Pattern Matching in Constrained Sequences

Yongwook Choi and Wojciech Szpankowski
Department of Computer Science
Purdue University
W. Lafayette, IN 47907 U.S.A.
Email: ywchoi@purdue.edu, spa@cs.purdue.edu

***Abstract*— Constrained sequences find applications in communication, magnetic recording, and biology. In this paper, we restrict our attention to the so-called (d,k) constrained binary sequences in which any run of zeros must be of length at least d and at most k, where $0 \leq d < k$. In some applications one needs to know the number of occurrences of a given pattern w in such sequences, for which we coin the term *constrained pattern matching*. For a given word w or a set of words $\mathcal{W}$, we estimate the (conditional) probability of the number of occurrences of w in a (d,k) sequence generated by a memoryless source. As a by-product, we enumerate asymptotically the number of (d,k) sequences with exactly r occurrences of a given word w, and compute Shannon entropy of (d,k) sequences with a given number of occurrences of w. Throughout this paper we use techniques of analytic information theory such as combinatorial calculus, generating functions, and complex asymptotics.**

I. Introduction

The main idea of constrained coding is to restrict the set of messages to a constrained set. In digital communication systems such as magnetic and optical recording, the main purpose of constrained encoding is to improve the performance by matching system characteristics to those of the channel. In biology, constrained sequences and constrained channels may be used to model Darwin selection and biodiversity. Indeed, biomolecular structures, species, and in general biodiversity, as they exist today, have gone through significant metamorphosis over eons through mutation and natural selection. One can argue that biodiversity is a consequence of *information transfer*, which occurs over time (*e.g.* inheritance) and across biological entities (*e.g.* symbiosis, predator-prey). To capture sources of variation and natural selection, one is tempted to introduce the so-called *Darwin channel* which is a combination of deletion/insertion channels and noisy constrained channels (cf. [4], [5]).

In this paper, we restrict our goal to study and understand some aspects of pattern matching in constrained sequences. Although our methods work for a large class of constrained systems, we further restrict our analysis to the so-called (d,k) sequences in which runs of zeros cannot be smaller than d and bigger than k, where $0 \leq d < k$. Such sequences have proved to be very useful for digital recording. Also, spike trains recorded from different neurons in the brain of an animal seem to satisfy structural constraints that exactly match the framework of (d,k) binary sequences. For example, refractoriness requires that a neuron cannot fire two spikes in too short a time; this precisely translates into the constraint that the induced binary spike train will need to contain at least a certain number of zeros (corresponding to no activity) between each two consecutive ones (corresponding to firing times).

In these applications, one often requires that some given words do not occur or occur only a few times in a (d,k) sequence. Therefore, we study here the following problem: given a word w or a set of words $\mathcal{W}$ how many times it occurs in a (d,k) sequence. For such a problem we coin the term *constrained pattern matching* as an extension of standard pattern matching [11], [16], [18]. We study this problem in a probabilistic framework, that is, we assume that a sequence is generated by a (biased) memoryless source and derive the (conditional) distribution of the number of occurrences of w in a (d,k) sequence. We need the conditional distribution since naturally only a small fraction of binary sequences satisfies the (d,k) constraints.

In the (standard) pattern matching problem, one asks for pattern occurrences in a binary string also known as *text* without any additional restrictions on the text. In a probabilistic framework, one determines the distribution of the number of pattern occurrences. The first analysis of such pattern matching goes back at least to Feller, and enormous progress in this area has been reported since then [2], [7], [11], [14], [18], [19]. For instance, Guibas and Odlyzko [7] (cf. also [8], [16]) revealed the fundamental role played by autocorrelation languages and their associated polynomials in the analysis of pattern matching. Régnier and Szpankowski [15], [16] established that the number of occurrences of a given pattern is asymptotically normal under a diversity of probabilistic models that include Markov chains. Nicodème, Salvy, and Flajolet [14] showed generally that the number of places in a random text at which a 'motif' (i.e., a general regular expression pattern) terminates is asymptotically normally distributed. Bender and Kochman [2] studied a generalized pattern occurrences using (in a nutshell) the de Bruijn graph representation that allowed the authors to establish the central limit theorem, but without explicit mean and variance. Recent surveys on pattern matching can be found in Lothaire [11] (Chaps. 6 and 7). To the best of our knowledge, none of these works deal with pattern matching in constrained sequences such as (d,k) sequences.

