ON THE APPROXIMATE PATTERN OCCURRENCES IN A TEXT*

December 14, 1996

Mireille Régnier†
INRIA
Rocquencourt
78153 Le Chesnay Cedex
France
Mireille.Regnier@inria.fr

Wojciech Szpankowski‡
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.
spa@cs.purdue.edu

Abstract

Consider a given pattern H and a random text T generated by a Bernoulli source. We study the frequency of approximate occurrences of the pattern H in a random text when overlapping copies of the approximate pattern are counted separately. We provide exact and asymptotic formulæ for mean, variance and probability of occurrence as well as asymptotic results including the central limit theorem and large deviations. Our approach is combinatorial: we first construct certain language expressions that characterize pattern occurrences which are translated into generating functions, and finally we use analytical methods to extract asymptotic behaviors of the pattern frequency. Applications of these results include molecular biology, source coding, synchronization, wireless communications, approximate pattern matching, games, and stock market analysis. These findings are of particular interest to information theory (e.g., second-order properties of the relative frequency), and molecular biology problems (e.g., finding patterns with unexpected high or low frequencies, and gene recognition).

Key Words: Approximate pattern occurrences, autocorrelation polynomials, combinatorics on words, languages, generating functions, matrix analysis, asymptotic analysis, large deviations.

*This research was supported by NATO Collaborative Grant CRG.950060. It was initiated at INRIA, Sophia Antipolis during the summer of 1996, and both authors are grateful to project MISTRAL for hospitality and support.

†This work was additionally supported by the ESPRIT III Program No. 7141 ALCOM II.

‡This research was additionally supported by NSF Grants NCR-9206315 and NCR-9415491.

1 Introduction

Repeated patterns and related phenomena in words (sequences, strings) are known to play a central role in many facets of computer science, telecommunications, and molecular biology. One of the most fundamental questions arising in such studies is the frequency of pattern occurrences in another string known as text. For applications even more important is to know how many times a given pattern *approximately* occurs in a (random) text. By approximate occurrence we mean that there exists a substring of the text within given distance from the (given) pattern. The definition of the distance is irrelevant in this paper. This problem is also more challenging that the exact pattern occurrence. Applications include wireless communications, approximate pattern matching (cf. [15]), molecular biology (cf. [29]), games, code synchronization, (cf. [9, 10, 11]), source coding (cf. [4], stock market analysis, and so forth.

We study the problem in a probabilistic framework in which the text is generated randomly according to the so called *Bernoulli model* in which every symbol of a finite alphabet $\mathcal{S}$ is created independently of the other symbols with different probabilities of symbol generations (if all the probabilities are the same, then the model is called *symmetric Bernoulli* model). Our approach to this problem is combinatorial: We construct certain languages that characterize approximate pattern occurrences in a text which are further translated into generating functions. This falls under the methodology of "combinatorics on words" (cf. [3, 10, 11, 18])

Pattern occurrences in a random string is a classical problem (cf. [6]). Several authors also contributed to this problem, however, the most important recent contributions belong to Guibas and Odlyzko, who in a series of papers (cf. [9, 10, 11]) laid the foundations for the *exact* pattern occurrence in the symmetric Bernoulli model. In particular, the authors of [11] computed the moment generating function for the number of strings of length n that do *not* contain any of the given set of patterns. Certainly, this suffices to estimate the probability of at least one pattern occurrence in a random string generated by the symmetric Bernoulli model. Fudos *et al.* [8] computed the probability of exactly r occurrences of a pattern in a random text in the *asymmetric* Bernoulli model, just directly extending the results of Guibas and Odlyzko. This was recently further generalized to Markovian model by us (cf. [24]) where "combinatorics on words" approach was used. In [24] we deal only with a *single* pattern while in this paper we consider a set of patterns or approximate pattern occurrences. The Markovian model was also tackled by Li [17], Chrysaphinou and Papastavridis [2] who extended the Guibas and Odlyzko result of no pattern occurrence to

Markovian texts. Recently, Prum *et al.* [23] (see also [26]) obtained the limiting distribution for the number of pattern occurrences in the Markovian model.

In this paper, we provide a complete characterization of the frequency of approximate pattern occurrences in a random text generated according to the Bernoulli model using a combinatorial approach that might be of interest to other problems on words. Let $O_n(\mathcal{H})$ denote the number of approximate occurrences of a given pattern H in a random text when *overlapping* approximate copies of the pattern are counted separately. In the above $\mathcal{H}$ is a set of all strings of length m which are within given distance from H. We compute exactly the generating function of O_n (cf. Theorem 2.1) which further provides the mean EO_n and the variance Var O_n (cf. Theorem 2.2). Evaluation of the variance is quite challenging since it depends on the internal structure of the patterns through the so called autocorrelation matrix introduced in this paper. In addition, we present several of asymptotic results concerning $\Pr\{O_n = r\}$ for different range of r. We consider $r = O(1)$, as well as the central limit and the large deviations range of r.

Our results should be of particular interest to information theory (e.g., relative frequency, code synchronization, source coding, etc.) and molecular biology. Two problems of molecular biology can benefit from these results. Namely: finding patterns with unexpected (high or low) frequencies (the so called contrast words) [29], and recognizing genes by statistical properties [29]. Statistical methods have been successfully used from the early 80's to extract information from sequences of DNA. In particular, identifying deviant short motifs, the frequency of which is either too high or too low, might point out unknown biological information (cf. [29] and others for the analysis of functions of contrast words in DNA texts). From this perspective, our results give estimates for the statistical significance of deviations of word occurrences from the expected values and allow a biologist to build a dictionary of contrast words in genetic texts.

