Department of Computer Science Technical Reports

Department of Computer Science

1995

On Pattern Occurrences in a Random Text

Ioannis Fudos

Evaggelia Pitoura

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

Report Number:
95-020

ON PATTERN OCCURRENCES IN A RANDOM TEXT

Ioannis Fudos
Evaggelia Pitoura
Wojciech Szpankowski

Department of Computer Science
Purdue University
West Lafayette, IN 47907

CSD-TR-95-020
March 1995

ON PATTERN OCCURRENCES IN A RANDOM TEXT

March 15, 1995

Ioannis Fudos, Evaggelia Pitoura, and Wojciech Szpankowski*
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.
{fudos,pitoura,spa}@cs.purdue.edu

Abstract

Consider a given pattern H and a random text T of length n. We assume that symbols in the text occur independently, and various symbols have different probabilities of occurrence (i.e., the so called *asymmetric Bernoulli model*). We are concerned with the probability of exactly r occurrences of H in the text T. We derive the generating function of this probability, and show that asymptotically it behaves as $\alpha n^r \rho_{\mathrm{H}}^{n-r-1}$, where α is an explicitly computed constant, and $\rho_{\mathrm{H}} < 1$ is the root of an equation depending on the structure of the pattern. We then extend these findings to random patterns.

Key Words: Pattern occurrence, Bernoulli model, autocorrelation polynomial, generating functions, asymptotic analysis.

*Research of this author was supported by NSF Grants CCR-9201078 and NCR-9206315.

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.

The goal of this paper is to study the number of occurrences of a *given* pattern in a *random* text of length n. More precisely, we compute the probability that a given pattern occurs exactly r times in a random text (overlapping copies of the pattern being counted separately). The text is generated according to the so called *asymmetric Bernoulli* model, that is, 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.

Studying the occurrence of patterns in a random string is a classical problem. Feller [4] already in 1968 suggested some solutions in his book. Several other authors also contributed to this problem: e.g., see [2, 3, 8, 10] and references there. However, the most important recent contributions belong to Guibas and Odlyzko, who in a series of papers (cf. [5, 6, 7]) laid the foundations of the analysis for the symmetric model. In particular, in [7] the authors 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 [7] 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 some of the results of [7]. 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. We also provide precise asymptotic results useful in some engineering computations.

Applications of these results range from wireless communications (cf. [1]) to approximate pattern matching (cf. [9, 14]), molecular biology (cf. [12]), games, codes (cf. [5, 6, 7]), and stock market analysis. In fact, this work was prompted by questions posed by E. Ukkonen and T. Imieliński concerning approximate pattern matching by q-grams (cf. [9]), and developing performance analysis models for database systems in wireless communications (cf. [1]), respectively.

In passing, we should point out that our findings can be a starting point for deriving moments and the limiting distribution for the frequency of pattern occurrences in a random text. We leave these problems for future research.

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.

2. MAIN RESULTS

Let us consider two strings, a pattern string $\mathtt{H} = \mathtt{h}_1\mathtt{h}_2\ldots\mathtt{h}_m$ and a text string $\mathtt{T} = \mathtt{t}_1\mathtt{t}_2\ldots\mathtt{t}_n$ 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 $\mathtt{T}$ is a realization of an independently, identically distributed sequence of random variables (i.i.d.), such that a symbol $\mathtt{s} \in \Sigma$ occurs with probability $P(\mathtt{s})$. In other words, the text is generated according to the asymmetric Bernoulli model.

Our main goal is to estimate the probability of multiple pattern occurrences in the text assuming the asymmetric Bernoulli model. More precisely, we compute the probability that the pattern $\mathtt{H}$ occurs exactly r times in $\mathtt{T}$, where overlapping copies of $\mathtt{H}$ are counted separately.

To present our main findings we adopt some notation from [6, 7] (cf. also [3, 8]). Below, we write $P(\mathtt{H}_i^j)$ for the probability of the substring $\mathtt{H}_i^j = \mathtt{h}_i \ldots \mathtt{h}_j$.