In the information theory community, (d,k) sequences were analyzed since Shannon with some recent contributions [3], [10], [12], [20]. Pattern matching in constrained sequences can in principle be analyzed by various versions of the de

Bruijn graph [2], [6] or automaton approach [2], [14]. This is an elegant and general approach but it sometimes leads to complicated analyses and is computationally extensive. In our constrained pattern matching, for example, one must build a de Bruijn graph over *all* strings of length equal to the longest string in the set $\mathcal{W}$. The (d,k) constraints are built into the graph as *forbidden strings* (i.e., runs of zeros of length smaller than d and larger than k) which result in forbidden edges of the graph. Based on this method, one represents the number of pattern occurrences as a product of a matrix representation of the underlying de Bruijn graph and hence its largest eigenvalue (cf. [2], [6]). In general, this matrix is of a large dimension and such a solution is not easily interpretable in terms of the original patterns.

In this paper, we take the view of combinatorics on words. We first construct languages representing (d,k) sequences containing exactly r occurrences of a given pattern w or a set of patterns $\mathcal{W}$. Using generating functions and complex asymptotics, we present simple and precise asymptotics for the mean, variance, and the central limit theorem for the number of occurrences. In particular, we estimate the probability that a randomly generated sequence is a (d,k) sequence. Furthermore, we enumerate (d,k) sequences that contain exactly r occurrences of w and compute Shannon entropy.

II. Main Results for a Single Pattern

In this section, we consider only *restricted* (d,k) sequences that start with 0 and end with 1; we later relax this assumption. Our goal is to derive the probability distribution of the number of occurrences of a given pattern w in a (d,k) sequence generated by a binary memoryless source. Here w is also a (d,k) sequence, and pattern overlapping is allowed.

A. Language Representation

Let us define

$$\mathcal{A}_{d,k} = \{\underbrace{0\ldots0}_{d},\cdots,\underbrace{0\ldots0}_{k}\}$$

as a set of runs of zeros of length between d and k. We also define the following set (known as the *extended alphabet* [12])

$$\mathcal{B}_{d,k} = \mathcal{A}_{d,k}\cdot\{1\} = \{\underbrace{0\ldots0}_{d}1,\cdots,\underbrace{0\ldots0}_{k}1\}.$$

In order to assure that we deal only with restricted (d,k) sequences, we build sequences and patterns over $\mathcal{B}_{d,k}$. As a consequence, occurrences of pattern w are also over $\mathcal{B}_{d,k}$ (e.g., $w=01$ does *not* occur in a $(1,4)$ sequence such as 0010001, which contains only two symbols over $\mathcal{B}_{d,k}$, namely 001 and 0001). We shall relax this assumption later on.

Let $w = w_1\ldots w_m \in \{0,1\}^m$ with $w_1=0$ and $w_m=1$, but over $\mathcal{B}_{d,k}$ we have $w=\beta_1\ldots\beta_{m'}$, where $\beta_i\in\mathcal{B}_{d,k}$ and $\sum_{i=1}^{m'}|\beta_i| = m$. Let $\mathcal{S}$ denote the *autocorrelation set* of w over $\mathcal{B}_{d,k}$, that is,

$$\mathcal{S} = \{\beta_{l+1}^{m'}:\ \beta_1^l = \beta_{m'-l+1}^{m'}\},\quad 1\le l\le m'$$

where $\beta_i^j = \beta_i\cdots\beta_j$ and $\beta_i^j=\epsilon$ if $i>j$.

As in [9], [16], we define four languages, $\mathcal{T}_r^{(d,k)}$, $\mathcal{R}^{(d,k)}$, $\mathcal{M}^{(d,k)}$ and $\mathcal{U}^{(d,k)}$ as follows:

(i) $\mathcal{T}_r^{(d,k)}$ as the set of all (d,k) sequences (over the extended alphabet $\mathcal{B}_{d,k}$) containing exactly r occurrences of w;

(ii) $\mathcal{R}^{(d,k)}$ as the set of all (d,k) sequences (over the extended alphabet $\mathcal{B}_{d,k}$) containing only one occurrence of w, located at the right end;

(iii) $\mathcal{U}^{(d,k)}$ defined as

$$\mathcal{U}^{(d,k)} = \{u:\ w\cdot u\in\mathcal{T}_1^{(d,k)}\},$$

that is, a word $u\in\mathcal{U}^{(d,k)}$ if u is a (d,k) sequence and $w\cdot u$ has exactly one occurrence of w at the left end of $w\cdot u$;

(iv) $\mathcal{M}^{(d,k)}$ defined as

$$\mathcal{M}^{(d,k)} = \{v:\ w\cdot v\in\mathcal{T}_2^{(d,k)} \text{ and } w \text{ occurs at the right end of } w\cdot v\},$$

that is, any word in $\{w\}\cdot\mathcal{M}^{(d,k)}$ has exactly two occurrences of w, one at the left and one at the right end.

