u}_{r'})}{\mu(u_{r',1})}.$$

Lastly, if $p \in \mathcal{P}(\underline{u}_r, \underline{u}_{r'})$ then

$$\begin{aligned}
\mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u}_r)\widetilde{\mathbb{I}}_{i+p}(\underline{u}_{r'}) &\leq \mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u}_r)\mathbb{I}_{i+p}(\underline{u}_{r'}) \\
&= \widetilde{\mu}(\underline{u}_r)\Pi(u_{r,\ell_r}, u_{r',\ell_r-p+1})\Pi(u_{r',\ell_r-p+1}, u_{r',\ell_r-p+2})\dots\Pi(u_{r',\ell_{r'}-1}, u_{r',\ell_{r'}}) \\
&= \widetilde{\mu}(\underline{u}_r)\mu(\underline{u}_{r'})\frac{1}{\mu(\underline{u}_{r'}^{(\ell_r-p)})}.
\end{aligned}$$

It follows that

$$\begin{aligned}
b_2'(\underline{u}_r, \underline{u}_{r'}) \leq (n-\ell_r+1)\Bigg(&\mu(\underline{u}_r)\widetilde{\mu}(\underline{u}_{r'})\left(\frac{\Omega_{r'}(u_{r',\ell_{r'}}, u_{r,1})}{\mu(u_{r,1})} + M(\underline{u}_{r'}, \underline{u}_r)\right) \\
&+\mu(\underline{u}_{r'})\widetilde{\mu}(\underline{u}_r)\left(\frac{\Omega_r(u_{r,\ell_r}, u_{r',1})}{\mu(u_{r',1})} + M(\underline{u}_r, \underline{u}_{r'})\right)\Bigg).
\end{aligned}$$

Therefore, with the notation (34), we obtain $b_2'(\underline{u}_r) \leq T_1(\underline{u}_r, \underline{u}_r)$ and $b_2'(\underline{u}_r, \underline{u}_{r'}) + b_2'(\underline{u}_{r'}, \underline{u}_r) \leq T_1(\underline{u}_r, \underline{u}_{r'}) + T_1(\underline{u}_{r'}, \underline{u}_r)$, leading to

$$b_2 \leq \sum_{1\leq r,r'\leq m} T_1(\underline{u}_r, \underline{u}_{r'}). \tag{51}$$

Bounding $\mathbf{b_3}$: This is straightforward, using the following decomposition

$$b_3 = \sum_{i\in I}\mathbb{E}|\mathbb{E}(\widetilde{Y}_i - \mathbb{E}\widetilde{Y}_i \mid \widetilde{\mathcal{Y}}_i)| = \sum_{r=1}^{m} b_3'(\underline{u}_r),$$

where

$$b_3'(\underline{u}_r) = \sum_{i=1}^{n-\ell_r+1}\mathbb{E}|\mathbb{E}(\widetilde{\mathbb{I}}_i(\underline{u}_r) - \mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u}_r) \mid \widetilde{\mathcal{Y}}_i)|.$$

Because of the choice of the neighborhood B_i, $\widetilde{\mathcal{Y}}_i \subset \sigma(X_1, \dots, X_{[i]-2\ell+1}, X_{[i]+2\ell-1}, \dots, X_n)$, and we can follow the different steps used in Section 3.1 to obtain

$$b_3'(\underline{u}_r) \leq \gamma_1(\ell_r, \ell, u_{r,\ell_r})(n-\ell_r+1)\widetilde{\mu}(\underline{u}_r)|\alpha|^\ell.$$

Therefore, we have

$$b_3 \leq |\alpha|^\ell \sum_{r=1}^{m}\gamma_1(\ell_r, \ell, u_{r,\ell_r})(n-\ell_r+1)\widetilde{\mu}(\underline{u}_r). \tag{52}$$

Combining (50), (51), (52) and Theorem 1 establishes the theorem.

B.2 Proof of Theorem 5

We bound b_1, b_2, b_3 given in (29), (30), (31) following the different steps described in Section A.

Bounding $\mathbf{b_1}$: We write

$$b_1 = \sum_{r=1}^{m} b_1'(\underline{u}_r) + \sum_{r=1}^{m}\sum_{r'\neq r} b_1'(\underline{u}_r, \underline{u}_{r'})$$

with

$$
\begin{aligned}
b_1'(\underline{u}_r) &= \sum_{i=1}^{n-\ell_r+1} \sum_{k\geq 1} \sum_{\substack{(j,k')\in B_{i,k}\\ j\in\{1,\dots,n-\ell_r+1\}}} \widetilde{\mathbb{EI}}_{i,k}(\underline{u}_r)\widetilde{\mathbb{EI}}_{j,k'}(\underline{u}_r)\\
b_1'(\underline{u}_r,\underline{u}_{r'}) &= \sum_{i=1}^{n-\ell_r+1} \sum_{k\geq 1} \sum_{\substack{(j,k')\in B_{i,k}\\ j\in I_{r'}}} \widetilde{\mathbb{EI}}_{i,k}(\underline{u}_r)\widetilde{\mathbb{EI}}_{[j],k'}(\underline{u}_{r'}).
\end{aligned}
$$