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

$$C_{\mathtt{FH}}(z) = \sum_{k \epsilon FH} P(\mathtt{H}_{k+1}^m) z^{k-1}, \tag{1}$$

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

Observe that in the Bernoulli model, $P(\mathtt{H}_i^j) = \prod_{k=i}^j P(\mathtt{h}_k)$. The following example illustrates the above definition. For a more comprehensive discussion of the correlation polynomial the reader is referred to [6, 7] and [3, 8].

Example 1. *Illustration to Definition 1*

Let $\Sigma = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$, and $P(\mathtt{a}) = 2/3$, $P(\mathtt{b}) = 1/6$, and $P(\mathtt{c}) = 1/6$. If we assume $\mathtt{F} =$ aabccaab and $\mathtt{H} =$ aabccaababc, then

$$C_{\mathtt{FH}}(z) = P(\mathtt{ccaababc})z^2 + P(\mathtt{abc})z^7 = \frac{1}{26244}z^2 + \frac{1}{54}z^7$$

for the Bernoulli model. □

We can now proceed to formulate our main results. In the sequel, we denote by $O_n(\mathtt{H})$ a random variable representing the number of occurrences of $\mathtt{H}$ in a random text $\mathtt{T}$ of size n. We also write $t_{r,n}(\mathtt{H}) = \Pr\{O_n(\mathtt{H}) = r\}$. Furthermore, following Guibas and Odlyzko we introduce in a non-standard way the probability generating function, namely: $T_r(z) = \sum_{n \geq 0} t_{r,n} z^{-n}$ for $|z| \geq 1$.

In the next section, we prove the following result.

Theorem 1. *Let* $\mathtt{H}$ *be a given pattern, and* $\mathtt{T}$ *be a random text generated according to the asymmetric Bernoulli model.*

(i) *For any* $r \geq 0$

$$T_r(z) = \frac{z^m P(\mathtt{H})\,(N_{\mathtt{H}}(z))^{r-1}}{(D_{\mathtt{H}}(z))^{r+1}} \tag{2}$$

where

$$N_{\mathtt{H}}(z) = P(\mathtt{H})z^{-1} + (z-1)(A_{\mathtt{H}}(z) - z^{m-1}) \;, \tag{3}$$

$$D_{\mathtt{H}}(z) = P(\mathtt{H}) + (z-1)A_{\mathtt{H}}(z) \;. \tag{4}$$

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

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

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

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

$$= a_{-r-1} n^r \rho_{\mathtt{H}}^{n-r-1} + O(n^{r-1}\rho_{\mathtt{H}}^n) \tag{7}$$

where

$$a_{-r-1} = \frac{\rho_{\mathtt{H}}^m P(\mathtt{H})\,(N_{\mathtt{H}}(\rho_{\mathtt{H}}))^{r-1}}{(D'_{\mathtt{H}}(\rho_{\mathtt{H}}))^{r+1}} \;, \tag{8}$$

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

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

with $j = 1, 2, \ldots r$. ■

Remark 1. In some applications, one is more interested in the probability of at least R occurrences of $\mathtt{H}$ in $\mathtt{T}$. Often R is small, and then we immediately have (i.e., for $R = O(1)$)

$$\begin{aligned} \Pr\{O_n(\mathtt{H}) > R\} &= 1 - \Pr\{O_n(\mathtt{H}) = 0\} - \ldots - \Pr\{O_n(\mathtt{H}) = R\} \\ &= 1 - a_{-R-1} n^R \rho_{\mathtt{H}}^{n-R-1} + O(n^{R-1}\rho_{\mathtt{H}}^n) \end{aligned}$$

where a_{-R-1} is given by (8).

To illustrate the above theorem, and in particular the generating function $T_r(z)$, we consider one example.

