Compound Poisson approximations for occurrences of multiple words

Gesine Reinert* Sophie Schbath†

April 11, 2001

Abstract

A compound Poisson process approximation for the number of occurrences of multiple words in a sequence of letters is derived, where the letters are assumed to be independent and identically distributed. Using the Chen-Stein method, a bound on the error in the approximation is provided. For rare words, this error tends to zero as the length of the sequence increases to infinity. As an application the efficiency of the approximation for the number of occurrences of rare stem-loop motifs in DNA sequences is illustrated.

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

MSC 1991 classification. 60F05, 92D20

*Department of Statistics, UCLA, Los Angeles, CA 90095, USA. This work was partially supported by NSF grant DMS-9505075.

†Unité de Biométrie, INRA, 78352 Jouy-en-Josas, France. This work was partially supported by NATO, and by NSF grant BIR-9504393.

1 Introduction

When searching a database for the occurrence of a combination of several words within a sequence, the typical Poisson approximation used by programs like BLAST is no longer valid, as overlapping words may be dependent on each other. Here a compound Poisson approximation for the multiple occurrences of short words within a sequence is derived. Using the Chen-Stein method for Poisson process approximation, an explicit error bound for the approximation is given, improving those obtained by Schbath (1995a) for a single rare word. The approximation error increases with the amount of overlap between the words. The results are applied to the occurrences of stem-loop motifs. Another application might be a set of words coding for the same amino-acid sequence.

In general, consider a finite sequence S of letters chosen independently from a finite alphabet $\mathcal{A}$. The main example will be the four-letter DNA alphabet $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ but the 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). A compound Poisson approximation is obtained in Arratia *et al.* (1990), Geske *et al.* (1995) and Schbath (1995a) for the number of occurrences of a rare word, whereas the number of clumps of a rare word is approximated by a Poisson variable (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). 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 conditions for asymptotic independence. The multidimensional approximation and in particular the asymptotic statistical independence for counts of multiple words is very useful to study statistical properties of any function of these counts. (Note that asymptotic statistical independence does not necessarily have a biological interpretation.)

Instead of using the Chen-Stein method for Poisson process approximation as stated in Arratia *et al.* (1990) and refined in Barbour *et al.* (1992b), 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 simple self-overlapping structure in a two-letter alphabet (Roos and Stark (1996)), but it is not adapted for a multidimensional approximation to multiple words.

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) give 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) proved the multidimensional Gaussian approximation for an m-tuple of renewal counts.

The general case where the letters are modeled using a stationary Markov chain is treated in Reinert and Schbath (1998); Reinert and Schbath (1998) also give a Poisson process approximation. The purpose of this paper is to treat the independent case only, meaning that the letters are assumed to be independent and identically distributed, as under this additional assumption the arguments and bounds in Reinert and Schbath (1998) simplify considerably.

To approximate the counts of occurrences of words, the "declumping" approach is used - first the number of clumps of occurrences is counted, and then the sizes of the clumps are determined. In Section 2, an occurrence of a word and an occurrence of a clump in a sequence are defined, as well as the number of occurrences of a word and the number of clumps of a word in a sequence. Moreover the decomposition of the count of a word with respect to the number of clumps is introduced; this decomposition is fundamental to proving the compound Poisson approximation via the Chen-Stein method given in Arratia *et al.* (1990). Next, in Section 3 the compound Poisson process approximations for counts of m words with not necessarily identical lengths are presented. As an illustration (Section 4),the count of some stem-loop motifs are studied, like `ATGGCNNNNGCCAT` (`N` denotes any letter in the four-letter DNA alphabet), in a model for 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 useful tool to approximate a collection of counts.

2 Preliminary notation and the Chen-Stein method

Cnsider a sequence of i.i.d. letters $\mathcal{M} = \{X_i\}_{i\in\mathbb{Z}}$ on a finite alphabet $\mathcal{A}$, where the letters $\{X_i\}_{i\in\mathbb{Z}}$ are chosen independently with probabilities $\mathbb{P}(X_i = x) = \mu(x), x \in \mathcal{A}$; assume that $\mu(x) > 0\ \forall x \in \mathcal{A}$. 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 $\mathbb{I}_i(\underline{u})$. 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

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

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

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

and its expectation is

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

2.1 Overlaps and clumps

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 as

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

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 example (2) below) are the ones which are not a nontrivial multiple of 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 $\underline{u} = \texttt{AAGAAGAA}$.

