WAITING TIMES FOR CLUMPS OF PATTERNS AND FOR STRUCTURED MOTIFS IN RANDOM SEQUENCES

VALERI T. STEFANOV[1], *The University of Western Australia*
STEPHANE ROBIN[2], *Institut National Agronomique Paris-Grignon, France*
SOPHIE SCHBATH[3], *Institut National de la Recherche Agronomique, France*

June 16, 2005

Keywords: pattern; clump; probability generating function; random sum; structured motif

AMS 2000 Subject Classification: Primary 60E10
Secondary 60K20

Abstract

This paper provides exact probability results for waiting times associated with occurrences of two types of motifs in a random sequence. First, we provide an explicit expression for the probability generating function of the interarrival time between two clumps of a pattern. It allows, in particular, to measure the quality of the Poisson approximation which is currently used for evaluation of the distribution of the number of clumps of a pattern. Second, we provide explicit expressions for the probability generating functions of both the waiting time until the first occurrence, and the interarrival time between consecutive occurrences, of a structured motif. Distributional results for structured motifs are of interest in genome analysis because such motifs are promoter candidates. As an application, we determine significant structured motifs in a data set of DNA regulatory sequences.

1 Introduction

Distributions associated with pattern occurrences in a random sequence of letters have been extensively studied in the literature. Genome analysis is a most popular application area for such results (see Reinert *et al.* (2000) for a recent survey). The exact distribution of the number of occurrences of a pattern is usually obtained through the distribution of the waiting time until the j-th occurrence of the pattern. The latter distribution is derived either by recursive formulas or through its probability generating function. The probability generating function approach leads to considering the

[1] Postal address: School of Mathematics and Statistics, The University of Western Australia, Crawley (Perth) 6009, W. A., Australia. Email address: stefanov@maths.uwa.edu.au

[2] ENGREF / INA PG / INRA unit of Applied Mathematics and Computer Sciences 16, rue Claude Bernard, F-75005, Paris, France. Email address: robin@inapg.inra.fr

[3] Unité MIG, INRA, 78352 Jouy-en-Josas, France. Email address: sophie.schbath@jouy.inra.fr

waiting time until the first occurrence of a pattern and the interarrival time between two consecutive occurrences of a pattern. Explicit and calculable expressions for the probability generating functions of these quantities in Markov sequences and for a single pattern or a set of patterns are found in Robin and Daudin (1999), Robin and Daudin (2001) and Stefanov (2003).

Pattern occurrences may overlap in a sequence, but one may be interested in counting nonoverlapping occurrences of a pattern. There are two ways for such counting (*i*) counting renewals or (*ii*) counting clumps of a pattern.

In Section 2 below we consider clumps. For renewals see Feller (1968), Régnier (2000), Chryssaphinou et al. (2001) and the references therein. A clump of a pattern is a maximal set of overlapping occurrences of the pattern in a sequence. Poisson approximation results exist for the distribution of the number of clumps (or declumped count) and these are theoretically valid when the sequence is long and the pattern is rare enough (Schbath (1995), Reinert and Schbath (1998)). There is no exact result for the distribution of the number of clumps in the literature. In Section 2, we provide an explicit expression for the probability generating function of the waiting time until the next clump occurrence (that is, the interarrival time between two consecutive clump occurrences). This leads to the exact evaluation of the distribution of the declumped count of a pattern.

In Section 3 below we study the waiting time until the first occurrence of a more complex pattern called a structured motif. A structured motif is composed of two patterns separated by a variable distance. The interest in this waiting time is due to the biological challenge of identifying promoter motifs along genomes. Programs to extract automatically structured motifs from DNA sequences exist (cf. Marsan and Sagot (2000), Eskin and Pevzner (2002), Morgante *et al.* (2004)). Only statistically significant motifs should be suggested to biologists as candidate promoters. The statistical significance of a motif in a sequence is identified through the probability that the sequence contains at least one occurrence of the motif. Robin *et al.* (2002) provides an approximation to this probability. In Section 3 we provide explicit expressions for the probability generating functions of (i) the first arrival time of a structured motif, and (*ii*) the intersite distance between consecutive occurrences of structured motifs. This leads to exact evaluation of the aforementioned probability. These are the first exact probability results on structured motifs in the literature. Note that our definition of a structured motif is slightly different from the usual one (cf. Robin *et al.* (2002)) but it accommodates all cases of structured motifs as long as the patterns involved in a structured motif do not appear too frequently in the considered sequences. This is usually the case in practice.

In Section 4 we provide two applications to DNA sequences.

2 Clumps of a pattern

Let $\{X(n)\}_{n\geq 0}$ be an ergodic finite-state Markov chain with discrete-time parameter, state space $\{1, 2, \ldots, N\}$, and one-step transition probabilities $p_{i,j}$, $i, j = 1, 2, \ldots, N$.

The pattern (word) of interest is $\mathbf{w} = w_1 w_2 \cdots w_k$, where $1 \leq w_i \leq N,\ i = 1, 2, \ldots, k$. For $j \in \{1, 2, \ldots, k\}$, denote the probability generating function[4] (p.g.f.) of the waiting time to reach the pattern $w_1 w_2 \cdots w_j$ from state s by $G_j^{(s)}(t)$ when we allow the initial state s to contribute to the pattern and by $\tilde{G}_j^{(s)}(t)$ when we do not allow s to contribute. Denote by $G_j^{(w_1, w_2, \ldots, w_r)}(t), 1 \leq r \leq j$, the p.g.f. of the waiting time to reach the pattern $w_1 w_2 \cdots w_j$, given the pattern $w_1 w_2 \cdots w_r$ has already been reached (note that $G_j^{(w_1, w_2, \ldots, w_j)}(t) = 1$). Introduce the indicator functions

$$Y_i = \mathbb{I}\{\text{an occurrence of } \mathbf{w} \text{ ends at position } i \text{ in the sequence}\}.$$

Denote by τ_k the first return time to the pattern $w_1 w_2 \cdots w_k$, that is

$$\tau_k = \inf\{n \geq 1 : Y_{i+n} = 1 | Y_i = 1\}.$$

Of course, τ_k represents the distance between two successive occurrences of the pattern (cf. figure 1). The possible values of τ_k are $1, 2, \ldots$. Let

$$c_i = P(\tau_k = i), \quad i = 1, 2, \ldots \tag{2.1}$$

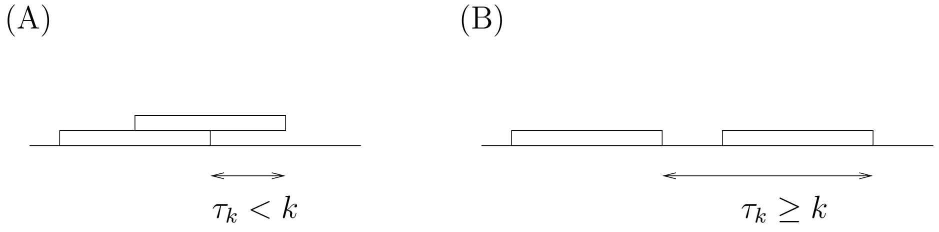


Figure 1: Waiting times τ_k between two successive occurrences of $w_1 w_2 \cdots w_k$: overlapping case (A), non-overlapping case (B).