Example 2: *Illustration to Theorem 1*

Let $\Sigma = \{\mathtt{a}, \mathtt{b}\}$, $P(\mathtt{a}) = \frac{1}{3}$ and $P(\mathtt{b}) = \frac{2}{3}$. We consider two different patterns:

(a) Let $\mathtt{H} = \mathtt{bb}$, then we obtain, $A_{\mathtt{H}}(z) = z + \frac{2}{3}$, $P(\mathtt{H}) = \frac{4}{9}$ and for $r = 1$ from (15) and (18) we arrive at,

$$T_1(z) = 36\frac{z^2}{(9z^2 - 3z - 2)^2} .$$

Thus,

$$t_{1,n}(\mathtt{H}) = \frac{4}{9}(n + \frac{1}{3})\left(\frac{-1}{3}\right)^n + \frac{4}{9}(n - \frac{1}{3})\left(\frac{2}{3}\right)^n$$

which can be checked by direct computations. For instance, for $n = 3$, the above formula gives $t_{1,3}(\mathtt{H}) = \frac{8}{27}$. Indeed, $t_{1,3}(\mathtt{H}) = P(\mathtt{abb}) + P(\mathtt{bba}) = \frac{8}{27}$. Similarly, for $n = 4$, the formula gives $t_{1,4}(\mathtt{H}) = \frac{28}{81}$, which is what we get from direct manipulations: $t_{1,4}(\mathtt{H}) = P(\mathtt{aabb}) + P(\mathtt{abba}) + P(\mathtt{babb}) + P(\mathtt{bbaa}) + P(\mathtt{bbab}) = \frac{28}{81}$.

(b) Let $\mathtt{H} = \mathtt{bab}$, then we have $A_{\mathtt{H}}(z) = z^2 + \frac{2}{9}$, $P(\mathtt{H}) = \frac{4}{27}$ and for $r = 2$ from (15) and (18) we obtain,

$$T_2(z) = 216\frac{z^2(3z^2 - 3z + 2)}{(27z^3 - 27z^2 + 6z - 2)^3} .$$

Thus, for $n = 5$, we get

$$t_{2,5}(\mathtt{H}) = \frac{8}{243}$$

which we can verify by direct computations, $t_{2,5}(\mathtt{H}) = P(\mathtt{babab}) = \frac{8}{243}$. □

We now consider the case of a random pattern $\mathtt{H}$ generated according to the same Bernoulli model as the text $\mathtt{T}$. Let O_n denote the number of occurrences of a pattern of length m in a text of length n. We also write $\tau_{r,n} = \Pr\{O_n = r\}$. Clearly, we have the following

$$\tau_{r,n} = \sum_{\mathtt{H}\in\mathcal{H}} t_{r,n}(\mathtt{H})P(\mathtt{H}) \tag{10}$$

where $\mathcal{H}$ is the set of all strings of length m over the alphabet Σ.

The next main finding is a direct consequence of Theorem 1 and formula (10).

Theorem 2. *Assume that the pattern* H *and the text* T *are random strings satisfying the Bernoulli model.*

(i) *For any* m

$$\tau_{r,n} = O(\max_{\mathtt{H}\in\mathcal{H}}\{t_{r,n}(\mathtt{H})\}) \ . \tag{11}$$

More precisely,

$$\max_{\mathtt{H}\in\mathcal{H}}\{t_{r,n}(\mathtt{H})P(\mathtt{H})\} \leq \tau_{r,n} \leq V^m \max_{\mathtt{H}\in\mathcal{H}}\{t_{r,n}(\mathtt{H})P(\mathtt{H})\} \ . \tag{12}$$

(ii) *Let* $\rho_* = \max_{\mathtt{H}\in\mathcal{H}}\{\rho_\mathtt{H}\}$, *and let* H^* *be the pattern for which the maximum of* $\rho_\mathtt{H}$ *is achieved. If:*

- m=$o(n)$, *then*

$$\lim_{n\to\infty} \frac{\log \tau_{r,n}}{n} = \log(\rho_*) \ , \tag{13}$$

- $m = O(1)$, *then*

$$\tau_{r,n} \sim a_{-r-1} n^r \rho_*^{n-r-1} P(\mathtt{H}^*) \ , \tag{14}$$

  where a_{-r-1} *is defined in (8).* ■

We should observe that the asymptotic formula (13) is not too useful if $\rho_* = 1$, which can happen quite often. In general, nevertheless, deriving asymptotics for $\tau_{r,n}$ is not too difficult since all terms in (10) are nonnegative. It is well known (cf. Odlyzko [11]) that the main contribution to the sum (10) comes from a few terms around $\max_{\mathtt{H}\in\mathcal{H}}\{t_{r,n}(\mathtt{H})P(\mathtt{H})\}$. For example, more careful analysis can provide asymptotics for $m = O(\log n)$, but we will not explore this issue any further in this note.