In order to study the occurrences of a word $\underline{u}$ the concept of clumps of a word $\underline{u}$ is introduced. 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 corresponding indicator random variable by $\widetilde{\mathbb{I}}_i(\underline{u})$; i.e.

$$\widetilde{\mathbb{I}}_i(\underline{u}) \quad = \quad \mathbb{I}_i(\underline{u}) \prod_{j=i-\ell+1}^{i-1} (1 - \mathbb{I}_j(\underline{u})).$$

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

$$\widetilde{\mu}(\underline{u}) := \mathbb{E}\widetilde{\mathbb{I}}_i(\underline{u}) = \mu(\underline{u}) - \sum_{p \in \mathcal{P}'(\underline{u})} \mu(\underline{u}^{(p)}\underline{u}), \tag{2}$$

where $\underline{u}^{(p)} = u_1 u_2 \cdots u_p$ is the word composed of the first p letters of $\underline{u}$.

Here is the sketch of the proof for equation (2). 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 (2) then easily follows.

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

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

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

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

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 sequence $S = X_1X_2\ldots X_n$ is observable, and the observable number of clumps, denoted by $\widetilde{N}^*(\underline{u})$, may be different from $\widetilde{N}(\underline{u})$. In the above example, in S there is a unique clump of $\underline{u}$ starting at position 2 and ending at position 15, and $X_1 = \mathtt{T}$ 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 $\mathtt{G}$. The quantity of interest is $\widetilde{N}^*(\underline{u})$, the observable number of clumps, but here we will work with the count $\widetilde{N}(\underline{u})$ instead, since it is easier and the boundary effect can be controlled. Indeed, $\mathbb{P}(\widetilde{N}^*(\underline{u}) \neq \widetilde{N}(\underline{u}))$ is an upper bound on the total variation distance between $\widetilde{N}^*(\underline{u})$ and $\widetilde{N}(\underline{u})$ (see, e.g., Barbour *et al.* (1992b)), and

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

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 $\widetilde{N}(\underline{u})$; as approximate $\widetilde{N}(\underline{u})$ is approximated, 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 $\widetilde{N}^*(\underline{u}) = \widetilde{N}(\underline{u}) = N(\underline{u})$.*

2.2 Clumps of different sizes

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 the above 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 example,

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

for $\underline{u} = \texttt{AAGAAGAA}$. Note that the length of two k-clumps may differ a lot, and that the size of $\mathcal{C}_k(\underline{u})$ increases exponentially with k whenever $\underline{u}$ has more than one principal period. In fact,

$$|\mathcal{C}_k(\underline{u})| = |\mathcal{P}'(\underline{u})|^{k-1},$$

see Schbath (1995b).

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 corresponding indicator random variable by $\widetilde{\mathbb{I}}_{i,k}(\underline{u})$; Schbath (1995a) proved that its expectation is

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

where

$$A(\underline{u}) = \sum_{p\in\mathcal{P}'(\underline{u})} \prod_{j=1}^{p} \mu(u_{j+1}) = \sum_{p\in\mathcal{P}'(\underline{u})} \frac{\mu(\underline{u}^{(p+1)})}{\mu(u_1)}. \tag{3}$$

The derivation of (3) is similar to the one of (2). Note that

$$\sum_{k\geq 1} \widetilde{\mu}_k(\underline{u}) = \widetilde{\mu}(\underline{u}), \tag{4}$$

$$\sum_{k\geq 1} k\widetilde{\mu}_k(\underline{u}) = \mu(\underline{u}). \tag{5}$$

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

$$\widetilde{N}_k(\underline{u}) := \sum_{i=1}^{n-\ell+1} \widetilde{\mathbb{I}}_{i,k}(\underline{u}), \tag{6}$$

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 $\check{N}(\underline{u})$ defined by

$$\check{N}(\underline{u}) := \sum_{k\geq 1} k\widetilde{N}_k(\underline{u}) \tag{7}$$

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

$$\mathbb{P}(N(\underline{u}) \neq \check{N}(\underline{u})) \leq 2(\ell-1) \sum_{p \in \mathcal{P}'(\underline{u})} \mu(\underline{u}^{(p)}\underline{u}) = 2(\ell-1)\big(\mu(\underline{u}) - \widetilde{\mu}(\underline{u})\big). \tag{8}$$

The counts $N(\underline{u})$ and $\check{N}(\underline{u})$ have the same expectation because of (5). Now focus on $\check{N}(\underline{u})$ 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 $\check{N}(\underline{u}) = \widetilde{N}_1(\underline{u}) = N(\underline{u})$.*

2.3 The Chen-Stein method

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

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

where B and h are assumed to be measurable. The Chen-Stein method for Poisson approximation has been developed by Chen (1975); a friendly exposition is in Arratia *et al.* (1989) and Arratia *et al.* (1990); an exhaustive description with many examples can be found in 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 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

$$\begin{aligned} b_1 &= \sum_{\alpha\in I}\sum_{\beta\in B_\alpha} \mathbb{E}Y_\alpha \mathbb{E}Y_\beta && (9)\\ b_2 &= \sum_{\alpha\in I}\sum_{\beta\in B_\alpha,\beta\neq\alpha} \mathbb{E}(Y_\alpha Y_\beta) && (10)\\ b_3 &= \sum_{\alpha\in I} \mathbb{E}|\mathbb{E}\{Y_\alpha - p_\alpha|\sigma(Y_\beta, \beta\notin B_\alpha)\}|\,. \end{aligned}$$

