Occurrence of structured motifs in random sequences: Arbitrary number of boxes

Valeri T. Stefanov\textsuperscript{a}, Stéphane Robin\textsuperscript{b}, Sophie Schbath\textsuperscript{c,∗}

\textsuperscript{a} School of Mathematics and Statistics, University of Western Australia, Crawley 6009, W.A., Australia
\textsuperscript{b} AgroParisTech/INRA UMR518, Unité Mathématique et Informatique Appliquée, 16 rue Claude Bernard, F-75005 Paris, France
\textsuperscript{c} INRA UR1077, Unité Mathématique, Informatique et Génome, F-78352 Jouy-en-Josas, France

\textbf{A B S T R A C T}

Structured motifs with arbitrary number of boxes are considered. In particular, such motifs are of interest in molecular biology for identifying gene promoters along genomes. Neat closed-form expressions for relevant distributions associated with occurrences of structured motifs are derived. Our methodology is based on developing a suitable semi-Markov embedding of the problem. A numerical example is also provided.

© 2011 Elsevier B.V. All rights reserved.

\section{1. Introduction}

A single pattern (word) on a finite alphabet is a finite string of letters. A compound pattern is a finite collection of distinct single patterns. The number of these single patterns is called a size of the compound pattern. Structured motifs (also called gapped patterns) are important special compound patterns whose sizes are usually huge. This paper deals with the distribution theory of occurrences of structured motifs on strings of letters generated by a Markov source. Currently, they are of interest in molecular biology for identifying gene promoter motifs along genomes. There are satisfactory results in the literature on exact distributions associated with occurrences of compound patterns if their sizes are small to moderate. Various tools and techniques are used in this area, such as combinatorial, Markov chain and Markov renewal embeddings, martingales, and exponential families. Relevant references on occurrence of patterns are found in the surveys [5,11].

A structured motif is a string of letters which is best visualized as a finite number of boxes, numbered from 1 to \(b\), where each two adjacent boxes are separated by a variable number of letters and each box, \(i\) say, stands for a fixed single pattern, \(w_i\) say. Note that the size of a structured motif grows exponentially with a linear growth of any of the variable distances between adjacent boxes. Structured motifs, as special compound patterns, allow the use of specific analytical tools for their treatment. For example, Robin et al. [8] provided an approximation for the distribution of occurrence of the simplest structured motifs consisting of two boxes, whereas Stefanov et al. [12] provided the first exact distributional result for the waiting time of the first occurrence of such a structured motif. More specifically, they derived an explicit, closed-form expression for the generating function of this waiting time in terms of well known distributional results for the simplest compound pattern consisting of only two single patterns. Recently, Nuel [3] and Pozdnyakov [4] derived results for structured motifs applying Markov chain embedding with automata and martingale techniques, respectively. On the other hand, there are still no satisfactory results to cover structured motifs with arbitrary number of boxes and arbitrary, however large, gaps between the boxes.

\* Corresponding author. Fax: +33 1 34 65 29 01.
E-mail addresses: stefanov@maths.uwa.edu.au (V.T. Stefanov), Stephane.Robin@agroparistech.fr (S. Robin), sophie.schbath@jouy.inra.fr (S. Schbath).

0166-218X/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2010.12.023
In this paper, we develop a general approach, based on a suitable semi-Markov embedding, which results in explicit, closed-form expressions for relevant generating functions on structured motifs with any number of boxes and arbitrary gaps between the corresponding boxes. Furthermore, the expressions are again in terms of well-known exact distribution results for the simplest compound pattern consisting of only two single patterns.

