Exact distribution of the distances between any occurrences of a set of words

S. Robin, J.-J. Daudin
INA-PG – INRA
16 rue Claude Bernard
75005 Paris, France

To appear in

Annals of the Institute of Stastiscal Mathematics

Running head: Exact distribution of distances between words

Abstract

The distribution of the distance between two (or more) successive occurrences of a specific word in a random sequence of letters is known under different models. In this paper, a more general problem is studied: the distribution of the distance between two (or more) successive occurrences of any word of a given set under a Markov model for the sequence. The generating function and a recurrence for obtaining the probabilities are given. These results are applied to study the distribution of the "CHI" motif in the genome sequence of *Haemophilus influenzae.*

keywords: distance between occurrences, genome sequence analysis, semi Markov process.

1 Introduction

Let $S = \{S_1 S_2, ...\}$ be a sequence of repeated trials with possible outcomes taken from an alphabet $\mathcal{A}$. Let w be a specific string of letters of $\mathcal{A}$, called a "word" or a "pattern". The exact distribution of the waiting time until w appears has been studied by many authors, often for $\mathcal{A}$ =$\{0,1\}$. Among recent references, Robin and Daudin (1999) have given the generating probability function and a recurrence relation for computing the probabilities under a first order Markovian model for S. A more complete source of references may be obtained in Koutras (1997) who gives the relation between the generating function of the waiting time and the generating function of the number of occurrences of w.

What happens when we consider more than one word?

Let $\mathcal{W} = \{\mathbf{w}_1, ...\mathbf{w}_m\}$ be a set of words of respective lengths $\{k_i\}_{i=1..m}$ not included in each other. In this paper, we study the waiting time until the first word of $\mathcal{W}$ appears. This problems occurs in some applications:

- In genome sequence analysis where the same biological function is fulfilled by different words. For example, a particular pattern of nucleotide called CHI (for *Crossover Hotspot Initiator*) possess a function of protection of the genome and is characteristic of each species. It is unknown for recently sequenced species but is expected to be frequent and regularly spaced. For some organism, the same function is fulfilled by few different words: for example, the CHI of *Haemophilus influenza* is (`gNtggtgg`) where `N` can be any of the four letters of the alphabet $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$. The knowledge of the exact distribution of the distance between the occurrences of several words would be of great help to study the longitudinal distribution of such a motif.

- A particular case of this problem is *the sooner and later waiting time problem* for success and failure runs where $\mathcal{A}$ =$\{0,1\}$ and $\mathcal{W} = \{1...1,\ 0...0\}$ with k_1 '1' and k_0 '0'. This problem is found in reliability and in psychology.

The general problem has been studied by Breen *et al.* (1985) using renewal theory and only in the non-overlapping and independent case and by Chryssaphinou and Papastavridis (90) in the overlapping and first order Markovian case. Mori (1991) gives a limit theorem for the waiting time till each of a given set of patterns of same length in a sequence of iid random variables distributed uniformly on any alphabet.

The *sooner and later waiting time problem* has been recently studied by Aki and Hirano (1993) and Uchida and Aki (1995) under a first order Markovian model, by Aki and Balakrishnan (1996) under a markovian model of order two. Koutras and Alexandrou (1997) have generalized the problem with a three letters alphabet under a general non homogeneous model.

The finite Markov chain imbedding technique provides a quite general method to calculate the distribution of the first occurrence of different motifs (see Fu and Koutras (94) or Koutras and Alexandrou (1997)) or of the counting of runs (Fu (1996)). However, for complex motifs, especially for overlapping words that are not simple runs, the transition matrix is not straightforward to obtain. Furthermore this matrix has dimensions proportional either to the total length of the motifs or to the length of the sequence. This point seems to make the method untractable for large scaled problems such as DNA analysis.

In this paper we give the exact distribution and the generating function of the waiting time in the case of any set of words with any alphabet and a first order Markov model.

We first present the model and the notations. In the second part, we derive the probability generating function. The third section is devoted to the law of the $r-$scans and the fourth presents an application to the CHI of *Haemophilus influenza.*

Set of words. Let us study the occurrences of a set of words $\mathcal{W} = \{\mathbf{w}_i\}_{i=1..m}$ not included in each other and of respective lengths $\{k_i\}$. $w_{i,u}$ denotes the $u-$th letter of $\mathbf{w}_i : \mathbf{w}_i = (w_{i,1}, ..., w_{i,k_i})$. As usual, the position of a word in the sequence is the position of its last letter.