We think of B_α as a neighborhood of strong dependence of Y_α. Intuitively, b_1 describes the contribution related to the size of the neighborhood and the weights of the random variables in that neighborhood; if all Y_α had the same probability of success, then b_1 would be directly proportional to the neighborhood size. The term b_2 accounts for the strength of the dependence inside the neighborhood; as it depends on the second moments, it can be viewed as a "second order interaction" term. Finally, b_3 is related to the strength of dependence of Y_α with random variables outside its neighborhood. In particular, $b_3 = 0$ if Y_α is independent of $\sigma(Y_\beta, \beta \notin B_\alpha\}$.

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 $|\mathbb{E}h(\underline{Y}) - \mathbb{E}h(\underline{Z})| \leq d_{\mathrm{TV}}(\mathcal{L}(\underline{Y}), \mathcal{L}(\underline{Z}))$. Thus, if $T(\underline{Y})$ is a test statistic then, for all $t \in \mathbb{R}$,

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

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

3 Occurrences of m words of different lengths

3.1 Notation

Now consider m different words $\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_m$ of length $\ell_1, \ell_2, \ldots, \ell_m$, respectively;

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

Assume that

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

Clumps of $\underline{u}_r$ and clumps of $\underline{u}_{r'}$ may overlap in the sequence. Assumption (A) guarantees that a clump of $\underline{u}_r$ and a clump of $\underline{u}_{r'}$ can overlap on at most $\max\{\ell_r, \ell_{r'}\} - 1$ letters. Heuristically, if $\underline{u}_r$ and $\underline{u}_{r'}$ do not satisfy Assumption (A), the approximation of their counts by independent Poisson variables should not be valid.

In order to describe the possible overlaps between two words $\underline{u}_r$ and $\underline{u}_{r'}$, define

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

Under Assumption (A) , 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). Note the lack of symmetry; for example, for $\underline{u} = \texttt{AAGAAGAA}$ and $\underline{v} = \texttt{AAGAATCA}$, it follows that $\mathcal{P}(\underline{u}, \underline{v}) = \{3, 6, 7\}$ and $\mathcal{P}(\underline{v}, \underline{u}) = \{7\}$.

For a bound on the error in the compound Poisson approximation, the following quantities, defined for all r and $r' \in \{1, \ldots, m\}$, are needed.

$$M(\underline{u}_r, \underline{u}_{r'}) = \mathbf{1}(r \neq r') \sum_{p \in \mathcal{P}(\underline{u}_r, \underline{u}_{r'})} \frac{1}{\mu(\underline{u}_{r'}^{(\ell_r - p)})} \tag{11}$$

$$R(\underline{u}_r, \underline{u}_{r'}) = (n - \ell_r + 1)\{(\ell_r - 1)\mu(\underline{u}_r)\widetilde{\mu}(\underline{u}_{r'}) + (\ell_{r'} - 1)\widetilde{\mu}(\underline{u}_r)\mu(\underline{u}_{r'}) + (2\ell_r + 2\ell_{r'} - 3)\widetilde{\mu}(\underline{u}_r)\widetilde{\mu}(\underline{u}_{r'})\} \tag{12}$$

$$T(\underline{u}_r, \underline{u}_{r'}) = (n - \ell_r + 1)\mu(\underline{u}_r)\mu(\underline{u}_{r'})\{2(\ell_r + \ell_{r'} - 2) + M(\underline{u}_r, \underline{u}_{r'}) + M(\underline{u}_{r'}, \underline{u}_r)\}. \tag{13}$$

The quantity $M(\underline{u}_r, \underline{u}_{r'})$ can be seen as a measure of the overlapping structure between $\underline{u}_r$ and $\underline{u}_{r'}$. If $\underline{u}_r$ and $\underline{u}_{r'}$ cannot overlap, $M(\underline{u}_r, \underline{u}_{r'})$ equals zero; otherwise, the more they can overlap from the right, the larger is $M(\underline{u}_r, \underline{u}_{r'})$. The quantities R and T correspond to the quantities R and T in Reinert and Schbath (1998). It will turn out that R is used to describe the "neighborhood size term" b_1, whereas T describes the "second order interaction term" b_2, when applying Theoren 1.