2. Model and structured motif

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

Let \( w_1, w_2, \ldots, w_b \) be \( b \) patterns of length \( k_1, k_2, \ldots, k_b \), respectively, on the finite alphabet \( \{1, 2, \ldots, N\} \). A structured motif \( m \) formed by these \( b \) patterns is any string of letters which (i) begins with pattern \( w_1 \) and ends with pattern \( w_b \); (ii) all remaining patterns \( w_2, \ldots, w_{b-1} \) appear in that order in the string (note that conditions (i) and (ii) do not preclude the appearance in the string of any of the \( w_i \) more than once); (iii) the initial pattern \( w_1 \) and the end pattern \( w_b \) together with a fixed appearance of the patterns \( w_2, \ldots, w_{b-1} \) in that order satisfy the following condition: for \( i = 1, 2, \ldots, b-1 \), and nonnegative integers \( d_i, D_i \) the number of letters separating patterns \( w_i \) and \( w_{i+1} \) is not smaller than \( d_i \) and not greater than \( D_i \). We denote a structured motif \( m \) by

\[
w_1(d_1 : D_1)w_2(d_2 : D_2)w_3 \ldots w_{b-1}(d_{b-1} : D_{b-1})w_b
\]

and, for \( i = 1, 2, \ldots, b-1 \), we denote by \( m_i \) the sub-structured motif \( w_1(d_1 : D_1)w_2(d_2 : D_2)w_3 \ldots w_{i-1}(d_{i-1} : D_{i-1})w_i \); of course \( m_0 \) is a prefix of \( m_b \).

Throughout the paper we assume that the following restrictions apply on structured motifs.

**Restriction 1.** Pattern \( w_1 \) appears only once in the structured motif \( m \);

**Restriction 2.** For each \( i = 1, 2, \ldots, b-1 \), pattern \( w_i+1 \) appears only once in the two-box sub-structured motif \( w_i(d_i : D_i)w_{i+1} \).

These restrictions are not strict in practice because the probability for \( w_1 \) to occur more than once within the structured motif or for some \( w_i+1 \) to occur more than once in \( w_i(d_i : D_i)w_{i+1} \) is relatively very small. Therefore, one would expect that identifying 'significant' structured motifs would be equally successful when counting only restricted motifs and using distributional results for them or counting unrestricted motifs and using distributional results for them. On the other hand, in this paper we manifest the advantage of imposing Restrictions 1 and 2 by providing neat explicit, closed-form expressions for relevant distributions on restricted structured motifs.

3. Main results

3.1. Notation and waiting times

Denote by

- \( W_{ij} \) the family \( \{w_i, w_j\} \) consisting of the two patterns \( w_i \) and \( w_j \);
- \( T_{ij} \) the waiting time to reach pattern \( w_i \) from pattern \( w_j \);
- \( T_i^{(s)} \) the waiting time to reach pattern \( w_i \) from state \( s \);
- \( T_{Wn}^{(s)} \) the waiting time to reach the family of patterns \( W \) from pattern \( w_n \);
- \( X_{i,j,n} \) the probability to reach pattern \( w_i \) before pattern \( w_j \), given one starts from pattern \( w_n \).

Of course \( r_{i,j,n} = P(X_{i,j,n} = T_{Wn}) \). Note that there are general results on patterns which provide explicit, closed-form solutions for the probability generating functions (p.g.f.'s) of all the random variables \( T_{ij}, T_i^{(s)}, T_{Wn}, X_{i,j,n} \), and also allow exact computation of the probabilities \( r_{i,j,n} \) (cf. [6,7,10,11,12]).

Recall that \( G_Y(t) \) denotes the p.g.f. of a random variable \( Y \). For \( i = 1, \ldots, b-1 \), introduce the following random variables:

\[
\begin{align*}
F_{i+1,1|i} & = (X_{i+1,1|i} \mid X_{i+1,1|i} < D_i + k_{i+1} \text{ or } X_{i+1,1|i} > D_i + k_{i+1}), \\
S_{i+1,1|i} & = (X_{i+1,1|i} \mid D_i + k_{i+1} \leq X_{i+1,1|i} \leq D_i + k_{i+1}).
\end{align*}
\]