To simplify our notation, we drop the upper index (d,k) unless it is necessary. It is easy to see that [16], [18]

$$\mathcal{T}_r = \mathcal{R}\cdot\mathcal{M}^{r-1}\cdot\mathcal{U}, \tag{1}$$

$$\mathcal{T}_0\cdot\{w\} = \mathcal{R}\cdot\mathcal{S}. \tag{2}$$

In order to find relationships between the languages $\mathcal{R}$, $\mathcal{M}$, and $\mathcal{U}$, we extend the approach from [16] to yield

$$\mathcal{M}^* = \mathcal{B}^*\cdot\{w\}+\mathcal{S}, \tag{3}$$

$$\mathcal{U}\cdot\mathcal{B} = \mathcal{M}+\mathcal{U}-\{\epsilon\}, \tag{4}$$

$$\{w\}\cdot\mathcal{M} = \mathcal{B}\cdot\mathcal{R}-(\mathcal{R}-\{w\}), \tag{5}$$

where $\mathcal{B}^*$ is the set of all restricted (d,k) sequences that start with 0 and end with 1, that is,

$$\mathcal{B}^* = \{\epsilon\}+\mathcal{B}+\mathcal{B}^2+\mathcal{B}^3+\cdots.$$

Similarly, $\mathcal{M}^* = \sum_{i=0}^{\infty}\mathcal{M}^i$, where $\mathcal{M}^0=\{\epsilon\}$.

B. Probability Generating Functions

At this point we need to set up the probabilistic framework. Throughout, we assume that a binary sequence is generated by a memoryless source with p being the probability of emitting a '0' and $q=1-p$. Among others, we compute the probability that a randomly generated sequence is a (d,k) sequence. We actually derive the conditional probability distribution of the number of occurrences of w in a (d,k) sequence.

We start by defining for a language $\mathcal{L}$ its *probability generating function* $L(z)$ as

$$L(z) = \sum_{u\in\mathcal{L}} P(u)z^{|u|},$$

where $P(u)$ is the probability of u. In particular, the *autocorrelation polynomial* $S(z)$ is the probability generating function for the autocorrelation language $\mathcal{S}$. In general, we write $[z^n]L(z)$ for the coefficient of $L(z)$ at z^n.

The language relationships (3)–(5) are translated into probability generating functions:

$$\frac{1}{1-M(z)} = \frac{1}{1-B(z)} \cdot z^m P(w) + S(z), \quad (6)$$

$$U(z) = \frac{M(z)-1}{B(z)-1}, \quad (7)$$

$$R(z) = z^m P(w) \cdot U(z), \quad (8)$$

where $P(w)$ is the probability of w, and

$$B(z) = p^d q z^{d+1} + p^{d+1} q z^{d+2} + \cdots + p^k q z^{k+1} = zq\frac{(zp)^d - (zp)^{k+1}}{1-zp}. \quad (9)$$

In particular, from (1)–(2) and above, one finds

$$T_0(z) = \frac{S(z)}{D(z)}, \quad (10)$$

$$T_r(z) = \frac{z^m P(w)(D(z)+B(z)-1)^{r-1}}{D(z)^{r+1}}, \quad (11)$$

$$\text{where} \quad D(z) = S(z)(1-B(z)) + z^m P(w). \quad (12)$$

C. Number of Occurrences

Let O_n be a random variable representing the number of occurrences of w in a (regular) binary sequence of length n. Then, the generating function $T_r(z)$ for (d,k) sequences is defined as follows

$$T_r(z) = \sum_{n \geq 0} P(O_n = r, \mathcal{D}_n) z^n,$$

where $\mathcal{D}_n$ is the event that a randomly generated binary sequence of length n is a (d,k) sequence. Let us also define the bivariate generating function $T(z,u)$ as

$$T(z,u) = \sum_{r \geq 0} T_r(z) u^r = \sum_{r \geq 0} \sum_{n \geq 0} P(O_n = r, \mathcal{D}_n) z^n u^r.$$