Moreover introduce the set of possible words of length $\ell - 1$ preceding a clump of $\underline{u}$;

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

Similarly, $\mathcal{D}(\underline{u})$ 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 $\underline{C}$ and $\underline{C}'$ in $\mathcal{C}_k(\underline{u})$ can occur simultaneously at position i. Therefore,

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

Note that

$$\sum_{k \geq 1} \sum_{\underline{g} \in \mathcal{G}(\underline{u}),\, \underline{C} \in \mathcal{C}_k(\underline{u})} \mu(\underline{gC}) = \sum_{k \geq 1} \sum_{k^* \geq k} \widetilde{\mu}_{k^*}(\underline{u}) = \mu(\underline{u}). \tag{14}$$

3.2 Results

Enumerate the elements of $\mathcal{C}_k(\underline{u}_r)$ from C_1 to $C_{|\mathcal{P}'(\underline{u}_r)|^{k-1}}$. Choose the index set

$$I = \{(i, r, k, c) : 1 \leq i \leq n, r = 1, \ldots, m, k = 1, 2, \ldots, c = 1, \ldots, |\mathcal{P}'(\underline{u}_r)|^{k-1}\}.$$

For each $(i, r, k, c) \in I$, define the Bernoulli process $\underline{\widetilde{\mathbb{Y}}} = (\widetilde{Y}_{(i,r,k,c)})_{(i,r,k,c) \in I}$ by

$$\widetilde{Y}_{(i,r,k,c)} \quad = \sum_{\underline{g} \in \mathcal{G}(\underline{u}_r), \underline{d} \in \mathcal{D}(\underline{u}_r)} \mathbb{I}_{i-\ell_r+1}(\underline{gC_cd}).$$

Thus $\widetilde{Y}_{(i,r,k,c)}$ equals 1 if and only if a specific clump $\underline{C}_c$ of size k of word $\underline{u}_r$ occurs at position i in the sequence. Furthermore define the Poisson process $\underline{\mathbb{Z}} = (Z_{i,r,k,c})_{(i,r,k,c)\in I}$ by having independent components and each component being Poisson distributed with mean $\mathbb{E}\widetilde{Y}_{(i,r,k,c)}$. We think of $Z_{i,r,k,c}$ approximating the indicator random variable $\widetilde{Y}_{(i,r,k,c)}$.

For $(i,r,k,c) \in I$, choose as neighborhood

$$B_{i,r,k,c} := \{(j,r',k',c') \in I : -|\underline{C}_{c'}| - \ell_{r'} - \ell_r + 3 \leq j - i \leq |\underline{C}_c| + \ell_r + \ell_{r'} - 3\}.$$

Theorem 2 *Under Assumption (A) and with the notation (12) and (13), we have*

$$d_{TV}\left(\mathcal{L}(\underline{\widetilde{\mathbb{Y}}}), \mathcal{L}(\underline{\mathbb{Z}})\right) \leq \sum_{1\leq r,r'\leq m} \{R(\underline{u}_r, \underline{u}_{r'}) + T(\underline{u}_r, \underline{u}_{r'})\}.$$

Let $(Z_k^{(r)})_{k\geq 1, r\in\{1,\dots,m\}}$ be independent Poisson variables with expectation $\mathbb{E}Z_k^{(r)} = (n - \ell_r + 1)\widetilde{\mu}_k(\underline{u}_r)$. From Theorem 2 and Equation (6), the following corollary is easily obtained.

Corollary 1 *Under Assumption (A) and with the notation (12) and (13), we have*

$$d_{TV}\Big(\mathcal{L}\big((\widetilde{N}_k(\underline{u}_r))_{k\geq 1, r\in\{1,\dots,m\}}\big), \mathcal{L}\big((Z_k^{(r)})_{k\geq 1, r\in\{1,\dots,m\}}\big)\Big) \quad \leq \quad \sum_{1\leq r,r'\leq m} \{R(\underline{u}_r, \underline{u}_{r'}) + T(\underline{u}_r, \underline{u}_{r'})\}.$$

Now let $CP^{(r)}$ denote the compound Poisson distribution of $\sum_{k\geq 1} kZ_k^{(r)}$. From Theorem 2, Equation (7) and Inequality (8), we have the following two corollaries.

Corollary 2 *Under Assumption (A) and with the notation (12) and (13) we have*

$$\begin{aligned} (i) \quad & d_{TV}\Big(\mathcal{L}\big(\check{N}(\underline{u}_1), \dots, \check{N}(\underline{u}_m)\big), \mathrm{CP}^{(1)} \otimes \cdots \otimes \mathrm{CP}^{(m)}\Big) && \leq \sum_{1\leq r,r'\leq m} \{R(\underline{u}_r, \underline{u}_{r'}) + T(\underline{u}_r, \underline{u}_{r'})\}, \\ (ii) \quad & d_{TV}\Big(\mathcal{L}\big(N(\underline{u}_1), \dots, N(\underline{u}_m)\big), \mathrm{CP}^{(1)} \otimes \cdots \otimes \mathrm{CP}^{(m)}\Big) && \leq \sum_{1\leq r,r'\leq m} \{R(\underline{u}_r, \underline{u}_{r'}) + T(\underline{u}_r, \underline{u}_{r'})\} \\ & && \quad +2\sum_{r=1}^{m} (\ell_r - 1)\big(\mu(\underline{u}_r) - \widetilde{\mu}(\underline{u}_r)\big). \end{aligned}$$

Let $(Z_k)_{k\geq 1}$ be independent Poisson variables with expectation $\mathbb{E}Z_k = \sum_{r=1}^{m}(n - \ell_r + 1)\widetilde{\mu}_k(\underline{u}_r)$, and let CP denote the compound Poisson distribution of $\sum_{k\geq 1} kZ_k$.