Actually, \( F_{i+1,1|i} \) is a random variable whose distribution equals the conditional distribution of the waiting time to reach \( w_{i+1} \) from the sub-structured motif \( m_i \), given the sub-structured motif \( m_{i+1} \) is not achieved. Likewise, the distribution of \( S_{i+1,1|i} \) equals that of the conditional distribution of the same waiting time, given the sub-structured motif \( m_{i+1} \) is achieved. Similarly to equations (3.2) and (3.3) in [12], the p.g.f.'s of \( F_{i+1,1|i} \) and \( S_{i+1,1|i} \) are given by:

\[
\begin{align*}
G_{F_{i+1,1|i}}(t) & = G_{X_{i+1,1|i}}(t) - \sum_{x=d_i+k_{i+1}}^{D_i+k_{i+1}} a_{i+1,1|i}(x) t^x \left( 1 - q_{i,1} \right)^{-1} \\
G_{S_{i+1,1|i}}(t) & = \sum_{x=d_i+k_{i+1}}^{D_i+k_{i+1}} a_{i+1,1|i}(x) t^x q_{i,1}^{-1}.
\end{align*}
\]
where
\[ a_{i+1,1j}(x) = P(X_{i+1,1j} = x) \]
and \( q_{5,i} \) is the probability of success \((w_{i+1} \text{ achieves the sub-structured motif } m_{i+1})\), that is, the probability of the event \( d_i + k_{i+1} \leq X_{i+1,1j} \leq D_i + k_{i+1} \). Namely, we have
\[ q_{5,i} = \sum_{x=d_i+k_{i+1}}^{D_i+k_{i+1}} a_{i+1,1j}(x). \] (3)
The values of the \( a_{i+1,1j}(x) \) are exactly computable through inversion of the p.g.f. of \( X_{i+1,1j} \).

3.2. The embedding semi-Markov process

In what follows we introduce a special semi-Markov process, \( Y_t \) say, with \( 2b \) states where the states are numbered from 0 through \( 2b - 1 \). As we will see in Theorem 1, this semi-Markov process models the evolution towards achieving the structured motif \( m_b \). Further, the states have the following interpretation:

- state 0: neutral initial state
- state 1: pattern \( w_1 \) is reached
- state \( i \): pattern \( w_i \) is reached from \( m_{i-1} \) without achieving \( m_i \), \( i = 2, 3, \ldots, b \)
- state \( b+1 \): the sub-structured motif \( m_{i+1} \) is achieved, \( i = 1, 2, \ldots, b - 1 \).

In other words, state \( 2b - 1 \) corresponds to the achievement of the structured motif \( m_b \).

The entries of the one-step transition probability matrix \( P \) of the embedded discrete-time Markov chain of \( Y_t \) are given by:

\[
\begin{align*}
p_{0,1} &= 1 \\
p_{1,1} &= r_{1,2}[1] \\
p_{1,2} &= r_{2,1}[1](1 - q_{5,1}) \\
p_{1,b+1} &= r_{2,1}[1] q_{5,1} \\
p_{1,i} &= 1, \quad i = 2, 3, \ldots, b, \\
p_{b+1,1} &= r_{1,i-2}[i-1] \\
p_{b+1,i} &= r_{1,i+2}[i+1] \quad i = 1, \ldots, b - 2, \\
p_{b+1,i+2} &= r_{2,i+2}[i+1](1 - q_{5,i+1}) \quad i = 1, \ldots, b - 2, \\
p_{b+1,i+1} &= r_{2,i+2}[i+1] q_{5,i+1} \quad i = 1, \ldots, b - 2, \\
p_{2b-1,2b-1} &= 1, \\
p_{i,j} &= 0, \quad \text{otherwise},
\end{align*}
\] (4)
where the \( r_{i,j}[n] \) and \( q_{5,i} \) have been introduced above. The state \( 2b - 1 \) is an absorbing one. The transition-state diagram is found in Fig. 1.

