1996

A Last Word on Pattern Frequency Occurences In A Markovian Sequence?

Mireille Régnier

Wojciech Szpankowski
Purdue University, spa@cs.purdue.edu

Report Number:

96-043

A LAST WORD ON PATTERN FREQUENCY OCCURENCES IN A MARKOVIAN SEQUENCE?

Mireille Regnier
Wojciech Szpankowski

CSD-TR 96-043
July 1996

A LAST WORD ON PATTERN FREQUENCY OCCURRENCES IN A MARKOVIAN SEQUENCE? *

July 23, 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 Markovian source of any order. We study the frequency of pattern occurrences in a random text when overlapping copies of the pattern are counted separately. We provide exact and asymptotic formulæ for all moments (including the variance), and probability of r pattern occurrences for three different regions of r, namely: (i) $r = O(1)$, (ii) central limit regime, and (iii) large deviations regime. Our approach is uniform and seems to be novel: We first construct some language expressions that characterize pattern occurrences which are later translated into generating functions. 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: Frequency of pattern occurrences, Markov source, empirical distribution, source coding, autocorrelation polynomials, languages, generating functions, asymptotic analysis, large deviations.

*This research was supported by NATO Collaborative Grant CRG.950060. Part of this work was done during authors visits at Purdue University and at INRIA, Rocquencourt.

†This work was additionally supported by the ESPRIT III Program No. 7141 ALCOM II and GdR 1029.

‡This research was additionally supported by NSF Grants CCR-9201078, 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. Applications of these results include wireless communications (cf. [1]), approximate pattern matching (cf. [21]), molecular biology (cf. [30]), games, code synchronization, (cf. [16, 17, 18]), source coding (cf. [7], stock market analysis, and so forth. In fact, this work and the one by Fudos *et al.* [12] was prompted by questions posed by E. Ukkonen, T. Imieliński and P. Pevzner concerning approximate pattern matching by q-grams (cf. [21]), developing performance analysis models for database systems in wireless communications (cf. [1]), and gene recognition in a DNA sequence (cf. [30]), respectively. Actually, one of the earliest application appears to be to code synchronization (cf. [16]).

We study the problem in a probabilistic framework in which the text is generated randomly either by a memoryless source (the so called *Bernoulli model*) or by a Markovian source (the so called *Markovian model*). In the former, every symbol of a finite alphabet $\mathcal{S}$ is created independently of the other symbols, and the probabilities of symbol generation are not the same (if all probabilities of symbol generation are the same, the model is called *symmetric Bernoulli* model). In the Markovian model, the next symbol depends on a finite number previous symbols.

Pattern occurrences in a random string is a classical problem. Feller [10] already in 1968 suggested some solutions in his book. Several other authors also contributed to this problem: e.g., see [3, 5, 20, 25] and references there. However, the most important recent contributions belong to Guibas and Odlyzko, who in a series of papers (cf. [16, 17, 18]) laid the foundations of the analysis for the symmetric Bernoulli model. In particular, the authors of [18] computed the moment generating function for the number of strings of length n that do *not* contain any one of a 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. Furthermore, Guibas and Odlyzko [18] in a passing remark also presented some basic results for several pattern occurrences in a random text for the symmetric Bernoulli model, and for the probability of no occurrence of a given pattern in the asymmetric model. Recently, Fudos *et al.* [12] 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. The Markovian model was tackled by Li [25], Chrysaphinou and Papastavridis [5] who extended the Guibas and Odlyzko result of no pattern occurrence to Markovian texts. Recently, Prum *et al.* [31] (see also [33]) obtained the limiting distribution for the number of pattern occurrences in the Markovian model. Some other contributions are [3, 14, 22, 23, 28, 30, 36].

In this paper, we provide a complete characterization of the frequency of pattern occurrences in a random text generated according either to the Bernoulli model or the Markovian model using a methodology that might be of interest to other problems on words. Our method treats uniformly both models, and therefore we concentrate on discussing the Markovian model. Let O_n denote the number of occurrences of a given pattern H in a random text when *overlapping* copies of the pattern are counted separately. We compute exactly the mean EO_n and the variance Var O_n. Evaluation of the variance was quite challenging n the past as pointed out in [30] and [31]. It turns out that the variance depends on the internal structure of the pattern through the so called autocorrelation polynomial. Actually, Prum *et al.* [31] suggested two quite sophisticated methods to estimate the variance, and this should be compared with our computations (cf. Theorem 2.2, and Section 3).

We also estimate asymptotically the probability of exact r occurrences of the pattern for three different ranges of r (cf. Theorem 2.2). Namely, (i) $r = O(1)$, (ii) $r = EO_n + x\sqrt{n}$ for $x = O(1)$ (i.e., central limit regime), and (iii) $r = (1+\delta)EO_n$ (i.e., large deviations regime). For our results to hold we assume that $nP(H) \to \infty$ (see [14] for other regimes of $nP(H)$). However, for a given pattern H it is natural to assume that the length of the pattern is constant with respect to n (and for simplicity of the presentation we adopt this assumption).

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) [13], and recognizing genes by statistical properties [9]. 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. [9] 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.

Another biological problem for which our results might be useful is the gene recognition. Most gene recognition techniques rely on the observation that statistics of patterns/motifs/codon usage in coding and non-coding regions are different. Our findings allow to estimate the statistical significance of such differences, and one can construct the confidence interval for pattern occurrences.

