1995

Frequency of Pattern Occurences in a (DNA) Sequence

Mireille Régnier

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

Report Number:
95-053

FREQUENCY OF PATTERN OCCURRENCES IN A (DNA) SEQUENCE

Mireille Regnier
Wojciech Szpankowski

CSD-TR-95-053
August 1995

FREQUENCY OF PATTERN OCCURRENCES IN A (DNA) SEQUENCE*

July 7, 1995

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 of length n. We assume that consecutive symbols in the text are generated either independently or with a Markovian dependency, i.e., we study both the so called *Bernoulli model* and the *Markovian model*. Our goal is to assess the limiting distribution of the frequency of the pattern occurrences in a random sequence. Overlapping copies of a pattern are counted separately! We prove that the number of pattern occurrences tends to a normal distribution, and we derive explicit and asymptotic formulas for the mean and the variance of the pattern occurrence. During the course of the derivation we compute the probability of exactly r occurrences of H in the text T. We derive the generating function of this probability, and using an analytical technique we derive in a uniform manner all results announced above. Applications of these results range from wireless communications to approximate pattern matching, molecular biology, games, codes, and stock market analysis. These findings are of particular interest to molecular biology problems such as finding patterns with unexpected (high or low) frequencies (the so called contrast words) and gene recognition.

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. Some notable applications include coding theory and data compression, formal language theory, finding repeated motifs of a DNA sequence, and the design and analysis of algorithms. 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 range from wireless communications (cf. [1]) to approximate pattern matching (cf. [14, 21]), molecular biology (cf. [18]), games, codes (cf. [10, 11, 12]), and

*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 and NCR-9206315.

stock market analysis. In fact, this work was prompted by questions posed by E. Ukkonen, T. Imieliński and P. Pevzner concerning approximate pattern matching by q-grams (cf. [14]), developing performance analysis models for database systems in wireless communications (cf. [1]), and gene recognition in a DNA sequence (cf. [18]), respectively.

Our results are obtained in a probabilistic framework, namely we assume that the text is generated randomly according to either *Bernoulli model* or *Markovian model*. In the former, every symbol of a finite alphabet Σ 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 the previous one, and this dependency is described by a transition matrix $\mathbf{P}$. We obtain the mean, variance and the limiting distribution of the number of pattern occurrences in a random text.

In this conference version of the paper, we concentrate mostly on the Bernoulli model. Most derivations are shown only for this model. In fact, our analysis turns out to be quite simple so an extension to a Markovian model is possible without major changes.

Studying pattern occurrences in a random string is a classical problem. Feller [7] already in 1968 suggested some solutions in his book. Several other authors also contributed to this problem: e.g., see [3, 4, 13, 16] and references there. However, the most important recent contributions belong to Guibas and Odlyzko, who in a series of papers (cf. [10, 11, 12]) laid the foundations of the analysis for the symmetric Bernoulli model. In particular, the authors of [12] 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 [12] 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. In this paper, we extend these results of [12]. In particular, we compute the probability of exactly r occurrences of a pattern (given or random) in a random text in the *asymmetric* Bernoulli model and Markovian model. Furthermore – which is our main contribution that does not follow from Guibas and Odlyzko results – we derive the limiting distribution of the number of pattern occurrences. Finally, we should mention that the existence of a limiting distribution follows from a martingale argument proposed in [19]. Mean and variance for the *symmetric* Bernoulli model were also computed in [18]. All of this is generalized and simplified in this paper. In addition, we provide the rate of convergence for the central limit theorem, convergence in moments, and large deviations results (cf. Theorem 2.2).

Our results are of particular interest to molecular biology problems such as finding pattern with unexpected (high or low) frequencies (the so called contrast words) [9] and used in recognizing genes by statistical properties [6]. 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. [6] 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 mentioned above for which our results might be useful is the gene recognition. The most gene recognition techniques rely on the observation that statistics of patterns/motifs/codon usage in coding and non-coding regions are different. Our paper provides a method to estimate the statistical significance of such differences.

In general, using our results we can construct confidence interval for pattern occurrences, and use it to recognize some regions in a DNA sequences. In fact, our approach can be also used to recognize statistical properties of an information source (e.g., DNA, image, text, etc.), and using this information one can tune up algorithms used in these problems. We leave further applications of our method to a journal version of the paper restricting ourselves to presenting the results and sketching our proofs which seem to be general (and leading to surprisingly simple results).