Denoting by \( H_{ij} \) the holding (sojourn) time in state \( i \), given the next state to be visited is state \( j \), we assume that

\[
\begin{align*}
H_{0,1} &\overset{d}{=} T_{1}^{(s)} \\
H_{1,1} &\overset{d}{=} X_{1,2}[1] \\
H_{1,2} &\overset{d}{=} F_{2,1}[1] \\
H_{1,b+1} &\overset{d}{=} S_{2,1}[1] \\
H_{i,1} &\overset{d}{=} T_{1}^{(s)} \quad i = 2, \ldots, b, \\
H_{b+1,1} &\overset{d}{=} X_{1,i+2}[i+1] \quad i = 1, \ldots, b - 2, \\
H_{b+1,i+2} &\overset{d}{=} F_{1,i+2}[i+1] \quad i = 1, \ldots, b - 2, \\
H_{b+1,i+1} &\overset{d}{=} S_{1,i+2}[i+1] \quad i = 1, \ldots, b - 2,
\end{align*}
\] (5)
where \( s \) is the initial letter and \( \overset{d}{=} \) stands for ‘equality in distribution’.

Denote by \( T_i \) the waiting time to reach the structured motif \( m_i \).

**Theorem 1.** The waiting time \( T_s \) for the first occurrence of the structured motif \( m_b \), given the initial letter is \( s \), is equal in distribution to the first passage time from state 0 to state \( 2b - 1 \) in the semi-Markov process \( Y_t \) which has been introduced above.

**Proof.** Recall from (5) that the holding time \( H_{0,1} \) at state 0 is equal in distribution to \( T_{1}^{(s)} \). Therefore, the waiting time until a transition from state 0 to state 1 occurs in \( Y_t \) corresponds to the waiting time for the first occurrence of pattern \( w_1 \), given the initial letter is \( s \). Upon reaching \( w_1 \) we are interested whether pattern \( w_2 \) occurs between \( d_1 \) and \( D_1 \) letters after \( w_1 \) with no further occurrence of \( w_1 \) during that time (recall Restriction 1 on structured motifs). Therefore, we look at which of the patterns \( w_1 \) and \( w_2 \) will be reached first, from \( w_1 \), and in case this is \( w_2 \) we check whether the distance is right (between \( d_1 \) and \( D_1 \) letters) for achieving the sub-structured motif \( m_2 \). Then one of the following three events occurs:
(i) the pattern $w_1$ is reached before pattern $w_2$; this happens with probability $p_{1,1}$ introduced in (4) above and the distribution of the time until reaching $w_1$, given that that event occurs, is equal to that of $H_{1,1}$ where the latter has been introduced in (5); in terms of the semi-Markov process $Y_t$, a transition from state 1 to itself has occurred.

(ii) the pattern $w_2$ is reached before pattern $w_1$, and the distance is the wrong one, that is the sub-structured motif $m_2$ has not been achieved; this happens with probability $p_{1,2}$ and the distribution of the time until reaching $w_2$, given that that event occurs, is equal to that of $H_{1,2}$; in terms of $Y_t$, a transition from state 1 to state 2 has occurred.

(iii) the pattern $w_0$ is reached before pattern $w_1$, and the distance is the right one, that is the sub-structured motif $m_0$ has been achieved; this happens with probability $p_{1,b+1}$ and the distribution of the time until reaching $w_2$, given that that event occurs, is equal to that of $H_{1,b+1}$. In terms of $Y_t$, state $b + 1$ is reached.

If event (ii) occurs then (see Restriction 2, introduced above, on structured motifs) there is no chance for the sub-structured motif $m_2$ to occur before first reaching pattern $w_1$; the waiting time for this is given by $T_{12}$ ($=H_{2,1}$). In terms of $Y_t$, a transition from state 2 to state 1 has occurred. In other words, both events (i) and (ii) get us ultimately to pattern $w_1$ (state 1 for $Y_t$).