One can also use these results to recognize statistical properties of various other information sources such as images, text, etc. In information theory, the *relative frequency* defined as $\Delta_n(\mathcal{H}) = O_n(\mathcal{H})/(n-m+1)$, where m is the length of the pattern, is often used to estimate the statistics of the information source. The relative frequency was mostly studied for *exact* pattern occurrence, while in this paper we extend it to approximate occurrence. Such an extension is relevant to some recent applications such as lossy extension of the Lempel-Ziv scheme (cf. [19, 20, 30]) and lossy extension of the shortest common superstring problem (cf. [7, 31]). It is well known [4, 21] that $\Delta_n(H)$ for the exact pattern occurrence converges almost surely to the probability $P(H)$ of the pattern H. Of course, the same holds for the approximate pattern occurrence if one replace $P(H)$ by $P(\mathcal{H})$. Recently, Marton and Shields [21] proved that $\Delta_n(H)$ for the exact pattern occurrence converges exponentially fast to $P(H)$ for sources satisfying the so called blow-up property (e.g., Markov sources, hidden Markov, etc). Our results extends Marton and Shields results to approximate pattern occurrences (for the Bernoulli model but our results from [24] suggest that extension to Markovian model is possible). Such a rate of convergence is needed in some applications (cf. [20]).

This paper is organized as follows. In the next section we present our main results and their consequences. The proofs are delayed until the last section. Our derivation in Section 3.1 use a combinatorial approach of languages. In Section 3.2 we translate language relationships into associated generating functions, and finally we use analytical tools in Section 3.3 to derive asymptotic results.

2 Main Results

Let us consider two strings, a *pattern* string $H = h_1h_2 \ldots h_m$ and a *text* string $T = t_1t_2 \ldots t_n$ of respective lengths equal to m and n over an alphabet $\mathcal{S}$ of size V. We shall write $\mathcal{S} = \{1, 2, \ldots, V\}$ to simplify the presentation. Throughout, we assume that the pattern string is *fixed* and given, while the text string is random. More precisely, the text string T is a realization of an independently, identically distributed sequence of random variables (i.i.d.), such that a symbol $s \in \mathcal{S}$ occurs with probability $P(s)$. This defines the so called *Bernoulli model.* We shall write $P(H[i,j]) = \Pr\{T[i+k, j+k] = H[i,j]\}$ for the probability of the substring $H[i,j] = h_i \ldots h_j$ occurring in the random text $T[i+k, j+k]$ between symbols $i+k$ and $j+k$ for any k. In particular, we denote $P(H) = P(H[1,m])$.

Our goal is to estimate the frequency of *overlapping approximate* pattern occurrences in the text generated by a Bernoulli source. More precisely, let $d(H,F)$ be a distance between patterns H and F (which are assumed to be of equal length). The distance $d(\cdot,\cdot)$ can be any distance such as the Hamming distance, the edit distance, etc. For the given pattern H, we define its D-neighbourhood $\mathcal{H} = \{H_1, \ldots, H_M\}$ such that for any $1 \leq i \leq M$ the following holds $d(H, H_i) \leq D$ or $d(H, H_i) = D$ for fixed $D > 0$. (In fact, our results hold when $\mathcal{H}$ is a set of *any* given patterns $H_1, \ldots, H_M$ such that none contains another as a substring, but in this paper we concentrate on the approximate pattern occurrence case.) By an approximate pattern occurrence we mean that there exists $1 \leq j \leq n$ such that $d(T[j, j+m-1], H) \leq D$, or in other words, there exists $H_i \in \mathcal{H}$ such that $T[j, j+m-1] = H_i[1,m]$ for some $1 \leq j \leq n$.

We find it convenient and useful to express our findings in terms of languages. A language $\mathcal{L}$ is a collection of words satisfying a certain property. We associate with every language $\mathcal{L}$ a generating function defined as below:

Definition 1 *For any language $\mathcal{L}$ we define its generating function $L(z)$ as*

$$L(z) = \sum_{w \in \mathcal{L}} P(w) z^{|w|} \tag{1}$$

where $P(w)$ is the probability of the word w, $|w|$ is the length of w, and we adopt the usual convention that $P(\epsilon) = 1$.

It turns out that several properties of pattern occurrences depend on the so called *correlation polynomial* that is defined next.

Definition 2 *Given two strings H and F of lengths $|H|$ and $|F|$, let HF be the set of positive integers such that for any $k \in HF$ the last k symbols of H are equal to the first k symbols of F, that is, the suffix of length k of H is equal to the prefix of the same length of F. Then the* **correlation polynomial** *$A_{HF}(z)$ is defined as:*

$$A_{HF}(z) = \sum_{k \in HF} P(H[k+1, |H|]) z^{|H|-k} \tag{2}$$

In particular, the **autocorrelation polynomial** *of H becomes*

$$A_{HH}(z) = \sum_{k \in HH} P(H[k+1, |H|]) z^{|H|-k} \quad . \tag{3}$$

In addition, we define the **autocorrelation matrix** *of $\mathcal{H}$ as $\mathbb{A}(z) = \{A_{H_i H_j}\}_{i,j=1,M}$.*

In the sequel, we denote by $O_n(\mathcal{H})$ (or simply by O_n) a random variable representing the number of approximate occurrences of H in T. Let $\mathcal{T}_r$ be a language of words that contains exactly r approximate occurrences of H (or more generally: r occurrences of patterns from an arbitrary set $\mathcal{H}$). We denote by $T^{(r)}(z)$ its generating function which becomes:

$$T^{(r)}(z) = \sum_{n \geq 0} \Pr\{O_n(\mathcal{H}) = r\} z^n \tag{4}$$

for $|z| \leq 1$. In addition, we introduce a bivariate generating function as follows:

$$T(z,u) = \sum_{r=1}^{\infty} T^{(r)}(z) u^r = \sum_{r=1}^{\infty} \sum_{n=0}^{\infty} \Pr\{O_n(\mathcal{H}) = r\} z^n u^r \quad . \tag{5}$$

We shall work with matrices and vectors, so we adopt the following convention. Bold upper-case letters are reserved for vectors which are assumed to be column vectors; e.g., $\mathbf{U}^t(z) = (U_1(z), \ldots, U_M(z))$ where $U_i(z)$ is the generating function of a language $\mathcal{U}_{H_i}$ (see next section), and the upper index "t" denotes transpose. We shall use blackboard bold letters for matrices (e.g., $\mathbb{A}(z) = \{A_{H_iH_j}(z)\}_{i,j=1,M}$). In particular, we write $\mathbb{I}$ for the identity matrix, and $\mathbf{1} = (1,\ldots,1)^t$ for the unit vector. Finally, we recall that $(\mathbb{I}-\mathbb{M})^{-1} = \sum_{r\geq 0}\mathbb{M}^k$ provided the inverse matrix exists (i.e., $\det(\mathbb{I}-\mathbb{M}) \neq 0$ or $\| \mathbb{M}(z) \| < 1$ where $\| \cdot \|$ is any matrix norm).