Model for the sequence. Let us consider a sequence $\{S_x\}_{x\geq 1}$ of letters taken from an alphabet $\mathcal{A}$. The sequence is assumed to be a homogeneous first order Markov chain (MC1) with transition probability Π with general term $\pi(a,b)$

$$\forall (a,b) \in \mathcal{A}\times\mathcal{A}, \forall x \geq 1, \quad \Pr\{S_{x+1} = b \,|\, S_x = a\} = \pi(a,b).$$

In the following, $\pi^{(n)}(a,b)$ shall denote the general term of Π^n and $\mu = [\mu(a)]_{a\in\mathcal{A}}$ shall denote the stationary distribution of Π which satisfies $\mu.\Pi = \mu$.

Notations. For any words $\mathbf{w}_i$ and $\mathbf{w}_j$ of $\mathcal{W}$, let us use the following notations.

- $\varepsilon_{ij}(u)$ is the overlapping indicator of $\mathbf{w}_j$ over $\mathbf{w}_i$ with u letters which equals one when the first u letters of $\mathbf{w}_j$ are the same as the last u letters of $\mathbf{w}_i$:

$$\varepsilon_{ij}(u) = \mathbb{I}\{(w_{i,k_i-u+1}...w_{i,k_i}) = (w_{j,1}...w_{j,u})\}$$

  where $\mathbb{I}\{A\}$ equals one if A is true and zero otherwise. The condition of non-inclusion of the words implies that $\varepsilon_{ij}(u) = 0$ for $u = \min(k_i, k_j)$ if $i \neq j$ and for $u > \min(k_i, k_j)$ in any case.

- $\tau_i(u,v)$ is the probability to observe $w_{i,u}$ to $w_{i,v}$ given $w_{i,u-1}$:

$$\tau_i(u,v) = \prod_{x=u}^{v} \pi(w_{i,x-1}, w_{i,x})$$

  with convention $\tau_i(u,1) = 1$.

- $d_{ij}(t)$ is the overlapping polynomial which does not take into account complete overlap:

$$d_{ij}(t) = \sum_{u=1}^{\min(k_i,k_j)-1} \frac{\varepsilon_{ij}(u)t^u}{\tau_j(2,u)}.$$

2 Semi-Markov model for the occurrences of the words

2.1 Model and notations

The way the words occur along the sequence is completely described by two processes

- the positions of the occurrences of the different words of $\mathcal{W}$: $\{X_n\}_{n\geq 1}$;
- the words occurring at each of these positions $\{I_n\}_{n\geq 1}$:

$$\{I_n = i\} \iff \{\text{the word occuring at } X_n \text{ is } \mathbf{w}_i\}.$$

This describes a semi-Markov process with states in $\mathcal{W}$. This process is assumed to be homogenous along the sequence.

Let $p_{ij}(y)$ denote the probability that the first word of $\mathcal{W}$ occurring after $\mathbf{w}_i$ is $\mathbf{w}_j$ and that it appears y positions after:

$$p_{ij}(y) = \Pr\{(X_{n+1} - X_n = y) \cap (I_{n+1} = j) \mid I_n = i\}.$$

$p_{ij}(y)$ does not depend of n thanks to the Markovian structure of the sequence. Let ϕ_{ij} be the generating function of the $p_{ij}(y)$'s:

$$\phi_{ij}(t) = \sum_{y\geq 1} p_{ij}(y)t^y.$$

Let $q_{ij}(y)$ denote the probability that the word $\mathbf{w}_j$ occurs y positions after $\mathbf{w}_i$:

$$q_{ij}(y) = \sum_{r\geq 1} \Pr\{(X_{n+r} - X_n = y) \cap (I_{n+r} = j) \mid I_n = i\}.$$

The difference between $p_{ij}(y)$ and $q_{ij}(y)$ is that $q_{ij}(y)$ concerns any occurrence of $\mathbf{w}_j$ after any number of renewals of any word of $\mathcal{W}$, while $p_{ij}(y)$ concerns only the first occurrence of $\mathbf{w}_j$ among the set $\mathcal{W}$. Let $f_{ij}(t)$ be the generating function of the $q_{ij}(y)$'s:

$$f_{ij}(t) = \sum_{y\geq 1} q_{ij}(y)t^y.$$

If the first word is distributed according to some distribution ν over $\mathcal{W}$, we shall denote $p_{\nu j}(y) = \sum_i \Pr\{(X_{n+1} - X_n = y) \cap (I_{n+1} = j) \mid I_n = i\}\nu(i)$ where $\nu(i) = \Pr\{I_n = i\}$. We shall denote $\phi_{\nu j}(t)$ the corresponding generating function.

At last, we shall denote $p_{i\bullet}(y)$ the probability that the first occurrence after $\mathbf{w}_i$ of any word of $\mathcal{W}$ occurs y position later : $p_{i\bullet}(y) = \sum_j p_{ij}(y)$, and $\phi_{i\bullet}(t)$ the corresponding probability generating function.