One can also use these results to recognize statistical properties of various other information sources such as images, text, etc. In information theory, *relative frequency* defined as $\Delta_n = O_n/(n-m+1)$, where m is the length of the pattern, is often used to estimate the information source. It is well known [7, 27] that Δ_n converges almost surely to the probability $P(H)$ of the pattern H, but less is known about second-order properties such as limiting distribution, large deviations, and rate of convergence. Rate of convergence to the source entropy – which is related to the rate of convergence of the relative frequency [27] – have recently appeared in the formulation of some results on data compression (cf. [26, 34, 35, 38]). Marton and Shields [27] proved that Δ_n converges exponentially fast to $P(H)$ for sources satisfying the so called blow-up property (e.g., Markov sources, hidden Markov, etc). Our results characterize precisely such a convergence in the central limit regime and the large deviations regime. Finally, results of this paper should shed some light on second-order properties of the powerful method of typical types [7].

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 language approach, thus is also valid for Markovian models since no probabilistic assumption is made. 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 $H = h_1 h_2 \ldots h_m$ and a text string $T = t_1 t_2 \ldots t_n$ 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:

(B) either 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 $P(s)$ (i.e., Bernoulli model)

(M) or the text is a realization of a *stationary* Markov sequence of order K, that is, probability of the next symbol occurrence depends on K previous symbols. In most derivations we deal only with the first order Markov chain, and then we define the transition matrix $\mathbf{P} = \{p_{i,j}\}_{i,j\in S}$ where $p_{i,j} = \Pr\{t_{k+1} = j | t_k = i\}$. By $\pi = (\pi_1, \ldots, \pi_V)$ we denote the stationary distribution satisfying $\pi\mathbf{P} = \pi$. For stationary Markov chains $\Pr\{t_k = i\} = \pi_i$ for all $k \geq 0$.

Our goal is to estimate the frequency of multiple pattern occurrences in the text assuming either Bernoulli or Markovian model. To present our main findings we adopt some notation (cf. also [3, 16, 17, 20]). Below, we write $P(H_i^j) = \Pr\{T_{i+k}^{j+k} = H_i^j\}$ for the probability of the substring $H_i^j = h_i \ldots h_j$ occurrence in the random text T_{i+k}^{j+k} between symbols $i+k$ and $j+k$ for any k.

We find it convenient and useful to express our findings in terms of languages. A language $\mathcal{L}$ is a collection of words satisfying some properties. We associate with a 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 stationary probability of the word w, $|w|$ is the length of w, and we adopt a usual convention that $P(\epsilon) = 1$.

We define its H*-conditional generating function as*

$$L_{\mathrm{H}}(z) = \sum_{w\in\mathcal{L}} P(w | w_{-m} = h_1 \cdots w_{-1} = h_m) z^{|w|} \tag{2}$$

where w_{-i} stands for a symbol preceding the first character of w at distance i.

It turns out that several properties of pattern occurrences depend on the so called *autocorrelation polynomial* that we define next for the above two probabilistic models.

Definition 2 (i) (Bernoulli Model) *Given a string H we define the* autocorrelation polynomial *$A(z)$, as follows:*

$$A(z) = \sum_{k\in HH} P(H_{k+1}^m) z^{m-k}, \tag{3}$$

where HH is the set of positions of H for which a prefix of H is equal to a suffix of H, e.g., $k \in HH$ means that the last k symbols of H are equal to the first k symbols of H.

(ii) (Markovian Model) *The autocorrelation polynomial in the Markov model becomes*

$$A_{\rm H}(z) = \sum_{k\epsilon HH} P(H_{k+1}^m | H_1^k) z^{m-k}. \tag{4}$$

We can now proceed to formulate our main results. In the sequel, we denote by $O_n(\mathrm{H})$ (or simply by O_n) a random variable representing the number of occurrences of H in a random text T of size n. We introduce the generating function of the language $\mathcal{T}_r$ of words that contain exactly r occurrences of H: $T^{(r)}(z) = \sum_{n\geq 0} \Pr\{O_n(\mathrm{H}) = r\} z^n$ for $|z| \leq 1$. We also define a bivariate generating function as follows:

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

Our main results for the Markovian model are summarized in the following two theorems. The first theorem presents exact formulas for the generating functions $T^{(r)}(z)$ and $T(z,u)$, and can be used to compute exactly all parameters related to the pattern occurrence $O_n(H)$. In the second theorem, we provide asymptotic formulas for $\Pr\{O_n(H) = r\}$ for three regimes of r, namely: (i) $r = O(1)$, (ii) $r = EO_n + x\sqrt{\mathrm{Var}\, O_n}$ when $x = O(1)$ (i.e., local central limit), (iii) $r = (1+\delta)EO_n$ for some δ (i.e., large deviations). All proofs are presented in the next section. The method of derivation is interesting of its own right. 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, and T be a random text of length n generated according to a stationary Markov chain (of any order) over a V-ary alphabet $\mathcal{S}$. The generating functions $T^{(r)}(z)$ and $T(z,u)$ can be computed as follows:*

$$T^{(r)}(z) = R(z) M_{\rm H}^{r-1}(z) U_{\rm H}(z) \ , \tag{6}$$

$$T(z,u) = R(z) \frac{u}{1 - u M_{\rm H}(z)} U_{\rm H}(z) \ , \tag{7}$$

where, after defining

$$D_H(z) = (1-z)(A_{\rm H}(z) + (P_H(H) - P(H))z^m) + z^m P(H) \ , \tag{8}$$

we derive,

$$M_{\rm H}(z) = 1 + \frac{z-1}{D_H(z)} \ , \tag{9}$$

$$U_{\rm H}(z) = \frac{1 - M_{\rm H}(z)}{1-z} = \frac{1}{D_H(z)} \ , \tag{10}$$

$$R(z) = z^m P(H) U_{\rm H}(z) \ . \tag{11}$$

In the above, $P(H) = P(w = H)$ and $P_H(H) = P(w = H | w_{-1}^{-m} = H)$.

