16

Occurrence of Patterns and Motifs in Random Strings

Valeri T. Stefanov

School of Mathematics and Statistics, University of Western Australia, Crawley, Australia

Abstract: Patterns and motifs on finite alphabets are of interest in many applied areas, such as computational molecular biology, computer science, communication theory, and reliability theory. The exact distribution theory associated with occurrences of patterns (single or compound) and motifs, in random strings of letters, is treated in this chapter. The strings are generated by a Markov source, and for the case of single patterns, they are generated by general discrete-time or continuous-time models. Here, the interest is in finding closed-form expressions for the distributions of the following quantities: (i) the waiting time until the first occurrence of a pattern (motif), (ii) the intersite distances between consecutive occurrences of such, and (iii) the count of occurrences of a pattern, or more generally, the weighted count of occurrences of a compound pattern, both within a finite time horizon. General exact distribution results are discussed. Also, a brief guide on various methodological tools used in the area is provided in the Introduction.

Keywords and phrases: Pattern, motif, waiting time, Markov chain, semi-Markov process

16.1 Introduction

Patterns and motifs on finite alphabets are of interest in many applied areas, such as computational molecular biology, computer science, communication theory, and reliability theory. A word on an alphabet is called a single pattern, and a set of distinct single patterns (words) is called a compound pattern. The strings (texts) of letters can be generated either by independent and identically distributed multinomial trials, or by general discrete-time or continuous-time models (Markov chains or semi-Markov processes). The main interest, from a

J. Glaz et al. (eds.), *Scan Statistics: Methods and Applications*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4749-0_16,

probabilistic/statistical point of view, is in finding practicable closed-form expressions for the distributions of the following quantities: the waiting time until the first occurrence of a pattern (single or compound) or motif, the intersite distances between consecutive occurrences of such, and the count(s) of occurrences of a pattern(s) or motif within a finite time horizon. Motifs are special cases of compound patterns which usually contain a huge number of distinct single patterns.

The theory on pattern occurrence attracted a variety of methodological tools. For example, the following methodologies have been widely used in the literature: combinatorial methods and classical probabilistic methods based on conditioning arguments, Markov chain embeddings, Markov renewal embeddings, exponential families, martingale techniques, and automata theory. The usefulness of these methodologies to the area is well illustrated in the sources which follow.

Runs are the simplest patterns. Feller (1950) showed how recurrent event theory can be used to solve problems about success runs. For a comprehensive account of the literature on runs see Balakrishnan and Koutras (2002). The key to handling complex patterns was provided by Conway's leading numbers, which account for the overlapping structure of a pattern. Guibas and Odlyzko (1981) derived results applying elementary methods, and Chryssaphinou and Papastavridis (1990) extended them to more general models [see also Robin and Daudin (1999, 2001), Rukhin (2002, 2006), Han and Hirano (2003), and Inoue and Aki (2007)]. Li (1980) introduced martingale techniques to the area, and Gerber and Li (1981) combined the latter with a relevant Markov chain embedding. Martingale tools have also been used in Pozdnyakov *et al.* (2005), Glaz *et al.* (2006), and Pozdnyakov (2008).

Markov chain embeddings have been widely used in the area for treating problems on pattern occurrence; a few relevant sources are Fu (1996), Chadjiconstantinidis, Antzoulakos, and Koutras (2000), Antzoulakos (2001), Fu and Chang (2002), and Fu and Lou (2003). Blom and Thorburn (1982) made connections with Markov renewal theory, and this was systematically exploited by Biggins and Cannings (1987) and Biggins (1987). Stefanov and Pakes (1997) introduced exponential family methodology, combined with a minimal Markov chain embedding, and Stefanov (2000) extended it in combination with suitable Markov renewal embeddings to handle some special compound patterns (sets of runs).

Nicodème, Salvy, and Flajolet (2002) used automata theory comprehensively. Nuel (2008) combined automata theory with Markov chain embeddings and elaborated on a route which leads, for any given pattern(s), to a minimal embedding Markov chain. Reinert, Schbath, and Waterman (2000) provided a survey on some probabilistic tools used in the theory of patterns, and Szpankowski (2001) treated problems on pattern occurrence associated with average case analysis of string searching algorithms. The first exact distributional results on structured motifs are found in Stefanov, Robin, and Schbath (2007) [cf. also Robin *et al.* (2002), Nuel (2008), and Pozdnyakov (2008)].

In this chapter, results are discussed which provide explicit, closed-form solutions for the distributions of the aforementioned random quantities associated with the occurrence of patterns and structured motifs. These results are derived using predominantly simple probabilistic tools. Also, for a given alphabet, they require a preliminary (easy) evaluation of a few basic characteristics, and then each pattern case is covered in an automated way.

In Sections 16.2 and 16.3 we discuss single patterns. The strings are generated by discrete- or continuous-time semi-Markov processes. The exact distribution of the waiting time until the first occurrence of a pattern, given any (fixed) portion of it has been reached, is found. Also joint distributional results are discussed. The method relies on the knowledge of basic characteristics associated with the underlying model used to generate the strings. These basic characteristics are the probability generating functions (pgf's) of the waiting times until another letter of the alphabet is reached. In other words, we need to know only the pgf's of the waiting times until the simplest special patterns consisting of a single letter from the alphabet are first reached. These pgf's can be evaluated using well-known analytical results if the underlying model is a discrete- or continuous-time finite-state semi-Markov process. In terms of these basic characteristics, simple recurrence relations are provided; these lead to exact evaluation of the relevant pgf's for any pattern. The results on single patterns, as provided in Sections 16.2 and 16.3, lead to an easy solution for compound patterns, which consist of a small to moderate number of distinct single patterns. This is discussed in Subsection 16.4.1. The distribution of the count, and more generally the weighted count, of a compound pattern within a finite time horizon is discussed in Subsection 16.4.2. A neat explicit expression is derived for this distribution in terms of the aforementioned waiting time distributions. The result in Subsection 16.4.2 has not appeared in the literature before. Structured motifs are covered in Subsection 16.4.3. It is shown that results on compound patterns, consisting of only two single patterns, are enough to derive exact distribution results on structured motifs.