2.2 Matrices of generating functions

Lemma 1 *The probabilities $p_{ij}(y)$'s and $q_{ij}(y)$'s satisfy the following relation*

$$q_{ij}(y) = p_{ij}(y) + \sum_{l=1}^{m} \sum_{z=1}^{y-1} q_{il}(z) p_{lj}(y-z). \tag{2.1}$$

Proof. The first term of the right hand side gives the probability that the first word of $\mathcal{W}$ occurring after $\mathbf{w}_i$ is $\mathbf{w}_j$ and that it appears y positions after. The second term calculates the probability of $\mathbf{w}_j$ y position after $\mathbf{w}_i$ is not the first one by conditioning on all possible occurrences of any word of $\mathcal{W}$ between positions 1 and $y-1$. ■

Lemma 2 *The probability $q_{ij}(y)$ is*

$$\begin{aligned} q_{ij}(y) = \; & \varepsilon_{ij}(k_j - y)\tau_j(k_j - y + 1, k_j)\mathbb{I}(y < k_j) \\ & + \pi^{(y-k_j+1)}(w_{i,k_i}, w_{j,1})\tau_j(2, k_j)\mathbb{I}(y \geq k_j). \end{aligned} \tag{2.2}$$

Proof. The first term of the right hand side considers the case where $\mathbf{w}_j$ overlaps $\mathbf{w}_i$ $(y < k_j)$ and the second term considers the case where it does not $(y \geq k_j)$. ■

Lemma 3 *The generating function $f_{ij}(t) = \sum_{y \geq 1} q_{ij}(y) t^y$ is*

$$f_{ij}(t) = t^{k_j} \tau_j(2, k_j) \left[d_{ij}\left(\frac{1}{t}\right) + \frac{g_{i,j}(t)}{t} \right]$$

where $g_{ij}(t) = \sum\limits_{u \geq 1} \pi^{(u)}(w_{i,k_i}, w_{j,1}) t^u$.

Proof. Multiplying the left side of equation (2.2) by t^y and summing over all positive y's, we get for $f_{ij}(t)$:

$\sum\limits_{y=1}^{k_j-1} \varepsilon_{ij}(k_j - y)\tau_j(k_j - y + 1, k_j) t^y + \sum\limits_{y \geq k_j} \pi^{(y-k_j+1)}(w_{i,k_i}, w_{j,1})\tau_j(2, k_j) t^y$

$= t^{k_j} \sum\limits_{z=1}^{k_j-1} \varepsilon_{ij}(z)\tau_j(z+1, k_j) t^{-z} + \tau_j(2, k_j) t^{k_j - 1} \sum\limits_{u \geq 1} \pi^{(u)}(w_{i,k_i}, w_{j,1}) t^u$

$= \; t^{k_j}\tau_j(2, k_j) \left[\sum\limits_{z=1}^{k_j-1} \dfrac{\varepsilon_{ij}(z)}{\tau_j(2, z)} t^{-z} + \dfrac{1}{t} \sum\limits_{u \geq 1} \pi^{(u)}(w_{i,k_i}, w_{j,1}) t^u \right]$ and the lemma is proved since $\varepsilon_{ij}(u) = 0$ for $u > \min(k_i, k_j)$. ■

If we consider the $(m \times m)$ matrices $\mathbf{F}(\mathbf{t})$, $\mathbf{D}(t)$ and $\mathbf{G}(t)$ with respective general terms $f_{ij}(t)$, $d_{ij}(t)$ and $g_{ij}(t)$, and the $m-$diagonal matrix $\mathbf{T}(t)$ with general term $\tau_j(2, k_j) t^{k_j}$ $(j = 1..m)$, we obtain the following matrix decomposition :

$$\mathbf{F}(t) = \left[\mathbf{D}\left(\frac{1}{t}\right) + \frac{1}{t}\mathbf{G}(t) \right] \mathbf{T}(t).$$

Considering that $\mathbf{H}(t) = \left[(\mathbf{I} - \Pi t)^{-1} - \mathbf{I}\right]$ is the generating matrix of the transition probabilities between any couple of letters of $\mathcal{A}$, $\mathbf{G}(t)$ is the $m \times m$ generating matrix of the transition probabilities between the last letter of each word of $\mathcal{W}$ and each first letter: $g_{ij}(t) = [\mathbf{H}(t)]_{w_{i,k_i}, w_{j,1}}$.

Theorem 4 *The generating matrices $\Phi(t) = [\phi_{ij}(t)]_{i,j=1..m}$ and $\mathbf{F}(t)$ satisfy*

$$\mathbf{F}(t) = \Phi(t)\,[\mathbf{I} + \mathbf{F}(t)]$$

where $\mathbf{I}$ is the $m \times m$ identity matrix.