Now, we are ready to summarize our main findings in the form of two following theorems. The first theorem presents exact formulæ for the generating functions $T^{(r)}(z)$ and $T(z,u)$, and can be used to compute exactly parameters related to the pattern occurrence $O_n(\mathcal{H})$. In the second theorem, we provide asymptotic results for the probability $\Pr\{O_n = r\}$ for various ranges of r. All proofs are presented in the next section. The method of derivation extends the method presented in [24]. The proof of Theorem 2.1 is presented in Section 3.2 while the proof of Theorem 2.2 can be found in Section 3.3.

Theorem 2.1 *Let H be a given pattern of size m, $\mathcal{H}$ be the D-neighbourhood of H, and T be a random text of length n generated according to the Bernoulli model. The generating functions $T^{(r)}(z)$ and $T(z,u)$ can be computed as follows:*

$$\begin{aligned} T^{(r)}(z) &= \mathbf{R}^t(z)\mathbb{M}(z)^{r-1}\mathbf{U}(z) && (6)\\ &= z^m\mathbf{H}^t(\mathbb{D}(z)+(z-1)\mathbb{I})^{r-1}[\mathbb{D}(z)]^{-(r+1)}\mathbf{1}\ , && (7)\\ T(z,u) &= \mathbf{R}^t(z)u(\mathbb{I}-u\mathbb{M}(z))^{-1}\mathbf{U}(z)\ , && (8)\end{aligned}$$

where

$$\begin{aligned} (\mathbb{I}-\mathbb{M}(z))^{-1} &= \mathbb{A}(z)+\frac{z^m}{1-z}\mathbf{1}\cdot\mathbf{H}^t\ , && (9)\\ \mathbb{D}(z) &= (1-z)(\mathbb{I}-\mathbb{M}(z))^{-1} = (1-z)\mathbb{A}(z)+z^m\mathbf{1}\cdot\mathbf{H}^t\ , && (10)\\ \mathbf{U}(z) &= \frac{1}{1-z}(\mathbb{I}-\mathbb{M}(z))\cdot\mathbf{1}\ , && (11)\\ \mathbf{R}^t(z) &= \frac{z^m}{1-z}\mathbf{H}^t\cdot(\mathbb{I}-\mathbb{M}(z))\ . && (12)\end{aligned}$$

In the above, $\mathbf{H} = (P(H_1),\ldots,P(H_M))^t$, and $\mathbb{A}(z) = \{A_{H_i,H_j}(z)\}_{i,j=1.M}$ is the matrix of the correlation polynomials of patterns from the set $\mathcal{H}$.

The above theorem is a key to the next asymptotic results. These results are derived in the next section using analytical tools.

Theorem 2.2 *Let the hypotheses of Theorem 2.1 be fulfilled, and in addition $nP(\mathcal{H}) \to \infty$ where $P(\mathcal{H}) = \sum_{H_i\in\mathcal{H}} P(H_i) = \mathbf{H}^t\cdot\mathbf{1}$.*

(i) MOMENTS. *We obtain*

$$
\begin{aligned}
EO_n(\mathcal{H}) &= (n-m+1)P(\mathcal{H}) \ , && (13)\\
\mathrm{Var}\ O_n(\mathcal{H}) &= (n-m+1)\Big(P(\mathcal{H}) + P^2(\mathcal{H}) + 2mP^2(\mathcal{H}) + 2\mathbf{H}^t(\mathbb{A}(1)-\mathbb{I})\mathbf{1}\Big) && (14)\\
&+ \ m(m-1)P^2(\mathcal{H}) - 2\mathbf{H}^t\dot{\mathbb{A}}(1)\cdot\mathbf{1} \ , && (15)
\end{aligned}
$$

where $\dot{\mathbb{A}}(1)$ *denotes the derivative of the matrix* $\mathbb{A}(z)$ *at* $z=1$.

(ii) DISTRIBUTION: CASE $r = O(1)$. *Let* $\rho_{\mathcal{H}}$ *be the smallest root of* $\det \mathbb{D}(z) = 0$ *outside the unit circle* $|z|<1$, *and let* $\rho > \rho_{\mathcal{H}}$. *Then:*

$$
\begin{aligned}
\Pr\{O_n(\mathcal{H}) = r\} &= (-1)^{r+1}\frac{a_{r+1}}{r!}(n)_r\rho_{\mathcal{H}}^{-(n-m+r+1)} && (16)\\
&+ \sum_{j=1}^{r}(-1)^j a_j\binom{n}{j-1}\rho_{\mathcal{H}}^{-(n+j)} + O(\rho^{-n}) \ ,
\end{aligned}
$$

where $(n)_r = n(n-1)\cdots(n-r+1)$ *and*

$$
a_{r+1} = \frac{\mathbf{H}^t\left(\mathbb{D}(\rho_{\mathcal{H}}) + (\rho_{\mathcal{H}}-1)\mathbb{I}\right)^{r-1}\left(\mathbb{D}^*(\rho_{\mathcal{H}})\right)^{r+1}\cdot\mathbf{1}}{\left(\det'\mathbb{D}(\rho_{\mathcal{H}})\right)^{r+1}} \ , \tag{17}
$$

where $\mathbb{D}^*(z)$ *is the adjoint matrix of* $\mathbb{D}(z)$, *and* $\det'\mathbb{D}(\rho_{\mathcal{H}})$ *denotes the derivative of* $\det\mathbb{D}(z)$ *at* $z=\rho_{\mathcal{H}}$. *The remaining coefficients* a_j *can be computed according to the following formula:*

$$
a_j = \frac{1}{(r+1-j)!}\lim_{z\to\rho_{\mathcal{H}}}\frac{d^{r+1-j}}{dz^{r+1-j}}\left(T^{(r)}(z)(z-\rho_{\mathcal{H}})^{r+1}\right) \tag{18}
$$

with $j = 1,2,\ldots r$.