The overlapping structure of the pattern dictates which of the $c_i,\ i \in \{1, 2, \ldots, k-1\}$, are nonzero. For instance, if $\mathbf{w} = 33133$ then only c_1 and c_2 are zero. Of course, if there is no proper prefix to be also a suffix of the pattern $w_1 w_2 \cdots w_k$ then $c_i = 0$, for all $i \in \{1, 2, \ldots k-1\}$ (cf. figure 1.B). The c_i's can be obtained recursively from Robin and Daudin (1999) or calculated after expanding in a series, up to k terms, the p.g.f. $G_{\tau_k}(t)$ of τ_k. An explicit expression for $G_{\tau_k}(t)$ can be found in the previous reference. Also one may derive such easily using the automated approach introduced in Stefanov (2003). Clearly the p.g.f. of τ_k is equal to $G_k^{(w_1, w_2, \ldots, w_J)}(t)$, where the $G_k^{(\cdot)}(\cdot)$ have been introduced a few lines earlier and J is the largest integer such that $w_1 w_2 \cdots w_J$ is both a proper prefix and suffix to the pattern $w_1 w_2 \cdots w_k$. For instance, if $\mathbf{w} = 33133$ then, $J = 2$. The integer $k - J$ is also called the minimal period of the pattern $w_1 w_2 \cdots w_k$ in the terminology introduced by Guibas and Odlyzko (1981).

[4] Recall that the probability generating function of a discrete random variable Y on $\{0, 1, 2, \ldots\}$ is defined by $G_Y(t) := \sum_{i=0}^{\infty} P(Y = i) t^i$.

Introduce the following indicator functions

$$\widetilde{Y}_i = \mathbb{I}\{\text{the first occurrence of } \mathbf{w} \text{ in a clump of } \mathbf{w} \text{ ends at position } i\}.$$

An occurrence of $\mathbf{w}$ is the first occurrence of $\mathbf{w}$ in a clump of $\mathbf{w}$ if and only if this occurrence is not overlapped by a previous occurrence of $\mathbf{w}$. That is,

$$\widetilde{Y}_i = Y_i \prod_{u=1}^{k-1} (1 - Y_{i-u}). \tag{2.2}$$

The position of a clump of $\mathbf{w}$ is defined to be the end position of the first occurrence of $\mathbf{w}$ in this clump. Therefore, we say there is a clump at position i if and only if $\widetilde{Y}_i = 1$.

2.1 Waiting time for the next clump occurrence

Denote by ν_k the interarrival time between two clumps of $\mathbf{w}$, i.e. the distance between the first occurrences of $\mathbf{w}$ in two successive clumps (cf. figure 2). More formally,

$$\nu_k = \inf\{n \geq 1 : \widetilde{Y}_{i+n} = 1 \mid \widetilde{Y}_i = 1\}.$$

Due to the strong Markov property and equation (2.2), conditioning on $\{\widetilde{Y}_i = 1\}$ in the above equation is equivalent to conditioning on $\{Y_i = 1\}$: $P(\widetilde{Y}_{i+n} = 1 \mid \widetilde{Y}_i = 1) = P(\widetilde{Y}_{i+n} = 1 \mid Y_i = 1)$. Therefore

$$\nu_k \overset{\mathcal{D}}{=} \inf\{n \geq 1 : \widetilde{Y}_{i+n} = 1 \mid Y_i = 1\},$$

where '$\overset{\mathcal{D}}{=}$' means equality in distribution. In other words, ν_k has the same distribution as the distance between any occurrence of $\mathbf{w}$ and the first occurrence of $\mathbf{w}$ in the next clump. We will call ν_k briefly the waiting time until the next occurrence of a clump of $\mathbf{w}$.

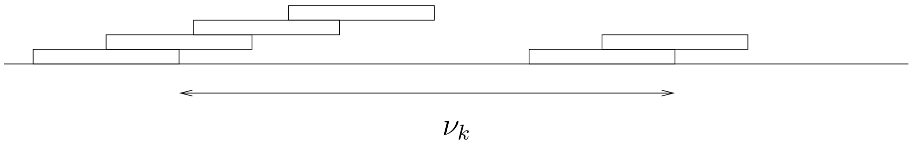


Figure 2: Waiting time ν_k between two successive clumps of $\mathbf{w}$.

The following theorem provides a simple formula for the p.g.f. of ν_k in terms of the p.g.f. $G_{\tau_k}(t) = G_k^{(w_1,w_2,\ldots,w_J)}(t)$ where J has been introduced a few lines earlier.

Theorem 1 *The probability generating function, $G_{\nu_k}(t)$, of ν_k - the waiting time until the next occurrence of a clump of the pattern $w_1 w_2 \cdots w_k$ - is given by:*

$$G_{\nu_k}(t) = \left(G_{\tau_k}(t) - \sum_{i=1}^{k-1} c_i t^i \right) \left(1 - \sum_{i=1}^{k-1} c_i t^i \right)^{-1}, \tag{2.3}$$

where the c_i's are given by (2.1).

Proof. Denote by ω_k the distance between two successive and overlapping occurrences of $w_1 w_2 \cdots w_k$:

$$\omega_k = (\tau_k \mid \tau_k < k)$$

and by ρ_k the distance between two successive and non-overlapping occurrences of $w_1 w_2 \cdots w_k$:

$$\rho_k = (\tau_k \mid \tau_k \geq k).$$

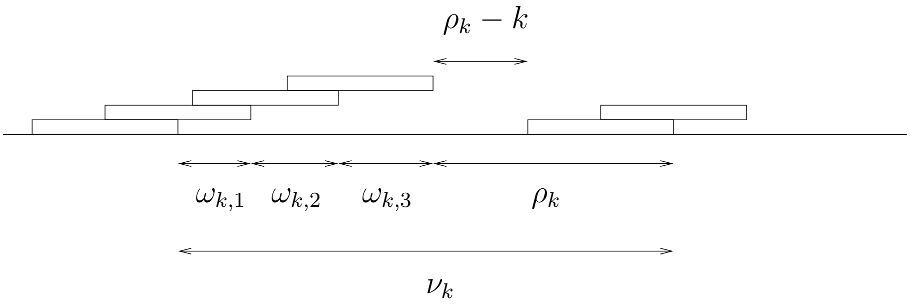


Figure 3: Decomposition of the waiting time ν_k between two successive clumps.

Note that ν_k can be decomposed as the following sum of independent random variables (cf. figure 3):

$$\nu_k = \sum_{g=1}^{L} \omega_{k,g} + \rho_k \tag{2.4}$$

where $\omega_{k,g}$'s are independent copies of ω_k and L has a geometric distribution with probability of success $(1-\alpha)$ (that is, $P(L=\ell) = \alpha^\ell(1-\alpha)$, $\ell \geq 0$), where α is the probability of self-overlap of the pattern $\mathbf{w}$, that is

$$\alpha = P(\tau_k < k) = \sum_{i=1}^{k-1} c_i. \tag{2.5}$$

Note that the support of the geometric distribution above is the set of all non-negative integers, that is $\{0, 1, 2, \dots\}$. Let $G_{\omega_k}(t)$, $G_{\rho_k}(t)$ and $G_{geom(p)}(t)$ denote the p.g.f.'s of ω_k, ρ_k and a geometric random variable with probability of success p, respectively. Then, in view of (2.4), we have

$$G_{\nu_k}(t) = G_{geom(1-\alpha)}\left[G_{\omega_k}(t)\right] G_{\rho_k}(t).$$

Since

$$G_{geom(p)}(t) = \sum_{i=0}^{\infty} p(1-p)^i t^i = p/[1-(1-p)t], \tag{2.6}$$

we get

$$G_{\nu_k}(t) = \frac{(1-\alpha)G_{\rho_k}(t)}{1-\alpha G_{\omega_k}(t)}. \tag{2.7}$$

It is straightforward to see that the p.g.f.'s of ω_k and ρ_k (recall that $\rho_k = (\tau_k \mid \tau_k \geq k)$) can be expressed in terms of the p.g.f. of τ_k, $G_{\tau_k}(t)$, as follows:

$$G_{\rho_k}(t) = \left(G_{\tau_k}(t) - \sum_{i=1}^{k-1} c_i t^i\right)(1-\alpha)^{-1}, \tag{2.8}$$

$$G_{\omega_k}(t) = \left(\sum_{i=1}^{k-1} c_i t^i\right)\alpha^{-1}. \tag{2.9}$$

Therefore, in view of (2.7), (2.8) and (2.9), the statement of Theorem 1 holds. ■

Two straightforward corollaries follow from Theorem 1. The first one is due to the fact that the number of letters separating two consecutive occurrences of clumps of the pattern $\mathbf{w}$ is equal to $\rho_k - k$ (cf. figure 3). The second one is due to the fact that the length of a clump, that is the number of letters that compose it, is equal to $\sum_{g=1}^{L} \omega_{k,g} + k$ (cf. figure 3).