Assume now that the sub-structured motif $m_i$, $i = 2, 3, \ldots, b - 1$, has been achieved. From that epoch of time, recalling Restrictions 1 and 2 on structured motifs, we wait until one of the patterns $w_1$ and $w_{i+1}$ is reached and in case this is $w_{i+1}$ we check whether the distance is right (between $d_i$ and $D_i$ letters) for achieving the sub-structured motif $m_{i+1}$. Analogously to the preceding case one of the following three events occurs:

(i*) the pattern $w_1$ is reached before pattern $w_{i+1}$; this happens with probability $p_{b+i-1,i}$ and the distribution of the time until reaching $w_1$, given that that event occurs, is equal to that of $H_{b+i-1,i}$;

(ii*) the pattern $w_{i+1}$ is reached before pattern $w_1$, and the distance is the wrong one, that is the sub-structured motif $m_{i+1}$ has not been achieved; this happens with probability $p_{b+i-1,i}$ and the distribution of the time until reaching $w_{i+1}$, given that that event occurs, is equal to that of $H_{b+i-1,i}$;

(iii*) the pattern $w_{i+1}$ is reached before pattern $w_1$, and the distance is the right one, that is the sub-structured motif $m_{i+1}$ has been achieved; this happens with probability $p_{b+i-1,b+i}$ and the distribution of the time until reaching $w_{i+1}$, given that that event occurs, is equal to that of $H_{b+i-1,b+i}$.

The same arguments, as those above, apply if event (ii*) occurs to conclude that one is ultimately in state 1 of the semi-Markov process $Y_t$ without being able to achieve the sub-structured motif $m_{i+1}$.

The proof of Theorem 1 is complete.

Denote by $\tau_{b}^{\text{intersite}}$ the intersite distance between two consecutive occurrences of the structured motif $m_b$. Assuming $w_b$ is not a proper subpattern of $w_1$ it is clear that $\tau_{b}^{\text{intersite}}$ is equal in distribution to the waiting time for the first occurrence of $m_b$ if starting from pattern $w_b$. Therefore, we get the following result for the intersite distance.

**Theorem 2.** If $w_b$ is not a proper subpattern of $w_1$, then the intersite distance $\tau_{b}^{\text{intersite}}$ is equal in distribution to the first passage time from state 0 to state $2b - 1$ in the semi-Markov process $Y_t$ which has been introduced above after the substitution of $H_{0,1}$ by $T_{1b}^{(s)}$ in (5) by $H_{0,1}$.

Note that in case $w_b$ is a proper subpattern of $w_1$ then we may introduce a new structured motif with $b + 1$ boxes derived from $m_b$ via splitting box 1 into two relevant boxes with a distance zero between them where the first box contains a prefix of $w_1$ which does not contain $w_b$ and the second box contains the remaining piece of pattern $w_1$. For the newly introduced structured motif the aforementioned assumption is satisfied and Theorem 2 is applicable.

Let $S_{ij}$ be the number of transitions from state $i$ to state $j$ until absorption of the semi-Markov process $Y_t$ in state $2b - 1$. Given the initial state is $k$, denote by $G_{S}^{(s)}(s)$ the joint p.g.f. of $S$ – the $2b \times 2b$ matrix whose $(i, j)$-entry is $S_{ij} = \text{where } s$ is an...
2b × 2b matrix whose (i, j)-entry is denoted by s_{ij}. That is,

\[ G^{(k)}_S(s) = E \left( \prod_{i=0}^{2b-1} s_{ij} | Y_0 = k \right), \quad k = 0, \ldots, 2b - 2. \]

Let \( G_S(s) \) be the vector whose coordinates are the \( G^{(k)}_S(s) \). An explicit solution for \( G_S(s) \) in terms of the transition probability matrix \( P \) which has been introduced in (4) above is given by (cf. [2] p.103).