This paper is organized as follows. In the next section we present our main results and their consequences. The proofs are delayed till 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 fact, we believe that this approach should be successful in finding palindrom occurrences in a DNA sequence (a problem of particular interest to molecular biologists since it allows to predict the secondary structure of DNA).

2 Main Results

Let us consider two strings, a pattern string $H = h_1 h_2 \dots h_m$ and a text string $T = t_1 t_2 \dots t_n$ of respective lengths equal to m and n over an alphabet Σ of size V. We assume that the pattern string is fixed and given, while the text string is random. More precisely, the text string T is:

(i) either a realization of an independently, identically distributed sequence of random variables (i.i.d.), such that a symbol $s \in \Sigma$ occurs with probability $P(s)$ (i.e., Bernoulli model)

(ii) or the text is a realization of a Markov sequence, that is, probability of the next symbol occurrence depends on the previous sequence. In this case, we define $\mathbf{P} = \{p_{s_1,s_2}\}_{s_1,s_2 \in \Sigma}$ where $p_{s_1,s_2} = \Pr\{t_{j+1} = s_2 | t_j = s_1\}$.

Our main goal is to estimate the frequency of multiple pattern occurrences in the text assuming the asymmetric Bernoulli model and the Markovian model. However, in this conference version of the paper we concentrate only on the Bernoulli model. As it turns out, our novel method of derivation can easily be extended to the Markovian case, thus we leave the technicalities of this case to a journal version. We claim that one of our main contribution is the method of analysis.

To present our main findings we adopt some notation from [11, 12] (cf. also [4, 13]). Below, we write $P(\mathrm{H}_i^j)$ for the probability of the substring $\mathrm{H}_i^j = \mathrm{h}_i \ldots \mathrm{h}_j$.

Definition 1 *For two strings F and H we define the correlation polynomial $C_{FH}(z)$, as follows*

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

where $k \in FH$ means that the last k symbols of F are equal to the first k symbols of H (i.e., the size k suffix of F is equal to the size k prefix of H). If $F = H$, then the correlation polynomial is called the **autocorrelation polynomial**, *and is denoted by $A_H(z) = C_{HH}(z)$.*

We can now proceed to formulate our main results. In the sequel, we denote by $O_n(\mathrm{H})$ a random variable representing the number of occurrences of H in a random text T of size n. We also write $t_{r,n}(\mathrm{H}) = \Pr\{O_n(\mathrm{H}) = r\}$. We introduce the probability generating function as: $T_r(z) = \sum_{n \geq 0} t_{r,n} z^n$ for $|z| \geq 1$. Finally, we define the bivariate generating function

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

Our first preliminary result for the Bernoulli model is summarized in the next theorem. The proof is presented in the next section. Its method of derivation is of its own interest, and may be applied to a larger class of problems on words.

Theorem 2.1 *Let H be a given pattern, and T be a random text generated according to the asymmetric Bernoulli model.*

(i) The generating function $T(z,u)$ (of a language of words containing at least one occurrence of a given pattern H) is:

$$T(z,u) = \frac{P(H) z^m}{D_{\mathrm{H}}(z)} \cdot \frac{u}{D_{\mathrm{H}}(z) - N_{\mathrm{H}}(z) u}$$

with

$$D_{\mathrm{H}}(z) = P(H) z^m + (1-z) A_H(z) \tag{2}$$

$$N_{\mathrm{H}}(z) = P(H) z^m + (1-z)(A_H(z) - 1) \tag{3}$$

Furthermore: For any $r \geq 1$

$$T_r(z) = \frac{z^m P(\mathrm{H}) (N_{\mathrm{H}}(z))^{r-1}}{(D_{\mathrm{H}}(z))^{r+1}} \tag{4}$$

(ii) Let ρ_{H} be the largest root in $|z| < 1$ of $D_{\mathrm{H}}(z) = 0$. Then, $0 < \rho_{\mathrm{H}} < 1$, and more precisely

$$\rho_{\mathrm{H}} = 1 - \frac{P(\mathrm{H})}{A_{\mathrm{H}}(1)} + O(P^2(\mathrm{H})) \ . \tag{5}$$