Proof. Let us multiply each side of equation (2.1) by t^y and sum over all positive y's: the left term gives $f_{ij}(t)$, the right term gives $\phi_{ij}(t) + \sum_{l=1}^{m} \phi_{il}(t) f_{lj}(t)$ and the theorem is proved. ■

A consequence of this theorem is that, if t is such that $[\mathbf{I} + \mathbf{F}(t)]$ is invertible, one has

$$\Phi(t) = \mathbf{F}(t).\,[\mathbf{I} + \mathbf{F}(t)]^{-1}. \tag{2.3}$$

It is hence of a great importance to know if $[\mathbf{I} + \mathbf{F}(t)]$ is invertible. Since $\mathbf{F}(t)$ is a matrix of rational functions, the determinant of $[\mathbf{I} + \mathbf{F}(t)]$ is also a rational function. Therefore it has a finite number of poles and roots and $[\mathbf{I} + \mathbf{F}(t)]$ is invertible for every t except a finite number. Note that $\mathbf{F}(0) = \mathbf{0}$ and $\mathbf{F}(1)$ is not defined.

Since $\Phi(t)$ is a matrix of rational functions, we can consider that it is characterized by all its coefficients which may be calculated by the aid of (2.3) on any open interval containing no pole and no root. From now on, we shall consider that $\Phi(t)$ is known.

The following corollary is a direct application.

Corollary 5 *The probability generating function $\phi_{i\bullet}(t)$ of the distance $Y_{i\bullet}$ between an occurrence of $\mathbf{w}_i$ and the next occurrence of any word of $\mathcal{W}$ is*

$$\phi_{i\bullet}(t) = \phi_i(t).\mathbf{1}$$

where $\phi_i(t)$ denotes the $i-th$ row of $\Phi(t)$.

Proof. $\phi_{i\bullet}(t)$ is defined by $\phi_{i\bullet}(t) = \sum_{y \geq 1} p_{i\bullet}(y) t^y$ where $p_{i\bullet}(y)$ is the probability that the next word of $\mathcal{W}$ occurs y letters after $\mathbf{w}_i$: $p_{i\bullet}(y) = \sum_{j=1}^{m} p_{ij}(y) = \Pr\{X_{n+1} - X_n = y \mid I_n = i\}$. It is clear that $\phi_{i\bullet}(t) = \sum_{j=1}^{m} \sum_{y \geq 1} p_{ij}(y) t^y = \sum_{j=1}^{m} \phi_{ij}(t) = \phi_i(t).\mathbf{1}$. ■

Corollary 6 *The generating function of the $p_{\nu j}(y)$ is*

$$\phi_{\nu j}(t) = \nu.\Phi_j(t)$$

where $\Phi_j(t)$ is the $j-th$ column of Φ.

The generating function of the $p_{\nu\bullet}(y)$ is

$$\phi_{\nu\bullet}(t) = \nu.\Phi(t).\mathbf{1}.$$

Proof. The first part of the corollary is obvious since $p_{\nu j}(y) = \sum_i \nu(i) p_{ij}(y)$. The second comes from $p_{\nu\bullet}(y) = \sum_j p_{\nu j}(y)$. ■

Sequence with a given beginning. Corollary **6** can be applied to the particular case where the sequence begins with one of the words of $\mathcal{W}$, say $\mathbf{w}_1$. In this case we shall consider the distribution $\nu = [1, 0, ...0]$.

If the sequence does not begin with a word of $\mathcal{W}$ but with some other word $\mathbf{w}_0$, we can apply the following result.

Corollary 7 *The generating vector $\phi_0(t) = [\phi_{01}(t), ...\phi_{0m}(t)]$ satisfies*

$$\phi_0(t) = \mathbf{f}_0(t) \left[\mathbf{I} - \Phi(t) \right].$$

Proof. For the generating functions, equation (2.1) is equivalent to $f_{ij}(t) = \phi_{ij}(t) + \sum_{l=1}^{m} f_{il}(t) \phi_{lj}(t)$. Here we consider $i = 0, ...m$ and $j = 1, ...m$ since $\mathbf{w}_0$ is only taken into account as a beginning word. Therefore, Theorem 4 can be extended as follows

$$\begin{bmatrix} \mathbf{f}_0(t) \\ \mathbf{F}(t) \end{bmatrix} = \begin{bmatrix} \phi_0(t) \\ \Phi(t) \end{bmatrix} + \begin{bmatrix} \mathbf{f}_0(t) \\ \mathbf{F}(t) \end{bmatrix} \Phi(t)$$

and the corollary is proved. ■

2.3 Probabilities and moments

Several elementary results can be derived from the preceding theorems and lemmas.