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 $nP(H) \to \infty$. The following results hold.*

(i) MOMENTS. *There exists $R > 1$ such that*

$$EO_n(H) = P(H)(n - m + 1) , \tag{12}$$

$$\mathrm{Var}\ O_n(H) = nP(H)c_1 + P(H)c_2 + O(R^{-n}) , \tag{13}$$

where

$$\begin{aligned} c_1 &= P(H)(2A_H(1) - 1 - (2m-1)P(H) + 2(P_H(H) - P(H))) , \\ c_2 &= P(H)((m-1)(3m-1)P(H) + (1-m)(2A_H(1) - 1) - 2A'_{\mathrm{H}}(1) \\ &- 2(2m-1)(P_H(H) - P(H)) . \end{aligned}$$

(ii) CASE $r = O(1)$. *Let ρ_H be the smallest root of $D_H(z) = 0$ outside the unit circle $|z| < 1$, and let $\rho > \rho_H$. Then:*

$$\Pr\{O_n(\mathrm{H}) = r\} = \sum_{j=1}^{r+1} (-1)^j a_j \binom{n}{j-1} \rho_{\mathrm{H}}^{-(n+j)} + O(\rho^{-n}) , \tag{14}$$

where

$$a_{r+1} = \frac{\rho_H^m P(\mathrm{H})\,(\rho_{\mathrm{H}} - 1)^{r-1}}{(D'_{\mathrm{H}}(\rho_{\mathrm{H}}))^{r+1}} , \tag{15}$$

and the remaining coefficients can be computed according to the standard formula, namely

$$a_j = \frac{1}{(r+1-j)!} \lim_{z \to \rho_{\mathrm{H}}} \frac{d^{r+1-j}}{dz^{r+1-j}} \left(T^{(r)}(z)(z - \rho_{\mathrm{H}})^{r+1} \right) \tag{16}$$

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

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

$$\Pr\{O_n(H) = r\} = \frac{1}{\sqrt{2\pi c_1 n}} e^{-\frac{1}{2}x^2} \left(1 + O\left(\frac{1}{\sqrt{n}} \right) \right) , \tag{17}$$

(iv) CASE $r = (1+\delta)EO_n$. *Let $a = 1 + \delta$ and $\delta \neq 0$. Define $\rho(t)$ to be the root of*

$$1 - e^t M_{\mathrm{H}}(e^{\rho}) = 0 , \tag{18}$$

and ω_a to be the root of

$$\rho'(\omega_a) = a \ . \tag{19}$$

Then:

$$\Pr\{O_n(H) = r\} = \frac{1}{\omega_a\sqrt{2\pi c_1 n}} e^{-((n-m+1)I(a)} \left(1 + O\left(\frac{1}{n}\right)\right) \tag{20}$$

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

As mentioned before, the above results have abundance of applications in information theory and molecular biology. Hereafter, we are concerned with the *relative frequency* defined as

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

Relative frequency appears in the definition of types and typical types (cf. [7]), and is often used to estimate information source statistics. As a corollary to Theorem 2.2, we obtain the following second-order characterization of $\Delta_n(H)$:

Corollary 2.1 *Under hypotheses of Theorem 2.2, the following holds:*
(i) *(*Central Limit regime*) For $x = O(1)$*

$$\Pr\{\Delta_n(H) = P(H) + x\sqrt{c_1/(n-m+1)}\} = \frac{1}{\sqrt{2\pi c_1 n}} e^{-\frac{1}{2}x^2} \left(1 + O\left(\frac{1}{\sqrt{n}}\right)\right) \ . \tag{21}$$

(ii) *(*Large deviations*) For $a = 1 + \delta$ with $\delta > 0$*

$$\Pr\{|\Delta_n(H) - P(H)| > \delta P(H)\} = \frac{1}{\omega_a\sqrt{2\pi c_1 n}} e^{-(n-m+1)I(a)} \left(1 + O\left(\frac{1}{n}\right)\right) \tag{22}$$

where ω_a and $I(a)$ are defined in Theorem 2.2 (iii).

The above results should be compared with first-order properties of $\Delta_n(H)$ discussed in [7, 27].

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 and works uniformly for both the Bernoulli model and the Markovian model. 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 on Certain Languages

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

We start with some definitions:

Definition 3 *Given a pattern* H*:*

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

(ii) We define $\mathcal{R}_{\mathrm{H}}$ and $\mathcal{L}_{\mathrm{H}}$ as languages containing only one occurrence of H *at the right and respectively left end of a word belonging to these languages. We also define $\mathcal{U}_{\mathrm{H}}$ as*

$$\mathcal{L}_{\mathrm{H}} = \mathrm{H} \cdot \mathcal{U}_H \tag{23}$$

where the operation $\cdot$ means concatenation of words. In other words a word $u \in \mathcal{U}_{\mathrm{H}}$ if Hu *has exactly one occurrence of* H *at the left end of* Hu.

(iii) Let $\mathcal{M}_{\mathrm{H}}$ be a language such that $H\mathcal{M}_H$ has exactly two occurrences of H *at the left and right end of a word from $\mathcal{M}_{\mathrm{H}}$, that is, $\mathcal{M}_{\mathrm{H}} = \{w :$* H$w$ has exactly two occurrences of H one at the right end and the other at the left end$\}$.