\[ G_S(s) = (I - T(s))^{-1} R(s), \quad (6) \]

where I is the \((2b - 1) \times (2b - 1)\) identity matrix; for \( i, j = 0, 1, \ldots, 2b - 2 \), the \((i, j)\)-entry of the \((2b - 1) \times (2b - 1)\) matrix \( T(s) \) is given by \( p_{ij} \) where the \( p_{ij} \) have been introduced in (4) above; for \( i = 0, 1, \ldots, 2b - 2 \) the \( i\)-th coordinate of the \((2b - 1)\)-dimensional column vector \( R(s) \) is given by \( p_{i,2b-1} \). Of course the first coordinate \( G^{(0)}_S(s) \) represents the joint p.g.f. of \( S \) given the initial state is 0.

**Proposition 1.** The probability generating function of the waiting time \( \tau_b \), \( G_{\tau_b}(u) \) say, for the first occurrence of the structured motif \( m_b \) is equal to \( G^{(0)}_S(h(u)) \) where \( h(u) \) is a \( 2b \times 2b \) matrix whose \((i,j)\)-entry, \( h_{ij} \), say, \( i = 0, 1, \ldots, 2b - 1 \), is the probability generating function of \( H_{ij} \) — the holding time in state \( i \), given the next state to be visited is state \( j \), for the semi-Markov process \( Y_t \).

**Proof.** It follows from Lemma 2.1 in [9] and the basic property of semi-Markov processes that the sojourns in states given a realization of the embedded discrete-time Markov chain form a sequence of independent random variables whose distributions depend only on the occupied state and the next state to be visited.

From our results above, and with the help of any symbolic algebra tool such as Maple, one can evaluate an explicit expression for the generating function \( G_{\tau_b}(t) \) for any \( b \).

**Example (Structured Motif with 3 Boxes).** The expression for the generating function \( G^{(0)}_S(s) \) is found using formula (6). Its explicit expression is given by

\[
\frac{p_{1,4}p_{4,5}s_{0,1}s_{1,4}s_{4,5}}{1 - p_{1,1}s_{1,1} - p_{1,2}s_{1,2} - p_{1,3}s_{1,3} - p_{1,4}s_{1,4} + p_{1,4}s_{3,1} - p_{1,4}s_{4,1} - p_{1,4}s_{4,3}s_{3,1}},
\]

where the \( p_{ij} \)'s are the one-step transition probabilities given in (4). Substituting the \( s_{ij} \)'s (see Proposition 1) by the p.g.f.'s \( h_{ij} \) of the holding times \( H_{ij} \) and plugging the expressions for the \( p_{ij} \)'s from (4) in the above expression for \( G^{(0)}_S(s) \), we get the following explicit expression for the p.g.f. \( G_{\tau_b}(t) \) of the waiting time to reach the structured motif \( m_3 \), given the initial letter is \( s \) (here \( q = 1 - q \)):

\[
\begin{align*}
&= r_{2,1|1}q_{5,1}r_{3,1|2}q_{5,2}G_{S,1}(t)G_{S,1}(t)G_{S,1}(t) + r_{1,2|1}q_{5,1}r_{1,2|1}q_{5,2}G_{S,1}(t)G_{S,1}(t)A(t),
\end{align*}
\]

where

\[
A(t) = r_{1,3|2}q_{5,2}G_{S,1}(t) + r_{3,1|2}q_{5,2}G_{S,1}(t)G_{S,1}(t).
\]