From (1), we find

$$T(z,u) = R(z)\frac{u}{1-uM(z)}U(z) + T_0(z). \quad (13)$$

Observe that $T(z,u)$ is *not* a bivariate *probability* generating function since $[z^n]T(z,1) \neq 1$. But we can easily make it a *conditional* probability generating function. First, define

$$P(\mathcal{D}_n) = [z^n]T(z,1)$$

as the probability that a randomly generated sequence of length n is a (d,k) sequence. We also introduce a short-hand notation $O_n(\mathcal{D}_n)$ for the conditional number of occurrences of w in a (d,k) sequence. More formally,

$$P(O_n(\mathcal{D}_n) = r) = P(O_n = r \,|\, \mathcal{D}_n).$$

Therefore, the probability generating function of $O_n(\mathcal{D}_n)$ is

$$\mathbf{E}[u^{O_n(\mathcal{D}_n)}] = \frac{[z^n]T(z,u)}{[z^n]T(z,1)}.$$

Thus, the expected value of $O_n(\mathcal{D}_n)$ is

$$\mathbf{E}[O_n(\mathcal{D}_n)] = \frac{[z^n]T_u(z,1)}{[z^n]T(z,1)},$$

where $T_u(z,1)$ is the derivative of $T(z,u)$ at $u=1$, and

$$\mathbf{E}[O_n(\mathcal{D}_n)(O_n(\mathcal{D}_n)-1)] = \frac{[z^n]T_{uu}(z,1)}{[z^n]T(z,1)}$$

is the second factorial moment.

D. Asymptotics

We first obtain asymptotic formulas for the mean and the variance of $O_n(\mathcal{D}_n)$.

Theorem 1: Let $\rho := \rho(p)$ be the smallest real root of $B(z) = 1$ where $B(z)$ is defined in (9), and let $\lambda = 1/\rho$. Then, for large n, the probability of generating a (d,k) sequence is asymptotically

$$P(\mathcal{D}_n) = \frac{1}{B'(\rho)}\lambda^{n+1} + O(\omega^n)$$

for some $\omega < \lambda$. Furthermore, the mean is

$$\mathbf{E}[O_n(\mathcal{D}_n)] = \frac{(n-m+1)P(w)}{B'(\rho)}\lambda^{-m+1} + O(1),$$

and the variance becomes

$$\mathbf{Var}[O_n(\mathcal{D}_n)] = (n-m+1)P(w)\left[\frac{(1-2m)P(w)}{B'(\rho)^2}\lambda^{-2m+2} + \frac{P(w)B''(\rho)}{B'(\rho)^3}\lambda^{-2m+1} + \frac{2S(\rho)-1}{B'(\rho)}\lambda^{-m+1}\right] + O(1).$$

Proof. From (6)-(13), we find

$$T(z,1) = \frac{1}{1-B(z)}, \quad T_u(z,1) = \frac{z^m P(w)}{(1-B(z))^2}, \quad \text{and}$$

$$T_{uu}(z,1) = \frac{2z^m P(w) M(z)}{U(z)(1-B(z))^3} = \frac{2z^m P(w) D(z)}{(1-B(z))^3} - \frac{2z^m P(w)}{(1-B(z))^2}.$$

By Cauchy's coefficient formula and Cauchy's residue theorem [18] we immediately obtain

$$P(\mathcal{D}_n) = [z^n]T(z,1) = [z^n]\frac{1}{1-B(z)} = \frac{1}{B'(\rho)}\lambda^{n+1} + O(\omega^n),$$

where ρ is the smallest real root of $B(z) = 1$, $\lambda = 1/\rho$, and $\omega < \lambda$. By elementary analysis we can prove such a root exists. To find moments, we proceed as follows.

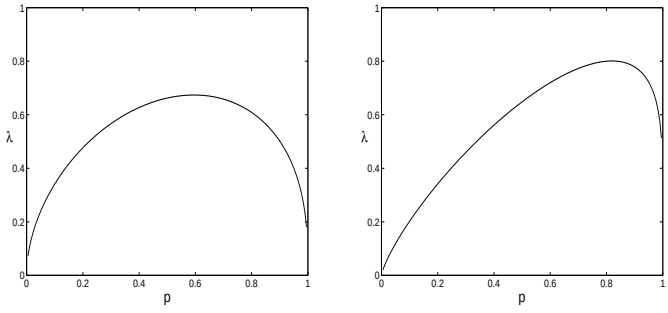
$$[z^n]T_u(z,1) = [z^n]\frac{z^m P(w)}{(1-B(z))^2} = \frac{P(w)}{B'(\rho)^2}\left((n-m+1)\lambda + \frac{B''(\rho)}{B'(\rho)}\right)\lambda^{n-m+1} + O(\omega^n).$$