Corollary 1 *The probability generating function of the number of letters separating two consecutive occurrences of clumps of the pattern $w_1 w_2 \cdots w_k$ is*

$$t^{-k}\left(G_{\tau_k}(t) - \sum_{i=1}^{k-1} c_i t^i\right)(1-\alpha)^{-1},$$

Corollary 2 *The probability generating function of the length of a clump of the pattern $w_1 w_2 \cdots w_k$ is*

$$t^k(1-\alpha)\left(1 - \sum_{i=1}^{k-1} c_i t^i\right)^{-1}.$$

2.2 Generalization to h-gap clusters

The method used in deriving Theorem 1 also covers a more general kind of clusters of a pattern. Recall that all consecutive occurrences of the pattern in a clump are overlapping. Define an $h-gap$ cluster of a pattern to be a string consisting of occurrences of a pattern with the property that there are no more than h symbols separating any two consecutive occurrences of the pattern and the string is not a substring of another string with that property (cf. figure 4). Clearly, it is a generalization of a clump of a pattern.

Let μ_k be the waiting time, starting from the beginning of an $h-gap$ cluster until reaching the beginning of the next $h-gap$ cluster. Following the same arguments as those used in the proof of Theorem 1, but considering $\omega'_k = (\tau_k \mid \tau_k < k+h+1)$, $\rho'_k = (\tau_k \mid \tau_k \geq k+h+1)$ and $L' \sim \text{Geom}(1 - \sum_{i=1}^{k+h} c_i)$, one gets the following.

Theorem 2 *The probability generating function, $G_{\mu_k}(t)$, of the waiting time μ_k until the next occurrence of an $h-gap$ cluster of the pattern $w_1 w_2 \cdots w_k$ is given by:*

$$G_{\mu_k}(t) = \left(G_{\tau_k}(t) - \sum_{i=1}^{k+h} c_i t^i\right)\left(1 - \sum_{i=1}^{k+h} c_i t^i\right)^{-1}, \tag{2.10}$$

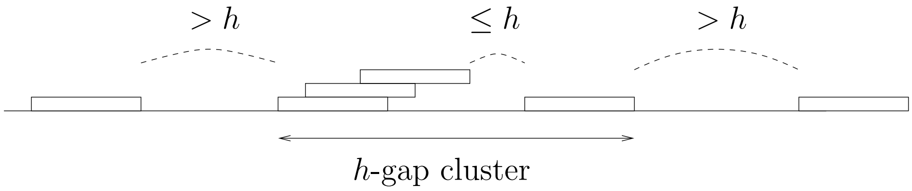


Figure 4: An h-gap cluster.

where the c_i's are given by (2.1).

Extensions of the above results to clumps and $h-gap$ clusters composed of exactly ℓ, or at least ℓ, occurrences are easily derivable in similar terms to those in the above theorems, and are therefore left to the reader.

2.3 Distribution of the number of clumps

Note that the distribution of the number of clumps, $\widetilde{N}(n)$, of the pattern $\mathbf{w}$ in a sequence of length n with initial state s, can be evaluated using the following identity

$$P\left(\widetilde{N}(n) \leq m \mid X(0) = s\right) = P\left(\widetilde{T}_m \geq n \mid X(0) = s\right),$$

where $\widetilde{T}_m$ is the position of the m-th clump of $\mathbf{w}$. More specifically, note that $\widetilde{T}_m = \widetilde{T}_1 + \sum_{i=2}^{m} (\widetilde{T}_i - \widetilde{T}_{i-1})$ where $\nu_{k,i} := (\widetilde{T}_i - \widetilde{T}_{i-1})$, $i \geq 2$, is the waiting time between the $(i-1)$-th and i-th clumps (cf. figure 5) and the $\nu_{k,i}$'s are independent copies of ν_k. Further, given the initial state is s, the position $\widetilde{T}_1$ of the first clump is the waiting time $\tau_k^{(s)}$ until the first occurrence of $\mathbf{w}$ from state s.

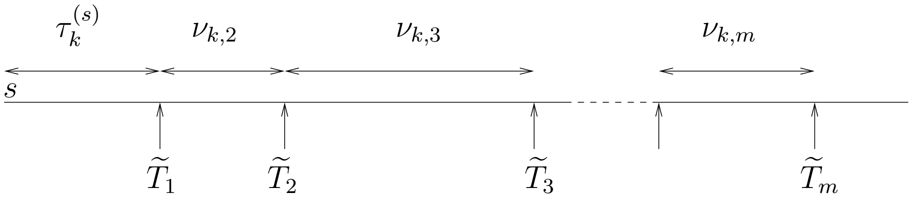


Figure 5: Positions of clumps and waiting times.

Therefore, we have

$$P\left(\widetilde{N}(n) \leq m | X(0) = s\right) = P\left(\tau_k^{(s)} + \sum_{i=2}^{m} \nu_{k,i} \geq n\right).$$

The right-hand term can be calculated by inverting the p.g.f. of the cumulative distribution function of $\tau_k^{(s)} + \sum_{i=2}^{m} \nu_{k,i}$. The latter equals $(G_{\nu_k}(t))^{m-1} G_k^{(s)}(t)/(1-t)$, where the p.g.f. $G_{\nu_k}(t)$ of ν_k is given in Theorem 1 above and the p.g.f. $G_k^{(s)}(t)$ of $\tau_k^{(s)}$ is given by Robin and Daudin (1999) or Stefanov (2003). The above renders the probability of the event $(\widetilde{N}(n) \leq m)$ for any initial distribution. In particular, if the initial distribution is the steady-state one, that is $\pi_s = P(X(0) = s)$, then

$$P\left(\widetilde{N}(n) \leq m\right) = \sum_{s=1}^{N} P\left(\widetilde{N}(n) \leq m | X(0) = s\right) \pi_s. \tag{2.11}$$

An algorithm for a rapid numerical inversion of p.g.f.'s, with any given accuracy, is provided in Abate and Whitt (1992). This algorithm was used for the numerical evaluation in the examples discussed in Section 4.1.

3 Structured motifs

3.1 Two boxes

Let $\mathbf{w}_1$ and $\mathbf{w}_2$ be two patterns of length k_1 and k_2, respectively. The alphabet size is finite and equals N. A structured motif $\mathbf{m}$ formed by the patterns $\mathbf{w}_1$ and $\mathbf{w}_2$, and denoted by $\mathbf{m} = \mathbf{w}_1(d_1 : d_2)\mathbf{w}_2$, is a string with the following property. Pattern $\mathbf{w}_1$ is a prefix and pattern $\mathbf{w}_2$ is a suffix to the string and the number of letters between the two patterns is not smaller than d_1 and not greater than d_2 (cf. figure 6). Also it is assumed that patterns $\mathbf{w}_1$ and $\mathbf{w}_2$ appear only once in the string. The usual definition of a structured motif in DNA sequence analysis does not impose the latter restriction. This is not a strong restriction in practice because the probability for $\mathbf{w}_1$ and $\mathbf{w}_2$ to occur more than once in a sequence smaller than $k_1 + d_2 + k_2$ letters is negligible. We will then get identical significance of the structured motifs, if they are counted with or without the above restriction.