16.2 Patterns: Discrete-Time Models

In this section we explain how to derive a closed-form expression for the pgf of the waiting time to reach a pattern (word) starting from either a given letter or an already-achieved portion of the pattern. The strings of letters are generated by a finite-state discrete-time Markov chain whose state space and states are also called alphabet and letters, respectively.

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 $p_{i,j}$, $i, j = 1, 2, \ldots, N$. Denote by $g_{i,j}(t)$ the pgf of the waiting time, $\tau_{i,j}$, to reach state j from state i, that is $g_{i,j}(t) = E(t^{\tau_{i,j}})$, and

$$\tau_{i,j} = \inf\{n : X(n) = j | X(0) = i\}.$$

We assume $\tau_{i,i} = 0$, and therefore $g_{i,i}(t) = 1$, for each i. The first return time to state i is denoted by $\tilde{\tau}_{i,i}$, that is,

$$\tilde{\tau}_{i,i} = \inf\{n > 0 : X(n) = i | X(0) = i\},$$

and its pgf is denoted by $\tilde{g}_{i,i}(t)$.

The pattern of interest is $\mathbf{w}_k = w_1 w_2 \ldots w_k$, where $1 \leq w_i \leq N$, $i = 1, 2, \ldots, k$. For $j < k$, the subpattern $\mathbf{w}_j$ is also called a prefix of $\mathbf{w}_k$. For each j, $j = 2, 3, \ldots, k-1$, and $r < j$, and each n, $n = 1, 2, \ldots, N$, denote by $I_{r,j,n}$ the indicator function which is equal to one if and only if none of the strings $w_i w_{i+1} \ldots w_j n$ for $i = 2, 3, \ldots, r$ is a prefix of $\mathbf{w}_k$ but $w_{r+1} w_{r+2} \ldots w_j n$ is. Also, the indicator function $I_{j,j,n}$ is equal to one if and only if none of the strings $w_i w_{i+1} \ldots w_j n$ for $i = 2, 3, \ldots, j$ is a prefix of $\mathbf{w}_k$.

Denote by $G_j^{(s)}(t)$ $(\tilde{G}_j^{(s)}(t))$, $j = 1, 2, \ldots, k$, the pgf of the waiting time to reach the pattern $\mathbf{w}_j$ from state s, allowing (not allowing) the initial state s to contribute to the pattern. Also, denote by $G_j^{(\mathbf{w}_r)}(t), 1 \leq r \leq j$, the pgf of the waiting time to reach the pattern $\mathbf{w}_j$, given that the pattern $\mathbf{w}_r$ has already been reached (note that $G_j^{(\mathbf{w}_j)}(t) = 1$). The following theorem provides a simple route for evaluating these pgf's knowing the pgf's, $g_{i,j}(t)$, of the transition times between the states of the original Markov chain $X(n)$. The expressions for the pgf's $g_{i,j}(t)$ are easily recoverable from well-known analytical results [see Theorem 2.19 on page 81 of Kijima (1997)], for any given finite-state Markov chain with not too large a state space.

Theorem 16.2.1 *Let the pattern of interest be $\mathbf{w}_k$. The following recurrence relations hold for each j, $j = 1, 2, \ldots, k-1$, and each r, $r = 1, 2, \ldots, j$ (with the convention $\sum_{i=1}^{0} = 0$):*

$$\tilde{G}_{j+1}^{(s)}(t) = \frac{p_{w_j, w_{j+1}} t \tilde{G}_j^{(s)}(t)}{1 - \displaystyle\sum_{n=1,\, n \neq w_{j+1}}^{N} p_{w_j, n} t \left(\sum_{i=1}^{j-1} I_{i,j,n} G_j^{(\mathbf{w}_{j-i+1})}(t) + I_{j,j,n} G_j^{(n)}(t) \right)},$$

$$G_{j+1}^{(\mathbf{w}_r)}(t) = \frac{p_{w_j, w_{j+1}} t G_j^{(\mathbf{w}_r)}(t)}{1 - \displaystyle\sum_{n=1,\, n \neq w_{j+1}}^{N} p_{w_j, n} t \left(\sum_{i=1}^{j-1} I_{i,j,n} G_j^{(\mathbf{w}_{j-i+1})}(t) + I_{j,j,n} G_j^{(n)}(t) \right)},$$

where

$$\begin{aligned}
\tilde{G}^{(s)}_{j+1}(t) &= G^{(s)}_{j+1}(t), \qquad \text{if } s \neq w_1,\\
\tilde{G}^{(w_1)}_{j+1}(t) &= \tilde{g}_{w_1,w_1}(t) G^{(w_1)}_{j+1}(t),\\
G^{(s)}_1(t) &= g_{s,w_1}(t),\\
\tilde{G}^{(s)}_1(t) &= \tilde{g}_{w_1,w_1}(t) = \sum_{n=1}^{N} p_{w_1,n} t g_{n,w_1}(t),
\end{aligned}$$

and the $g_{i,j}(t)$ and the indicator functions $I_{i,j,n}$ are as above.

The pgf of the intersite distance between consecutive occurrences of the pattern $\mathbf{w}_k$ is given by $G^{(\mathbf{w}_j)}_k(t)$, where j is the largest integer such that $\mathbf{w}_j$ is a proper prefix as well as a suffix of the pattern $\mathbf{w}_k$. Also, the pgf of the waiting time until the r-th occurrence of the pattern $\mathbf{w}_k$, given the initial state i, is equal to $G^{(i)}_k(t)\left(G^{(\mathbf{w}_j)}_k(t)\right)^{r-1}$, where j has the same property as above.