(iii) DISTRIBUTION: CASE $r = EO_n + x\sqrt{\mathrm{Var}\ O_n}$. *Let* $x = O(1)$. *Then:*

$$
\Pr\{O_n(\mathcal{H}) = r\} = \frac{1}{\sqrt{2\pi\mathrm{Var}\ O_n}}e^{-\frac{1}{2}x^2}\left(1 + O\left(\frac{1}{\sqrt{n}}\right)\right) \ , \tag{19}
$$

(iv) DISTRIBUTION: CASE $r = (1+\delta)EO_n$ *with* $\delta \neq 0$. *Define* $\tau(t)$ *to be the root of*

$$
\det(\mathbb{I} - e^t\mathbb{M}(e^\tau)) = 0 \ , \tag{20}
$$

and ω_a *to be the root of*

$$
\tau'(\omega_a) = a \tag{21}
$$

where $a = 1+\delta$. *Then:*

$$
\Pr\{O_n(\mathcal{H}) = r\} = \frac{1}{\omega_a\sqrt{2\pi\mathrm{Var}\ O_n}}e^{-((n-m+1)I(a)}\left(1 + O\left(\frac{1}{n}\right)\right) \tag{22}
$$

where $I(a) = a\omega_a - \tau(\omega_a)$.

As mentioned before, the above results have abundance of applications in information theory and molecular biology. For example, they can be used to estimate the *relative frequency* defined as

$$\Delta_n(\mathcal{H}) = \frac{O_n(\mathcal{H})}{n-m+1} \ .$$

Relative frequency appears in the definition of types and typical types (cf. [4]), and is often used to estimate information source statistics. The reader is referred to [24] for more details.

3 Analysis

The key element of our analysis is a derivation of the generating function $T(z,u)$ presented in Theorem 2.1. The first part of below derivation is quite general. It is based on constructing some special languages and finding relationships among them. Later in Section 3.2 we translate them into generating functions.

3.1 Combinatorial Relationships Between Certain Languages

A collection of words sharing a given property is usually called a *language*. This section is devoted to present some combinatorial relationships between certain languages that are crucial to derive our results. In this section we do not make any probabilistic assumptions.

We start with some definitions:

Definition 3 *Let $\mathcal{H}$ be a set of patterns $\mathcal{H} = \{H_i\}_{i\in\{1,\dots,M\}}$:*

(i) Let $\mathcal{T}$ be a language of words containing at least one occurrence from $\mathcal{H}$, and for any integer r, let $\mathcal{T}_r$ be the language of words containing exactly r occurrences from $\mathcal{H}$.

(ii) For $i, j \in \{1,\dots,M\}$, we define for $r \geq 1$ the language $\mathcal{M}_{i,j}^{(r-1)}$ as

$$\mathcal{M}_{i,j}^{(r-1)} = \{w:\ H_i w \in \mathcal{T}_r \text{ and } H_j \text{ occurs at the right end of } w\} \ . \tag{23}$$

We write $\mathcal{M}_{i,j} = \mathcal{M}_{i,j}^{(1)}$.

(iii) The language $\mathcal{R}_i$ is the set of words containing only one occurrence of H_i, located at the right end. We also define $\mathcal{U}_i$ as

$$\mathcal{U}_i = \{u:\ H_i u \in \mathcal{T}_1\} \ . \tag{24}$$

In other words a word $u \in \mathcal{U}_i$ if the only occurrence from $\mathcal{H}$ in $H_i u$ is H_i.

(iv) *Finally, we define the sets $\mathcal{A}_{i,j}$ associated with the correlation of H_i and H_j, for $i, j \in \{1, \ldots, M\}$, that is:*

$$\mathcal{A}_{i,j} = \{H_j[k+1, m] : \quad k \in H_i H_j\} \ ,$$

where $H_i H_j$ is the autocorrelation sequence introduced in Definition 2.

Remark:

(i) When H_i dees not overlap on its right end with H_j, the set $\mathcal{A}_{i,j}$ is empty and $A_{i,j}(z) = 0$.

(ii) It is worth noting that ϵ belongs to $\mathcal{A}_{i,j}$ if and only if $i = j$. In other words, H_i coincides with H_j on its first character if and only if $i = j$. Hence, the constant term in $A_{i,j}(z)$ is 0 when $i \neq j$ and 1 when $i = j$.

We now can describe the languages $\mathcal{T}$ and $\mathcal{T}_r$ in terms of the languages just introduced. This will further lead to a simple formula for the generating function of $O_n(\mathcal{H})$. We prove below the following:

Theorem 3.1 *The language $\mathcal{T}_r$ can be represented for any $r \geq 1$ as follows:*

$$\mathcal{T}_r = \sum_{i,j \in \{1,\ldots,M\}} \mathcal{R}_i \mathcal{M}_{i,j}^{(r-1)} \mathcal{U}_j \ . \tag{25}$$

The language $\mathcal{T}$ satisfies the fundamental equation:

$$\mathcal{T} = \sum_{r \geq 1} \sum_{i,j \in \{1,\ldots,M\}} \mathcal{R}_i \mathcal{M}_{i,j}^{(r-1)} \mathcal{U}_j \ . \tag{26}$$

Proof: We first prove (25) and obtain our decomposition of $\mathcal{T}_r$ as follows. Let the first occurrence of $\mathcal{H}$ in a word belonging to $\mathcal{T}_r$ be, say, H_i; it determines a prefix p of this word that is in $\mathcal{R}_i$. Then, one concatenates a non-empty word w that produces the second occurrence of $\mathcal{H}$, say H_k. Hence, w is in some $\mathcal{M}_{i,k}$. This process is repeated $r-1$ times and we may assume the last occurrence is H_j; e.g. the word concatenated to the right of p is in $\mathcal{M}_{i,j}^{(r-1)}$. Finally, one adds after the last $\mathcal{H}$ occurrence a suffix u that does not produce a new occurrence of $\mathcal{H}$. Equivalently, u is in $\mathcal{U}_j$, and w is a proper subword of $H_j u$. Finally, a word belongs to $\mathcal{T}$ if it belongs to $\mathcal{T}_r$ for some $r \geq 1$. ∎

We now prove the following result that summarizes relationships between the languages introduced in Definition 3. Below, we use the following notation: We define $\oplus$, $\ominus$ and $\cdot$ as disjoint union, subtraction and concatenation of languages. For sake of clarity, we assimilate below a singleton $\{w\}$ to its unique element w.