(iv) *Finally we defined a set $\mathcal{A}_{\mathrm{H}}$ associated with the autocorrelation of* H*, that is:*

$$\mathcal{A}_{\mathrm{H}} = \{\mathrm{H}_{k+1}^m : \quad k \in HH\} ,$$

where HH *is the autocorrelation sequence introduced in Definition 2.*

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

Theorem 3.1 *The language $\mathcal{T}$ satisfies the fundamental equation:*

$$\mathcal{T} = \mathcal{R}_{\mathrm{H}} \cdot \mathcal{M}_{\mathrm{H}}^* \cdot \mathcal{U}_H \ . \tag{24}$$

Notably, the language $\mathcal{T}_r$ can be represented for any $r \geq 0$ as follows:

$$\mathcal{T}_r = \mathcal{R}_{\mathrm{H}} \cdot \mathcal{M}_{\mathrm{H}}^{r-1} \cdot \mathcal{U}_H \ . \tag{25}$$

Proof: We first prove (25) and obtain our decomposition of $\mathcal{T}_r$ as follows: The first occurrence of H in a word belonging to $\mathcal{T}_r$ determines a prefix p that is in $\mathcal{R}_{\mathrm{H}}$. Then, one concatenates a non-empty word w that creates the second occurrence of H. Hence, w is in $\mathcal{M}_{\mathrm{H}}$. This process is repeated $r-1$ times. Finally, one adds after the last H occurrence a suffix w that does not create a new occurrence of H. Equivalently, Hu is in $\mathcal{L}_{\mathrm{H}}$, which means that u is in $\mathcal{U}_{\mathrm{H}}$, and w is a proper subword of Hu. Finally, a word belongs to $\mathcal{T}$ if for some $1 \leq r < \infty$ it belongs to $\mathcal{T}_r$. The set union $\bigcup_{r=1}^{\infty} \mathcal{M}_{\mathrm{H}}^{r-1}$ yields precisely $\mathcal{M}_{\mathrm{H}}^{*}$. ∎

We now prove the following result that summarizes relationships between the languages introduced in Definition 3.

Theorem 3.2 *The sets $\mathcal{M}_{\mathrm{H}}$, $\mathcal{U}_{\mathrm{H}}$ and $\mathcal{R}_{\mathrm{H}}$ satisfy:*

$$\bigcup_{k\geq 1} \mathcal{M}_{\mathrm{H}}^{k} = \mathcal{W} \cdot \mathrm{H} + \mathcal{A}_{\mathrm{H}} - \{\epsilon\} \,, \tag{26}$$

$$\mathcal{U}_{\mathrm{H}} \cdot \mathcal{S} = \mathcal{M}_{\mathrm{H}} + \mathcal{U}_{\mathrm{H}} - \{\epsilon\} \,, \tag{27}$$

$$\mathrm{H} \cdot \mathcal{M}_{\mathrm{H}} = \mathcal{S} \cdot \mathcal{R}_{\mathrm{H}} - (\mathcal{R}_{\mathrm{H}} - \mathrm{H}) \,, \tag{28}$$

where $\mathcal{W}$ is the set of all words, $\mathcal{S}$ is the alphabet set, ϵ is the empty word and $\oplus$ and $\ominus$ are disjoint union and subtraction of languages. In particular, a combination of (27) and (28) gives

$$H \cdot \mathcal{U}_H \cdot (\mathcal{S} - \epsilon) = (\mathcal{S} - \epsilon)\mathcal{R}_H \ . \tag{29}$$

Additionally, we have:

$$\mathcal{T}_0 \cdot \mathrm{H} = \mathcal{R}_{\mathrm{H}} \cdot \mathcal{A}_{\mathrm{H}} \ . \tag{30}$$

Proof: All the above relations are proved in a similar fashion. We first deal with (26). Let k be the number of H occurrences in $W \cdot \mathrm{H}$. By definition, $k \geq 1$ and the last occurrence is on the right: this implies that $W \cdot \mathrm{H} \subseteq \bigcup_{k\geq 1} \mathcal{M}_{\mathrm{H}}^{k}$. Furthermore, a word w in $\bigcup_{k\geq 1} \mathcal{M}_{\mathrm{H}}^{k}$ is not in $W \cdot \mathrm{H}$ iff its size $|w|$ is smaller than $|\mathrm{H}|$. Then, the second H occurrence in Hw overlaps with H, which means that w is in $\mathcal{A}_{\mathrm{H}}$.

Let us turn now to (27). When one adds a character s right after a word u from $\mathcal{U}_{\mathrm{H}}$, two cases may occur. Either Hus still does not contain a second occurrence of H, which means that us is a non-empty word of $\mathcal{U}_{\mathrm{H}}$. Or a new H appears, clearly at the right end. Then, us is in $\mathcal{M}_{\mathrm{H}}$. Furthermore, the whole set $\mathcal{M}_{\mathrm{H}} + (\mathcal{U}_{\mathrm{H}} - \epsilon)$ is attained, i.e., a strict prefix of $\mathcal{M}_{\mathrm{H}}$ cannot contain a new H occurrence. Hence, it is in $\mathcal{U}_{\mathrm{H}}$, and a strict prefix of a $\mathcal{U}_{\mathrm{H}}$-word is in $\mathcal{U}_{\mathrm{H}}$.

We now prove (28). Let $x = sw$ be a word in $\mathrm{H} \cdot \mathcal{M}_{\mathrm{H}}$ where s is a symbol from $\mathcal{S}$. As x contains exactly two occurrences of H located at its left and right ends, w is in $\mathcal{R}_{\mathrm{H}}$ and x is in $\mathcal{S} \cdot \mathcal{R}_{\mathrm{H}} - \mathcal{R}_{\mathrm{H}}$. Reciprocally, if a word $sw\mathrm{H}$ from $\mathcal{S} \cdot \mathcal{R}_{\mathrm{H}}$ is not in $\mathcal{R}_{\mathrm{H}}$, then $sw\mathrm{H}$ contains a second H occurrence starting in sw. As $w\mathrm{H}$ is in $\mathcal{R}_{\mathrm{H}}$, the only possible position is on the left end, and then x is in $\mathrm{H} \cdot \mathcal{M}_{\mathrm{H}}$. We now rewrite:

$$\mathcal{S} \cdot \mathcal{R}_{\mathrm{H}} - \mathcal{R}_{\mathrm{H}} = \mathcal{S} \cdot \mathcal{R}_{\mathrm{H}} - (\mathcal{R}_{\mathrm{H}} \cap \mathcal{S} \cdot \mathcal{R}_{\mathrm{H}}) = \mathcal{S} \cdot \mathcal{R}_{\mathrm{H}} - (\mathcal{R}_{\mathrm{H}} - \mathrm{H})$$

which yields $\mathrm{H} \cdot \mathcal{M}_{\mathrm{H}} - \mathrm{H} = (\mathcal{S} - \epsilon) \cdot \mathcal{R}_{\mathrm{H}}$.