Transition probabilities between words. Since $\sum_{y \geq 1} p_{ij}(y) = \Pr\{I_{n+1} = j \mid I_n = i\}$, $\phi_{ij}(1)$ is the probability that $\mathbf{w}_j$ is the first word of $\mathcal{W}$ to occur after $\mathbf{w}_i$. It is therefore the transition probability from $\mathbf{w}_i$ to $\mathbf{w}_j$:

$$s(i, j) = \Pr\{I_{n+1} = j \mid I_n = i\}$$

and the transition matrix $\mathbf{S} = [s(i, j)]_{i,j=1..m}$ is

$$\mathbf{S} = \Phi(1).$$

The stationary distribution of $\mathbf{S}$ can produce a particular distribution ν.

Moments of the distances. It is well known that the expectation and variance of a random variable with generating function ψ are respectively $\psi'(1)$ and $\psi''(1) + \psi'(1)[1 - \psi'(1)]$.

The generating function of the distance between an occurrence of $\mathbf{w}_i$ and the next occurrence of $\mathbf{w}_j$ given that $\mathbf{w}_j$ is the first word of $\mathcal{W}$ to occur after $\mathbf{w}_i$ is $\phi_{ij}(t)/s(i, j)$. These conditional expectation and variance are easy to derive from $\phi'_{ij}(t)$ and $\phi''_{ij}(t)$.

The moments of $Y_{\nu\bullet}$ can also be derived using Corollary **6**.

Computation of the probabilities. The probabilities $p(y)$'s can be computed using the Taylor expansion of $\phi(t)$ which needs the computation of the y-th derivative of $\phi(t)$ at $t=0$. A more efficient way is to obtain a finite recurrence relation between the $p(y)$'s using the fact that $\phi(t)$ is a rational function. We have used this method to compute $p(y)$ for $y=1$ to more than 30 000 in a few minutes on a PC.

3 $r-$scans

r -scans have been used by Karlin and Macken (1991) in a genome analysis context to study the homogeneity of the distribution of a motif along a sequence. It is defined as the distance between an occurrence and the $r-$th next:

$$Y^r = X_{n+r} - X_n.$$

It is clear that its distribution depends on n only through I_n.

3.1 Exact distribution

Let us now consider the r-scan Y_i^r between an occurrence of $\mathbf{w}_i$ and the $r-$th next occurrence of any element of $\mathcal{W}$. Let us denote

$$p_{ij}^r(y) = \Pr\{(X_{n+r} - X_n = y) \cap I_{n+r} = j \mid I_n = i\}$$

and the corresponding generating function $\phi_{ij}^r(t) = \sum_{y \geq 1} p_{ij}^r(y) t^y$.

Theorem 8 *The generating function matrix $\Phi^r(t)$ of general term $\phi_{ij}^r(t)$ is*

$$\Phi^r(t) = [\Phi(t)]^r .$$

Proof. Let us proceed by recurrence and assume that the theorem is true up to $r-1$. One has $p_{ij}^r(y) = \sum_{l=1}^m \sum_{z=1}^{y-1} p_{il}(z) p_{lj}^{r-1}(y-z)$ so $\phi_{ij}^r(t)$ is equal to $\sum_{l=1}^m \left[\sum_{z\geq 1} p_{il}(z) t^z\right] \left[\sum_{y \geq z+1} p_{lj}^{r-1}(y-z) t^{y-z}\right] = \sum_{l=1}^m \phi_{il}(t) \phi_{lj}^{r-1}(t)$ so $\Phi^r(t) = \Phi(t)\Phi^{r-1}(t)$. ■

The following two corollaries come straightforward.

Corollary 9 *The probability generating function $\phi_{i\bullet}^r(t)$ of the r-scan Y_i^r is*

$$\phi_{i\bullet}^r(t) = \phi_i^r(t).\mathbf{1}$$

where $\phi_i^r(t)$ is the $i-$th row of $\Phi^r(t)$.

Corollary 10 *Let ν be a distribution over $\mathcal{W}$. If the first word is distributed according to ν, then the probability generating function of the $p_{\nu\bullet}^r(y)$ is*

$$\phi_{\nu\bullet}^r(t) = \nu.\Phi^r(t).\mathbf{1}.$$

First word $\mathbf{w}_0$ given. Using the notation of Corollary **7** and the properties of the generating functions of sums of independent random variables, the following corollary is obvious.

Corollary 11 *If the first word is $\mathbf{w}_0$, the generating function of the $p^r_{0\bullet}(y)$ is*

$$\phi^r_{0\bullet}(t) = \phi_0(t).\Phi^{r-1}(t).\mathbf{1}$$

Computation of the probabilities. The probabilities $p^r(y)$'s could be computed using $\phi^r(t)$ but this appears to lead to numerical instability so it seems better to use the r-th convolution of the $p(y)$.