Corollary 3 *Under Assumption (A) and with the notation (12) and (13), we have*

$$d_{TV}\left(\mathcal{L}\left(\sum_{r=1}^{m} N(\underline{u}_r)\right), \mathrm{CP}\right) \quad \leq \quad \sum_{1\leq r,r'\leq m} \{R(\underline{u}_r,\underline{u}_{r'}) + T(\underline{u}_r,\underline{u}_{r'})\} + 2\sum_{r=1}^{m}(\ell_r-1)\big(\mu(\underline{u}_r)-\widetilde{\mu}(\underline{u}_r)\big).$$

Note that the compound Poisson distributions defined in Corollaries 2 and 3 reduce to simple Poisson distributions if every $\underline{u}_r$, $r=1,\dots,m$, is a non self-overlapping word.

Poisson approximations should be good for rare events. Here, rare words mean that $\mathbb{E}N(\underline{u}_r)$ is bounded away from 0 and ∞ for $r=1,\dots,m$; we use the notation $\mathbb{E}N(\underline{u}_r)\asymp 1$. For a fixed alphabet, this asymptotic framework is equivalent to $\mu(\underline{u}_r)\asymp\frac{1}{n}$, and $\ell_r\asymp\log n$ because

$$\mu(u_1)(\min_{x\in\mathcal{A}}\mu(x))^{\ell_r-1}\leq\mu(\underline{u}_r)\leq\mu(u_1)(\max_{x\in\mathcal{A}}\mu(x))^{\ell_r-1}.$$

If $\mathbb{E}N(\underline{u}_r)\asymp 1$, then $T(\underline{u}_r,\underline{u}_{r'})\asymp n^{-1}\log n+(M(\underline{u}_r,\underline{u}_{r'})+M(\underline{u}_{r'},\underline{u}_r))n^{-1}$, and

$$d_{\mathrm{TV}}\left(\mathcal{L}(\widetilde{\mathbb{Y}}),\mathcal{L}(\underline{\mathbb{Z}})\right)\leq \mathrm{O}\left(n^{-1}\log n\right)+\mathrm{O}\left(\sum_{r\neq r'}(M(\underline{u}_r,\underline{u}_{r'})+M(\underline{u}_{r'},\underline{u}_r))n^{-1}\right).$$

Therefore, if $\underline{u}_r$ and $\underline{u}_{r'}$ cannot overlap too much for $r\neq r'$ (this is measured by $M(\underline{u}_r,\underline{u}_{r'})+M(\underline{u}_{r'},\underline{u}_r)$), the error bound is very small for large n. In the extreme case where $\underline{u}_1=\mathtt{AAA}\cdots\mathtt{AAA}$ and $\underline{u}_2=\mathtt{TAA}\cdots\mathtt{AAA}$, both of length ℓ, we have $M(\underline{u}_2,\underline{u}_1)=\sum_{s=1}^{\ell-1}\mu(\mathtt{A})^{-s}\asymp\mu(\mathtt{A})^{-\log n}$, 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 $\underline{u}_1$ and $\underline{u}_2$ should not be independent even asymptotically.

3.3 Proof of Theorem 2

Our task consists now in bounding b_1 and b_2 given in (9) and (10). For any $\underline{C}\in\mathcal{C}_k(\underline{u}_r)$ it follows that $|\underline{C}|\leq k(\ell_r-1)$. This motivates the introduction of the set

$$B_{i,k,k',r,r'}:=\{j\in\{1,\dots,n\}: i-(k'+1)(\ell_{r'}-1)+\ell_r-1\leq j\leq i+(k+1)(\ell_r-1)+\ell_{r'}-1\};\quad (15)$$

for each fixed i,r,r',k,k',c,c', thus $\{j:(j,r',k',c')\in B_{i,r,k,c}\}\subset B_{i,k,k',r,r'}$.

