Discrete Applied Mathematics 159 (2011) 826–831

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

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

Valeri T. Stefanov[a], Stéphane Robin[b], Sophie Schbath[c,*]

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

ARTICLE INFO

Article history:
Received 17 August 2009
Received in revised form 16 December 2010
Accepted 29 December 2010
Available online 22 January 2011

Keywords:
Pattern
Structured motif
Semi-Markov process
Probability generating function
DNA sequence

ABSTRACT

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.

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, $\mathbf{w}_i$ 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 $\pi_{i,j}, i, j = 1, 2, \ldots, N$.

Let $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_b$ be b patterns of length $k_1, k_2, \ldots, k_b$, respectively, on the finite alphabet $\{1, 2, \ldots, N\}$. A structured motif $\mathbf{m}_b$ formed by these b patterns is any string of letters which (i) begins with pattern $\mathbf{w}_1$ and ends with pattern $\mathbf{w}_b$; (ii) all remaining patterns $\mathbf{w}_2, \ldots, \mathbf{w}_{b-1}$ 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 $\mathbf{w}_i$ more than once); (iii) the initial pattern $\mathbf{w}_1$ and the end pattern $\mathbf{w}_b$ together with a fixed appearance of the patterns $\mathbf{w}_2, \ldots, \mathbf{w}_{b-1}$ 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 $\mathbf{w}_i$ and $\mathbf{w}_{i+1}$ is not smaller than d_i and not greater than D_i. We denote a structured motif $\mathbf{m}_b$ by

$$\mathbf{w}_1(d_1 : D_1)\mathbf{w}_2(d_2 : D_2)\mathbf{w}_3 \ldots \mathbf{w}_{b-1}(d_{b-1} : D_{b-1})\mathbf{w}_b$$

and, for $i = 1, 2, \ldots, b-1$, we denote by $\mathbf{m}_i$ the sub-structured motif $\mathbf{w}_1(d_1 : D_1)\mathbf{w}_2(d_2 : D_2)\mathbf{w}_3 \ldots \mathbf{w}_{i-1}(d_{i-1} : D_{i-1})\mathbf{w}_i$; of course $\mathbf{m}_i$ is a prefix of $\mathbf{m}_b$.

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

Restriction 1. Pattern $\mathbf{w}_1$ appears only once in the structured motif $\mathbf{m}_b$;

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

These restrictions are not strong in practice because the probability for $\mathbf{w}_1$ to occur more than once within the structured motif or for some $\mathbf{w}_{i+1}$ to occur more than once in $\mathbf{w}_i(d_i : D_i)\mathbf{w}_{i+1}$ 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
$\mathbf{W}_{i,j}$ — the family $\{\mathbf{w}_i, \mathbf{w}_j\}$ consisting of the two patterns $\mathbf{w}_i$ and $\mathbf{w}_j$;
$T_{i|j}$ — the waiting time to reach pattern $\mathbf{w}_i$ from pattern $\mathbf{w}_j$;
$T_i^{(s)}$ — the waiting time to reach pattern $\mathbf{w}_i$ from state s;
$T_{\mathbf{W}|n}$ — the waiting time to reach the family of patterns $\mathbf{W}$ from pattern $\mathbf{w}_n$;
$X_{i,j|n}$ — the waiting time to reach the family of patterns $\mathbf{W}_{i,j}$ from pattern $\mathbf{w}_n$ given the reached pattern is $\mathbf{w}_i$;
$r_{i,j|n}$ — the probability to reach pattern $\mathbf{w}_i$ before pattern $\mathbf{w}_j$, given one starts from pattern $\mathbf{w}_n$.

Of course $r_{i,j|n} = P(X_{i,j|n} = T_{\mathbf{W}_{i,j}|n})$. 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 $T_{i|j}, T_i^{(s)}, T_{\mathbf{W}|n}, X_{i,j|n}$, and also allow exact computation of the probabilities $r_{i,j|n}$ (cf. [6,7,10,11,1]).

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:

$$F_{i+1,1|i} = (X_{i+1,1|i} \mid X_{i+1,1|i} < d_i + k_{i+1} \text{ or } X_{i+1,1|i} > D_i + k_{i+1}),$$
$$S_{i+1,1|i} = (X_{i+1,1|i} \mid d_i + k_{i+1} \leq X_{i+1,1|i} \leq D_i + k_{i+1}).$$

Actually, $F_{i+1,1|i}$ is a random variable whose distribution equals the conditional distribution of the waiting time to reach $\mathbf{w}_{i+1}$ from the sub-structured motif $\mathbf{m}_i$, given the sub-structured motif $\mathbf{m}_{i+1}$ 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 $\mathbf{m}_{i+1}$ 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:

$$G_{F_{i+1,1|i}}(t) = \left(G_{X_{i+1,1|i}}(t) - \sum_{x=d_i+k_{i+1}}^{D_i+k_{i+1}} a_{i+1,1|i}(x)t^x\right)(1 - q_{S,i})^{-1} \tag{1}$$

$$G_{S_{i+1,1|i}}(t) = \left(\sum_{x=d_i+k_{i+1}}^{D_i+k_{i+1}} a_{i+1,1|i}(x)t^x\right) q_{S,i}^{-1}, \tag{2}$$