Available online at www.sciencedirect.com

ScienceDirect

Discrete Applied Mathematics 155 (2007) 868–880

DISCRETE APPLIED MATHEMATICS

www.elsevier.com/locate/dam

Waiting times for clumps of patterns and for structured motifs in random sequences

V.T. Stefanov[a], S. Robin[b], S. Schbath[c,*]

[a] *School of Mathematics and Statistics, The University of Western Australia, Crawley (Perth) 6009, W.A., Australia*
[b] *ENGREF / INA PG / INRA unit of Applied Mathematics and Computer Sciences, 16, rue Claude Bernard, 75005, Paris, France*
[c] *Unité Mathématique, Informatique & Génome, Institut National de la Recherche Agronomique, 78352 Jouy-en-Josas, France*

Received 29 June 2004; received in revised form 1 July 2005; accepted 5 July 2005
Available online 16 October 2006

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.

MSC: 60E10; 60J20; 60K20

Keywords: Pattern; Clump; Probability generating function; Random sum; Structured motif; DNA sequence

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 [12] or [6], Chapter 6, for recent surveys). 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 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 [13,14,17].

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.

* Corresponding author.
E-mail addresses: stefanov@maths.uwa.edu.au (V.T. Stefanov), robin@inapg.inra.fr (S. Robin), sophie.schbath@jouy.inra.fr (S. Schbath).

0166-218X/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2005.07.016

In Section 2 below we consider clumps. For renewals see [2,4,10] 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 [11,16]. 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. [3,7,8]). 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. [15] 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. [15]) 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 $\{X(n)\}_{n\geqslant 0}$ 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 $1 \leqslant w_i \leqslant N$, $i = 1, 2, \ldots, k$. For $j \in \{1, 2, \ldots, k\}$, denote the probability generating function[1] (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 $G_j^{(w_1,w_2,\ldots,w_r)}(t)$, $1 \leqslant r \leqslant j$, 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

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

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

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

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. Fig. 1(B)). The c_i's can be obtained recursively from Robin and Daudin [13] 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 [17]. 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 [5].

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