Bounding b_1: We have

$$\begin{aligned} b_1 &= \sum_{(i,r,k,c)\in I}\sum_{(j,r',k',c')\in B_{i,r,k,c}} \mathbb{E}\widetilde{Y}_{i,r,k,c}\mathbb{E}\widetilde{Y}_{j,r',k',c'} \\ &\le \sum_{r,r'=1}^{m}\sum_{k,k'\ge 1}\sum_{i=1}^{n-\ell_r+1}\sum_{j\in B_{i,k,k',r,r'}} \widetilde{\mu}_k(\underline{u}_r)\widetilde{\mu}_{k'}(\underline{u}_{r'}) \\ &\le \sum_{r,r'=1}^{m}(n-\ell_r+1)\sum_{k,k'\ge 1}\widetilde{\mu}_k(\underline{u}_r)\widetilde{\mu}_{k'}(\underline{u}_{r'})((k+2)(\ell_r-1)+(k'+2)(\ell_{r'}-1)+1), \end{aligned}$$

where $B_{i,k,k',r,r'}$ is defined in (15). Now use (4) and (5) to obtain

$$\begin{aligned} b_1' &\le \sum_{r,r'=1}^{m}(n-\ell_r+1)\{(\ell_r-1)\mu(\underline{u}_r)\widetilde{\mu}(\underline{u}_{r'})+(\ell_{r'}-1)\widetilde{\mu}(\underline{u}_r)\mu(\underline{u}_{r'})+(2\ell_r+2\ell_{r'}-3)\widetilde{\mu}(\underline{u}_r)\widetilde{\mu}(\underline{u}_{r'})\} \\ &= \sum_{r,r'=1}^{m} R(\underline{u}_r,\underline{u}_{r'}). \end{aligned}$$

Bounding b_2: To bound b_2 write

$$b_2 = \sum_{(i,r,k,c)\in I}\sum_{(j,r',k',c')\in B_{i,r,k,c}\setminus\{(i,r,k,c)\}} \mathbb{E}\widetilde{Y}_{i,r,k,c}\widetilde{Y}_{j,r',k',c'} = \sum_{r,r'=1}^{m} b_2'(\underline{u}_r,\underline{u}_{r'}),$$

with

$$b_2'(\underline{u}_r,\underline{u}_{r'}) = \sum_{i=1}^{n-\ell_r+1}\sum_{k,k'\ge 1}\sum_{j:(j,r',k',c')\in B_{i,r,k,c}\setminus\{(i,r,k,c)\}} \mathbb{E}\widetilde{Y}_{i,r,k,c}\widetilde{Y}_{j,r',k',c'}.$$

Now distinguish two cases: The first one when the k-clump starting at i and the k'-clump starting at j overlap in $\mathcal{M}$, and the second one when the clumps do not overlap but the "enlarged clumps" – including the (ℓ_r-1) preceding letters and the (ℓ_r-1) following letters of the clumps – overlap. Write $\underline{C}=\underline{C}_c$, and $\underline{C}'=\underline{C}_{c'}$, for convenience.

1. First consider the case when $\underline{C}$ and $\underline{C}'$ overlap in the sequence, that is,

$$j\in\{i+|\underline{C}|-\ell_r,\dots,i+|\underline{C}|-1\}\cup\{i-|\underline{C}'|+1,\dots,i-|\underline{C}'|+\ell_{r'}\}.$$

   Let $b_{21}'(\underline{u}_r,\underline{u}_{r'})$ denote the quantity corresponding to this case. Since two clumps of $\underline{u}_r$ cannot overlap in the sequence, the composed words $\underline{C}$ and $\underline{C}'$ starting at i and j cannot overlap. Therefore we may restrict ourselves to the case $r\ne r'$. First focus on $j>i$. If $\underline{C}$ and $\underline{C}'$ overlap, Assumption

(A) ensures that only the last occurrence of $\underline{u}_r$ in $\underline{C}$ overlaps with the first occurrence of $\underline{u}_{r'}$ in $\underline{C}'$. The last occurrence of $\underline{u}_r$ in $\underline{C}$ starts at position $i+|\underline{C}|-\ell_r$. An occurrence of $\underline{u}_{r'}$ starting at position j may overlap $\underline{u}_r$ at $i+|\underline{C}|-\ell_r$ only if $j=i+|\underline{C}|-\ell_r+p$ with $p\in\mathcal{P}(\underline{u}_r,\underline{u}_{r'})$. Therefore,

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

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

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

Finally, using (14) and applying the same reasoning to $j<i$ yields

$$b'_{21}(\underline{u}_r,\underline{u}_{r'})\leq(n-\ell_r+1)\mu(\underline{u}_r)\mu(\underline{u}_{r'})(M(\underline{u}_r,\underline{u}_{r'})+M(\underline{u}_{r'},\underline{u}_r)),\tag{16}$$

where M is given by (11).