3. ANALYSIS

We first prove Theorem 1(i), that is, we derive formula (2) for the generating function $T_r(z) = \sum_{n\geq 0} t_{r,n} z^{-n}$. Following Guibas and Odlyzko [7], we introduce a new probability, namely $s_r(n)$ representing the probability of H appearing exactly $r+1$ times in a random string T, where one of the occurrences of H is located at the *very end* of the string. Let $S_r(z) = \sum_{n=0}^{\infty} s_r(n) z^{-n}$.

First, we will derive $T_0(z)$ and $S_0(z)$. From Theorem 3.3 of [7] we have

$$\begin{aligned} (z-1)T_0(z) + z\,S_0(z) &= z \\ P(\mathtt{H})\,T_0(z) - z\,A_\mathtt{H}(z)\,S_0(z) &= 0 \end{aligned}$$

By solving for $T_0(z)$, $S_0(z)$ we get,

$$\begin{aligned} S_0(z) &= \frac{P(\mathtt{H})}{(z-1)\,A_{\mathtt{H}}(z)+P(\mathtt{H})}\,, \\ T_0(z) &= \frac{z\,A_{\mathtt{H}}(z)}{(z-1)\,A_{\mathtt{H}}(z)+P(\mathtt{H})} \end{aligned} \tag{15}$$

To illustrate the proof, we will use the analog of die-throwing, i.e., we consider that the text T is generated by throwing a V-sided die n times. We observe that the probability $t_r(n)$,[1] that H appears exactly r times by the n-th throw is equal to the sum of the probabilities of all possible events at the $(n+1)$-th throw, given that by the n-th throw we have exactly r appearances of H. At the $(n+1)$-th throw we can either have one more appearance of H at the end of the string (an event having probability $s_r(n+1)$ to occur) or we can have no more appearances of H. The second event appears with probability P_1, where P_1 is the probability of having exactly r occurrences of the pattern in a text of length $n+1$, where there is no pattern occurrence at the very end of the text, and thus $t_r(n+1) = P_1 + s_{r-1}(n+1)$. By adding the probabilities of the two events we get,

$$t_r(n) = t_r(n+1) + s_r(n+1) - s_{r-1}(n+1), \qquad r \geq 0,\, n \geq 0 \tag{16}$$

Let k be the position of the last occurrence of H in T. Then, the probability $t_{r+1}(n)$ that we will have $r+1$ appearances of H by the n-th throw can be written as the sum of the products $s_r(k)u(n-k)$, where $u(n-k)$ is the probability of a string of length $n-k$ that it does not itself contain H and if appended to H does not form any additional H patterns. Note, that in the Bernoulli model, $s_0(n-k+m) = P(H)u(n-k)$. Thus,

$$t_{r+1}(n) = \sum_{k=0}^{n-m} s_r(k)\frac{s_0(n-k+m)}{P(\mathtt{H})}, \qquad r \geq 0,\, n \geq 0\ . \tag{17}$$

By multiplying both (16) and (17) by z^{-n} and summing on n we obtain the following system,

$$\begin{aligned} S_r(z) &= S_{r-1}(z) + \frac{1-z}{z}T_r(z) \\ T_{r+1}(z) &= \frac{1}{P(\mathtt{H})}S_r(z)S_0(z)z^m \end{aligned}$$

Solving now for $T_r(z)$ we get,

$$T_r(z) = \left(1 + \frac{1-z}{z\,P(\mathtt{H})}S_0(z)z^m\right)^{r-1}\frac{1}{P(\mathtt{H})}S_0^2(z)z^m \tag{18}$$

[1] For the simplicity of presentation, in this section we rather write $t_r(n)$ instead of $t_{r,n}(\mathtt{H})$.

Finally, by substituting $S_0(z)$ from (15) we get,

$$T_r(z) = z^m P(\mathrm{H}) \frac{\left(P(\mathrm{H})z^{-1} + (z-1)(A_{\mathrm{H}}(z) - z^{m-1})\right)^{r-1}}{(P(\mathrm{H}) + (z-1)A_{\mathrm{H}}(z))^{r+1}} \tag{19}$$

which proves formula (2) of Theorem 1(i).