Deriving (30) is only a little more intricate. Let t be some word in $\mathcal{T}_0$. We consider the factorization $t = w_1 w_2$ such that w_2 is the largest suffix that also is a $(m-k)$-prefix of H, with $k \in HH$ and $m = |H|$. In other words, w_2 is the largest suffix satisfying the equation $w_2 \cdot H = H \cdot a$, where a is in $\mathcal{A}_{\mathrm{H}}$. If $w_1\mathrm{H}$ were not in $\mathcal{R}_{\mathrm{H}}$, a second occurrence of H would occur in $w_1\mathrm{H}$ starting in w_1. As $w_1\mathrm{H}a = w_1 w_2 \mathrm{H}$, this contradicts the maximal property of w_2. Therefore, $\mathcal{T}_0 \cdot H \subseteq \mathcal{R}_H \cdot \mathcal{A}_H$. Finally, we consider a word $w_1 H a$ in $\mathcal{R}_{\mathrm{H}} \cdot \mathcal{A}_{\mathrm{H}}$. We may rewrite it as $H \cdot a = w_2 \cdot H$. It suffices now to show that $w_1 w_2 \in \mathcal{T}_0$. Indeed, since $|w_2| < |H|$, any occurrence of H would go across w_1 and $w_1 H$ would contain two occurrences of H, which is contradicts the definition fo $\mathcal{R}_H$. This proves $R_H \cdot \mathcal{A}_H \subseteq \mathcal{T}_0 \cdot H$, and completes the proof of Theorem 3.2. ■

3.2 Associated Generating Functions

In the previous section we did not make any probabilistic assumptions. Thus, Theorem 3.2 is true for any model, including Bernoulli and Markovian ones. In this section, we translate the language relationships into generating functions. Therefore, we need back our probabilistic assumptions. Most of our derivations deal with the Markovian model.

To transfer our language relations 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. We start with the following simply property holding in both probabilistic models:

(P1) Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be two arbitrary languages with generating functions (cf. (1)) $L_1(z)$ and $L_2(z)$, respectively. Then, the union language $\mathcal{L} = \mathcal{L}_1 \oplus \mathcal{L}_2$ is transferred into the generating function $L(z)$ such that

$$L(z) = L_1(z) + L_2(z) .$$

To translate the concatenation operation, one needs to consider the Bernoulli and the Markovian models separately. We start with the **Bernoulli model**:

(P2) Let us now consider a new language $\mathcal{L}$ that is, constructed from the concatenation of two other languages, say $\mathcal{L}_1$ and $\mathcal{L}_2$, that is $\mathcal{L} = \mathcal{L}_1 \cdot \mathcal{L}_2$. In the *Bernoulli model*, the generating function $L(z)$ of $\mathcal{L}$ becomes

$$L(z) = L_1(z)L_2(z)$$

since $P(wv) = P(w)P(v)$ for $w \in \mathcal{L}_1$ and $v \in \mathcal{L}_2$. In particular, the generating function $L(z)$ of $\mathcal{L} = \mathcal{S} \cdot \mathcal{L}_1$ is $L(z) = zL_1(z)$, where $\mathcal{S}$ is the alphabet set.

In the **Markovian model** $P(wv) \neq P(w)P(v)$, thus property (P2) is not any longer true. We have to replace it by a more sophisticated one. We have to condition $\mathcal{L}_2$ on symbols preceding a word from $\mathcal{L}_2$ (i.e., belonging to $\mathcal{L}_1$). In general, for a K order Markov chain, one must distinguish V^K ending states for $\mathcal{L}_1$ and V^K initial states for for $\mathcal{L}_2$. For simplicity of presentation, we only consider first-order Markov chains (i.e., $K = 1$), and we write $\ell(w)$ for the last symbol of a word w. We need the following definitions:

Definition 4 *Given a language $\mathcal{L}$, we define:*

$$L_i^j(z) = \sum_{w \in \mathcal{L}} P(w, \ell(w) = j | w_1 = i) z^{|w|} \ . \tag{31}$$

Additionally:

$$L_i(z) = \sum_{j \in \mathcal{S}} L_i^j(z) \ .$$

The following is a simple consequence of our previous definitions:

Corollary 3.1 *Let $\mathcal{L}$ be a language that does not contain the empty string. Its two generating functions defined respectively in (1) and (2) satisfy:*

$$L(z) = \sum_{k \in \mathcal{S}} \pi_k L_k(z) \tag{32}$$

$$L_{\mathrm{H}}(z) = \sum_{k \in \mathcal{S}}^{V} p_{\ell(\mathrm{H}),k} L_k(z) \tag{33}$$

where, we recall, $L_H(z)$ represents a language whose words are preceded by H.

Now, we can present the corresponding property (P2) for the Markovian model.