4. A numerical example

Consider the following structured motif

\[
m = TTAAGC(1:5)TTTTAA(10:14)TATAAT
\]

which is a putative promotor in the bacterium *Helicobacter pylori* [13]. For assessing the significance of this motif in the genome of *H. pylori*, one evaluates the probability to reach the motif \( m \) before a given position \( t \) in a random string generated by a Markov chain on the four letter DNA alphabet \( \mathcal{A} = \{A, C, G, T\} \). The transition probabilities \( \pi_{ab}, a, b \in \mathcal{A} \), are estimated using the maximum likelihood method, that is \( \pi_{ab} = N(ab)/\sum_{b \in \mathcal{A}} N(ab) \), where \( N(ab) \) is the number of occurrences of the pair \( ab \) in the *H. pylori* genome. Then the transition matrix is:

\[
\Pi = \begin{pmatrix}
0.4170192 & 0.1862611 & 0.1314293 & 0.2652904 \\
0.2943887 & 0.1792699 & 0.2258409 & 0.3005005 \\
0.2618678 & 0.2271056 & 0.3064068 & 0.2046198 \\
0.2221892 & 0.1858492 & 0.1718480 & 0.4201127
\end{pmatrix}
\]

The p.g.f. of the waiting time to reach motif \( m \) is evaluated from formula (7). All terms involved in this formula, including the p.g.f. expressions of \( T^{(0)}_{ij} \) and \( X_{ij,t}^{(0)} \), are evaluated using the results in [10,11]. The cumulative probabilities \( F(t) := P(t_m + 1 \leq t) \) have been computed by Taylor expansion for \( t \) from 30 to 30,000 assuming that the first letter is A; 1 is added to \( t_m \) because the waiting time does not count the initial letter as a step \( F(t) = 0 \) for \( 1 \leq t \leq 29 \). Note that \( P(t_m + 1 \leq t) \) represents the probability to observe at least one occurrence of \( m \) in a sequence of length \( t \).
Table 1
Cumulative probabilities \( F(t) \) for the waiting time to reach motif \( m \) in a sequence of length \( t \).

<table>
<thead>
<tr>
<th>( t )</th>
<th>( F(t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>( 2.774 \times 10^{-10} )</td>
</tr>
<tr>
<td>100</td>
<td>( 5.415 \times 10^{-7} )</td>
</tr>
<tr>
<td>500</td>
<td>( 3.782 \times 10^{-6} )</td>
</tr>
<tr>
<td>1,000</td>
<td>( 7.832 \times 10^{-6} )</td>
</tr>
<tr>
<td>5,000</td>
<td>( 4.023 \times 10^{-5} )</td>
</tr>
<tr>
<td>10,000</td>
<td>( 8.073 \times 10^{-5} )</td>
</tr>
<tr>
<td>20,000</td>
<td>( 1.617 \times 10^{-4} )</td>
</tr>
<tr>
<td>30,000</td>
<td>( 2.427 \times 10^{-4} )</td>
</tr>
</tbody>
</table>

Table 2
Quantiles for the distribution of the waiting time to reach motif \( m \) for various level \( \alpha \).

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>( t )</th>
<th>( \alpha )</th>
<th>( t )</th>
<th>( \alpha )</th>
<th>( t )</th>
<th>( \alpha )</th>
<th>( t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 10^{-9} )</td>
<td>31</td>
<td>( 10^{-8} )</td>
<td>34</td>
<td>( 10^{-7} )</td>
<td>46</td>
<td>( 10^{-6} )</td>
<td>157</td>
</tr>
<tr>
<td>( 10^{-5} )</td>
<td>1268</td>
<td>( 10^{-4} )</td>
<td>12,380</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1 gives cumulative probability values for some length \( t \) whereas Table 2 gives the longest length \( t \) such that \( P(\tau_m + 1 \leq t) \leq \alpha \) for some level \( \alpha \).

In order to give an idea how Restrictions 1 and 2 on the structured motif affect the cumulative probabilities we run a few million simulations and derived empirical cumulative probabilities \( \phi(t) \) for the motif without Restrictions 1 and 2. We get: \( \phi(10,000) = 8.7 \times 10^{-5}, \phi(20,000) = 1.6 \times 10^{-4} \) and \( \phi(30,000) = 2.5 \times 10^{-4} \). Note that the values of \( F(t) \) are very close to those of \( \phi(t) \). Such comparisons for smaller values of \( t \) would require much larger numbers of simulations.

References