Theorem 3.2 *The languages $\mathcal{M}_{i,j}$, $\mathcal{U}_i$ and $\mathcal{R}_j$ satisfy, for $i,j \in \{1,\ldots,M\}$:*

$$\bigcup_{k\geq 1} \mathcal{M}_{i,j}^{(k)} = \mathcal{W}\cdot H_j \oplus \mathcal{A}_{i,j} \ominus \{\epsilon\} \ , \tag{27}$$

$$\mathcal{U}_i \cdot \mathcal{S} = \bigcup_j \mathcal{M}_{i,j} \oplus \mathcal{U}_i \ominus \{\epsilon\} \ , \tag{28}$$

$$\mathcal{S}\cdot\mathcal{R}_j - (\mathcal{R}_j - H_j) = \bigcup_i H_i \mathcal{M}_{i,j} \ , \tag{29}$$

where $\mathcal{W}$ is the set of all words, $\mathcal{S}$ is the alphabet set, ϵ is the empty word.

Proof: All the above relations are proved in a similar fashion. We first deal with (27). Let w be in $\mathcal{W}\cdot\mathcal{H}$ and $k+1$ be the number of subwords of $H_i \cdot w$ that are in $\mathcal{H}$. Certainly, this number is greater than or equal to 2 and the last occurrence, say H_j, is on the right of $H_i w$: This implies that $w \in \mathcal{M}_{i,j}^{(k)}$. Furthermore, a word w in $\bigcup_{k\geq 1}\mathcal{M}_{i,j}^{(k)}$ is not in $\mathcal{W}\cdot H_j$ iff its size $|w|$ is smaller than $|H_j|$. Then, the right $\mathcal{H}$ occurrence in $H_i w$ overlaps with H_i, which means that w is in $\mathcal{A}_{i,j}$. Reciprocally, any word in $\mathcal{A}_{i,j}$ qualifies, but the empty word, when it belongs to it. Although ϵ is not in $\mathcal{A}_{i,j}$ when $i \neq j$, our set expression remains correct; e.g. $\mathcal{A}_{i,j} - \{\epsilon\} = \mathcal{A}_{i,j}$ when $\epsilon \notin \mathcal{A}_{i,j}$.

Let us turn now to (28). When one adds a character s right after a word u from $\mathcal{U}_i$, two cases may occur. Either $H_i us$ still does not contain a second occurrence of $\mathcal{H}$, which means that us is a non-empty word of $\mathcal{U}_i$. Or a new element of $\mathcal{H}$ appears, say H_j, clearly at the right end. Then, us is in $\mathcal{M}_{i,j}$ and we get the left inclusion. Furthermore, any non-empty word of $\mathcal{U}_i - \{\epsilon\}$ is in $\mathcal{U}_i\cdot\mathcal{S}$, and a strict prefix of a word w in $\mathcal{M}_{i,j}$ cannot contain any $\mathcal{H}$-occurrence; hence, this prefix is in $\mathcal{U}_i$ and w is in $\mathcal{U}_i\cdot\mathcal{S}$.

We now prove (29). Let $x = sw$ be a word in $H_i \cdot \mathcal{M}_{i,j}$ where s is a symbol from $\mathcal{S}$. As x contains exactly two occurrences of $\mathcal{H}$, H_i located at its left end, and H_j located at its right end, w is in $\mathcal{R}_j$ and x is in $\mathcal{S}\cdot\mathcal{R}_j - \mathcal{R}_j$. Reciprocally, if a word swH_j from $\mathcal{S}\cdot\mathcal{R}_j$ is not in $\mathcal{R}_j$, then swH_j contains a second $\mathcal{H}$ occurrence, say H_i. As $w\mathcal{H}$ is in $\mathcal{R}_j$, the only possible position is on the left end, and then x is in $H_i\cdot\mathcal{M}_{i,j}$. We now rewrite:

$$\mathcal{S}\cdot\mathcal{R}_j - \mathcal{R}_j = \mathcal{S}\cdot\mathcal{R}_j - (\mathcal{R}_j \cap \mathcal{S}\cdot\mathcal{R}_j) = \mathcal{S}\cdot\mathcal{R}_j - (\mathcal{R}_j - H_j)$$

which completes the proof. ∎

3.2 Associated Generating Functions

In this section, we translate the language relationships into generating functions. We need a few rules associated with two operations on languages. Namely: the disjoint union $\oplus$ and concatenation $\cdot$ become the sum operation $+$ and the multiplication operation on generating functions. Namely, the union language $\mathcal{L} = \mathcal{L}_1 \oplus \mathcal{L}_2$ is transferred into the generating function $L(z) = L_1(z) + L_2(z)$, whenever $\mathcal{L}_1 \cap \mathcal{L}_2 = \emptyset$. The generating function of $\mathcal{L} = \mathcal{L}_1 \cdot \mathcal{L}_2$ is $L(z) = L_1(z)L_2(z)$ for the Bernoulli model (cf. [24] for extension to Markov model).

Lemma 3.1 *The generating functions associated with languages $\mathcal{M}_{i,j}$,$\mathcal{U}_i$ and $\mathcal{R}_i$ satisfy the following matrix equations:*

$$(\mathbb{I} - \mathbb{M}(z))^{-1} = \mathbb{A}(z) + \frac{z^m}{1-z} \cdot \mathbf{1} \cdot \mathbf{H}^t \ , \tag{30}$$

$$(\mathbb{I} - \mathbb{M}(z))^{-1}\mathbf{U}(z) = \frac{1}{1-z}\mathbf{1} \ , \tag{31}$$

$$\mathbf{R}^t(z)(\mathbb{I} - \mathbb{M}(z))^{-1} = \frac{z^m}{1-z}\mathbf{H}^t \ , \tag{32}$$

that are defined for $|z| < 1$.