3.2 Chen-Stein approximation for extremal $r-$scans

Dembo and Karlin (1992) have obtained results (Theorem 1) on the distribution of minimum (and maximum) r-scans using a Poisson approximation. Let $N^-(a)$ denote the number of r-scans shorter than (or equal to) a in a sequence of n distances: they approximate the distribution of $N^-(a)$ by a Poisson distribution with parameter $\lambda(a)$:

$$\lambda(a) = (n - r + 1)\Pr\{Y_r \leq a\}.$$

Using the Chen-Stein method, assuming that the distances Y are i.i.d., they give an upper bound $b(a)$ for distance in total variation between the distribution of $N^-(a)$ and a Poisson distribution $\mathcal{P}[\lambda(a)]$:

$$b(a) = \left(1 - e^{-\lambda(a)}\right)\left[2(r-1)\Pr\{Y_r \leq a\} + 2\sum_{i=1}^{r-1}\Pr\{Y_i \leq a\}\right].$$

Using the exact distribution of $Y_{\nu\bullet}$ given above, we can calculate $\Pr\{Y_r \leq a\}$ for any r and any a, and then calculate the exact values of $\lambda(a)$. r-scan can be used to detect region with a high (or low) density of a family of words. Usually, one computes the threshold a^* such that $\exp[-\lambda(a^*)] = \alpha$ (where α is a given risk). If the smallest r-scan is smaller than a^*, it is said significantly small.

Case of a renewal process. The i.i.d. hypothesis required for the validity of the Chen-Stein bound does not hold in the general case. This hypothesis is true if the semi-Markovian process is a renewal process, e.g. when the distribution of Y_{ij} does not depend on i, as in the example proposed in section **4.2**. In this case, we may calculate the Chen-Stein bound $b(a)$ and derive two more thresholds:

$$a^*_{\text{inf}} : \exp[-\lambda(a)] + b(a) = \alpha, \qquad a^*_{\text{sup}} : \exp[-\lambda(a)] - b(a) = \alpha.$$

It is only possible to arrive at a conclusion when the smallest r-scan is either smaller than a^*_{inf} or larger than a^*_{sup}. When $b(a) > \exp[-\lambda(a)]$ for some $a \geq a^*$, it may be impossible to obtain a^*_{sup} which is not defined in this case. According to our experience, this happens fairly often, especially when the ratio between the total number of occurrences and r is not sufficiently high.

4 Applications

4.1 Sooner waiting time problem

The sooner waiting problem has been extensively studied. For example Koutras and Alexandrou (1997) have generating function in a sequence of trinary trials. This problem considers 'pure words' i.e. words made of repetitions of the same letter. The most often studied case is the success and failure runs where $\mathcal{A} = \{0,1\}$ and $\mathcal{W} = \{(1...1),\ (0...0)\}$ with $|\mathbf{w}_1| = k_1$ and $|\mathbf{w}_0| = k_0$. Our results can be applied to deal with any size of alphabet. To calculate the generating function we only need the transition matrix Π, the overlapping polynomial matrix $\mathbf{D}(t)$ and the diagonal matrix $\mathbf{T}(t)$.

For example, in the simplest case of success and failure runs, we get

$$\mathbf{D}(t) = \begin{bmatrix} \pi(0,0)\dfrac{1-[\pi(0,0)t]^{-k_0}}{1-[\pi(0,0)t]^{-1}} & 0 \\ 0 & \pi(1,1)\dfrac{1-[\pi(1,1)t]^{-k_1}}{1-[\pi(1,1)t]^{-1}} \end{bmatrix}$$

and

$$\mathbf{T}(t) = \begin{bmatrix} t^{k_0}\pi(0,0)^{k_0-1} & 0 \\ 0 & t^{k_1}\pi(1,1)^{k_1-1} \end{bmatrix}.$$

4.2 Genome analysis: CHI of *Haemophilus influenza*

We study here the Crossover Hotspot Initiator of *Haemophilus influenza* which is known to protect the genome against restriction enzymes and is therefore expected to be particularly frequent and regularly spaced.

There are $m = 4$ versions of this pattern which are all $k_i \equiv 8$ letters long:

$$\mathcal{W} = \{\texttt{gatggtgg}, \texttt{gctggtgg}, \texttt{ggtggtgg}, \texttt{gttggtgg}\}.$$

It can be noted that $\mathbf{w}_3$ overlaps itself and the other words with 1, 2 and 5 letters. The other words overlap each other with only 1 letter. The overlapping structure does not depend on the first word, the overlapping polynomials $d_{ij}(t)$ and the generating functions $\phi_{ij}(t)$ do not depend on i so $\phi_{i\bullet}(t) = \phi_{\nu\bullet}(t)$ and the probability distribution function (pdf) of the distance and of the r-scans do not depend on the first word. Hence the distances are i.i.d. and we are in the renewal case described in section **3.2**. For the same reason, the generating matrix $\Phi(t)$ and the transition matrix $\mathbf{S}$ have all their rows equal; any row of $\mathbf{S}$ gives its steady distribution ν.