For large n and fixed r the following asymptotic formula holds for some $\rho < \rho_{\mathrm{H}}$

$$t_{r,n}(\mathrm{H}) = \sum_{j=1}^{r+1} a_{-j} n^{j-1} \rho_{\mathrm{H}}^{n-j} + O(\rho^n) \tag{6}$$

$$= \frac{1}{r!} a'_{-r-1} n^r \rho_{\mathrm{H}}^{n+m-r-1} + O(n^{r-1}\rho_{\mathrm{H}}^n) \tag{7}$$

where

$$a'_{-r-1} = \frac{P(\mathrm{H})\left(N_{\mathrm{H}}(\rho_{\mathrm{H}})\right)^{r-1}}{\left(D'_{\mathrm{H}}(\rho_{\mathrm{H}})\right)^{r+1}} , \tag{8}$$

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

$$a_{-j} = \frac{1}{(r-j+1)!} \lim_{z \to \rho_{\mathrm{H}}} \frac{d^{r+1-j}}{dz^{r+1-j}} \left(T_r(z)(z-\rho_{\mathrm{H}})^{r+1} \right) \tag{9}$$

with $j = 1, 2, \ldots r$. ■

Using the above result, in particular (2.1), we prove in Section 3 our main result concerning the frequency of pattern occurrences.

Theorem 2.2 *(i) Under the same assumptions as in Theorem 2.1 we obtain*

$$EO_n(\mathrm{H}) = P(\mathrm{H})(n-m+1) , \tag{10}$$

$$\mathrm{Var}\ O_n(\mathrm{H}) = nP(\mathrm{H})c_1 + P(\mathrm{H})c_2 \tag{11}$$

where

$$c_1 = 2A_{\mathrm{H}}(1) - (2m-1)P(\mathrm{H}) + 1 ,$$

$$c_2 = m(3m-2)P(\mathrm{H}) - 2mA_{\mathrm{H}}(1) - m - 2A'_{\mathrm{H}}(1)$$

(ii) *For large n*

$$\frac{O_n(\mathrm{H}) - EO_n(\mathrm{H})}{\sqrt{\mathrm{Var}\ O_n(\mathrm{H})}} \xrightarrow{d} N(0,1) \tag{12}$$

where $\xrightarrow{d}$ mean "in distribution", and $N(0,1)$ denotes the standard normal distribution. The convergence above also holds in moments. More precisely, for a complex t we have

$$e^{-t\frac{\mu_n}{\sigma}} T_n\left(e^{\frac{t}{\sigma_n}}\right) = e^{\frac{t^2}{2}} \left(1 + O\left(\frac{1}{\sqrt{n}}\right)\right) \tag{13}$$

where $\mu_n = EO_n(H)$ and $\sigma_n = \mathrm{Var}\ O_n(H)$ and $T_n(u) = Eu^{O_n(H)}$.

The above results find several applications in molecular biology and other areas, as mentioned in the introduction. In particular, Theorem 2.2 can be used to construct a confidence interval for the frequency count. We leave a complete discussion of our consequences to the journal version. We point out only that our result gives also the rate of convergence that might be crucial for some applications.

3 Analysis

The key element of our analysis is a derivation of the generating function $T(z,u)$ presented in Theorem 2.1. The main part of the below derivation is general enough that it works for Bernoulli as well as for Markovian models. It is based on constructing special languages and finding relationships between them. Later in this section, when we work with generating functions, we restrict ourselves to the Bernoulli model. A similar approach, however, works without any significant change (except that calculations become more cumbersome) for the Markovian model.

We proceed in two stages. First, we use language algebra and derive associated functional equations for generating functions (cf. Theorem 2.1). Once this is done, we use standard analytical tools to derive the asymptotic properties of the frequency count $O_n(H)$ (cf. Theorem 2.2).

3.1 General Relationships on Certain Languages

We start with some definitions:

Definition 2 *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 conatining only one occurrence of H *at the right and respectively left end of a word from these languages. We also define $\mathcal{U}_{\mathrm{H}}$ as*

$$\mathcal{L}_{\mathrm{H}} = \mathrm{H} \cdot U_H \tag{14}$$

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 that 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 the 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 defined in Definition 1.*

Using the above definition, we easily prove the following result that summarizes relationships between the languages introduced above.