(P2') Let $\mathcal{L} = \mathcal{W} \cdot \mathcal{V}$. Then, according to definition (31) we have

$$L_k^l(z) = \sum_{i,j \in \mathcal{S}} p_{ji} W_k^j(z) V_i^l(z) \ . \tag{34}$$

To prove this, let $w \in \mathcal{W}$ and $v \in \mathcal{V}$. Observe that

$$
\begin{aligned}
P(wv) &= \sum_{j\in\mathcal{S}} P(wv,\ \ell(w)=j) \\
&= \sum_{j\in\mathcal{S}} P(w,\ \ell(w)=j)P(v|\ell(w)=j) \\
&= \sum_{j\in\mathcal{S}}\sum_{i\in\mathcal{S}} P(w,\ \ell(w)=j)p_{ji}P(v|v_1=i)\ .
\end{aligned}
$$

After conditioning on the first symbol of $\mathcal{W}$ and the last symbol of $\mathcal{V}$, we prove (34).

Now, we are ready to translate our basic relations from Theorems 3.1 and 3.2 into associated generating functions. Before proceeding with it, let us observe that one actually must deal only with two kinds of words. Namely, (i) we have words w for which no assumption is made on the preceding words (e.g., these are the words in $\mathcal{R}_{\mathrm{H}}$ with generating function $R(z)$); (ii) the only assumption we ever made on the preceding word is that it admits H as a suffix (e.g., words in $\mathcal{U}_H$ and $\mathcal{M}_H$ whose generating functions are $U_H(z)$ and $M_H(z)$, respectively). We also recall that $P(H) = P(w = H)$ and $P_H(H) = P(w = H|w_{-1}^{-m} = H)$.

Lemma 3.1 *The generating functions associated with languages $\mathcal{M}_{\mathrm{H}}, \mathcal{U}_{\mathrm{H}}$ and $\mathcal{R}_{\mathrm{H}}$ satisfy:*

$$
\begin{aligned}
\frac{1}{1-M_{\mathrm{H}}(z)} &= \frac{P(H)z^m}{1-z} + A_{\mathrm{H}}(z) + (P(H) - P_H(H))z^m\ , && (35)\\
U_{\mathrm{H}}(z) &= \frac{M_{\mathrm{H}}(z)-1}{z-1}\ , && (36)\\
R(z) &= P(H)z^m \cdot U_{\mathrm{H}}(z)\ , && (37)
\end{aligned}
$$

provided the underlying Markov chain is stationary.

Proof: We first prove (36). Interestingly, it needs no stationarity assumption. Let us consider the language relationship (27) from Theorem 3.2 which we re-write as $\mathcal{U}_{\mathrm{H}} \cdot (\mathcal{S} - \epsilon) = \mathcal{M}_H - \epsilon$. Observe that the left side of this equation, after conditioning on a left occurrence of H, yields:

$$
\sum_{i\in\mathcal{S}} U_{\mathrm{H}}^i(z)\Big(\sum_{j\in\mathcal{S}} p_{i,j}z - 1\Big) = \sum_{i\in\mathcal{S}} U_{\mathrm{H}}^i(z)\cdot(z-1) = U_{\mathrm{H}}(z)\cdot(z-1)\ .
$$

Of course, $\mathcal{M}_H - \epsilon$ translates into $M_H(z) - 1$, and (36) is proved.

We now turn our attention to (37). By (28), we observe that $\mathcal{S}\cdot\mathcal{R}_{\mathrm{H}} - \mathcal{R}_{\mathrm{H}}$ can be translate as follows (no assumption is made on H occurring on the left):

$$
\sum_{j\in\mathcal{S}} \pi_j z \cdot \sum_{i\in\mathcal{S}} p_{j,i}R_i(z) - \sum_{i\in\mathcal{S}} \pi_i R_i(z)\ .
$$

But, due to the stationarity of the underlying Markov chain

$$\sum_j \pi_j p_{j,i} = \pi_i \ ,$$

which yields $(z-1)\sum_i \pi_i R_i(z)$, and since $\mathcal{R}_{\rm H}$ does not contain an empty string, we finally obtain $(z-1)R(z)$. Furthermore, $H \cdot \mathcal{M}_H$ translates into $P(H)z^m \cdot (M_{\rm H}(z)-1)$. But, by (36), this becomes $P(H)z^m \cdot U_{\rm H}(z)(z-1)$, and after a simplification, we prove (37).

Finally, we deal with (35), and prove it using (26) of Theorem 3.2. The left-hand side of (26) involves the language $\mathcal{M}_H$, hence we must condition on the left occurrence of H. In particular, $\bigcup_{r\geq 1}\mathcal{M}_H^r + \epsilon$ of (26) translates into $\frac{1}{1-M_{\rm H}(z)}$. Now we deal with $\mathcal{W}\cdot H$ of the right-hand side of (26). *Conditioning* on the left occurrence of H, we have

$$\sum_{n\geq 1}\sum_{|w|=n} z^{n+m}P(wH|w_{-1}=\ell({\rm H})) = \sum_{n\geq 1}\sum_{|w|=n} z^n P(wh_1|w_{-1}=\ell(H))P(H|H_1=h_1)z^m \ .$$

Due to the stationarity, left conditioning disappears, and for $n\geq 1$ we obtain:

$$\sum_{|w|=n} P(wh_1|w_{-1}=\ell({\rm H})) = \sum_{|w|=n} P(w|w_{-1}=\ell({\rm H}))\pi_{h_1} = \pi_{h_1} \ ,$$

where, we recall, $\ell(H)$ is the last character of H. Hence, the language $(\mathcal{W}-\{\epsilon\})\cdot {\rm H}$ contributes $\frac{z}{1-z}P(H)z^m$, while the languages $\{{\rm H}\}\oplus\mathcal{A}_{\rm H}-\{\epsilon\}$ introduces $P_H(H)z^m + A_{\rm H}(z) - \epsilon$. This completes the proof of the theorem. ■

Finally, the next result completes the proof of Theorem 2.1.