The proof of Theorem 16.2.1 is based on the following simple idea. Let $\tau_{\mathbf{w}_j|\bar{\mathbf{w}}_{j+1}}$ be the waiting time for the first return (strictly positive) from pattern $\mathbf{w}_j$ to itself given that the pattern $\mathbf{w}_{j+1}$ is not achieved. Of course, the pattern $\mathbf{w}_{j+1}$ is not achieved if the first state visited is not state w_{j+1}. Therefore, the pgf of $\tau_{\mathbf{w}_j|\bar{\mathbf{w}}_{j+1}}$ is equal to

$$g_{\tau_{\mathbf{w}_j|\bar{\mathbf{w}}_{j+1}}}(t) = \sum_{n=1,\, n\neq w_{j+1}}^{N} \frac{p_{w_j,n}t}{1-p_{w_j,w_{j+1}}} \left(\sum_{i=1}^{j-1} I_{i,j,n} G^{(\mathbf{w}_{j-i+1})}_j(t) + I_{j,j,n} G^{(n)}_j(t) \right).$$

Then, the waiting time to reach pattern $\mathbf{w}_{j+1}$ starting from state s is equal to one plus a geometric sum of independent random variables, $Y_1, Y_2, \ldots$, say, such that Y_1 has the distribution of the waiting time to reach subpattern $\mathbf{w}_j$ from state s and the remaining Y_n have the distribution of $\tau_{\mathbf{w}_j|\bar{\mathbf{w}}_{j+1}}$. This implies that

$$\tilde{G}^{(s)}_{j+1}(t) = \frac{p_{w_j,w_{j+1}} t \tilde{G}^{(s)}_j(t)}{1 - \displaystyle\sum_{n=1,\, n\neq w_{j+1}}^{N} p_{w_j,n} t \left(\sum_{i=1}^{j-1} I_{i,j,n} G^{(\mathbf{w}_{j-i+1})}_j(t) + I_{j,j,n} G^{(n)}_j(t) \right)}.$$

A detailed proof of Theorem 16.2.1 is found in Stefanov (2003).

16.3 Patterns: General Discrete-Time and Continuous-Time Models

In this section, extensions of the result from the preceding section are presented. Finite-state semi-Markov processes, with either discrete- or continuous-time parameters, are the underlying models for generating the strings. Also, joint distributions of the waiting time to reach a pattern, together with the associated counts of occurrences of each letter, are of interest.

16.3.1 Waiting times

The notation from the preceding section is further used here for identifying the counterparts of similar quantities. For example, $g_{i,j}(t)$ will again denote the pgf of the waiting time to reach state j from state i in the more general discrete- or continuous-time model considered here.

Let $\{X(u)\}_{u \geq 0}$ (the time parameter u may be either discrete or continuous) be a semi-Markov process whose associated embedded discrete-time Markov chain has a finite state space $\{1, 2, \ldots, N\}$ and one-step transition probabilities $p_{i,j}$, $i, j = 1, 2, \ldots, N$. For a formal definition of a semi-Markov process see Çinlar (1975). Denote by $\phi_{i,j}(t)$ the pgf of the holding (sojourn) time in state i, given that the next state to be visited is state j (if the holding time distributions are discrete, then the time parameter is discrete). We denote by $g_{i,j}(t)$ the pgf of the waiting time, $\tau_{i,j}$, to reach state j from state i; that is, $g_{i,j}(t) = E(t^{\tau_{i,j}})$, where

$$\tau_{i,j} = \inf\{u : X(u) = j | X(0) = i\}.$$

We assume $\tau_{i,i} = 0$, and therefore $g_{i,i}(t) = 1$, for each i. The first return time to state i is denoted by $\tilde{\tau}_{i,i}$ and its pgf by $\tilde{g}_{i,i}(t)$. Of course, if $X(u)$ is a discrete-time Markov chain,

$$\tilde{\tau}_{i,i} = \inf\{u > 0 : X(u) = i | X(0) = i\},$$

and if $X(u)$ is a continuous-time Markov chain,

$$\tilde{\tau}_{i,i} = \inf\{u > 0 : X(u) = i, X(u-) \neq i | X(0) = i\}.$$

If $X(u)$ is a general semi-Markov process, then $\tilde{\tau}_{i,i}$ is understood to be the waiting time to reach state i from itself given that at least one transition has been made in the associated embedded discrete-time Markov chain. This clarifies the interpretation of $\tilde{\tau}_{i,i}$ in case one-step transitions are allowed from a state to itself in the embedded discrete-time Markov chain.

Again, as in the preceding section, the pattern of interest is denoted by $\mathbf{w}_k$. Denote by $G_j^{(s)}(t)$ $(\tilde{G}_j^{(s)}(t))$, $j = 1, 2, \ldots, k$, the pgf of the waiting time to reach the pattern $\mathbf{w}_j$ from state s, allowing (not allowing) the initial state s to contribute to the pattern. Also denote by $G_j^{(\mathbf{w}_r)}(t), 1 \leq r \leq j$, the pgf of the waiting time to reach the pattern $\mathbf{w}_j$, given that the pattern $\mathbf{w}_r$ has already been reached (note that $G_j^{(\mathbf{w}_j)}(t) = 1$). The following theorem provides a simple route for evaluating these pgf's in terms of the following characteristics of the original semi-Markov process $X(u)$: the pgf's, $g_{i,j}(t)$, of the transition times between the states, the pgf's, $\phi_{i,j}(t)$, of the holding time distributions, and the transition probabilities, $p_{i,j}$, of the embedded discrete-time Markov chain.

Theorem 16.3.1 *Let the pattern of interest be $\mathbf{w}_k$. The following recurrence relations hold for each j, $j = 1, 2, \ldots, k-1$, and each r, $r = 1, 2, \ldots, j$ (with the convention $\sum_{i=1}^{0} = 0$):*

$$\tilde{G}_{j+1}^{(s)}(t) = \frac{p_{w_j,w_{j+1}}\phi_{w_j,w_{j+1}}(t)\tilde{G}_j^{(s)}(t)}{1 - \sum\limits_{\substack{n=1,\\ n \neq w_{j+1}}}^{N} p_{w_j,n}\phi_{w_j,n}(t)\left(\sum\limits_{i=1}^{j-1} I_{i,j,n} G_j^{(\mathbf{w}_{j-i+1})}(t) + I_{j,j,n} G_j^{(n)}(t)\right)},$$

$$G_{j+1}^{(\mathbf{w}_r)}(t) = \frac{p_{w_j,w_{j+1}}\phi_{w_j,w_{j+1}}(t) G_j^{(\mathbf{w}_r)}(t)}{1 - \sum\limits_{\substack{n=1,\\ n \neq w_{j+1}}}^{N} p_{w_j,n}\phi_{w_j,n}(t)\left(\sum\limits_{i=1}^{j-1} I_{i,j,n} G_j^{(\mathbf{w}_{j-i+1})}(t) + I_{j,j,n} G_j^{(n)}(t)\right)},$$

where

$$\begin{aligned}
\tilde{G}_{j+1}^{(s)}(t) &= G_{j+1}^{(s)}(t), \qquad \text{if } s \neq w_1,\\
\tilde{G}_{j+1}^{(w_1)}(t) &= \tilde{g}_{w_1,w_1}(t) G_{j+1}^{(w_1)}(t),\\
G_1^{(s)}(t) &= g_{s,w_1}(t),\\
\tilde{G}_1^{(w_1)}(t) &= \tilde{g}_{w_1,w_1}(t) = \sum_{n=1}^{N} p_{w_1,n}\phi_{w_1,n}(t) g_{n,w_1}(t).
\end{aligned}$$