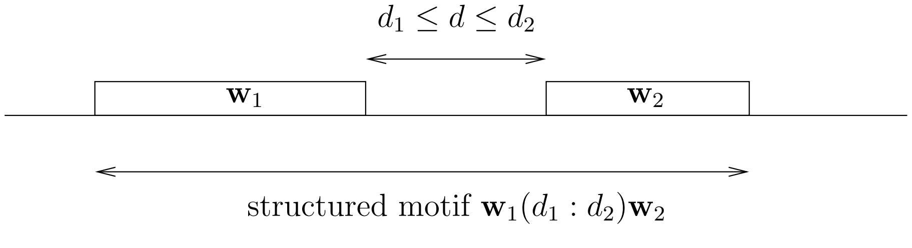


Figure 6: A structured motif $\mathbf{w}_1(d_1 : d_2)\mathbf{w}_2$.

The aim is then to determine the p.g.f. of the waiting time $\tau_{\mathbf{m}}^{(s)}$ to reach for the first time the structured motif $\mathbf{m}$ from state s in a text generated by the Markov chain introduced at the beginning of Section 2.
Some notation follow. Denote by

$\mathcal{W}$ - the pattern family consisting of the two patterns $\mathbf{w}_1$ and $\mathbf{w}_2$, that is, $\mathcal{W} = \{\mathbf{w}_1, \mathbf{w}_2\}$; the waiting time to reach a pattern family is conventionally called a 'sooner time';

T_{ij}, $i, j \in \{1, 2\}$ - the waiting time to reach pattern $\mathbf{w}_j$ from pattern $\mathbf{w}_i$;

$T_j^{(s)}$ - the waiting time to reach pattern $\mathbf{w}_j$ from state s;

r_{ij}, $i, j \in \{1, 2\}$ - the probability that the first pattern from the family $\mathcal{W}$ to be reached is w_j, given we start from pattern $\mathbf{w}_i$, that is $r_{ij} = P(T_{ij} = \inf_{1 \leq u \leq 2}(T_{iu}))$.

Further, denote by X_{ij}, $i, j \in \{1, 2\}$, the interarrival time between two consecutive occurrences of the pattern family $\mathcal{W}$, given the starting pattern is $\mathbf{w}_i$ and the reached pattern is $\mathbf{w}_j$. That is,

$$X_{ij} = (T_{ij} \mid T_{ij} = \inf_u(T_{iu})).$$

Let

$$a_{ij}(x) = P(X_{ij} = x). \tag{3.1}$$

Recall that $G_Y(t)$ denotes the p.g.f. of a random variable Y.

In order to reach the structured motif $\mathbf{m}$, we need to reach first the pattern $\mathbf{w}_1$ and, from this occurrence of $\mathbf{w}_1$, to reach the pattern $\mathbf{w}_2$ such that $d_1+k_2 \leq X_{12} \leq d_2+k_2$. Introduce the following random variables:

$$\begin{aligned} F_{12} &= (X_{12} \mid X_{12} < d_1 + k_2 \text{ or } X_{12} > d_2 + k_2), \\ S_{12} &= (X_{12} \mid d_1 + k_2 \leq X_{12} \leq d_2 + k_2). \end{aligned}$$

F_{12} corresponds to an occurrence of $\mathbf{w}_2$ that fails to achieve the structured motif, whereas for S_{12}, $\mathbf{w}_2$ achieves the structured motif. Similarly to (2.8), it is easy to see that the p.g.f.'s of F_{12} and S_{12} are given by:

$$G_{F_{12}}(t) = \left(G_{X_{12}}(t) - \sum_{x=d_1+k_2}^{d_2+k_2} a_{12}(x)t^x \right) (1 - q_S)^{-1} \tag{3.2}$$

$$G_{S_{12}}(t) = \left(\sum_{x=d_1+k_2}^{d_2+k_2} a_{12}(x)t^x \right) q_S^{-1}, \tag{3.3}$$

where q_S is the probability of success ($\mathbf{w}_2$ achieves the structured motif), i.e. the probability that $d_1 + k_2 \leq X_{12} \leq d_2 + k_2$. Namely, we have

$$q_S = \sum_{x=d_1+k_2}^{d_2+k_2} a_{12}(x), \tag{3.4}$$

where the $a_{12}(x)$ are defined in (3.1).

The following theorem provides explicit and calculable expressions for the p.g.f.'s of the waiting times to reach for the first time the structured motif $\mathbf{m} = \mathbf{w}_1(d_1 : d_2)\mathbf{w}_2$ from either state s or from pattern $\mathbf{w}_2$.

Theorem 3 *The probability generating function $G_{\mathbf{m}}^{(s)}(t)$ of the first arrival time of a structured motif $\mathbf{m}$ starting from state s, and the probability generating function*

$G_{\mathbf{m}}^{(\mathbf{w}_2)}(t)$ *of the first arrival time of a structured motif* $\mathbf{m}$ *starting from pattern* $\mathbf{w}_2$, *admit the following explicit expressions*

$$G_{\mathbf{m}}^{(s)}(t) = \frac{r_{12}\, q_S\, G_{T_1^{(s)}}(t)\, G_{S_{12}}(t)}{(1-(1-r_{12})G_{X_{11}}(t))\left(1-(1-q_S)\left(\frac{r_{12}\, G_{T_{21}}(t)\, G_{F_{12}}(t)}{1-(1-r_{12})G_{X_{11}}(t)}\right)\right)}, \tag{3.5}$$

$$G_{\mathbf{m}}^{(\mathbf{w}_2)}(t) = \frac{r_{12}\, q_S\, G_{T_{21}}(t)\, G_{S_{12}}(t)}{(1-(1-r_{12})G_{X_{11}}(t))\left(1-(1-q_S)\left(\frac{r_{12}\, G_{T_{21}}(t)\, G_{F_{12}}(t)}{1-(1-r_{12})G_{X_{11}}(t)}\right)\right)}, \tag{3.6}$$

where $G_{F_{12}}(t), G_{S_{12}}(t)$, *and* q_S *are given in (3.2), (3.3), and (3.4).*

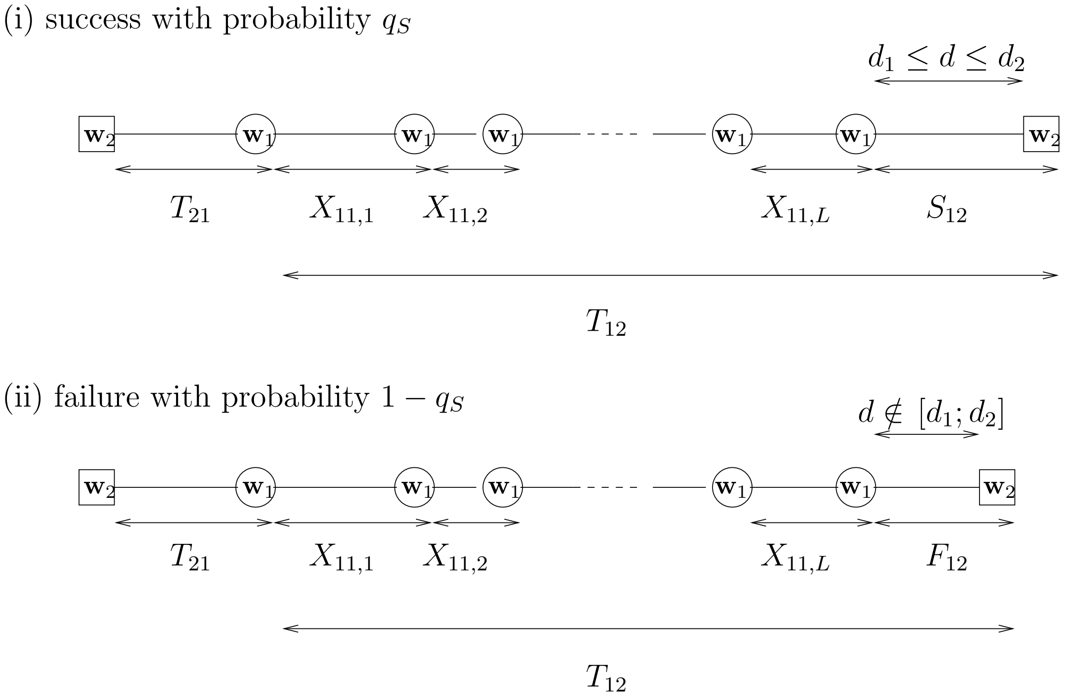
The quantities r_{12}, $G_{X_{11}}(t)$ and $G_{T_{21}}(t)$ are provided in Robin and Daudin (2001) whereas $G_{T_1^{(s)}}$ is given in Robin and Daudin (1999). These quantities can also be calculated from the results in Stefanov (2003).