In view of (41) and the structure of the neighborhood $B_{i,k}$, we obtain

$$
b_1'(\underline{u}_r) \leq (n-\ell_r+1)\Big(2(\ell-1)\widetilde{\mu}(\underline{u}_r)\mu(\underline{u}_r) + (6\ell-5)\widetilde{\mu}(\underline{u}_r)^2\Big) = T_2(\underline{u}_r,\underline{u}_r).
$$

In the same way, we have

$$
b_1'(\underline{u}_r,\underline{u}_{r'}) \leq (n-\ell_r+1)\Big((\ell-1)(\widetilde{\mu}(\underline{u}_r)\mu(\underline{u}_{r'}) + \mu(\underline{u}_r)\widetilde{\mu}(\underline{u}_{r'})) + (6\ell-5)\widetilde{\mu}(\underline{u}_r)\widetilde{\mu}(\underline{u}_{r'})\Big) = T_2(\underline{u}_r,\underline{u}_r').
$$

It follows that

$$
b_1 \leq \sum_{1\leq r,r'\leq m} T_2(\underline{u}_r,\underline{u}_r'). \tag{53}
$$

Bounding $\mathbf{b_2}$: To bound b_2 we write

$$
b_2 = \sum_{r=1}^{m} b_2'(\underline{u}_r) + \sum_{r=1}^{m}\sum_{r'\neq r} b_2'(\underline{u}_r,\underline{u}_{r'}),
$$

with

$$
\begin{aligned}
b_2'(\underline{u}_r) &= \sum_{i=1}^{n-\ell_r+1} \sum_{k\geq 1} \sum_{\substack{(j,k')\in B_{i,k}\setminus\{(i,k)\}\\ j\in\{1,\dots,n-\ell_r+1\}}} \widetilde{\mathbb{EI}}_{i,k}(\underline{u}_r)\widetilde{\mathbb{I}}_{j,k'}(\underline{u}_r)\\
b_2'(\underline{u}_r,\underline{u}_{r'}) &= \sum_{i=1}^{n-\ell_r+1} \sum_{k\geq 1} \sum_{\substack{(j,k')\in B_{i,k}\\ j\in I_{r'}}} \widetilde{\mathbb{EI}}_{i,k}(\underline{u}_r)\widetilde{\mathbb{I}}_{[j],k'}(\underline{u}_{r'}).
\end{aligned}
$$

Here again, we decompose $b_2'(\underline{u}_r)$ into two terms $b_{21}'(\underline{u}_r) + b_{22}'(\underline{u}_r)$ corresponding to whether or not the enlarged k-clump starting at position $i-\ell_r+1$ and the enlarged k'-clump starting at position $j-\ell_r+1$ overlap (they can overlap at most on $2(\ell_r-1)$ letters). $b_{21}'(\underline{u}_r)$ is bounded as in (44) with ℓ changed to ℓ_r, and with u_1, u_r changed to $u_{r,1}$ and $u_{r,\ell}$. To bound $b_{22}'(\underline{u}_r)$, we use the method that provided (45), observing that now the sums over j have at most $k(\ell-2)+3\ell-3\ell_r+1$ terms. Finally, we obtain

$$
b_2'(\underline{u}_r) \leq 2(n-\ell_r+1)\frac{\mu^2(\underline{u}_r)}{\mu(u_{r,1})}\Omega_r(u_{r,\ell_r},u_{r,1}) + \frac{2(n-\ell_r+1)\widetilde{\mu}(\underline{u}_r)}{\mu_{\min}}\big((\ell-2)\mu(\underline{u}_r) + (3\ell-3\ell_r+1)\widetilde{\mu}(\underline{u}_r)\big).
$$

Now we decompose $b_2'(\underline{u}_r,\underline{u}_{r'})$ into three terms $b_{21}'(\underline{u}_r,\underline{u}_{r'}) + b_{22}'(\underline{u}_r,\underline{u}_{r'}) + b_{23}'(\underline{u}_r,\underline{u}_{r'})$ as in Section A. The third term comes from the fact that a clump of $\underline{u}_r$ can overlap a clump of $\underline{u}_{r'}$ (but at most on