Thus

$$\mathbf{E}[O_n(\mathcal{D}_n)] = \frac{[z^n]T_u(z,1)}{[z^n]T(z,1)} = \frac{(n-m+1)P(w)}{B'(\rho)}\lambda^{-m+1} + O(1)$$

and

$$\mathbf{Var}[O_n(\mathcal{D}_n)] = \frac{[z^n]T_{uu}(z,1)}{[z^n]T(z,1)} + \mathbf{E}[O_n(\mathcal{D}_n)] - \mathbf{E}[O_n(\mathcal{D}_n)]^2.$$



(a) (1,2) sequences (b) (3,6) sequences

Fig. 1. λ versus p.

After some algebra, we establish the theorem. ∎

Remark 1. In Figure 1 we plot $\lambda = 1/\rho$ versus p for various (d,k) sequences. Observe that the probability $P(\mathcal{D}_n) \asymp \lambda^n$ is asymptotically maximized for some $p \neq 0.5$ (biased source) which may be used to design a better run-length coding (cf. [1]).

Our expressions for the bivariate generating functions allow us to estimate asymptotically the probability of r occurrences of w for various ranges of r.

Theorem 2: (i) Let $\tau := \tau(p,w)$ be the smallest real root of $D(z) = 0$ (cf. (12)) and $\rho := \rho(p)$ be the smallest real root of $B(z) = 1$. Then for $r = O(1)$ we have

$$P(O_n(\mathcal{D}_n) = r) \sim \frac{P(w)B'(\rho)(1-B(\tau))^{r-1}}{D'(\tau)^{r+1}\tau^{r-m}} \cdot \binom{n-m+r}{r} \left(\frac{\rho}{\tau}\right)^{n+1}$$

for large n and $r \geq 1$.

(ii) Let $r = \mathbf{E}[O_n(\mathcal{D}_n)] + x\sqrt{\mathbf{Var}[O_n(\mathcal{D}_n)]}$ for $x = O(1)$. Then

$$\frac{O_n(\mathcal{D}_n) - \mathbf{E}[O_n(\mathcal{D}_n)]}{\sqrt{\mathbf{Var}[O_n(\mathcal{D}_n)]}} \xrightarrow{d} N(0,1),$$

where $N(0,1)$ is the standard normal distribution.

Proof. In this conference version, we only prove part (i). By Cauchy's coefficient formula and Cauchy's residue theorem,

$$P(O_n = r, \mathcal{D}_n) = [z^{n-m}]\frac{P(w)(D(z)+B(z)-1)^{r-1}}{D(z)^{r+1}}$$

$$= \sum_{j=1}^{r+1}(-1)^j a_j \binom{n-m+j-1}{j-1}\left(\frac{1}{\tau}\right)^{n-m+j} + O(t^n)$$

where $\tau < t^{-1}$ is the smallest real root of $D(z) = 0$, and $a_{r+1} = \dfrac{P(w)(B(\tau)-1)^{r-1}}{D'(\tau)^{r+1}}$. It is easy to prove that such a root of $D(z)$ exists. Finally, we find

$$P(O_n(\mathcal{D}_n) = r) = \frac{P(O_n = r, \mathcal{D}_n)}{P(\mathcal{D}_n)}$$

$$\sim \frac{P(w)B'(\rho)(1-B(\tau))^{r-1}}{D'(\tau)^{r+1}}\binom{n-m+r}{r}\frac{\rho^{n+1}}{\tau^{n-m+r+1}}.$$

This proves part (i). ∎

Remark 2. When the binary source is unbiased ($p = q = \frac{1}{2}$), we can count the number, $N_n(r)$, of (d,k) sequences of length n that contain w exactly r times, by computing $[z^n]T_r(2z)$. In fact, $N_n(r) = 2^n P(O_n = r, \mathcal{D}_n)$ and one finds asymptotics of $N_n(r)$ from Theorem 2 for $r = O(1)$. In particular, Shannon entropy is

$$C(r) = \lim_{n\to\infty}\frac{\log N_n(r)}{n} = \log\left(\frac{2}{\tau}\right)$$

where $\tau = \tau(1/2, w)$ is defined in Theorem 2 for $p = 1/2$.