The model. Assuming that the sequence is generated according to a first order Markov model (adjusted to the sequence) with transition matrix

$$\Pi = \begin{pmatrix} 0.3827 & 0.1547 & 0.1639 & 0.2988 \\ 0.3429 & 0.1874 & 0.2156 & 0.2541 \\ 0.2693 & 0.2641 & 0.1974 & 0.2692 \\ 0.2302 & 0.1595 & 0.2202 & 0.3901 \end{pmatrix}$$

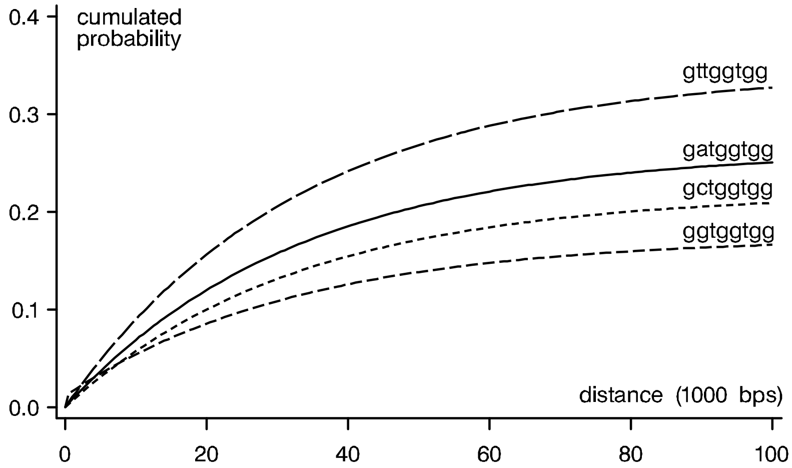

Figure 1: Distribution of the distances Y_{ij}'s between the different versions of the CHI motif of *H. influenza*.

the expected distance between two occurrences is $\mathbb{E}(Y_{\nu\bullet}) = 32\,535.8$ and the standard deviation is 32 916.6. Under this model, the generating functions $\phi_{ij}(t)$ are all ratios of two polynomials with degree 10 and the transition probabilities between words $s(i,j)$ are simply given by the steady distribution ν :

$$\nu = \begin{pmatrix} 0.2632 & 0.2195 & 0.1738 & 0.3435 \end{pmatrix}.$$

The cumulative probabilities $\sum_{z=1..y} p_{ij}(z)$ are given in figure **1**. Remember that they are not cumulative distribution functions since they do not reach 1: the asymptotic values of this distributions are equal to $s(i,j)$ which equal in this case to $\nu(j)$. We see that the $p_{i3}(y)$'s are superior to all other $p_{ij}(y)$'s for very small y: this is due to the fact that $\mathbf{w}_3$ has more chance to overlap other word.

The data. The complete genome of *H. influenza* is 1,830,022 base pairs (bps) long. The number of occurrences of $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \mathbf{w}_4$ are respectively 28, 56, 76 and 63 so there are 223 occurrences and $n = 222$ distances between them. The mean distance is $8,087.4$ bps, the standard deviation 9 544.1 bps, the smallest distance 3 bps (observed 8 times) and the greatest 70 722 bps. The distance 3 is necessarily obtained by an overlap of $\mathbf{w}_3$ on another word.

Use of r-scans The heterogeneity of the CHI motif in the genome can be checked using 3-scans (see figure **2**): the 5% threshold of the simple Poisson approximation is $a^* = 2744$ ($r/a^* = 1.09\,10^{-3}$) but, taking the error bound $b(a)$ into account we get $a^*_{\text{inf}} = 2537$ ($r/a^*_{\text{inf}} = 1.18\,10^{-3}$) and $a^*_{\text{sup}} = 3071$ ($r/a^*_{\text{sup}} = 0.98\,10^{-3}$). We observe 4 peaks exceeding the upper bound for the density r/a^*_{inf} and one candidate for which we can not decide.