Proof: We first prove (31). We rewrite the language relationship (28) from Theorem 3.2 as $\mathcal{U}_i \cdot \mathcal{S} - (\mathcal{U}_i - \epsilon) = \cup_{j=1}^{M} \mathcal{M}_{i,j}$ for any $i \in \{1, \ldots, M\}$. The left side of this equation yields $U_i(z) \cdot (z-1) + 1$. The right-hand side is the sum of the terms of the i-th row of matrix $\mathbb{M}(z)$, or, equivalently, the i-th row of $\mathbb{M}(z)$. As the result holds for any i, we get the equation (28) between two column vectors.

We now turn our attention to (32). The left-hand side of (29), i.e., $\mathcal{S} \cdot \mathcal{R}_j - (\mathcal{R}_j - H_j)$, translates into $(z-1)R_j(z) + P(H_j)z^m$, while $\sum_{i=1}^{M} H_i \cdot \mathcal{M}_{i,j}$ translates into $\sum_{i=1}^{M} P(H_i)z^m M_{i,j}(z)$. These are the j-th elements of the row vectors $(z-1)\mathbf{R}^t(z) + z^m\mathbf{H}^t$ and $z^m\mathbf{H}^t \cdot \mathbb{M}(z)$. Grouping the results for all j yields the equation (32) between row vectors.

Finally, we deal with (30). In the left-hand side of (27) all languages $\mathcal{M}_{i,j}^{(k)}$ are disjoint and the generating function of $\mathcal{M}_{i,j}^{(k)}$ is the (i,j)-element of matrix $\mathbb{M}^k(z)$. As the elements of $\mathbb{M}(z)$ are probability generating functions, one has $||\mathbb{M}(z)|| < 1$ for $|z| < 1$; hence the series $\sum_{k=0}^{\infty} \mathbb{M}^k(z)$ converges, and $(\mathbb{I} - \mathbb{M}(z))^{-1}$ is well defined for $|z| < 1$. Moreover, $\sum_{k=0}^{\infty} \mathbb{M}^k(z)$ is $\mathbb{M}(z) \cdot (\mathbb{I} - \mathbb{M}(z))^{-1}$. Now, the right-hand side, $\mathcal{A}_{i,j} - \{\epsilon\}$ translates into $\mathbb{A} - \mathbb{I}$. As $\mathcal{W} \cdot H_j$ translates into $\frac{1}{1-z} \cdot z^m$ (cf. [24]), the associated matrix is $\frac{z^m}{1-z}\mathbf{1} \cdot \mathbf{H}^t$.

Finally, (8) in Theorem 2.1 is a direct consequence of (26) using Theorem 3.2 and Lemma 3.1. ∎

3.3 Moments and Limiting Distribution

In this final subsection, we derive the first two moments of O_n as well as asymptotics for $\Pr\{O_n = r\}$ for different ranges of r, that is, we prove Theorem 2.2. Actually, we should mention that using general results from renewal theory one immediately guesses that the limiting distribution must be normal for $r = EO_n + O(\sqrt{n})$. However, here the challenge is to estimate precisely the variance. Our approach offers an easy, uniform, and precise derivation all of moments, including the variance, as well as local limit distributions (including the convergence rate) for the central and large deviations regimes.

A. Moments

First of all, from Theorem 2.1 we shall conclude that

$$T'(z,1) = \frac{z^m \mathbf{H}^t \cdot \mathbf{1}}{(1-z)^2} = \frac{\sum_{H_i \in \mathcal{H}} P(H_i) z^m}{(1-z)^2} , \tag{33}$$

$$\begin{aligned} T''(z,1) &= \frac{2z^m (\mathbf{H}^t \cdot \mathbf{1})^2 z^m}{(1-z)^3} + \frac{2z^m \mathbf{H}^t (\mathbb{A}(z) - \mathbb{I})\mathbf{1}}{(1-z)^2} \\ &= \frac{2\left(\sum_{H_i \in \mathcal{H}} P(H_i) z^m\right)^2}{(1-z)^3} + \frac{2z^m \left(\sum_{i,j} P(H_i) A_{i,j}(z) - \sum_{H_i \in \mathcal{H}} P(H_i)\right)}{(1-z)^2}. \end{aligned} \tag{34}$$

Indeed, we observe that

$$T'(z,u) = \mathbf{R}^t(z)(\mathbb{I} - u\mathbb{M}(z))^{-2}\mathbf{U}(z) \ ,$$

and then by (9)–(12) we directly prove

$$T'(z,1) = \mathbf{R}^t(z) \cdot (\mathbb{I} - \mathbb{M}(z))^{-2}\mathbf{U}(z) = \frac{z^m}{(1-z)^2}\mathbf{H}^t \ \cdot \mathbf{1}$$

which leads to (33).

To establish (34) we need a little more algebra. First, we derive from (8)

$$T''(z,u) = 2\mathbf{R}^t(z)\mathbb{M}(z)(\mathbb{I} - u\mathbb{M}(z))^{-3}\mathbf{U}(z)$$

which further yields from (10) – (12)

$$\begin{aligned} T''(z,1) &= 2\mathbf{R}(z)^t \mathbb{M}(z)(\mathbb{I} - \mathbb{M}(z))^{-3}\mathbf{U}(z) \\ &= \frac{2}{(1-z)^3}\mathbf{R}^t(z) \cdot \mathbb{M}(z)\mathbb{D}(z) \cdot \mathbb{D}(z)\mathbf{1} \\ &= \frac{2z^m}{(1-z)^3}\mathbf{H}^t \mathbb{D}(z)^{-1}(\mathbb{D}(z) + (z-1)\mathbb{I})\mathbb{D}(z) \cdot \mathbf{1} \\ &= \frac{2z^m}{(1-z)^3}\mathbf{H}^t(\mathbb{D}(z) + (z-1)\mathbb{I})\mathbf{1} \\ &= \frac{2z^m}{(1-z)^2}\mathbf{H}^t[(\mathbb{I} - \mathbb{M}(z))^{-1} - \mathbb{I}]\mathbf{1} \\ &= \frac{2z^m}{(1-z)^2}\mathbf{H}^t\left(\mathbb{A}(z) + \frac{z^m}{1-z}\mathbf{1} \cdot \mathbf{H}^t - \mathbb{I}\right)\mathbf{1} \ . \end{aligned}$$

In the above we often use $(\mathbb{I}-\mathbb{M})^{-1}=(1-z)^{-1}\mathbb{D}(z)$ (cf. (9)). Then (34) follows.