Remark 3. We considered only *restricted* (d,k) sequences. A small modifications can extend this analysis to *all* (d,k) sequences. Let $\mathcal{T}_r^{all}$ be the set of all (d,k) sequences containing exactly r occurrences of w. Then

$$\mathcal{T}_r^{all} = \{\epsilon, 1\}\cdot\mathcal{T}_r\cdot(\{\epsilon\} + \mathcal{A}_{d,k})$$

and one can easily derive generating functions and asymptotic expressions from the above.

Remark 4. We counted the occurrences of the pattern w over the alphabet $\mathcal{B}_{d,k}$. We can extend this analysis to count the occurrences over a binary alphabet (e.g., $w = 01$ occurs twice in a $(1,4)$ sequence 0010001). Again, let $w = w_1\ldots w_m \in \{0,1\}^m$ with $w_1 = 0$ and $w_m = 1$, and w be represented over $\mathcal{B}_{d,k}$, that is, $w = \beta_1\ldots\beta_{m'}$ where $\beta_i \in \mathcal{B}_{d,k}$. Then the autocorrelation set $\mathcal{S}_2$ over the binary alphabet $\{0,1\}$ is defined as

$$\mathcal{S}_2 = \{w_{l+1}^m :\ w_1^l = w_{m-l+1}^m\},\quad 1 \leq l \leq m.$$

Using the languages $\mathcal{T}_r$, $\mathcal{R}$, $\mathcal{M}$, and $\mathcal{U}$ defined above, we find

$$\begin{aligned}
\mathcal{T}_r &= \mathcal{R}\cdot\mathcal{M}^{r-1}\cdot\mathcal{U},\\
\mathcal{T}_0\cdot\mathcal{Z}\cdot\{w\} &= \mathcal{R}\cdot\mathcal{S}_2,\\
\mathcal{M}^* &= \mathcal{B}^*\cdot\mathcal{Z}\cdot\{w\} + \mathcal{S}_2,\\
\mathcal{U}\cdot\mathcal{B} &= \mathcal{M} + \mathcal{U} - \{\epsilon\},\\
\mathcal{Z}\cdot\{w\}\cdot(\mathcal{M} - \{\epsilon\}) &= \mathcal{B}\cdot\mathcal{R} - \mathcal{R},
\end{aligned}$$

where $\mathcal{Z} = \{\epsilon, 0, 00, \cdots, 0^{k+1-|\beta_1|}\}$ and 0^k denotes a run of zeros of length k. Applying the same techniques as above we can derive the generating functions and asymptotic results.

III. Main Result for a Set of Patterns

In this section, we are given a set of patterns $\mathcal{W} = \{w_1, w_2, \ldots, w_K\}$ such that w_i $(1 \leq i \leq K)$ is not a substring of another pattern w_j $(1 \leq j \leq K, i \neq j)$ over alphabet $\mathcal{B}_{d,k}$. Now we count the number of occurrences of $\mathcal{W}$ in a (d,k) sequence of length n.

We need to extend our definition of languages. In particular, for any given two strings u and v, let

$$\mathcal{S}_{u,v} = \{v_{k+1}^{|v|} :\ u_{|u|-k+1}^{|u|} = v_1^k\},\quad 1 \leq k \leq \min\{|u|,|v|\}.$$

be the *correlation set*. Now we define a correlation set over $\mathcal{B}_{d,k}$ for patterns in $\mathcal{W}$. Let $w_i = \beta_{i_1}\ldots\beta_{i_s}$ and $w_j =$

$\beta_{j_1} \dots \beta_{j_m}$. Then $\mathcal{S}_{ij}$, the correlation set for w_i and w_j over $\mathcal{B}_{d,k}$, is defined as

$$\mathcal{S}_{ij} = \{\beta^{j_m}_{j_{\ell+1}} : \ \beta^{i_s}_{i_{s-\ell+1}} = \beta^{j_\ell}_{j_1}\}, \quad 1 \le \ell \le \min\{s, m\}.$$

For $1 \le i, j \le K$, we introduce new languages as follows (again, we drop the upper index (d, k)):

(i) $\mathcal{R}_i$ as the set of all (d, k) sequences (over the alphabet $\mathcal{B}_{d,k}$) containing only one occurrence of w_i, located at the right end;