Two interpretations of this result can be given:

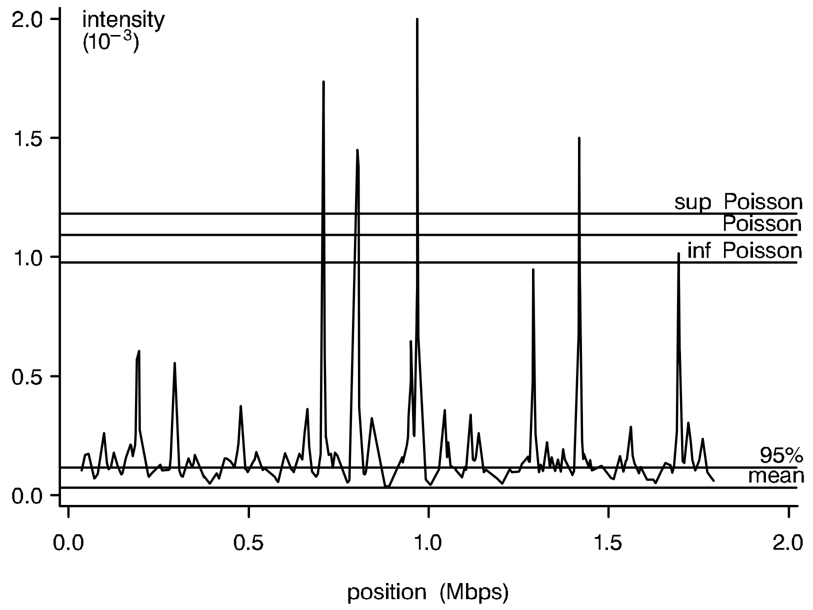

Figure 2: Density of the CHI motif using 3-scans in the genome of *H. influenza*: mean $= 1/\mathbb{E}Y$, 95% $= r/y^r_{95\%}$ where $\Pr\{Y^r < y^r_{95\%}\} = 95\%$, Poisson $= r/a^*$, inf Poisson $= r/a^*_{\text{sup}}$, sup Poisson $= r/a^*_{\text{inf}}$ (the highest peak at position 0.97 Mbps is truncated, its real height is $4.8\ 10^{-3}$).

- The peaks reveal real rich regions for the CHI motif.
- The general frequency of the CHI motif is much higher than expected under the M1 model. In this case, the peaks may be simple consequences of the departure from the model.

In this case, the second interpretation seems more valid since the density is everywhere higher than expected. The mean and variance of a 222-scan under the M1 model are respectively $\mu = 7\,223\,007$ and $\sigma = 490\,445$; the distance between the first and the last occurrence of the motif is $y^{222} = 1\,795\,407$. Using the central limit theorem for Y^{222}, we get a gaussian score of -11.1, which is highly significant.

The upper bound a^*_{sup} can not be computed for the 4-scan because $e^{-\lambda(a)} - b(a)$ turns out to be negative.

Acknowledgment We thank an anonymous referee for his helpful technical remarks.

References

Aki, S. and Balakrishnan, S., N.and Mohanty. (1996). Sooner and later waiting times problems for success and failure runs in higher order markov dependent trials. *Ann. Inst. Statist. Math.* **48 (4)** 773–787.

AKI, S. and HIRANO, K. (1993). Discrete distributions related to succession events in a two state markov chain. In *Statistical Sciences and Data Analysis*, (K. Matusita, M. L. Puri, and T. Hayakawa, ed.), 467–474. VSP Publishers, Amsterdam.

BREEN, S., WATERMAN, M. S. and ZHANG, N. (1985). Renewal theory for several patterns. *J. Appl. Prob.* **22** 228–234.

CHRYSSAPHINOU, O. and PAPASTAVRIDIS, S. (90). The occurrence of sequence patterns in repeated dependent experiment. *Theory Prob. Appl.* **35 (1)** 145–152.

DEMBO, A. and KARLIN, S. (1992). Poisson approximations for r-scan processes. *Ann. Appl. Prob.* **2 (2)** 329–357.

FU, C. J. (1996). Distribution of runs and patterns associated with a sequence of multi-state trials. *Statistica Sinica.* **6** 957–974.

FU, C. J. and KOUTRAS, M. V. (94). Distribution theory of runs: A markov chain approach. *J. Amer. Statist. Assoc.* **427** 1050–1058.

KARLIN, S. and MACKEN, C. (1991). Some statistical problems in the assessment of inhomogeneities of DNA sequence data. *J. Amer. Statist. Assoc.* **86** 27–35.

KOUTRAS, M. (1997). Waiting times and number of appearances of events in a sequence of discrete random variables. In *Advances in combinatorial methods and applications to probability and statistics*, (N. Balakrishnan, ed.), 253–261, Statistics and Industry and Technology Series. Birkhauser, Boston.

KOUTRAS, M. and ALEXANDROU, V. (1997). Sooner waiting time problems in a sequence of trinary trials. *J. Appl. Prob.* **34** 593–609.

MORI, T. (1991). On the waiting time til each of some given patterns occurs as a run. *Prob. Theory Rel. Fields.* **67** 313–323.

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

UCHIDA, M. and AKI, S. (1995). Sooner or later waiting time problems in a two-state markov chain. *Ann. Inst. Statist. Math.* **47** 415–433.