Theorem 3.1 *The sets $\mathcal{T}_0, \mathcal{R}_{\rm H}$ and $\mathcal{L}_{\rm H}$ satisfy:*

$$\mathcal{T}_0 \cdot {\rm H} = \mathcal{R}_{\rm H} \cdot \mathcal{A}_{\rm H} \quad (15)$$

$$\mathcal{R}_{\rm H} = \mathcal{T}_0 \cdot (\mathcal{S} \ominus \{\epsilon\}) \oplus \{c\} \quad (16)$$

$${\rm H} \cdot \mathcal{U}_H = (\mathcal{S} \ominus \epsilon) \cdot \mathcal{T}_0 \oplus \{c\} \quad (17)$$

where $\mathcal{S}$ is the alphabet set, ϵ is the empty word, and $\oplus$ and $\ominus$ are disjoint union and substraction of languages.

Proof: All three relations are proved in a similar fashion. Let us start with (16). We consider a word t in $\mathcal{T}_0$ and add some character x. Clearly, $t \cdot x$ is not an empty sequence. Two cases must be considered: either $t \cdot x$ does not contain H (and then must be in $\mathcal{T}_0 \ominus c$) or it contains H and thus the word is in $\mathcal{R}_{\rm H}$. This completes the proof. A similar reasoning yields (17). Finally, relation (15) follows directly from the definitions of $\mathcal{R}_{\rm H}$ and $\mathcal{A}_{\rm H}$. ■

The above relations allow us to describe the language $\mathcal{T}_r$, that further leads to the generating function of $O_n({\rm H})$. We prove below the following:

Theorem 3.2 *The language $\mathcal{T}_r$ satisfies the fundamental equation:*

$$\mathcal{T}_r = \mathcal{R}_{\rm H} \cdot \mathcal{M}_{\rm H}^{r-1} \cdot \mathcal{U}_H \quad (18)$$

where the set $\mathcal{M}_{\rm H}$ defined in Definition 2 can be represented as

$$\mathcal{M}_{\rm H} = \mathcal{U}_H \cdot \mathcal{S} \ominus (\mathcal{U}_H \ominus \{c\}) \ . \quad (19)$$

Also, the language $\mathcal{T}$ satisfies:

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

Proof: We get our decomposition as follows. The first occurrence of H in a word in $\mathcal{T}_r$ determines a prefix that is in $\mathcal{R}_{\rm H}$. Then, one concatenates a non-empty word $w \cdot x$ that creates the second occurrence of H. Hence, w is in $\mathcal{U}_H$ while $w \cdot x$ is not. Equivalently, $w \cdot x$ ranges over $\mathcal{U}_H \cdot \Lambda \ominus (\mathcal{U}_H - \{\epsilon\})$. This process is repeated $r-1$ times. Finally, one adds a suffix chosen from $\mathcal{U}_H$. ■

3.2 Generating Functions for the Bernoulli Model

In the previous section we did not make any probabilistic assumptions. Thus the above results are good for any model, including the Markovian one. Now, we start deriving generating functions, and therefore we restrict ourselves for simplicity of presentation to the Bernoulli model.

We start with a definition:

Definition 3 *For any language $\mathcal{L}$ we define its generating function $L(z)$ as*

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

where $P(w)$ is the probability of the word w, and $|w|$ is the length of w.

To transfer our language relations into generating functions, we need few rules that we discuss next: The two operations on our languages, namely: the disjoint union $\oplus$ and concatenation $\cdot$ become the sum operation $+$ and the multiplication operation on generating functions. Indeed, it is easy to prove the following two properties:

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

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

(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 $\mathcal{L} = \mathcal{L}_1 \cdot \mathcal{L}_2$. Then, the generating function $L(z)$ of $\mathcal{L}$ becomes

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

In particular, the generating function $L(z)$ of $\mathcal{L} = \mathcal{S} \cdot \mathcal{L}_1(z)$ where $\mathcal{S}$ is the alphabet is $L(z) = zL_1(z)$.

Now we are ready to translate our basic relations from Theorems 3.1 and 3.2 into appropriate generating functions. A direct application of properties (P1) and (P2) leads to the following result that immediately implies Theorem 2.1.

Lemma 3.1 *The generating function*

$$T(z,u) = \sum_{r=1}^{\infty} u^r T_r(z) = \sum_{r=1}^{\infty} u^r \sum_{w \in \mathcal{T}_r} P(w) z^{|w|}$$

associated with the language $\mathcal{T}$ (at least once occurrence of H) becomes

$$T(z,u) = R_H(z) \frac{u}{1 - uM_H(z)} U_H(z) \tag{20}$$

where

$$T_0(z) = \frac{A_H(z)}{P(\mathrm{H})z^m + (1-z)A_{\mathrm{H}}(z)} \tag{21}$$

$$R_{\mathrm{H}}(z) = \frac{P(\mathrm{H})z^m}{P(\mathrm{H})z^m + (1-z)A_{\mathrm{H}}(z)} \tag{22}$$

$$U_H(z) = \frac{1}{P(\mathrm{H})z^m + (1-z)A_{\mathrm{H}}(z)} \tag{23}$$

$$M_H(z) = (z-1)U_H(z) + 1 . \tag{24}$$

As mentioned above, this lemma directly implies part (i) of Theorem 2.1. Part (ii) follows after some simple asymptotic analysis as in [8].