(ii) $\mathcal{U}_i$ is defined as $\{u : \ w_i \cdot u \in \mathcal{T}_1\}$, that is, a word $u \in \mathcal{U}_i$ if u is a (d, k) sequence and $w_i \cdot u$ has exactly one occurrence of w_i at the left end of $w_i \cdot u$;

(iii) $\mathcal{M}^{[r]}_{ij}$ defined as, for $r \ge 1$,

$$\mathcal{M}^{[r]}_{ij} = \{v : \ w_i \cdot v \in \mathcal{T}_{r+1} \text{ and } w_j \text{ occurs at the right end of } w_i \cdot v\},$$

that is, any word in $\{w_i\} \cdot \mathcal{M}^{[r]}_{ij}$ is a (d, k) sequence and has one occurrence of w_i at the left end, one occurrence of w_j at the right end, and $r - 1$ occurrences from $\mathcal{W}$ elsewhere. We write $\mathcal{M}_{ij} = \mathcal{M}^{[1]}_{ij}$.

We can see that $\mathcal{T}_r (r \ge 1)$ and $\mathcal{T}_0$ are represented as follows

$$\mathcal{T}_r = \sum_{1 \le i,j \le K} \mathcal{R}_i \cdot \mathcal{M}^{[r-1]}_{ij} \cdot \mathcal{U}_j, \tag{14}$$

$$\mathcal{T}_0 \cdot \{w_j\} = \mathcal{R}_j + \sum_{1 \le i \le K} \mathcal{R}_i \cdot (\mathcal{S}_{ij} - \{\epsilon\}) \tag{15}$$

for $1 \le j \le K$. The languages $\mathcal{M}_{ij}$, $\mathcal{U}_i$, and $\mathcal{R}_j$ satisfy following relationships [15] for $1 \le i, j \le K$

$$\sum_{k \ge 1} \mathcal{M}^{[k]}_{ij} = \mathcal{B}^* \cdot \{w_j\} + \mathcal{S}_{ij} - \{\epsilon\}, \tag{16}$$

$$\mathcal{U}_i \cdot \mathcal{B} = \sum_{1 \le j \le K} \mathcal{M}_{ij} + \mathcal{U}_i - \{\epsilon\}, \tag{17}$$

$$\mathcal{B} \cdot \mathcal{R}_j - (\mathcal{R}_j - \{w_j\}) = \sum_{1 \le i \le K} \{w_i\} \cdot \mathcal{M}_{ij}. \tag{18}$$

As before, the language relationships (16)–(18) are translated into generating functions [15]:

$$(\mathbf{I} - \mathbf{M}(z))^{-1} = \mathbf{S}(z) + \frac{1}{1 - B(z)} \overrightarrow{1} \cdot \overrightarrow{W}^t(z),$$

$$\overrightarrow{U}(z) = \frac{1}{1 - B(z)} (\mathbf{I} - \mathbf{M}(z)) \cdot \overrightarrow{1},$$

$$\overrightarrow{R}^t(z) = \frac{1}{1 - B(z)} \overrightarrow{W}^t(z) \cdot (\mathbf{I} - \mathbf{M}(z)),$$

where $\mathbf{M}(z)$ and $\mathbf{S}(z)$ are $K \times K$ matrices such that $M_{ij}(z)$ and $S_{ij}(z)$ are the (i, j)-elements in $\mathbf{M}(z)$ and $\mathbf{S}(z)$, respectively. Furthermore, $\mathbf{I}$ is the $K \times K$ identity matrix, and $\overrightarrow{W}(z)$, $\overrightarrow{R}(z)$, $\overrightarrow{U}(z)$, and $\overrightarrow{1}$ are column vectors of length K such that $\overrightarrow{W}(z) = (z^{|w_1|} P(w_1), \cdots, z^{|w_K|} P(w_K))^t$, $\overrightarrow{R}(z) = (R_1(z), \cdots, R_K(z))^t$, $\overrightarrow{U}(z) = (U_1(z), \cdots, U_K(z))^t$, and $\overrightarrow{1} = (1, \cdots, 1)^t$.

From (14)–(15) and above, we can find

$$T_0(z) = \frac{\overrightarrow{R}^t(z) \cdot \mathbf{S}(z) \cdot \overrightarrow{1}}{\overrightarrow{W}^t(z) \cdot \overrightarrow{1}},$$

$$T_r(z) = \overrightarrow{W}^t(z) \cdot (\mathbf{D}(z) + (B(z) - 1)\mathbf{I})^{r-1} \cdot \mathbf{D}(z)^{-(r+1)} \cdot \overrightarrow{1},$$

where $\mathbf{D}(z) = \overrightarrow{1} \cdot \overrightarrow{W}^t(z) + (1 - B(z))\mathbf{S}(z)$.