Now, we observe that both expressions admit as a numerator a function that is entire beyond the unit circle. This allows for a very simple computation of the expectation and variance, based on the following basic formula:

$$[z^n](1-z)^{-p}=\frac{\Gamma(n+p)}{\Gamma(p)\Gamma(n+1)} \tag{35}$$

To obtain EO_n we proceed as follows:

$$EO_n=[z^n]T'(z,1)=\sum_{H_i\in\mathcal{H}}P(H_i)[z^{n-m}](1-z)^2=(n-m+1)\sum_{H_i\in\mathcal{H}}P(H_i)\ .$$

Computation of the variance is a little more intricate. To simplify our computations, let

$$\begin{aligned}\Phi_1(z) &= 2(\mathbf{H}^t\cdot\mathbf{1})^2z^m\ ,\\ \Phi_2(z) &= 2\mathbf{H}^t(\mathbb{A}(z)-\mathbb{I})\mathbf{1}\ .\end{aligned}$$

Using Cauchy's theorem, we also observe that

$$\begin{aligned}[z^{n-m}]\Phi_1(z)(1-z)^{-3} &= \Phi_1(1)\frac{(n-m+2)(n-m+1)}{2}+\Phi_1'(1)(n-m+1)+\frac{1}{2}\Phi_1''(1)\ ,\\ [z^{n-m}]\Phi_2(z)(1-z)^{-2} &= \Phi_2(1)(n-m+1)-\Phi_2'(1)\ .\end{aligned}$$

Then, a simple algebra leads to the formula on the variance (cf. (15) of Theorem 2.2).

B. Asymptotic Results

We now establish Theorem 2.2, that is, we compute $\Pr\{O_n=r\}$ for different ranges of r. Our derivation is along the lines of our previous paper [24], hence we skip most of the details referring the reader to the above paper.

We start with $r=O(1)$, and turn our attention to formula (6) of Theorem 2.1, that is:

$$T^{(r)}(z)=z^m\mathbf{H}^t(\mathbb{D}(z)+(z-1)\mathbb{I})^{r-1}[\mathbb{D}(z)]^{-(r+1)}\mathbf{1}$$

where $\mathbb{D}(z)$ is given by (10). To establish an asymptotic expression for $\Pr\{O_n=r\}$ one needs to extract the coefficient at z^n of $T^{(r)}(z)$. By Hadamard's theorem (cf. [25]) we conclude that the asymptotics of the coefficients of $T^{(r)}(z)$ depend on the singularities of $T^{(r)}(z)$. In our case, the generating function is a rational function. Indeed, we first observe that (cf. [12])

$$[\mathbb{D}(z)]^{-1}=\frac{\mathbb{D}^*(z)}{\det\mathbb{D}(z)}$$

where $\mathbb{D}^*(z)$ is the adjoint matrix of $\mathbb{D}(z)$. Thus, all singularities of $T^{(r)}(z)$ are contained in the set of roots of $\det\mathbb{D}(z)$. But since every entry in $\mathbb{A}(z)$ is a polynomial, we conclude that

$\det \mathbb{D}(z)$ is a polynomial. Thus, there exists the smallest root $\rho_{\mathcal{H}}$ of $\det \mathbb{D}(z) = 0$ outside $|z| > 1$, and it is of multiplicity $r+1$. In particular, $\det \mathbb{D}(z) = (z-\rho_{\mathcal{H}}) \det' \mathbb{D}(\rho_{\mathcal{H}}) + O((z-\rho_{\mathcal{H}})^2)$. The rest of the derivation follows exactly our footsteps from [24], so we refer the reader to it for details.

Now, we deal with $r = EO_n + x\sqrt{\mathrm{Var}\ O_n}$ when $x = O(1)$ (the so called central limit regime). Let $\mu_n = EO_n(\mathcal{H})$ and $\sigma_n^2 = \mathrm{Var}\ O_n(\mathcal{H})$. Thus, we consider formula (8) on $T(z,u)$ for complex z (actually, we assume $z = e^\tau$ with $\tau = t\mu_n/\sigma_n \to 0$ for fixed complex t). To establish normality of $(O_n(\mathcal{H}) - \mu_n)/\sigma_n$, it suffices, according to Levy's theorem, to prove the following

$$\lim_{n\to\infty} e^{-t\mu_n/\sigma_n} T_n(e^{t/\sigma_n}) = e^{t^2/2} \tag{36}$$

for some complex t around zero. In the above, we write $T_n(u) = Eu^{O_n}$ (i.e., the probability generating function for O_n) for $u = e^{t/\sigma_n}$. The computations are standard and go as below. The equation

$$\det(\mathbb{I} - e^t \mathbb{M}(e^\tau)) = 0 \tag{37}$$

implicitly defines in some neighbourhood of $t = 0$ a unique C^∞ function $\tau(t)$, satisfying $\tau(0) = 0$. Then, an elementary application of the residue theorem leads for some $R > 1$ to

$$T_n(e^t) = C(t)e^{(n+1-m)\tau(t)} + O(R^{-n}) \tag{38}$$

$C(t)$ is a polynomial in t, and one has, uniformly in t, $\tau(t) = t\tau'(0) + \tau''(0)t^2/2 + O(t^3)$. From the cumulant formula, it appears that $EO_n(\mathcal{H}) = [t]\log T_n(t) \sim n\tau'(0)$ as well as $\mathrm{Var}\ O_n \sim n\tau''(0)$, where $[t^r]T(t)$ denotes the the coefficient of $T(t)$ at t^r.

After some algebra, this leads (cf. [1]) to

$$\begin{aligned} e^{-t\mu_n/\sigma_n} T_n(e^{t/\sigma_n}) &= \exp\left(\frac{t^2}{2} + O(nt^3/\sigma^3)\right) \\ &= e^{t^2/2}\left(1 + O(1/\sqrt{n})\right) \end{aligned}$$

which completes the proof of the result.