The proof is based on the same idea as that used to prove Theorem 16.2.1. Similarly to the preceding section, denote by $\tau_{\mathbf{w}_j|\bar{\mathbf{w}}_{j+1}}$ the waiting time to reach $\mathbf{w}_j$ from itself given that the pattern $\mathbf{w}_{j+1}$ is not achieved. Then one may notice that the waiting time to reach pattern $\mathbf{w}_{j+1}$ starting from state s is equal to the sum of two independent random variables, where the first has a pgf which equals $\phi_{w_j,w_{j+1}}(t)$ and the second one is a geometric sum of independent random variables, $Y_1, Y_2, \ldots$, say, such that Y_1 has the distribution of the waiting time to reach subpattern $\mathbf{w}_j$ from state s and the remaining Y_n have the distribution of $\tau_{\mathbf{w}_j|\bar{\mathbf{w}}_{j+1}}$.

16.3.2 Joint generating functions associated with waiting times

In this subsection we consider the same general semi-Markov model $X(u)$ that has been introduced in the preceding subsection. Recall that its embedded discrete-time Markov chain has N states. Throughout this subsection these states will be called ‘symbols’. Again the notation from the preceding subsections is further used in this subsection for identifying the counterparts of similar quantities (such as $G_j^{(s)}(\cdot)$, etc.). Note that basic quantities of the underlying model, such as $\tau_{i,j}$ and $\phi_{i,j}$, have the same meaning as that in the preceding subsection.

Let $C_i(u)$ be the count of occurrences of symbol i up to time u, and let $g_{i,j}(\underline{\mathbf{t}})$, where $\underline{\mathbf{t}} = (t_0, t_1, \ldots, t_N)$, be the joint pgf of $(\tau_{i,j}, C_1(\tau_{i,j}), \ldots, C_N(\tau_{i,j}))$, where the $\tau_{i,j}$ have been introduced in the preceding subsection. Likewise, let $\tilde{g}_{i,i}(\underline{\mathbf{t}})$ be the joint pgf of $(\tilde{\tau}_{i,i}, C_1(\tilde{\tau}_{i,i}), \ldots, C_N(\tilde{\tau}_{i,i}))$, where again the $\tilde{\tau}_{i,i}$ have been introduced in the preceding subsection. Note that $g_{i,i}(\underline{\mathbf{t}}) = 1$. Denote by $\nu_j^{(s)}$ the waiting time to reach the pattern $\mathbf{w}_j$ from state s. Let $G_j^{(s)}(\underline{\mathbf{t}})$, $(\tilde{G}_j^{(s)}(\underline{\mathbf{t}}))$, be the joint pgf of $\nu_j^{(s)}, C_1(\nu_j^{(s)}), \ldots, C_N(\nu_j^{(s)})$, allowing (not allowing) the first symbol to contribute to the pattern. Further, let $\nu_j^{(\mathbf{w}_r)}$ be the waiting time to reach the pattern $\mathbf{w}_j$ from the already-reached prefix $\mathbf{w}_r$, and let $G_j^{(\mathbf{w}_r)}(\underline{\mathbf{t}})$ be the joint pgf of $\nu_j^{(\mathbf{w}_r)}, C_1(\nu_j^{(\mathbf{w}_r)}), \ldots, C_N(\nu_j^{(\mathbf{w}_r)})$. Note that the methodology introduced in Stefanov (2000; see Section 3) yields explicit expressions for the pgf’s $g_{i,j}(\underline{\mathbf{t}})$ associated with any given semi-Markov process, whose embedded discrete-time Markov chain has a relatively small number of states. Therefore, the recurrence relations in the following theorem provide a simple route for explicit evaluation of the joint pgf’s of the waiting time to reach, or the intersite distance between two consecutive occurrences of, a pattern and the associated counts of occurrences of the corresponding symbols (letters).

Theorem 16.3.2 *Let the pattern of interest be $\mathbf{w}_k$. The following recurrence relations hold for each j, $j = 1, 2, \ldots, k-1$, and each r, $r = 1, 2, \ldots, j$:*

$$\tilde{G}_{j+1}^{(s)}(\underline{\mathbf{t}}) = \frac{p_{w_j,w_{j+1}} t_{w_{j+1}} \phi_{w_j,w_{j+1}}(t_0) \tilde{G}_j^{(s)}(\underline{\mathbf{t}})}{1 - \sum\limits_{\substack{n=1,\\ n \neq w_{j+1}}}^{N} p_{w_j,n} t_n \phi_{w_j,n}(t_0) \left(\sum\limits_{i=1}^{j-1} I_{i,j,n} G_j^{(\mathbf{w}_{j-i+1})}(\underline{\mathbf{t}}) + I_{j,j,n} G_j^{(n)}(\underline{\mathbf{t}}) \right)},$$

$$G_{j+1}^{(\mathbf{w}_r)}(\underline{\mathbf{t}}) = \frac{p_{w_j,w_{j+1}} t_{w_{j+1}} \phi_{w_j,w_{j+1}}(t_0) G_j^{(\mathbf{w}_r)}(\underline{\mathbf{t}})}{1 - \sum\limits_{\substack{n=1,\\ n \neq w_{j+1}}}^{N} p_{w_j,n} t_n \phi_{w_j,n}(t_0) \left(\sum\limits_{i=1}^{j-1} I_{i,j,n} G_j^{(\mathbf{w}_{j-i+1})}(\underline{\mathbf{t}}) + I_{j,j,n} G_j^{(n)}(\underline{\mathbf{t}}) \right)},$$

where

$$
\begin{aligned}
\tilde{G}^{(s)}_{j+1}(\underline{\mathbf{t}}) &= G^{(s)}_{j+1}(\underline{\mathbf{t}}), \qquad \text{if} \;\; s \neq w_1, \\
\tilde{G}^{(w_1)}_{j+1}(\underline{\mathbf{t}}) &= \tilde{g}_{w_1,w_1}(\underline{\mathbf{t}}) G^{(w_1)}_{j+1}(\underline{\mathbf{t}}), \\
G^{(s)}_{1}(\underline{\mathbf{t}}) &= g_{s,w_1}(\underline{\mathbf{t}}), \\
\tilde{G}^{(w_1)}_{1}(\underline{\mathbf{t}}) &= \tilde{g}_{w_1,w_1}(\underline{\mathbf{t}}) = \sum_{n=1}^{N} p_{w_1,n} t_n \phi_{w_1,n}(t_0) g_{n,w_1}(\underline{\mathbf{t}}).
\end{aligned}
$$

The proof of this theorem is found in Stefanov (2003).