Lemma 3.2 *The generating function $T(z,u)$ of the language $\mathcal{T}$ of words containing at least one occurrence of H becomes*

$$T(z,u) = R_H(z)\frac{u}{1-uM_{\rm H}(z)}U_{\rm H}(z) \ , \tag{38}$$

$$T^{(r)}(z) = R(z)M_{\rm H}^{r-1}(z)U_{\rm H}(z) \ , \tag{39}$$

where $R_H(z)$, $M_{\rm H}(z)$ and $U_{\rm H}(z)$ are expressed as in Lemma 3.1.

Proof. The proof is a direct consequence of (34) and Theorems 3.2 and 3.1. ■

Remark. The generating functions $T^j(z)$ of $\mathcal{T}_0^j$ in the Markov case were previously derived by Chrysaphinou and Papastavridis in [5]. We avoid here such a tedious computation since they are unnecessary to derive our results. A simply derivation of $T_0(z)$ follows from (30) 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 on Markov chains and 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 conclude that

$$\begin{aligned} T'(z,1) &= \frac{z^m P(H)}{(1-z)^2}, \\ T''(z,1) &= \frac{2z^n P(H) M_{\mathrm{H}}(z) D_{\mathrm{H}}(z)}{(1-z)^3}. \end{aligned}$$

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:

$$[z^n](1-z)^{-p} = \frac{\Gamma(n+p)}{\Gamma(p)\Gamma(n+1)} \tag{40}$$

To obtain EO_n we proceed as follows:

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

Denoting

$$\phi(z) = 2z^m P(H) M_{\mathrm{H}}(z) D_{\mathrm{H}}(z)$$

we get

$$EO_n(O_n-1) = [z^n]T''(z,1) = \phi(1)\frac{(n+2)(n+1)}{2} + \phi'(1)(n+1) + \frac{1}{2}\phi''(1)$$

Observing that $M_{\mathrm{H}}(z)D_{\mathrm{H}}(z) = D_{\mathrm{H}}(z) + (1-z)$, we use MAPLE to obtain a precise formula on the variance (cf. (13) of Theorem 2.2).

B. CASE $r = O(1)$

Now, we prove part (ii) of Theorem 2.2, that is, we estimate $\Pr\{O_n = r\}$ for $r = O(1)$. We first re-write the formula on $T^{(r)}(z)$ as follows:

$$T^{(r)}(z) = \frac{z^m P(H)(D_H(z) + z - 1)^{r-1}}{D_H^{r+1}(z)} \ . \tag{41}$$

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. [32]) 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, thus we can only expect poles (which cause the denominator $D_H(z)$ to vanish). The next lemma establishes the existence of at least one such a pole.

Lemma 3.3 *The equation $D_{\mathrm{H}}(z) = 0$ has at least one solution; the solution of smallest modulus, ρ_H, is real positive and satisfies $\rho_H > 1$. All the other solutions ρ satisfy $\rho > \rho_H$ iff* H *is not periodic.*

Proof: The roots of D_{H} are the poles of $\frac{1}{1-M_{\mathrm{H}}(z)}$. As it is the generating function of a language, it has no pole in $|z| \leq 1$ and all the coefficients are real and positive. Hence, the root of smallest modulus, ρ_H, is real and positive. Moreover, there is only one root of modulus ρ_H iff D_H is not a function of z^d for some $d \geq 1$, e.g., if H is not periodic. ∎

In view of the above, we can expand the generating function $T^{(r)}(z)$ around $z = \rho_{\mathrm{H}}$ in the following Laurent's series (cf. [32, 37]):

$$T^{(r)}(z) = \sum_{j=1}^{r+1} \frac{a_j}{(z-\rho_{\mathrm{H}})^j} + \tilde{T}^{(r)}(z) \tag{42}$$

where $\tilde{T}^{(r)}(z)$ is analytical in $|z| \leq \rho_{\mathrm{H}}$. The term $\tilde{T}^{(r)}(z)$ contributes only to the lower terms in the asymptotic expansion of $T^{(r)}(z)$. Actually, it is easy to see that for $\rho > \rho_{\mathrm{H}}$ we have $\tilde{T}^{(r)}(z) = O(\rho^{-n})$ (cf. [37]). The constants a_j can be computed according to (16) with the leading constant a_{-r-1} having the explicit formula (15).

We need an asymptotic expansion for the first terms in (41). This is rather a standard computation (cf. [37]), but for the completeness we provide a short proof. The following chain of indentities is easy to justify for any $\rho > 0$:

$$\begin{aligned}
\sum_{j=1}^{r+1} \frac{a_j}{(z-\rho)^j} &= \sum_{j=1}^{r+1} \frac{a_j(-1)^j}{\rho^j(1-(z/\rho)^j} \\
&= \sum_{j=1}^{r+1} (-1)^j a_j \rho^{-j} \sum_{n=0}^{\infty} \binom{n+j-1}{n} \left(\frac{z}{\rho}\right)^n \\
&= \sum_{n=1}^{\infty} z^n \sum_{j=1}^{\min\{r+1,n\}} (-1)^j a_j \binom{n}{j-1} \rho^{-(n+j)} .
\end{aligned}$$

After some algebra, we prove part (ii) of Theorem 2.2.