Finally, we consider a large deviations result, that is $r = (1+\delta)EO_n = aEO_n$ for $a > 1$. From (38) we conclude that

$$\lim_{n\to\infty} \frac{\log T_n(e^t)}{n} = \tau(t) \ .$$

Thus, directly from Gärtner-Ellis theorem [5] we prove that

$$\lim_{n\to\infty} \frac{\log \Pr\{O_n > na\}}{n} = -I(a) \ ,$$

where, after defining ω_a as a solution of $\tau'(t) = a$, we obtain

$$I(a) = a\omega_a - \tau(\omega_a) \ .$$

The detailed computations are exactly the same as in [24], thus left for the reader.

References

[1] E. Bender, Central and Local Limit Theorems Applied to Asymptotic Enumeration, *J. Combin. Theory, Ser. A*, 15, 91-111, 1973.

[2] C. Chrysaphinou, and S. Papastavridis, The Occurrence of Sequence of Patterns in Repeated Dependent Experiments, *Theory of Probability and Applications*, 167-173, 1990.

[3] M. Crochemore and W. Rytter, *Text Algorithms*, Oxford University Press, New York 1995.

[4] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York 1981.

[5] R. Ellis, Large Deviations for a General Class of Random Vectors, *Ann. Probab.*, 1-12, 1984.

[6] W. Feller, *An Introduction to Probability and its Applications*, Vol. 1, John Wiley & Sons, New York 1968.

[7] A. Frieze and W. Szpankowski, Greedy Algorithms for the Shortest Common Superstring That Are Asymptotically Optimal, *Proc. European Symposium on Algorithms*, Springer LNCS, No. 1136, 194-207, Barcelona 1996.

[8] I. Fudos, E. Pitoura and W. Szpankowski, On Pattern Occurrences in a Random Text, *Information Processing Letters*, 57, 307-312, 1996.

[9] L. Guibas and A. Odlyzko, Maximal Prefix-Synchronized Codes, *SIAM J. Appl. Math*, 35, 401-418, 1978.

[10] L. Guibas and A. Odlyzko, Periods in Strings, *J. Combin.Theory Ser. A*, 30, 19-43, 1981.

[11] L. Guibas and A. W. Odlyzko, String Overlaps, Pattern Matching, and Nontransitive Games, *J. Combin.Theory Ser. A*, 30, 183-208, 1981.

[12] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge 1985.

[13] H-K. Hwang, *Théorèmes Limites Pour les Structures Combinatoires et les Fonctions Arithmétiques*, Thèse de Doctorat de l'Ecole Polytechnique, 1994.

[14] P. Jacquet and W. Szpankowski, Autocorrelation on Words and Its Applications. Analysis of Suffix Trees by String-Ruler Approach, *J. Combin.Theory Ser. A*, 66, 237-269, 1994.

[15] P. Jokinen and E. Ukkonen, Two Algorithms for Approximate String Matching in Static Texts, *Proc. MFCS 91*, *Lecture Notes in Computer Science* 520, 240-248, Springer Verlag 1991.

[16] D.E. Knuth, *The Art of Computer Programming: Fundamental Algorithms*, vol. 1., Addison-Wesley, Reading 1973 .

[17] S. R. Li, A Martingale Approach to the Study of Occurrences of Sequence Patterns in Repeated Experiments, *Ann. Probab.*, 8, 1171-1176, 1980.

[18] Lothaire, M., *Combinatorics on Words*, Addison Wesley, Reading, Mass. 1982.

[19] T. Luczak and W. Szpankowski, A Lossy Data Compression Based on String Matching: Preliminary Analysis and Suboptimal Algorithms, *Proc. Combinatorial Pattern Matching*, Asilomar, LNCS 807, 102-112, Springer-Verlag, 1994.

[20] T. Luczak and W. Szpankowski, A Suboptimal Lossy Data Compression Based on Approximate Pattern Matching, *1996 International Symposium on Information Theory*, Whistler 1996; also Purdue University CSD-TR-94-072, 1994.

[21] K. Marton and P. Shields, The Positive-Divergence and Blowing-up Properties, *Israel J. Math*, 80, 331-348 (1994).

[22] A. Odlyzko, Asymptotic Enumeration, in *Handbook of Combinatorics*, Vol. II, (Eds. R. Graham, M. Götschel and L. Lovász), Elsevier Science, 1995.

[23] B. Prum, F. Rodolphe, and E. Turckheim, Finding Words with Unexpected Frequencies in Deoxyribonucleic Acid Sequence, *J.R. Stat. Soc. B*, 57, 205-220, 1995.

[24] M. Regnier and W. Szpankowski, A Last Word on Pattern Frequency Occurrence in a Markovian Sequence?, Purdue University, CSD-TR-96-043, 1996.

[25] R. Remmert, *Theory of Complex Functions*, Springer Verlag, New York 1991.

[26] S. Schbath, *Etude Asymptotique du Nombre d'Occurrences d'un mot dans une Chaîne de Markov et Application à la Recherche de Mots de Fréquence Exceptionnelle dans les Séquences d'ADN*, Thèse Université René Descartes Paris V, 1995.

[27] W. Szpankowski, Asymptotic Properties of Data Compression and Suffix Trees, *IEEE Trans. Information Theory*, 39, 1647-1659, 1993.

[28] W. Szpankowski, A Generalized Suffix Tree and Its (Un)Expected Asymptotic Behaviors, *SIAM J. Computing*, 22, 1176-1198 (1993).

[29] M. Waterman, *Introduction to Computational Biology*, Chapman & Hall, New York 1995.

[30] E.H. Yang, and J. Kieffer, On the Performance of Data Compression Algorithms Based upon String Matching, preprint (1995).

[31] Z. Zhang and E. Yang, An On-Line Universal Lossy Data Compression Algorithm via Continuous Codebook Refinement – Part II: Optimality for Phi-Mixing Source Models, *IEEE Trans. Information Theory*, 42, 822-836, 1996.