16.4 Compound Patterns

Throughout this section we assume that the strings are generated by discrete-time Markov chains.

16.4.1 Compound patterns containing a small number of single patterns

Denote by $\mathbf{W}$ a compound pattern which consists of k distinct single patterns, $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \ldots, \mathbf{w}^{(k)}$. The latter may have different lengths, and it is assumed that none of them is a proper substring of any of the others. Let $\mathbf{a}$ be an arbitrary pattern; in particular, if $\mathbf{a}$ has length 1, that is, it is equal to a particular letter, s say, then we will denote $\mathbf{a}$ by s. Introduce the following quantities.

$T_{\mathbf{a},\mathbf{W}}$ — the waiting time, starting from pattern $\mathbf{a}$, to reach for the first time the compound pattern $\mathbf{W}$; if $\mathbf{a}$ equals one of the $\mathbf{w}^{(i)}$, then this waiting time is assumed to be greater than 0;

$T_{\mathbf{a},\mathbf{W}|\mathbf{w}^{(j)}}$ — the waiting time, starting from pattern $\mathbf{a}$, to reach for the first time the compound pattern $\mathbf{W}$, given that $\mathbf{W}$ is reached via $\mathbf{w}^{(j)}$;

$T_{\mathbf{a},\mathbf{b}}$ — the waiting time to reach pattern $\mathbf{b}$ starting from pattern $\mathbf{a}$;

$X_{i,j}$ — the interarrival time between two consecutive occurrences of pattern $\mathbf{W}$, given that the starting pattern is $\mathbf{w}^{(i)}$ and the reached pattern is $\mathbf{w}^{(j)}$;

$r_{i,j}$ — the probability that the first reached pattern from $\mathbf{W}$ is $\mathbf{w}^{(j)}$, given that the starting pattern is $\mathbf{w}^{(i)}$.

Of course, $X_{i,j} = T_{\mathbf{w}^{(i)},\mathbf{W}|\mathbf{w}^{(j)}}$. Introduce the following pgf's:

$$G_{\mathbf{a},\mathbf{W},j}(t) = \sum_{n=1}^{\infty} P\left(T_{\mathbf{a},\mathbf{W}} = T_{\mathbf{a},\mathbf{W}|\mathbf{w}^{(j)}} = n\right) t^n, \qquad j = 1, 2, \ldots, k,$$

and recall that by $G_Y(t)$ we denote the pgf of a random variable Y. Clearly,

$$r_{i,j} = P\left(T_{\mathbf{w}^{(i)},\mathbf{W}} = T_{\mathbf{w}^{(i)},\mathbf{W}|\mathbf{w}^{(j)}}\right) = G_{\mathbf{w}^{(i)},\mathbf{W},j}(1).$$

Also, it is easy to see that

$$G_{X_{i,j}}(t) = \frac{G_{\mathbf{w}^{(i)},\mathbf{W},j}(t)}{r_{i,j}}\ .$$

Therefore, both the $r_{i,j}$ and the pgf's $G_{X_{i,j}}(t)$ can be recovered from the pgf's $G_{\mathbf{w}^{(i)},\mathbf{W},j}(t)$. The following theorem [see Chryssaphinou and Papastavridis (1990) and Gerber and Li (1981)] provides, for each pattern $\mathbf{a}$, a system of linear equations from which one can recover the pgf's $G_{\mathbf{a},\mathbf{W},j}(t)$ and $G_{T_{\mathbf{a},\mathbf{W}}}(t)$ in terms of the pgf's $G_{T_{\mathbf{w}^{(i)},\mathbf{w}^{(j)}}}(t)$. The $G_{T_{\mathbf{w}^{(i)},\mathbf{w}^{(j)}}}(t)$ are derived from the results in Section 16.2.

Theorem 16.4.1 *The following identities hold:*

$$\begin{aligned} G_{T_{\mathbf{a},\mathbf{W}}}(t) &= \sum_{j=1}^{k} G_{\mathbf{a},\mathbf{W},j}(t), \\ G_{T_{\mathbf{a},\mathbf{w}^{(i)}}}(t) &= \sum_{j=1}^{k} G_{T_{\mathbf{w}^{(i)},\mathbf{w}^{(j)}}}(t) G_{\mathbf{a},\mathbf{W},j}(t), \qquad i = 1, 2, \ldots, k. \end{aligned}$$

In particular, we get the following explicit expressions for the $G_{\mathbf{w}^{(i)},\mathbf{W},j}(t)$ in terms of the $G_{T_{\mathbf{w}^{(i)},\mathbf{w}^{(j)}}}(t)$ if the compound pattern $\mathbf{W} = \{\mathbf{w}^{(1)}, \mathbf{w}^{(2)}\}$ consists of two patterns. For brevity, $G_{T_{i,j}}$ below stands for $G_{T_{\mathbf{w}^{(i)},\mathbf{w}^{(j)}}}(t)$.