$(\min(\ell_r, \ell_{r'}) - 1)$ letters, assuming (A3)). To bound each of the three terms, it suffices to follow the lines detailed in Section A. This results in

$$\begin{aligned} b'_2(\underline{u}_r, \underline{u}_{r'}) \quad \le \quad & (n - \ell_r + 1)\mu(\underline{u}_r)\mu(\underline{u}_{r'}) \left(\frac{\Omega_{r,r'}(u_{r,\ell_r}, u_{r',1})}{\mu(u_{r',1})} + \frac{\Omega_{r',r}(u_{r',\ell'_r}, u_{r,1})}{\mu(u_{r,1})} \right) \\ & + (n - \ell_r + 1)\frac{\widetilde{\mu}(\underline{u}_r)}{\mu_{\min}} ((\ell - 2)\mu(\underline{u}_{r'}) + (3\ell - \ell_r - 2\ell_{r'} + 1)\widetilde{\mu}(\underline{u}_{r'})) \\ & + (n - \ell_r + 1)\frac{\widetilde{\mu}(\underline{u}_{r'})}{\mu_{\min}} ((\ell - 2)\mu(\underline{u}_r) + (3\ell - 2\ell_r - \ell_{r'} + 1)\widetilde{\mu}(\underline{u}_r)) \\ & + (n - \ell_r + 1)\mu(\underline{u}_r)\mu(\underline{u}_{r'})(M(\underline{u}_r, \underline{u}_{r'}) + M(\underline{u}_{r'}, \underline{u}_r)). \end{aligned}$$

Therefore, with the notation (34), we obtain $b'_2(\underline{u}_r) \le T_3(\underline{u}_r, \underline{u}_r)$ and $b'_2(\underline{u}_r, \underline{u}_{r'}) \le T_3(\underline{u}_r, \underline{u}_{r'})$, leading to

$$b_2 \le \sum_{1 \le r, r' \le m} T_3(\underline{u}_r, \underline{u}_{r'}). \tag{54}$$

Bounding $\mathbf{b_3}$: We decompose b_3 as $b_3 = \sum_{r=1}^m b'_3(\underline{u}_r)$, where

$$\begin{aligned} b'_3(\underline{u}_r) \quad & = \quad \sum_{i=1}^{n-\ell_r+1} \sum_{k \ge 1} \mathbb{E}|\mathbb{E}(\widetilde{\mathbb{I}}_{i,k}(\underline{u}_r) - \mathbb{E}\widetilde{\mathbb{I}}_{i,k}(\underline{u}_r) \mid \widetilde{\mathcal{Y}}_{i,k})|. \\ & \le \quad (n - \ell_r + 1)\gamma_2(\ell_r, \ell)\widetilde{\mu}(\underline{u}_r)|\alpha|^\ell. \end{aligned}$$

The last inequality is obtained in the same spirit as for (49). It leads to

$$b_3 \le |\alpha|^\ell \sum_{r=1}^m \gamma_2(\ell_r, \ell)(n - \ell_r + 1)\widetilde{\mu}(\underline{u}_r);$$

combining this inequality with (53), (54) and Theorem 1 completes the proof.

Acknowledgment: This work was carried out while S. Schbath was visiting the Center for Computational Biology, University of Southern California. The authors would like to thank Michael Waterman for his encouragement and support, and Terry Speed for suggesting the stem-loop example. Moreover the authors would like to thank one of the referees for many helpful comments. G. Reinert was partially supported by NSF grant DMS-9505075. S. Schbath was partially supported by NSF Grant BIR 95-04393 and NATO.

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.

BARBOUR, A. D. and UTEV, U. (1997). Compound Poisson approximation in total variation. Preprint

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

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.* **32** 877–892.

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

KARLIN, S. and Taylor, H. M. (1975). *A first course in stochastic processes.* Academic Press, 2^{nd} ed.

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

LUNDSTROM, R. (1990). Stochastic models and statistical methods for DNA sequence data. Ph.D. Thesis, Department of Mathematics, University of Utah.

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.

ROOS, M. and STARK, D. (1996). Compound Poisson approximation of the number of visits to a small set in a Markov chain. Preprint.

SCHBATH, S. (1995a). Compound Poisson approximation of word counts in DNA sequences. *ESAIM: Probability and Statistics.* **1** 1–16.(http://www.emath.fr/ps/).

SCHBATH, S. (1995b). *Étude asymptotique du nombre d'occurrences d'un mot dans une chaîne de Markov et application á 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.

SENOUSSI, R. (1990). Statistique asymptotique presque-sûre de modèles statistiques convexes, *Ann. Inst. Henri Poincaré* **26** "19–44".

TANUSHEV, M. (1996). Central limit theorem for several patterns in a Markov chain sequence of letters. Preprint.

WATERMAN, M. S. (1995). *Introduction to computational biology.* Chapman & Hall.