Now, we can wrestle with part (ii) of Theorem 1, that is, extract an asymptotic behavior of $t_{r,n}$ from its generating function $T_r(z)$. By Hadamard's theorem (cf. [13]) 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 $D_{\mathrm{H}}(z)$ to vanish). The next lemma establishes the existence of at least one such pole.

Lemma. *The equation $D_{\mathrm{H}}(z) = 0$ has at least one solution in $|z| < 1$. The largest solution inside the circle $|z| < 1$ is denoted by ρ_{H}.*

Proof. The proof is based on the Rouché theorem, and it is only a slight modification of Theorem 11 in [8], thus the details are left for the interested reader. ∎

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. [13, 15]):

$$T_r(z) = \sum_{j=1}^{r+1} \frac{a_{-j}}{(z - \rho_{\mathrm{H}})^j} + \widetilde{T}_r(z) \tag{20}$$

where $\widetilde{T}_r(z)$ is analytical in $|z| > \rho_{\mathrm{H}}$, thus it contributes only to the lower terms in the asymptotic expansion of $\mathrm{T}_r(z)$. In fact, it is easy to see that for $\rho < \rho_{\mathrm{H}}$ we have $\widetilde{T}_r(z) = O(\rho^n)$ (cf. [15]). The constants a_{-j} can be computed according to (9) with the leading constant a_{-r-1} having the explicit formula (8). Finally, the asymptotic expansion of the root ρ_{H}, as presented in (5), follows directly from [8], however, a simple substitution of (5) into $D_{\mathrm{H}}(\rho_{\mathrm{H}}) = 0$ also proves its validity.

We need an asymptotic expansion for the first terms in (20). This is a rather standard computation (cf. [15]), but since we use z^{-n} instead of z^n, we present below a short derivation for the reader's convenience. The following chain of indentities is easy to justify for any $\rho > 0$:

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

Thus, the nth coefficient of the first term of (20) finally becomes ($n > r$)

$$[z^{-n}]\left(\sum_{j=1}^{r+1}\frac{a_{-j}}{(z-\rho_{\mathtt{H}})^j}\right)=\sum_{j=1}^{r+1}a_{-j}\binom{n-1}{j-1}\rho_{\mathtt{H}}^{n-j}\ . \tag{21}$$

The above completes the proof of Theorem 1(ii) after noting that $\binom{n-1}{j-1} = n^{j-1}(1+O(1/n))$. Thus, Theorem 1 has been proved.

Finally, we prove Theorem 2, which concerns the case where both the pattern and the text are random. Observe that the inequality (12) follows directly from the basic equation (10). To prove the first asymptotics, namely (13), we proceed as follows. Let q and $p < q$ be the largest and the smallest probability of symbols occurrence from the alphabet Σ. Then, (12) becomes

$$p^m \max_{\mathtt{H}\in\mathcal{H}}\{t_{r,n}(\mathtt{H})\} \le \tau_{r,n} \le (Vq)^m \max_{\mathtt{H}\in\mathcal{H}}\{t_{r,n}(\mathtt{H})\}\ .$$

Taking the logarithm of both sides of the above, and noting that $m/n = o(1)$ one proves (13). In a similar fashion we can prove (14), and this completes the proof of Theorem 2.

ACKNOWLEDGMENT

This work was prompted by a question concerning pattern occurrences in a random string posed by Professor T. Imieliński (Rutgers University) and Professor E. Ukkonen (University of Helsinki) to the third author. We thank them for this initiative. We also thank Dr. Odlyzko (AT&T), Prof. Waterman (University of Southern California) and Prof. Pevzner (University of Pennsylvania) for valuable discussions related to this problem.

References

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

[2] R. Benevento, The Occurrence of Sequence Patterns in Ergodic Markov Chains, *Stochastic Processes and Applications*, 17, 369-373, 1984.

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

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

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

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

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

[8] 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.

[9] 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.

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

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

[12] 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.

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

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

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