$$\begin{aligned} G_{\mathbf{w}^{(1)},\mathbf{W},1}(t) &= \frac{G_{T_{1,1}} G_{T_{2,2}} - G_{T_{1,2}}^2}{G_{T_{1,1}} G_{T_{2,2}} - G_{T_{1,2}} G_{T_{2,1}}}\ , \\ G_{\mathbf{w}^{(1)},\mathbf{W},2}(t) &= \frac{G_{T_{1,1}} G_{T_{1,2}} - G_{T_{1,1}} G_{T_{2,1}}}{G_{T_{1,1}} G_{T_{2,2}} - G_{T_{1,2}} G_{T_{2,1}}}\ , \\ G_{\mathbf{w}^{(2)},\mathbf{W},1}(t) &= \frac{G_{T_{2,1}} G_{T_{2,2}} - G_{T_{1,2}} G_{T_{2,2}}}{G_{T_{1,1}} G_{T_{2,2}} - G_{T_{1,2}} G_{T_{2,1}}}\ , \\ G_{\mathbf{w}^{(2)},\mathbf{W},2}(t) &= \frac{G_{T_{1,1}} G_{T_{2,2}} - G_{T_{2,1}}^2}{G_{T_{1,1}} G_{T_{2,2}} - G_{T_{1,2}} G_{T_{2,1}}}\ . \end{aligned}$$

16.4.2 Weighted counts of compound patterns

A quantity of interest is the count of occurrences of a compound pattern, $\mathbf{W}$ say (as introduced in Subsection 16.4.1), within a finite time horizon. A more general quantity is the weighted count of pattern occurrences which attaches a weight, h_i say, to each occurrence of a single pattern, $\mathbf{w}^{(i)}$, from $\mathbf{W}$. More specifically, introduce

$$H_{\mathbf{W}}(t) = \sum_{i=1}^{k} h_i N_{\mathbf{w}^{(i)}}(t),$$

where $N_{\mathbf{w}^{(i)}}(t)$ is the count of occurrences of pattern $\mathbf{w}^{(i)}$ within a time interval of length t. Recall the meaning of the $r_{i,j}$, $X_{i,j}$, and $T_{s,\mathbf{W}|\mathbf{w}^{(j)}}$ which are introduced in Subsection 16.4.1. Of course, the occurrence of $\mathbf{W}$ can be modelled by a k-state semi-Markov process, where an entry to state i identifies an occurrence of pattern $\mathbf{w}^{(i)}$. The one-step transition probabilities of the embedded discrete-time Markov chain of this semi-Markov process are the $r_{i,j}$. The holding time at state i, given that the next state to be visited is state j, is identified by the random variable $X_{i,j}$. For each initial letter, s say, we augment this semi-Markov process with one initial state, 0 say, and relevant one-step transition probabilities and holding times as follows (we denote the probability to move from state 0 to state j by $r_{0,j}$):

$$r_{0,0} = 0, \quad r_{0,j} = G_{s,\mathbf{W},j}(1), \quad j = 1, 2, \ldots, k,$$

and the holding time at state 0, given that the next state to be visited is state j, is identified by $T_{s,\mathbf{W}|\mathbf{w}^{(j)}}$, where the latter and $G_{s,\mathbf{W},j}(t)$ are introduced in Subsection 16.4.1. Now consider the semi-Markov processes, Y_t say, derived from that above as follows. The state space has $(k+1)^2$ states, identified by the pairs $(i, j),\ i, j = 0, 1, \ldots, k$. The process Y_t enters state (i, j) if pattern $\mathbf{w}^{(i)}$ is reached, given that the next occurrence of $\mathbf{W}$ is via pattern $\mathbf{w}^{(j)}$. The initial states are the states $(0, j)$ for $j = 1, 2, \ldots, k$, and the initial probabilities are the $r_{0,j}$. Clearly, the holding time distributions for this new semi-Markov process do not depend on the next state visited. Also, the holding time in state (i, j) is identified by the random variable $X_{i,j}$, and that in state $(0, j)$ by $T_{s,\mathbf{W}|\mathbf{w}^{(j)}}$. Then the weighted count $H_{\mathbf{W}}(t)$, introduced above, is equal to

$$H_{\mathbf{W}}(t) = \sum_{i=0}^{k}\sum_{j=0}^{k} h_i N_{(i,j)}(t),$$

where $N_{(i,j)}(t)$ counts the number of visits of Y_t to state (i, j) within a time interval of length t. Denote by $\nu_{(i_1,j_1),(i_2,j_2)}$ the first passage time of Y_t from state (i_1, j_1) to state (i_2, j_2) and by $L_{H_{\mathbf{W}}}^{\nu_{(i_1,j_1),(i_2,j_2)}}(s_1, s_2)$ the joint Laplace transform of the random variables $\nu_{(i_1,j_1),(i_2,j_2)}$ and $H_{\mathbf{W}}\left(\nu_{(i_1,j_1),(i_2,j_2)}\right)$, that is,

$$L_{H_{\mathbf{W}}}^{\nu_{(i_1,j_1),(i_2,j_2)}}(s_1, s_2) = E\left(\exp\left(-s_1\nu_{(i_1,j_1),(i_2,j_2)} - s_2 H_{\mathbf{W}}(\nu_{(i_1,j_1),(i_2,j_2)})\right)\right).$$

Closed-form expressions for the $L_{H_{\mathbf{W}}}^{\nu(i_1,j_1),(i_2,j_2)}(s_1, s_2)$ are derivable in terms of the $r_{i,j}$ and the Laplace transforms of the $X_{i,j}$ and the $T_{s,\mathbf{W}|\mathbf{w}^{(j)}}$, as explained in Stefanov (2006) for general reward functions on semi-Markov processes. Let

$$L_{t,H_{\mathbf{W}}}^{(s)}(s_1, s_2) = \int_0^\infty \int_0^\infty e^{-s_1 t - s_2 x} P\left(H_{\mathbf{W}}(t) \leq x | \text{ the initial letter is } s\right) dxdt$$

The following theorem follows from a general result on reward functions for semi-Markov processes [see Theorem 2.1 in Stefanov (2006)]. It provides an explicit, closed-form expression for the Laplace transform, $L_{t,H_{\mathbf{W}}}^{(s)}(s_1, s_2)$, of the weighted count of $\mathbf{W}$ occurrences within a time interval of length t, in terms of the $r_{i,j}$, the Laplace transforms, $\mathcal{L}[X_{i,j}](\cdot)$, of the interarrival times $X_{i,j}$ of the compound pattern $\mathbf{W}$, and the Laplace transforms, $\mathcal{L}[T_{s,\mathbf{W}|\mathbf{w}^{(j)}}](\cdot)$, of the waiting time to reach $\mathbf{W}$ from an initial letter s, for $s = 1, 2, \ldots, N$.