C. Case $r = EO_n + xO(\sqrt{n})$ for $x = O(1)$

We now establish part (iii) of Theorem 2.2, that is, we compute $\Pr\{O_n = r\}$ for $r = EO_n + x\sqrt{\text{Var } O_n}$ when $x = O(1)$ (the so called central limit regime). Let $\mu_n = EO_n(H)$ and $\sigma_n^2 = \text{Var } O_n(H)$. To establish normality of $(O_n(H) - \mu_n)/\sigma_n$, it suffices, according to Levy's theorem, to prove the following

$$\lim_{n\to\infty} e^{-t\mu_n/\sigma_n} T_n(e^{t/\sigma_n}) = e^{t^2/2} \tag{43}$$

for some complex t around zero. The computations are standard and go as below. The equation

$$1 - e^t M_{\mathrm{H}}(e^\rho) = 0 \tag{44}$$

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

$$T_n(e^t) = C(t)e^{(n+1-m)\rho(t)} + O(R^{-n}) \tag{45}$$

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

After some algebra, this leads (cf. [2]) 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.

Actually, we can proceed as in Greene and Knuth [15] or Hwang [19] to obtain much more refined local limit result. For example, direct application of results from [15] (cf. Chp. 4.3.3) leads to the following for $x = o(n^{1/6})$

$$\Pr\{O_n = EO_n + x\sqrt{nc_1}\} = \frac{1}{\sqrt{2\pi n c_1}} e^{-\frac{1}{2}x^2}\left(1 - \frac{\kappa_3}{2c_1^{3/2}\sqrt{n}}\left(x - \frac{x^3}{3}\right)\right) + O(n^{-3/2}) , \tag{46}$$

where κ_3 a constant (i.e., the third cumulant).

D. Case $r = (1+\delta)EO_n$ – Large deviations

Finally, we consider a large deviations result. From (45) we conclude that

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

Thus, directly from Gärtner-Ellis theorem [4, 8] 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 $\rho'(t) = a$, we obtain

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

But, due to our precise asymptotics for $T_n(e^t)$ we can do much better, as already suggested in [4, 15, 19]. We only sketch the approach. As in the central limit regime, we could use Cauchy's formula to compute the probability $\Pr\{O_n = r\}$ for $r = EO_n + xO(\sqrt{n})$. But, formula (46) is only good for $x = O(1)$. To expand its validity, we follow Greene and Knuth [15], and apply the so called "shift of mean", that is, we shift the mean of the generating function $T_n(u)$ to a new value, say $m = an$, so we can again apply the central limit formula (46) around the new mean. To accomplish this, we introduce a new parameter α such that

$$[z^m]T(u) = \frac{T(\alpha)}{\alpha^m}[z^m]\left(\frac{T(\alpha u)}{T(\alpha)}\right) .$$

The point to observe is that the new generating function $T(\alpha u)/T(\alpha)$ has a new mean at $\alpha T'(\alpha)/T(\alpha)$. Selection of α is easy. For example, for $T(u)$ given by (45) we compute α according to

$$\frac{\alpha\rho'(\alpha)}{\rho(\alpha)} = \frac{m}{n}$$

for $m = an$. The details of the computation can be found in [19], and for our specific case are reported in part (iv) of Theorem 2.2. This also completes the proof of the whole Theorem 2.2.

ACKNOWLEDGEMENT

It is our pleasure to acknowledge several discussion with A. Dembo, A. Odlyzko and P. Pevzner on the topic of this paper.

References

[1] D.Barbara, and T.Imielinski, Sleepers and Workoholics - Caching in Mobile Wireless Environments, *Proc. ACM SIGMOD*, 1-15, Minneapolis 1994

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

[3] S. Breen, M. Waterman and N. Zhang, Renewal Theory for Several Patterns, J. Appl. Prob., 22, 228-234, 1985.

[4] J. Bucklew, and J. Sadowsky, A Contribution to the Theory of Chernoff Bounds, *IEEE Trans. Information Theory*, 39, 249-254, 1993.

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

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

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

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

[9] J. Fickett, Recognition of Protein Coding Regions in DNA Sequences, *Nucleic Acids Res.*, 10, 5303-5318, 1982.

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

[11] P. Flajolet and M. Soria, General Combinatorial Schemas: Gaussian Limit Distributions and Exponential Tails, *Discrete Mathematics*, 114, 159-180, 1993.

[12] I. Fudos, E. Pitoura and W. Szpankowski, On Pattern Occurrences in a Random Text, *Information Processing Letters*, 1996.

[13] M.S. Gelfand, Prediction of Function in DNA Sequence Analysis, *J. Comput. Biol.*, 2, 87-117, 1995.

[14] M. Geske, A. Godbole, A. Schafner, A. Skolnick, G. Wallstrom, Compound Poisson Approximations for World Patterns Under Markovian Hypotheses, *J. Appl. Prob.*, 32, 877-892, 1995.

[15] D. Greene and D. E. Knuth, *Mathematics for the Analysis of Algorithms*, Birkhäuser, Boston 1990.

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

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

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

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

[20] 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.

[21] 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.

[22] S. Karlin, C. Bruge and A. Campbell, Statistical Analysis of Counts and Distributions of Restriction Sites in DNA Sequences, *Nucl. Acids Res.*, 20, 1363-1370, 1992.

[23] S. Karlin and F. Ost, Counts of Long Aligned Word Matches Among Random Letter Sequences, *Ann. Prob.*, 19, 293-351, 1987.

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

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

[26] 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.

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

[28] P. T. Nielsen, On the Expected Duration of a Search for Fixed Pattern in Random Data, *IEEE Trans. Information Theory*, 702-704, 1973.

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

[30] P. Pevzner, M. Borodovsky, and A. Mironov, Linguistic of Nucleotide Sequences: The Significance of Deviations from Mean: Statistical Characteristics and Prediction of the Frequency of Occurrence of Words, *J. Biomol. Struct. Dynam.*, 6, 1013-1026, 1991.

[31] 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.

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

[33] 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.

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

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

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

[37] H. Wilf, *generatingfunctionology*, Academic Press, Boston 1990.

[38] 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.