Proof. Consider first the waiting time $\tau_{\mathbf{m}}^{(\mathbf{w}_2)}$ to reach for the first time a structured motif $\mathbf{m} = \mathbf{w}_1(d_1 : d_2)\mathbf{w}_2$, starting from pattern $\mathbf{w}_2$. Then in order to reach a structured motif one should first reach pattern $\mathbf{w}_1$. The p.g.f. of the waiting time to reach for the first time $\mathbf{w}_1$ from $\mathbf{w}_2$ is equal to $G_{T_{21}}(t)$. After $\mathbf{w}_1$ has been reached, one is waiting for an occurrence of pattern $\mathbf{w}_2$. Of course there are two cases at the time we reach pattern $\mathbf{w}_2$ after an elapsed time T_{12}. Either (i) an occurrence of the structured motif $\mathbf{m}$ is reached i.e. there is a correct intersite distance between $\mathbf{w}_1$ and $\mathbf{w}_2$ and no occurrences of $\mathbf{w}_1$ has occurred in between (this happens with probability q_S) or (ii) no occurrence of the structured motif is reached. Now we use arguments similar to those used in the proof of Theorem 2.1 of Stefanov (2003) to show that, due to the strong Markov property, the p.g.f. of T_{12}, conditioned on either a 'success' (that is, a structured motif has been reached) or a 'failure', is equal to (i) $G_{geom(r_{12})}[G_{X_{11}}(t)]G_{S_{12}}(t)$ in the 'success' case, and (ii) $G_{geom(r_{12})}[G_{X_{11}}(t)]G_{F_{12}}(t)$ in the 'failure' case. This is illustrated in figure 7. The elapsed time T_{12} conditioned on either a 'success' or a 'failure' can be decomposed with respect to the occurrences of $\mathbf{w}_1$ within that elapsed time as follows:

$$\begin{aligned} T_{12}|'success' &\overset{\mathcal{D}}{=} \sum_{a=1}^{L} X_{11,a} + S_{12} \quad \text{case (i)} \\ T_{12}|'failure' &\overset{\mathcal{D}}{=} \sum_{a=1}^{L} X_{11,a} + F_{12} \quad \text{case (ii)}, \end{aligned}$$

where the $X_{11,a}$'s are independent copies of X_{11}, and L is geometrically distributed with parameter r_{12} (L counts the number of times $\mathbf{w}_1$ re-occurs before $\mathbf{w}_2$).
Therefore, the p.g.f. of T_{12} conditioned on either a 'success' or a 'failure' is given by

$$\begin{aligned} G_{T_{12}|'success'}(t) &= G_{geom(r_{12})}[G_{X_{11}}(t)]\, G_{S_{12}}(t) \quad \text{case (i)} \\ G_{T_{12}|'failure'}(t) &= G_{geom(r_{12})}[G_{X_{11}}(t)]\, G_{F_{12}}(t) \quad \text{case (ii)}. \end{aligned}$$


(i) success with probability q_S

(ii) failure with probability $1 - q_S$

Figure 7: Reaching $\mathbf{w}_1$ from $\mathbf{w}_2$, then $\mathbf{w}_2$ from $\mathbf{w}_1$. (i) $\mathbf{w}_2$ achieves the structured motif $\mathbf{w}_1(d_1 : d_2)\mathbf{w}_2$, (ii) $\mathbf{w}_2$ fails to achieve the structured motif.

In the failure case (ii), we reached pattern $\mathbf{w}_2$ without reaching a structured motif yet, so we have to wait again for the next $\mathbf{w}_1$ and then the next $\mathbf{w}_2$, and so on until the success of case (i). The structured motif will finally be reached after L' failures (cf. figure 8), where L' is geometrically distributed with parameter q_S (L' counts the number of times $\mathbf{w}_1$ is followed by $\mathbf{w}_2$ but not at a valid distance to reach the structure motif).

Finally, in view of the strong Markov property and the preceding arguments, it is clear that the waiting time $\tau_{\mathbf{m}}^{(\mathbf{w}_2)}$ may be decomposed as follows:

$$\tau_{\mathbf{m}}^{(\mathbf{w}_2)} \overset{\mathcal{D}}{=} T_{21} + \sum_{b=1}^{L'} \left(\sum_{a=1}^{L_1} X_{11,ab} + F_{12,b} + T_{21,b} \right) + \sum_{c=1}^{L_2} X_{11,c} + S_{12},$$

where all random variables in the right hand side are mutually independent, the $T_{21,b}$'s are independent copies of T_{21}, the $F_{12,b}$'s are independent copies of F_{12}, the $X_{11,ab}$'s are independent copies of X_{11}, and L_1 and L_2 are geometrically distributed with parameter r_{12}. Therefore, we get the following explicit expression for $G_{\mathbf{m}}^{(\mathbf{w}_2)}(t)$:

$$G_{\mathbf{m}}^{(\mathbf{w}_2)}(t) = G_{T_{21}}(t) G_{geom(q_S)} \left[G_{geom(r_{12})}[G_{X_{11}}(t)] G_{F_{12}}(t) G_{T_{21}}(t) \right] G_{geom(r_{12})}[G_{X_{11}}(t)] G_{S_{12}}(t).$$

Using the expression (2.6) for the p.g.f. of the geometric distribution one gets that (3.6) holds.

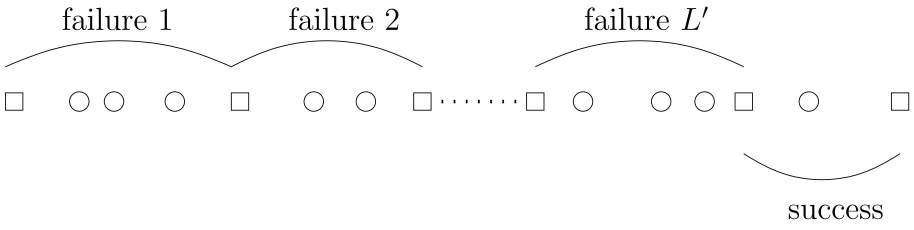


Figure 8: Reaching the structured motif $\mathbf{w}_1(d_1 : d_2)\mathbf{w}_2$ from $\mathbf{w}_2$. The squares stand for the occurrences of $\mathbf{w}_2$, and the circles stand for the occurrences of $\mathbf{w}_1$ like in figure 7.