Theorem 16.4.2 *The following identity holds for the Laplace transform* $L_{t,H_{\mathbf{W}}}^{(s)}$:

$$L_{t,H_{\mathbf{W}}}^{(s)}(s_1, s_2) = \sum_{m=1}^{k} r_{0,m} \sum_{i,j=1}^{k} \frac{\left(1 - \mathcal{L}[X_{i,j}](s_1 + s_2 h_i)\right) L_{H_{\mathbf{W}}}^{\nu(0,m),(i,j)}(s_1, s_2)}{s_2(s_1 + s_2 h_i)\left(1 - L_{H_{\mathbf{W}}}^{\nu(i,j),(i,j)}(s_1, s_2)\right)},$$

where the joint Laplace transforms $L_{H_{\mathbf{W}}}^{\nu(i_1,j_1),(i_2,j_2)}(s_1, s_2)$ *have been introduced above.*

16.4.3 Structured motifs

Structured motifs are special compound patterns, usually containing a huge number of single patterns. In this subsection we consider both the waiting time until the first occurrence, and the intersite distance between consecutive occurrences, of a structured motif. The interest in these waiting times is due to the biological challenge of identifying promoter motifs along genomes. A structured motif is composed of several patterns separated by a variable distance. If the number of patterns is n, then the structured motif is said to have n boxes. The formal definition of a structured motif with 2 boxes follows. Let $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ be two patterns of length k_1 and k_2, respectively. The alphabet size equals N, and the strings are generated by the Markov chain introduced in Section 16.2. 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 of the string, and the number of letters between the two patterns is not smaller than d_1 and not greater than d_2. Also, it is assumed that patterns $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ appear only once in the string. The pgf's of both the waiting time, $\tau_{\mathbf{m}}^{(s)}$, to reach for the first time the structured motif

$\mathbf{m}$ from state s, and the intersite distance, $\tau_{\mathbf{m}}^{(intersite)}$, between two consecutive occurrences of $\mathbf{m}$, are of interest.

Let $\mathbf{W} = \{\mathbf{w}^{(1)}, \mathbf{w}^{(2)}\}$ be a compound pattern consisting of two patterns. For brevity, denote by $T_{i,j}$, $i, j \in \{1, 2\}$, the waiting time to reach pattern $\mathbf{w}^{(j)}$ from pattern $\mathbf{w}^{(i)}$, and by $T_j^{(s)}$ the waiting time to reach pattern $\mathbf{w}^{(j)}$ from state s. The quantities $r_{i,j}$ and $X_{i,j}$, $i, j \in \{1, 2\}$, are introduced in Subsection 16.4.1. Let

$$a_{i,j}(x) = P(X_{i,j} = x).$$

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 $d_1 + k_2 \le X_{1,2} \le d_2 + k_2$. Introduce the following random variables:

$$\begin{aligned} F_{12} &= (X_{1,2} \mid X_{1,2} < d_1 + k_2 \ \text{ or } \ X_{1,2} > d_2 + k_2), \\ S_{12} &= (X_{1,2} \mid d_1 + k_2 \le X_{1,2} \le d_2 + k_2). \end{aligned}$$

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. One may notice that the pgf's of F_{12} and S_{12} are given by

$$\begin{aligned} G_{F_{12}}(t) &= \left(G_{X_{12}}(t) - \sum_{x=d_1+k_2}^{d_2+k_2} a_{1,2}(x) t^x \right) (1 - q_S)^{-1} \\ G_{S_{12}}(t) &= \left(\sum_{x=d_1+k_2}^{d_2+k_2} a_{1,2}(x) t^x \right) q_S^{-1}, \end{aligned}$$

where q_S is the probability of 'success' ($\mathbf{w}^{(2)}$ achieves the structured motif), i.e., the probability that $d_1 + k_2 \le X_{1,2} \le d_2 + k_2$. Namely, we have

$$q_S = \sum_{x=d_1+k_2}^{d_2+k_2} a_{1,2}(x).$$

The following theorem provides explicit and calculable expressions for the pgf's of both the waiting time to reach for the first time the structured motif $\mathbf{m} = \mathbf{w}^{(1)}(d_1 : d_2)\mathbf{w}^{(2)}$ from state s, and the intersite distance between two consecutive occurrences of $\mathbf{m}$.

Theorem 16.4.3 *The pgf, $G_{\mathbf{m}}^{(s)}(t)$, of the waiting time to reach for the first time a structured motif* $\mathbf{m}$ *starting from state s, and the pgf, $G_{\mathbf{m}}^{(intersite)}(t)$, of the intersite distance between two consecutive occurrences of* $\mathbf{m}$, *admit the following explicit expressions:*

$$G_{\mathbf{m}}^{(s)}(t) = \frac{r_{1,2}\, q_S\, G_{T_1^{(s)}}(t)\, G_{S_{12}}(t)}{(1 - (1 - r_{1,2}) G_{X_{1,1}}(t)) \left(1 - (1 - q_S) \left(\frac{r_{1,2}\, G_{T_{2,1}}(t)\, G_{F_{12}}(t)}{1 - (1 - r_{1,2}) G_{X_{1,1}}(t)} \right) \right)},$$

$$G_{\mathbf{m}}^{(intersite)}(t) = \frac{r_{1,2}\, q_S\, G_{T_{2,1}}(t)\, G_{S_{12}}(t)}{\left(1-(1-r_{1,2})G_{X_{1,1}}(t)\right)\left(1-(1-q_S)\left(\frac{r_{1,2}\, G_{T_{2,1}}(t)\, G_{F_{12}}(t)}{1-(1-r_{1,2})G_{X_{1,1}}(t)}\right)\right)},$$

where $G_{F_{12}}(t), G_{S_{12}}(t)$, and q_S are given above.

The proof of this theorem is found in Stefanov, Robin, and Schbath (2007). Note that, in view of this theorem, the availability of the pgf's $G_{X_{i,j}}(t),\ i,j = 1,2$, is enough to calculate explicit, closed-form expressions for $G_{\mathbf{m}}^{(s)}(t)$ and $G_{\mathbf{m}}^{(intersite)}(t)$. Explicit expressions for the $G_{X_{i,j}}(t)$, in terms of the $G_{T_{\mathbf{w}^{(i)},\mathbf{w}^{(j)}}}(t)$, are derived from the identities at the end of Subsection 16.4.1. Also, recall that the $G_{T_{\mathbf{w}^{(i)},\mathbf{w}^{(j)}}}(t)$ are calculated from Theorem 16.2.1 in Section 16.2.