2. $\underline{gCd}$ and $\underline{g'C'd'}$ overlap in the sequence (at most on $\ell_r+\ell_{r'}-2$ letters), but $\underline{C}$ and $\underline{C}'$ do not overlap, that is,

$$j\in\{i-|\underline{C}'|-\ell_{r'}-\ell_r+3,\ldots,i-|\underline{C}'|\}\cup\{i+|\underline{C}|,\ldots,i+|\underline{C}|+\ell_r+\ell_{r'}-3\}.$$

Denote the corresponding quantity by $b'_{22}(\underline{u}_r, \underline{u}_{r'})$. For the case that $j > i$ it follows

$$\begin{aligned}
&\sum_{i=1}^{n-\ell_r+1} \sum_{k,k'\geq 1} \sum_{\substack{\underline{g}\in\mathcal{G}(\underline{u}_r),\, \underline{d}\in\mathcal{D}(\underline{u}_r) \\ \underline{C}\in\mathcal{C}_k(\underline{u}_r)}} \sum_{j=i+|\underline{C}|}^{i+|\underline{C}|+\ell_r+\ell_{r'}-3} \mathbb{E}\mathbb{I}_{i-\ell_r+1}(\underline{gCd})\widetilde{\mathbb{I}}_{j,k'}(\underline{u}_{r'}) \\
\leq\; & \sum_{i=1}^{n-\ell+1} \sum_{k,k'\geq 1} \sum_{\substack{\underline{g}\in\mathcal{G}(\underline{u}_r) \\ \underline{C}\in\mathcal{C}_k(\underline{u}_r)}} \sum_{j=i+|\underline{C}|}^{i+|\underline{C}|+\ell_r+\ell_{r'}-3} \mathbb{E}\mathbb{I}_{i-\ell_r+1}(\underline{gC})\mathbb{I}_{j,k'}(\underline{u}_{r'}) \\
=\; & \sum_{i=1}^{n-\ell_r+1} \sum_{k\geq 1} \sum_{\substack{\underline{g}\in\mathcal{G}(\underline{u}) \\ \underline{C}\in\mathcal{C}_k(\underline{u})}} \sum_{j=i+|\underline{C}|}^{i+|\underline{C}|+\ell_r+\ell_{r'}-3} \mu(\underline{gC})\mu(\underline{u}_{r'}).
\end{aligned}$$

The case $j < i$ is treated analogously. Summing and using (14) leads to

$$b'_{22}(\underline{u}_r, \underline{u}_{r'}) \quad \leq \quad 2(n-\ell_r+1)(\ell_r+\ell_{r'}-2)\mu(\underline{u}_r)\mu(\underline{u}_{r'}). \tag{17}$$

Combining (17), and (16) gives $b'_2(\underline{u}, \underline{v}) \leq T(\underline{u}, \underline{v})$. ■

4 Application

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

$$a_1 a_2 \cdots a_r (\mathtt{N})_s \overline{a_r} \cdots \overline{a_2}\,\overline{a_1}, \tag{18}$$

where the integers r and s are fixed, and $a_i \in \mathcal{A}$ for $i = 1, \ldots, r$. Here N represents any letter in the alphabet $\mathcal{A}$, $(\mathtt{N})_s$ denotes s consecutive, possibly different letters N, and for each $i = 1, \ldots, r$, $\overline{a_i}$ is the complementary letter of a_i (A is the complementary letter of T and vice versa, and C is the complementary letter of G and vice versa). For example, the motif AGGCNNGCCT is such a motif, 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 $a_1 a_2 \cdots a_r$ 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

To assess its extent, the number of occurrences of these motifs in a DNA sequence are approximated. 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 AGGCabGCCT, $a, b \in \mathcal{A}$:

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

The expected count of AGGCNNGCCT in a sequence is

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

where the right-hand terms are given in (1). Using Corollary 3, we calculate the error bound for approximating the count $N(\mathtt{AGGCNNGCCT})$ by a compound Poisson variable $\sum_{k\geq 1} kZ_k$ such that the Z_k's are independent Poisson variables with expectation

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

where A is given in (3).

In the application below, a sequence of length $n = 48502$ on the alphabet $\mathcal{A}$ is considered, with the letter probabilities

$$\mu(A) = .2544 \quad \mu(C) = .2342 \quad \mu(G) = .2643 \quad \mu(T) = .2471\,. \tag{19}$$

These values correspond approximately to the genome of the bacteriophage *Lambda*.

Table 1 gives the expected counts in the sequence of the motifs AGGCGCCT, ATGCGCAT, ATGGCGCCAT, and ATTGGCGCCAAT; the first two nonzero digits are given. Note that inserting N's in the word does not change their expected count, so that the expected count of AGGC$(\mathtt{N})_3$GCCT, for example, equals the expected count of AGGCGCCT, which is 0.72. All the above motifs have a very small expected count, which is in agreement with the rare word condition. Naturally, the expected counts decrease with increasing sequence length.

Table 1

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 3, even though relevant biological loop sizes are slightly larger. This result is particularly interesting since increasing s means enormously increasing the number of words in our family, which is of course penalizing.

Comparing the results for the motifs $\texttt{AGGC(N)}_s\texttt{GCCT}$ and $\texttt{ATGC(N)}_s\texttt{GCAT}$, the bounds for $\texttt{ATGC(N)}_s\texttt{GCAT}$ are larger 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 in the motif.