Likewise, one gets the following expression for $G_{\mathbf{m}}^{(s)}(t)$:

$$G_{\mathbf{m}}^{(s)}(t) = G_{T_1^{(s)}}(t)G_{geom(q_S)}\left[G_{geom(r_{12})}[G_{X_{1,1}}(t)]G_{F_{12}}(t)G_{T_{21}}(t)\right]G_{geom(r_{1,2})}[G_{X_{11}}(t)]G_{S_{12}}(t).$$

This yields the explicit expression given in (3.5). The proof of Theorem 3 is complete. ■

Remark 1 *Note that the p.g.f. of the intersite distance between two consecutive occurrences of the structured motif* $\mathbf{w}_1(d_1 : d_2)\mathbf{w}_2$ *is equal to* $G_{\mathbf{m}}^{(\mathbf{w}_2)}(t)$*.*

Remark 2 *As we did in section 2.3 for the number of clumps, the above theorem can be used to get the p.g.f. of the number of occurrences of a structured motif in a random sequence.*

3.2 More than two boxes

Our methodology can be extended to structured motifs consisting of more than two boxes of patterns. A few comments on such extensions follow.

Consider a structured motif consisting of three boxes. That is, let $\mathbf{w}_1, \mathbf{w}_2$, and $\mathbf{w}_3$ be three patterns of length k_1, k_2, and k_3, respectively. A structured motif $\mathbf{m}$ formed by the patterns $\mathbf{w}_1, \mathbf{w}_2$, and $\mathbf{w}_3$, and denoted by $\mathbf{m} = \mathbf{w}_1(d_1 : d_2)\mathbf{w}_2(d_3 : d_4)\mathbf{w}_3$, is a string with the following property. Patterns $\mathbf{w}_1$ and $\mathbf{w}_3$ are a prefix and a suffix, respectively, to the string; pattern $\mathbf{w}_2$ appears between the patterns $\mathbf{w}_1$ and $\mathbf{w}_3$, and the number of letters separating it from $\mathbf{w}_1$ is not smaller than d_1 and not greater than d_2, and the number of letters separating it from pattern $\mathbf{w}_3$ is not smaller than d_3 and not greater than d_4. Further, it is assumed that a structured motif does not contain occurrences of the patterns $\mathbf{w}_1$ and $\mathbf{w}_2$ between the first two boxes and also it does not contain occurrences of the patterns $\mathbf{w}_1$ and $\mathbf{w}_3$ between the last two boxes. In other words, in a structure motif, according to this definition,

(i) pattern $\mathbf{w}_1$ appears only once (at the beginning of the motif), (*ii*) pattern $\mathbf{w}_2$ appears in the second box and does not appear between the first two boxes while it is allowed to appear arbitrarily between the second and third boxes,

(*iii*) pattern $\mathbf{w}_3$ appears in the third box and does not appear between the last two boxes while it is allowed to appear arbitrarily between the first and second boxes.

Here again, these technical restrictions will have a negligible effect on the *significance* of structured motifs in DNA sequence analysis because a box is unlikely to occur more than once in a very short sequence, where short sequence is meant to be a small portion of the structured motif.

A very careful scrutiny of the details in the proof of our results above reveals that our method can be extended to the case of structured motifs consisting of three boxes if the distributions of the following quantities are available:

(*i*) the waiting time to reach for the first time the structured motif $\mathbf{w}_1(d_1 : d_2)\mathbf{w}_2$,

(*ii*) the waiting time to reach for the first time the structured motif $\mathbf{w}_1(d_1 : d_2)\mathbf{w}_2$, given pattern $\mathbf{w}_1$ has been reached,

(*iii*) the waiting time to reach the pattern family consisting of the two patterns $\mathbf{w}_1$ and $\mathbf{w}_3$, given the sequence has pattern $\mathbf{w}_2$ as a prefix.

The distributions of (*i*) and (*ii*) are found from our results above on structured motifs consisting of two boxes. The distribution of (*iii*) can be recovered using the results in Stefanov (2003). In other words, an extension to structured motifs consisting of three boxes relies on availability of relevant results for structured motifs consisting of two boxes (as provided in this paper), and applying similar arguments to those used in the proofs above.

Likewise, our method can be extended to structured motifs consisting of four or more boxes along similar lines to those above. That is, results on motifs with b boxes will be derivable using results for motifs with $(b-1)$ boxes and applying similar arguments to those introduced in this paper.

4 Applications to DNA sequences

4.1 Exceptional number of clumps

The aim of this section is to measure the quality of the Poisson approximation which is currently used for distribution evaluation concerning the number of clumps of a pattern (Schbath (1995)). For this, we consider the complete genome of the phage *Lambda* ($n = 48,502$) whose estimated transition matrix on the $\{\mathtt{a},\mathtt{c},\mathtt{g},\mathtt{t}\}$ alphabet is given by

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

Denote by $p(\mathbf{w})$ the p-value $P(\widetilde{N}(\mathbf{w}) \geq \widetilde{N}^{\text{obs}}(\mathbf{w}))$, where $\widetilde{N}^{\text{obs}}(\mathbf{w})$ is the observed number of clumps of the word $\mathbf{w}$ in the *Lambda* genome. We have first computed, for all the words of length 3, 4, 5 and 6, an approximation $\widetilde{p}(\mathbf{w})$ of the p-value by approximating the number of clumps by a Poisson variable with parameter $\mathbb{E}(\widetilde{N}(\mathbf{w}))$. The expected number of clumps of $\mathbf{w} = w_1 w_2 \cdots w_k$ is given by

$$\mathbb{E}(\widetilde{N}(\mathbf{w})) = (1-\alpha)\mathbb{E}(N(\mathbf{w}))$$

where α is the probability of self-overlap given by (2.5), $N(\mathbf{w})$ is the number of occurrences of $\mathbf{w}$, and the expected count $\mathbb{E}(N(\mathbf{w}))$ is:

$$\mathbb{E}(N(\mathbf{w})) = (n-k+1)\pi_{w_1} \prod_{i=1}^{k-1} p_{w_i, w_i+1},$$

where π is the stationary distribution of the Markov chain. For the most exceptional words, i.e. words with an approximate p-value close to zero (clumps significantly frequent) or close to one (clumps significantly rare), we have also calculated their exact p-values using the method introduced in Section 2 above. The exact and approximated p-values are listed in Tables 1 to 4. The observed numbers of clumps are also displayed.

exceptionally frequent clumps					exceptionally rare clumps			
$\mathbf{w}$	$\widetilde{N}(\mathbf{w})$	$\widetilde{p}(\mathbf{w})$	$p(\mathbf{w})$		$\mathbf{w}$	$\widetilde{N}(\mathbf{w})$	$1-\widetilde{p}(\mathbf{w})$	1-$p(\mathbf{w})$
tag	217	$9.843\,10^{-42}$	$1.715\,10^{-43}$		cag	1131	$1.283\,10^{-47}$	$1.002\,10^{-48}$
ttg	734	$2.684\,10^{-26}$	$4.507\,10^{-29}$		ctg	1169	$5.448\,10^{-34}$	$4.721\,10^{-37}$
caa	697	$1.467\,10^{-19}$	$1.854\,10^{-21}$		tat	742	$9.581\,10^{-14}$	$1.525\,10^{-14}$
cta	287	$3.240\,10^{-17}$	$5.848\,10^{-19}$		ccg	884	$1.527\,10^{-13}$	$1.902\,10^{-14}$
cga	629	$1.390\,10^{-9}$	$3.005\,10^{-10}$		cgg	963	$2.039\,10^{-11}$	$2.913\,10^{-12}$
tcg	581	$3.088\,10^{-9}$	$8.100\,10^{-10}$		acc	679	$2.112\,10^{-6}$	$1.037\,10^{-6}$
ggg	468	$5.824\,10^{-8}$	$1.720\,10^{-8}$		tga	1091	$3.045\,10^{-5}$	$1.164\,10^{-5}$
aat	838	$9.032\,10^{-8}$	$1.534\,10^{-8}$		tca	855	$2.567\,10^{-4}$	$1.507\,10^{-4}$
ccc	346	$2.279\,10^{-5}$	$1.332\,10^{-5}$		ttc	842	$3.397\,10^{-4}$	$1.890\,10^{-4}$
ctt	603	$2.938\,10^{-5}$	$1.576\,10^{-5}$		aac	853	$1.823\,10^{-3}$	$1.229\,10^{-3}$