3.3 Moments and Limiting Distribution

We first wrestle with the moments of $O_n(H)$, that is, $EO_n(H) = [z^n]T'(z,1)$ and $EO_n(H)(O_n(H)-1) = [z^n]T''(z,1)$ where $[z^n]f(z)$ denotes the coefficient of $f(z)$ at z^n, and $T'(z,1)$, $T''(z,1)$ are the first and the second derivative of $T(z,u)$ at $u=1$. Below, we only show computation for the mean $EO_n(H)$ since the latter is similar but more cumbersome.

After using the identity $u = (uN_H(z) - D_H(z)/N_H(z) + D_h(z)/N_H(z)$ the calculation become very simply, and we find

$$T'_u(z,u) = \frac{z^m P(H)}{(D_H(z) - uN_H(z))^2} ,$$

and

$$T'(z,1) = \frac{P(H)}{z^{m-2}(1-z)^2} .$$

The formula (10) for the mean follows immediately from this. For the variance we use the following formula

$$T''_{uu}(z,1) = \frac{2P(H)N_H(z)}{z^{2m-3}(1-z)^3} ,$$

and it suffices to extract the coefficients at $1/(1-z)$ for $i = 1,2,3$.

The limiting distribution is not much more difficult! Let $\rho(u) < 1$ be the largest root of

$$D_H(z) - N_H(z)u = 0 ,$$

or equivalently

$$1 - uM_H(z) = 0 \tag{25}$$

where $M_H(z)$ is given above. Then, an elementary application of the residue theorem leads for any $R \gg 1$ to

$$T_n(u) = C(u)\rho(u)^{-(n+1-m)}(u) + O(R^{-n}) \tag{26}$$

where

$$C(u) = \frac{P(H)}{D'_H(1-\rho(u)) + D_H(\rho(u))} .$$

The form of (26) suggest to use Bender's result [2], but we rather derive usue Goncharov's theorem [15] to avoid checking all assumptions of the Bender theorem. Observe that $C(1) = 1$, and $EO_n(H) \sim -(n-m+1)\rho'(1)$ as well as Var $O_n \sim -(n-m+1)(\rho''(1) - [\rho'(u)]^2)$. We now use Goncharov's theorem (cf. [15]) to prove the limiting distribution. Let $\mu_n = EO_n(H)$ and $\sigma_n^2 =$ Var $O_n(H)$. Then, to establish normality of $(O_n(H) - \mu_n)/\sigma_n$ we must prove the following

$$\lim_{n\to\infty} e^{-t\mu_n/\sigma_n} T_n(e^{t/\sigma_n}) = e^{t^2/2} \tag{27}$$

for some complex t around zero. Using our expression (26) after some algebra we prove that (cf. [2])

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

and this completes the proof of our results.

Finally, we should mention that the above analysis can be relatively easily extened to the Markov case. In particular, (20) should hold, and thus the function $\rho(u)$ is the smallest root of equation (25), with an appropriate interpretation of $M_H(z)$. The algebra involved is more intricate, but conceptually we are in the same framework.

ACKNOWLEDGEMENT

It is our pleasure to acknowledge several discussion with Prof. 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] R. Benevento, The Occurrence of Sequence Patterns in Ergodic Markov Chains, *Stochastic Processes and Applications*, 17, 369-373, 1984.

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

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

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

[8] I. Fudos, E. Pitoura and W. Szpankowski, On Pattern Occurrences in a Random Text, Purdue University, CSD-TR-95-020, 1995.

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

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

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

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

[13] 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.

[14] 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.

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

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

[17] A. Odlyzko, Asymptotic Enumeration Methods, in *Handbook of Combinatorics*, 1995.

[18] 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.

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

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

[21] E. Ukkonen, Approximate String-Matching with q-grams and Maximal Matches, *Theoretical Computer Science*, 92, 191-211, 1992.

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