Figure 2

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 $\{5\}$.

Figure 3

A motif composed of words with many overlaps between themselves will produce large quantities M and hence large quantities T (given by (13)). The overlapping structure of a motif has an important influence on the error bound through the terms T. This explains the larger bounds for $\texttt{ATGC(N)}_s\texttt{GCAT}$.

However, Table 2 shows that adding just one letter to the stem of the motifs $\texttt{ATGC(N)}_s\texttt{GCAT}$, in such a way that the high overlapping structure of the motif is preserved, for instance yielding $\texttt{ATGGC(N)}_s\texttt{GCCAT}$, is sufficient to reduce the global error bound considerably to the order of 10^{-6}. Adding another letter leads to an error bound of order 10^{-8} for $\texttt{ATTGGC(N)}_s\texttt{GCCAAT}$. This illustrates that the approximation improves with increasing word length, i.e. with decreasing expected count of the motif (see Table 1).

Now focus on the weight of each of the three terms appearing in the global error bound given by Corollary 3. The three terms correspond, respectively, to the bounds of b_1, b_2, and the boundary effect calculated in Section 3. Table 2 gives the bound for each term and for the motifs $\texttt{AGGC(N)}_s\texttt{GCCT}$, $\texttt{ATGC(N)}_s\texttt{GCAT}$, $\texttt{ATGGC(N)}_s\texttt{GCCAT}$, and $\texttt{ATTGGC(N)}_s\texttt{GCCAAT}$, s varying from 1 to 3. It is obvious that the boundary effect is negligible and decreases smoothly as s increases. On the other hand, b_1 and b_2 are the main terms, and their bounds 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 $\texttt{AGGC(N)}_s\texttt{GCCT}$, $s = 1, \ldots, 3$; the explanation is that these motifs are composed of non-self-overlapping words (see (8)).

Table 2

In Reinert and Schbath (1998), corresponding bounds are calculated for the case that the sequence is not composed of i.i.d. letters but generated by a stationary Markov case. In comparison, the bounds for the independent model are orders of magnitude smaller.

Acknowledgment: The first author would like to thank the organizers for the opportunity to present this paper. Moreover we would like to thank Michael Waterman and Terry Speed for very helpful discussions. Finally the anonymous referees deserve thanks for many useful suggestions.

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.

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.

REINERT, G. and Schbath, S. 1998. Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains. Preprint.

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.

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.

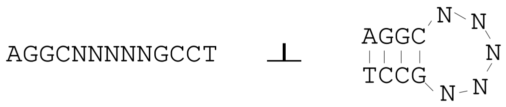

Figure 1: Stem-loop structure

A	T	G	C	a	b	c	G	C	A	T									
									A	T	G	C	a'	b'	c'	G	C	A	T
									$\uparrow$										
									9										
A	T	G	C	a	A	T	G	C	A	T									
					A	T	G	C	A	T	b	G	C	A	T				
					$\uparrow$														
					5														

Figure 2: Self-overlaps of the family ATGCNNNGCAT

A	G	G	C	a	A	G	G	C	C	T					
					A	G	G	C	C	T	b	G	C	C	T
					$\uparrow$										
					5										

Figure 3: Self-overlaps of the family AGGCNNNGCCT

	expected count
AGGCGCCT	0.72
ATGCGCAT	0.73
ATGGCGCCAT	4.5e-02
ATTGGCGCCAAT	2.8e-03

Table 1: Expected counts of some stem-loop motifs, in a sequence of length 48,502 of i.i.d. letters generated by (19)

	b_1	b_2	boundary effect	global bound
AGGCNGCCT	4.4e-04	3.4e-04	0	7.8e-04
AGGCNNGCCT	5.4e-04	4.3e-04	0	9.8e-04
AGGCNNNGCCT	5.4e-04	7.7e-04	0	1.3e-03
ATGCNGCAT	4.5e-04	6.2e-04	1.4e-08	1.0e-03
ATGCNNGCAT	5.1e-04	7.3e-04	7.0e-08	1.2e-03
ATGCNNNGCAT	5.6e-04	1.1e-03	1.1e-09	1.7e-03
ATGGCNGCCAT	2.1e-06	2.7e-06	6.9e-11	4.8e-06
ATGGCNNGCCAT	2.3e-06	3.1e-06	1.9e-11	5.5e-06
ATGGCNNNGCCAT	2.5e-06	4.7e-06	5.2e-12	7.3e-06
ATTGGCNGCCAGT	1.0e-08	1.2e-08	3.3e-13	2.2e-08
ATTGGCNNGCCAGT	1.1e-08	1.3e-08	9.0e-14	2.4e-08
ATTGGCNNNGCCAGT	1.1e-08	1.4e-08	2.4e-14	2.6e-08

Table 2: 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 of i.i.d. letters generated by (19)