Table 1: 3-letter words with the most exceptional number of clumps in the *Lambda* genome.

exceptionally frequent clumps				exceptionally rare clumps			
$\mathbf{w}$	$\widetilde{N}(\mathbf{w})$	$\widetilde{p}(\mathbf{w})$	$p(\mathbf{w})$	$\mathbf{w}$	$\widetilde{N}(\mathbf{w})$	$1-\widetilde{p}(\mathbf{w})$	1-$p(\mathbf{w})$
ctag	14	$1.043\,10^{-27}$	$6.037\,10^{-28}$	ccgg	328	$3.473\,10^{-27}$	$5.752\,10^{-28}$
ttgg	126	$4.956\,10^{-21}$	$1.204\,10^{-21}$	gcag	392	$6.670\,10^{-25}$	$8.328\,10^{-26}$
tagg	33	$7.907\,10^{-21}$	$4.433\,10^{-21}$	cagc	328	$1.091\,10^{-17}$	$3.269\,10^{-18}$
taga	44	$3.247\,10^{-16}$	$2.084\,10^{-16}$	caga	305	$2.655\,10^{-17}$	$8.352\,10^{-18}$
caag	106	$3.313\,10^{-16}$	$1.390\,10^{-16}$	ctga	334	$4.380\,10^{-17}$	$1.124\,10^{-17}$
cttg	115	$7.053\,10^{-16}$	$3.127\,10^{-16}$	gctg	392	$7.292\,10^{-17}$	$1.569\,10^{-17}$
ttag	62	$2.378\,10^{-12}$	$1.567\,10^{-12}$	ctgg	325	$1.746\,10^{-16}$	$5.484\,10^{-17}$
cgag	95	$2.392\,10^{-11}$	$1.534\,10^{-11}$	tcag	279	$2.639\,10^{-15}$	$1.034\,10^{-15}$
caat	162	$3.894\,10^{-10}$	$1.793\,10^{-11}$	tatc	229	$7.375\,10^{-15}$	$3.539\,10^{-15}$
ccaa	125	$8.295\,10^{-10}$	$4.479\,10^{-11}$	cagg	280	$4.446\,10^{-13}$	$2.165\,10^{-13}$

Table 2: 4-letter words with the most exceptional number of clumps in the *Lambda* genome.

exceptionally frequent clumps				exceptionally rare clumps			
$\mathbf{w}$	$\widetilde{N}(\mathbf{w})$	$\widetilde{p}(\mathbf{w})$	$p(\mathbf{w})$	$\mathbf{w}$	$\widetilde{N}(\mathbf{w})$	$1-\widetilde{p}(\mathbf{w})$	1-$p(\mathbf{w})$
aattg	32	$5.194\,10^{-12}$	$3.828\,10^{-12}$	gcaga	141	$1.240\,10^{-20}$	$6.617\,10^{-21}$
ttggg	20	$5.752\,10^{-11}$	$4.810\,10^{-11}$	ccgga	112	$1.840\,10^{-18}$	$1.134\,10^{-18}$
ttgga	21	$6.154\,10^{-11}$	$5.185\,10^{-11}$	tccgg	100	$1.168\,10^{-15}$	$8.014\,10^{-16}$
acttg	13	$2.900\,10^{-10}$	$2.578\,10^{-10}$	gccgg	114	$6.098\,10^{-13}$	$4.331\,10^{-14}$
taggg	3	$6.652\,10^{-10}$	$6.222\,10^{-10}$	ctgaa	124	$1.021\,10^{-12}$	$6.856\,10^{-13}$
tcgag	9	$1.465\,10^{-9}$	$1.339\,10^{-9}$	gctgg	124	$7.483\,10^{-12}$	$5.249\,10^{-13}$
ctagc	3	$1.609\,10^{-9}$	$1.490\,10^{-9}$	gccag	104	$1.084\,10^{-11}$	$8.508\,10^{-12}$
gctag	5	$3.240\,10^{-9}$	$2.979\,10^{-9}$	cggtg	108	$1.844\,10^{-11}$	$1.369\,10^{-12}$
ttgcg	36	$1.031\,10^{-8}$	$8.359\,10^{-9}$	ctgac	89	$2.909\,10^{-10}$	$2.299\,10^{-11}$
tctag	2	$1.201\,10^{-8}$	$1.149\,10^{-7}$	cagca	108	$3.747\,10^{-10}$	$3.010\,10^{-11}$

Table 3: 5-letter words with the most exceptional number of clumps in the *Lambda* genome.

exceptionally frequent clumps				exceptionally rare clumps			
$\mathbf{w}$	$\widetilde{N}(\mathbf{w})$	$\widetilde{p}(\mathbf{w})$	$p(\mathbf{w})$	$\mathbf{w}$	$\widetilde{N}(\mathbf{w})$	$1-\widetilde{p}(\mathbf{w})$	1-$p(\mathbf{w})$
ttgggc	1	$2.059\,10^{-7}$	$1.990\,10^{-7}$	gccgga	55	$4.009\,10^{-17}$	$3.279\,10^{-17}$
cgcgcg	1	$6.523\,10^{-7}$	$6.263\,10^{-7}$	tgccgg	46	$7.197\,10^{-10}$	$6.479\,10^{-10}$
cgaatt	2	$3.083\,10^{-6}$	$3.008\,10^{-6}$	gccagc	42	$1.134\,10^{-9}$	$1.070\,10^{-9}$
cacaat	1	$3.475\,10^{-6}$	$3.379\,10^{-6}$	ccggac	32	$2.143\,10^{-9}$	$1.986\,10^{-9}$
ggcgcc	1	$3.997\,10^{-6}$	$3.884\,10^{-6}$	gcagaa	47	$2.453\,10^{-9}$	$2.262\,10^{-9}$
gccggc	1	$4.244\,10^{-6}$	$4.125\,10^{-6}$	tatcag	30	$3.864\,10^{-9}$	$3.685\,10^{-9}$
gaattg	5	$5.720\,10^{-6}$	$5.523\,10^{-6}$	atcacc	32	$1.144\,10^{-8}$	$1.080\,10^{-8}$
tttgcg	5	$7.964\,10^{-6}$	$7.670\,10^{-6}$	ctgacc	30	$6.092\,10^{-8}$	$5.730\,10^{-8}$
gcatgc	6	$9.011\,10^{-6}$	$8.510\,10^{-6}$	ccggtg	34	$7.187\,10^{-8}$	$6.735\,10^{-8}$
gacttg	1	$2.223\,10^{-5}$	$2.204\,10^{-5}$	cagaaa	42	$6.636\,10^{-8}$	$6.228\,10^{-8}$

Table 4: 6-letter words with the most exceptional number of clumps in the *Lambda* genome.

Note that for 6-letter words, the approximate p-values are of the same order of magnitude as the exact p-values. This indicates that the Poisson approximation is good for rare words. As the words become shorter, and then more frequent, the difference between the two quantities often increases. For some 4- and 5-letter words, the approximate p-values (or 1-$\widetilde{p}(\mathbf{w})$ for exceptionally rare clumps) can be ten times larger than the exact p-values (or 1-$p(\mathbf{w})$). For some 3-letter words, one may observe a factor of 10^3 between the two probabilities.