In this conference version, we only present asymptotics for the first two moments.

Theorem 3: For large n

$$\mathbf{E}[O_n(\mathcal{D}_n)] = \sum_{i=1}^{K} \frac{(n - |w_i| + 1)P(w_i)}{B'(\rho)} \lambda^{-|w_i|+1} + O(1),$$

$$\mathbf{Var}[O_n(\mathcal{D}_n)] = n\alpha + O(1),$$

where $\rho := \rho(p)$ is the smallest real root of $B(z) = 1$, $\lambda = 1/\rho$, and α is an explicitly computable constant.

ACKNOWLEDGMENT

The work of W. Szpankowski was supported in part by the NSF Grants CCF-0513636, and DMS-0503742, the NIH Grant R01 GM068959-01, and AFOSR Grant 073071.

REFERENCES

[1] S. Aviran, P. Siegel, and J. Wolf, Optimal Parsing Trees for Run-Length Coding of Biased Data, *IEEE Int. Symposium on Information Theory*, 1495-1499, Seattle, 2006.

[2] E. Bender, and F. Kochman, The distribution of subword counts is usually normal. *European Journal of Combinatorics*, **14**, 265–275, 1993.

[3] A. Dembo and I. Kontoyiannis. Source coding, large deviations, and approximate pattern matching. *IEEE Trans. Inform. Theory*, 48, 1590–1615, 2002.

[4] E. Drinea and M. Mitzenmacher, On Lower Bounds for the Capacity of Deletion Channels, *IEEE Transactions on Information Theory*, 52, 4648-4657, 2006.

[5] J. Fan, T. L. Poo, and B. Marcus, Constraint Gain, *Trans. Information Theory*, 50, 1989-1999, 2004.

[6] P. Flajolet, W. Szpankowski, and B. Vallée, Hidden Word Statistics, *Journal of the ACM*, 53, 1-37, 2006.

[7] L. Guibas and A. M. Odlyzko. Periods in strings. *J. Combinatorial Theory*, 30:19–43, 1981.

[8] P. Jacquet and W. Szpankowski, Autocorrelation on words and its applications. A analysis of suffix trees by string-ruler approach, *J. Combinatorial Theory*, Ser. A. 66, 237-269, 1994.

[9] P. Jacquet and W. Szpankowski, On (d,k) Sequences Not Containing a Given Word, *IEEE Intl. Symposium on Information Theory*, 1486-1489, Seattle, 2006.

[10] V. Kolesnik and V. Krachkovsky, Generating Functions and Lower Bounds on Rates for Limited Error-Correcting Codes, *Trans. Information Theory*, 37, 778-788, 1991.

[11] M. Lothaire, (Editor) *Applied Combinatorics on Words*, Cambridge University Press, 2005.

[12] B. Marcus, R. Roth and P. Siegel, Constrained Systems and Coding for Recording Channels, Chap. 20 in *Handbook of Coding Theory* (eds. V. S. Pless and W. C. Huffman), Elsevier Science, 1998.

[13] B. Moision, A. Orlitsky and P. Siegel, On codes that avoid specific differences, *Trans. Information Theory*, 47, 433-442, 2001.

[14] P. Nicodème, B. Salvy, and P. Flajolet, Motif Statistics, *European Symposium on Algorithms*, *Lecture Notes in Computer Science*, No. 1643, 194–211, 1999.

[15] M. Régnier and W. Szpankowski, On the Approximate Pattern Occurrences in a Text, *Proc. Compression and Complexity of SEQUENCE'97*, IEEE Computer Society, 253–264, Positano, 1997.

[16] M. Régnier and W. Szpankowski. On pattern frequency occurrences in a Markovian sequence. *Algorithmica*, 22:631–649, 1998.

[17] G. Reinert, S. Schbath, and M. Waterman, Probabilistic and statistical properties of words: an overview. *J. Comput. Biol.*, 7, 1-46, 2000.

[18] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, John Wiley & Sons, New York, 2001.

[19] M. Waterman. *Introduction to Computational Biology*. Chapman & Hall, London, 1995.

[20] E. Zehavi and J. Wolf, On runlength codes, *Trans. Information Theory*, 34, 45-54, 1988.