Neat closed-form expressions for the relevant pgf's associated with structured motifs with n boxes are found in Stefanov, Robin, and Schbath (2009).

References

1. Antzoulakos, D. L. (2001). Waiting times for patterns in a sequence of multistate trials. *Journal of Applied Probability*, **38,** 508–518.

2. Balakrishnan, N. and Koutras, M. (2002). *Runs and Scans with Applications.* Wiley, New York.

3. Biggins, J. D. (1987). A note on repeated sequences in Markov chains. *Advances in Applied Probability*, **19,** 739–742.

4. Biggins, J. D. and Cannings, C. (1987). Markov renewal processes, counters and repeated sequences in Markov chains. *Advances in Applied Probability*, **19,** 521–545.

5. Blom, G. and Thorburn, D. (1982). How many random digits are required until given sequences are obtained? *Journal of Applied Probability*, **19,** 518–531.

6. Chadjiconstantinidis, S., Antzoulakos, D. L. and Koutras, M. V. (2000). Joint distributions of successes, failures and patterns in enumeration problems. *Advances in Applied Probability*, **32,** 866–884.

7. Chryssaphinou, O. and Papastavridis, S. (1990). The occurrence of a sequence of patterns in repeated dependent experiments. *Theory of Probability and Its Applications*, **35,** 167–173.

8. Çinlar, E. (1975). *Introduction to Stochastic Processes.* Prentice-Hall, Englewood Cliffs, NJ.

9. Feller, W. (1950). *An Introduction to Probability Theory and Its Applications*, Vol. 1. Wiley, New York.

10. Fu, J. C. (1996). Distribution theory of runs and patterns associated with a sequence of multistate trials. *Statistica Sinica*, **6,** 957–974.

11. Fu, J. C. and Chang, Y. M. (2002). On probability generating functions for waiting time distributions of compound patterns in a sequence of multistate trials. *Journal of Applied Probability*, **39,** 70–80.

12. Fu, J. C. and Lou, W. Y. W. (2003). *Distribution Theory of Runs and Patterns and its Applications*, World Scientific, Hackensack, NJ.

13. Gerber, H. and Li, S-Y. R. (1981). The occurrence of sequence patterns in repeated experiments and hitting times in a Markov chain. *Stochastic Processes and Their Applications*, **11,** 101–108.

14. Glaz, J., Kulldorff, M., Pozdnyakov, V. and Steele, J. M. (2006). Gambling teams and waiting times for patterns in two-state Markov chains. *Journal of Applied Probability*, **43,** 127–140.

15. Guibas, L. J. and Odlyzko, A. M. (1981). String overlaps, pattern matching, and nontransitive games. *Journal of Combinatorial Theory. Series A*, **30,** 183–208.

16. Han, Q. and Hirano, K. (2003). Sooner and later waiting time problems for patterns in Markov dependent trials. *Journal of Applied Probability*, **40,** 73–86.

17. Inoue, K. and Aki, S. (2007). On generating functions of waiting times and numbers of occurrences of compound patterns in a sequence of multistate trials. *Journal of Applied Probability* **44,** 71–81.

18. Kijima, M. (1997). *Markov Processes for Stochastic Modeling.* Chapman & Hall, London.

19. Li, S-Y. R. (1980). A martingale approach to the study of occurrence of sequence patterns in repeated experiments. *Annals of Probability*, **8,** 1171–1176.

20. Nicodème, P., Salvy, B. and Flajolet, P. (2002). Motif statistics. *Theoretical Computer Science*, **287,** 593–617.

21. Nuel, G. (2008). Pattern Markov chains: optimal Markov chain embedding through deterministic finite automata. *Journal of Applied Probability*, **45,** 226–243.

22. Pozdnyakov, V. (2008). A note on occurrence of gapped patterns in i.i.d. sequences. *Discrete Applied Mathematics* **156,** 93–102.

23. Pozdnyakov, V., Glaz, J., Kulldorff, M. and Steele, J. M. (2005). A martingale approach to scan statistics. *Annals of the Institute of Statistical Mathematics*, **57,** 21–37.

24. Reinert, G., Schbath, S. and Waterman, M. (2000). Probabilistic and statistical properties of words: an overview. *Journal of Computational Biology*, **7,** 1–46.

25. Robin, S. and Daudin, J. (1999). Exact distribution of word occurrences in a random sequence of letters. *Journal of Applied Probability*, **36,** 179–193.

26. Robin, S. and Daudin, J. (2001). Exact distribution of the distances between any occurrences of a set of words. *Annals of the Institute of Statistical Mathematics*, **36**, 895–905.

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

28. Rukhin, A. (2002). Distribution of the number of words with a prescribed frequency and tests of randomness. *Advances in Applied Probability*, **34,** 775–797.

29. Rukhin, A. (2006). Correlation matrices of chains for Markov sequences, and testing for randomness. (Russian) *Teoriya Veroyatnostei i ee Primeneniya*, **51,** 712–731.

30. Stefanov, V. T. (2000). On some waiting time problems. *Journal of Applied Probability*, **37,** 756–764.

31. Stefanov, V. T. (2003). The intersite distances between pattern occurrences in strings generated by general discrete- and continuous-time models: an algorithmic approach. *Journal of Applied Probability*, **40,** 881–892.

32. Stefanov, V. T. (2006). Exact distributions for reward functions on semi-Markov and Markov additive processes. *Journal of Applied Probability*, **43,** 1053–1065.

33. Stefanov, V. T. and Pakes, A. G (1997). Explicit distributional results in pattern formation. *Annals of Applied Probability*, **7,** 666–678.

34. Stefanov, V. T., Robin, S. and Schbath, S. (2007). Waiting times for clumps of patterns and for structured motifs in random sequences. *Discrete Applied Mathematics*, **155**, 868–880.

35. Stefanov, V. T., Robin, S. and Schbath, S. (2009). Occurrence of structured motifs in random sequences: arbitrary number of boxes. (in preparation).

36. Szpankowski, W. (2001). *Average Case Analysis of Algorithms on Sequences.* John Wiley & Sons, New York.