Furthermore, the approximate p-values (or 1-$\widetilde{p}(\mathbf{w})$) happen to be always larger than the exact p-values (or 1-$p(\mathbf{w})$) which indicates that the Poisson approximation is conservative. It is indeed easier and much faster to compute the approximate p-values with the Poisson approximation rather than the exact ones. For instance, if the calculations are executed on Apple PowerMac G4 then it takes on average 850 seconds to calculate an exact p-value for a k-letter word ($k = 3, 4, 5, 6$) whereas the corresponding approximate p-value is calculated almost instantaneously.

It is also important to note that the words are mainly ranked in the same order with respect to their exact p-values or their approximate p-values.

4.2 Significance of structured motifs

We have considered the same data set as that considered in Robin *et al.* (2002). It is composed of a set $\mathcal{S}$ of 130 sequences of length $n = 100$ located just before 130 genes of the bacterium *B. subtilis*, and 71 structured motifs of the form $\mathbf{m} = \mathbf{w}_1(16 : 18)\mathbf{w}_2$ $(6 \le k_1, k_2 \le 7)$ which are good candidates as promoter motifs for the bacterium. Promoters are usually located on the DNA sequences in front of genes; they are recognized by the RNA polymerase to bind to the DNA and to start the gene transcription. Because promoters have to be in such regulatory sequences, a challenging question is to find motifs that are present in a significant number of sequences. Therefore, for each of the 71 structured motifs $\mathbf{m}$, we have counted the number Q^{obs} of sequences from $\mathcal{S}$ containing at least one occurrence of $\mathbf{m}$. To know if this number Q^{obs} is significant, we calculate the p-value $P(Q \geq Q^{\text{obs}})$ where Q is the random number of sequences amongst a set of $|\mathcal{S}|$ random sequences of length n which contain $\mathbf{m}$. The random sequences are drawn according to the first-order Markov chain whose parameters are estimated from the sequence resulting of the concatenation of the 130 observed sequences.

Denote by $\gamma_n(\mathbf{m})$ the probability for the motif $\mathbf{m}$ to occur in a random sequence of length n. It is calculated by using the p.g.f. of $\tau_{\mathbf{m}}$ given by Theorem 3 above (the initial state s is selected according to the stationary distribution of the chain), that is

$$\gamma_n(\mathbf{m}) = P(\tau_{\mathbf{m}} \leq n).$$

The random variable Q is then distributed according to the binomial distribution $\mathcal{B}(|\mathcal{S}|, \gamma_n(\mathbf{m}))$; the p-values can then be easily calculated. There were 3277 seconds required to calculate the 71 p-values on IBM F80 computer (RS64III processor, 450MHz), i.e. 46 seconds on average per structured motif. Only motifs with a p-value less than 10^{-3} are listed in Table 5.

These exceptional structured motifs are close to the known consensus $\mathbf{w}_1 =$`ttgaca` and the so-called tata-box for $\mathbf{w}_2$ (Record *et al.* (1996)).

Note that the observed counts Q^{obs} are smaller than those in Robin *et al.* (2002) because we have not allowed errors in $\mathbf{w}_1$ and $\mathbf{w}_2$.

5 Conclusion

We provided the exact distributions (in terms of probability generating functions) of relevant quantities related to occurrences of clumps of patterns or structured motifs. Only approximations of these distributions have been proposed and used in practice so far. Our exact distributional result on the number of clumps demonstrates that the usual Poisson approximation of this count is very efficient for rare words. Moreover, we provided a powerful technique, based on random sums, for treating occurrences of complex motifs in Markovian sequences.

In a forthcoming paper we will present two extensions of our results. These concern structured motifs consisting of more than two boxes of patterns (a brief comment on such an extension has been made in section 3.2 above) and structured motifs with degenerated boxes (that is, errors in reading the patterns in the boxes are allowed). These extensions could not be included in the present paper because they are too technical. Note that the probabilistic problems for structured motifs are much more sophisticated than designing pattern matching algorithms to count structured motifs.

$\mathbf{w}_1$	$(d1:d2)$	$\mathbf{w}_2$	Q^{obs}	$\gamma_n(\mathbf{m})$	p-value
ttgactt	(16:18)	ataataa	3	$1.16\,10^{-5}$	$5.77\,10^{-10}$
tgactt	(16:18)	ataataa	3	$3.02\,10^{-5}$	$1.00\,10^{-8}$
ttgactt	(16:18)	atactaa	2	$4.01\,10^{-6}$	$1.37\,10^{-7}$
tgactt	(16:18)	atactaa	2	$1.04\,10^{-5}$	$9.18\,10^{-7}$
ttgaca	(16:18)	tataatg	2	$1.60\,10^{-5}$	$2.18\,10^{-6}$
ttgaca	(16:18)	tatatta	2	$2.36\,10^{-5}$	$4.75\,10^{-6}$
ttgact	(16:18)	tatact	2	$2.38\,10^{-5}$	$4.81\,10^{-6}$
ttgaca	(16:18)	tataata	2	$2.48\,10^{-5}$	$5.23\,10^{-6}$
ttgaca	(16:18)	atataat	2	$2.74\,10^{-5}$	$6.39\,10^{-6}$
tgacttt	(16:18)	taataa	2	$3.63\,10^{-5}$	$1.12\,10^{-5}$
gacttt	(16:18)	taataa	2	$1.06\,10^{-4}$	$9.52\,10^{-5}$
gttgaca	(16:18)	tataata	1	$3.89\,10^{-6}$	$5.09\,10^{-4}$
gttgaca	(16:18)	atataat	1	$4.30\,10^{-6}$	$5.63\,10^{-4}$
ttgacac	(16:18)	ataataa	1	$4.88\,10^{-6}$	$6.39\,10^{-4}$
gttgac	(16:18)	ctataat	1	$4.88\,10^{-6}$	$6.39\,10^{-4}$

Table 5: The most significant structured motifs (p-value $< 10^{-3}$).

References

ABATE, J. and WHITT, W., Numerical inversion of probability generating functions, Oper. Res. Letters 12 (1992) 245–251.

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

ESKIN, E. and PEVZNER, P. A., Finding composite regulatory patterns in DNA sequences, Bioinfomatics 18 (2002) S354-S363.

FELLER, W., An Introduction to Probability Theory and Its Applications (vol. 1, Wiley, 3rd edition, 1968).

GUIBAS, L. J. and ODLYZKO, A. M., Periods in strings, J. Combinatorial Theory A. 30 (1981) 14–42.

MARSAN, L. and SAGOT, M.-F., Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification, J. Comp. Biol. 7 (2000) 345–360.

MORGANTE, M., POLICRITI, A, VITACOLONNA, N. and ZUCCOLO, A, Structured motifs search, in: RECOMB'04, Proceedings of the Eighth Annual International Conference on Computational Molecular Biology, San Diego, CA, USA, 2004, 133–139.

RECORD, M. T., REZNIKOFF, W. S., CRAIG, M. L., MCQUADE, K. L. and SCHLAX, P. J., Escherichia coli RNA polymerase σ^{70} promoters, and the kinetics of the steps of transcription initiation, in F. C. Neidhardt, ed., volume 1 (ASM Press, 1996).

RÉGNIER, M., A unified approach to word occurrence probabilities, Discrete Applied Mathematics 104 (2000) 259–280.

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

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

ROBIN, S. and DAUDIN, J.-J., Exact distribution of word occurrences in a random sequence of letters J. Appl. Prob. 36 (1999) 179–193.

ROBIN, S. and DAUDIN, J.-J., Exact distribution of the distances between any occurences of a set of words, Ann. Inst. Statist. Math. 36 (2001) 895–905.

Robin, S., Daudin, J.-J., Richard, H., Sagot, M.-F. and Schbath, S., Occurrence probability of structured motifs in random sequences, J. Comp. Biol. 9 (2002) 761–773.

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

Stefanov, V.T., The intersite distances between pattern occurrences in strings generated by general discrete- and continuous-time models: an algorithmic approach, J. Appl. Probab. 40 (2003) 881-892.