A probabilistic approach to pattern matching with mismatches

Mikhail J. Atallah, Philippe Jacquet, Wojciec Szpankowski

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél.:(1) 39 63 55 11

Rapports de Recherche

N° 1354

Programme 1
Programmation, Calcul Symbolique
et Intelligence Artificielle

A PROBABILISTIC APPROACH TO PATTERN MATCHING WITH MISMATCHES

Mikhail J. ATALLAH
Philippe JACQUET
Wojciech SZPANKOWSKI

Décembre 1990

RR-1354

A PROBABILISTIC APPROACH TO PATTERN MATCHING WITH MISMATCHES

Mikhail J. Atallah, Philippe Jacquet and Wojciech Szpankowski

Abstract

The study and comparison of strings of symbols from a finite or an infinite alphabet is relevant to various areas of science, notably molecular biology, speech recognition, and computer science. In particular, the problem of finding the minimum "distance" between two strings (in general, two blocks of data) is of great importance. In this paper we investigate the (string) pattern matching problem in a probabilistic framework. Given a text string **a** and a pattern string **b** of minor length, we call "optimal matching of **b** over **a**" the maximum number of matches between **b** and all substrings of **a**. We consider the probabilistic model where strings are random and independent over a finite alphabet. Our aim is to precisely evaluate the behaviour of the optimal matching of **b** over **a** when the respective lengths of the strings both tends to infinity with a polynomial dependence between them.

UNE APPROCHE PROBABILISTE A LA RECONNAISSANCE DE MOTIFS AVEC ERREUR

Résumé

L'étude et la comparaison de séquences de symboles jouent des rôles importants dans de nombreux domaines scientifiques comme la biologie moléculaire et l'informatique. Dans cette note nous analysons une reconnaissance de motifs particulière, dite avec erreur. Etant données deux séquences **a** et **b**, de longueur inférieure, écrites dans un même alphabet de taille finie, on appelle le "maximum d'accords de **b** sur **a**", le nombre maximum d'accords, symbole par symbole, entre **b** et tous les sous mots de **a**. Nous nous intéressons au cas où les deux séquences sont construites de manière aléatoire et indépendante à partir de l'alphabet commun. Notre propos est de déterminer avec précision le comportement du maximum d'accords de **b** sur **a**, lorsque les longueurs respectives des deux séquences tendent vers l'infini en gardant entre elles une dépendance polynomiale.

A PROBABILISTIC APPROACH TO PATTERN MATCHING WITH MISMATCHES

November 28, 1990

Mikhail J. Atallah*
Dept. of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.

Philippe Jacquet†
INRIA
Rocquencourt
78153 Le Chesnay Cedex
France

Wojciech Szpankowski‡
Dept. of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.

Abstract

The study and comparison of strings of symbols from a finite or an infinite alphabet is relevant to various areas of science, notably molecular biology, speech recognition, and computer science. In particular, the problem of finding the minimum "distance" between two strings (in general, two blocks of data) is of great importance. In this paper we investigate the (string) pattern matching problem in a probabilistic framework. Given a text string **a** of length n and a pattern string **b** of length m, let $M_{m,n}$ be the maximum number of matches between **b** and all m-substrings of **a**. Our main probabilistic result shows that for $\log m/\log n \to \alpha$ (i.e., $m = \Theta(n^\alpha)$) with $\alpha \leq \beta$ for some $\beta \leq 1$ we have $M_{m,n} = mP + \sqrt{m\tau}$ where $\tau/\log n$ converges *in probability* to $2(P - P^2)$, and P is the probability of a match between any two symbols of these strings. The parameter β depends on the distribution of symbols from the alphabet. This result suggests an $O(n)$ algorithm that with high probability will compute $M_{m,n}$ for two strings **a** and **b**, even if the detailed probabilistic characteristics of the alphabet are not known.

*This author's research was supported by the Office of Naval Research under Grants N0014-84-K-0502 and N0014-86-K-0689, and in part by AFOSR Grant 90-0107, and the NSF under Grant DCR-8451393, and in part by Grant R01 LM05118 from the National Library of Medicine.

†This research was primary supported by NATO Collaborative Grant 0057/89.

‡This author's research was supported by AFOSR Grant 90-0107 and NATO Collaborative Grant 0057/89, and, in part by the NSF Grant CCR-8900305, and by Grant R01 LM05118 from the National Library of Medicine.

1. INTRODUCTION

Pattern matching is one of the most fundamental problems in computer science. The version of this problem we consider here is the following one. Consider two strings, a text string $\mathbf{a} = a_1 a_2 ... a_n$ and a pattern string $\mathbf{b} = b_1 b_2 ... b_m$ of lengths n and m respectively, such that symbols a_i and b_j belong to a V-ary alphabet $\Sigma = \{1, 2, ..., V\}$. Let C_i be the number of positions at which the string $a_{i+1} a_{i+2} ... a_{i+m}$ agrees with the pattern $\mathbf{b}$ (indices are modulo n). That is, $C_i = \sum_{j=1}^{m} equal(a_{i+j}, b_j)$ where $equal(x, y)$ is one if $x = y$, zero otherwise. We are interested in the quantity $M_{m,n} = \max_{1 \le i \le n}\{C_i\}$. This problem, posed by Galil [11] for the case of more than one mismatch, of course has a linear time solution for the case of zero mismatch (i.e., $k = m$). For the case of a single mismatch (i.e., $k = m - 1$) a linear time solution is also known (attributed to Vishkin in [11]). The best known time bound for the general case of arbitrary k is $O(n\sqrt{m}\text{polylog}(m))$ and is due to Abrahamson [1].

We analyze $M_{m,n}$ and propose an $O(m + n)$ time algorithm for estimating it, under the following probabilistic assumption: *symbols from the alphabet Σ are generated independently, and symbol σ from the alphabet Σ occurs with probability p_σ.* This probabilistic model is known as the *Bernoulli model.* The algorithm we propose does not assume that we know the probabilities p_σ, it just assumes that they exist.

Our linear time algorithm is a simple consequence of our main probabilistic results, which provide a tight estimate on $M_{n,m}$. More precisely, we show that for $m = \Theta(n^\alpha)$ (when $\alpha \le \beta \le 1$ and β depends on the probabilistic nature of the alphabet), $M_{m,n} = mP + \sqrt{m\tau}$, where $\tau / \log n$ converges *in probability* to $2(P - P^2)$, and P is the probability of a match between any two symbols of the text and pattern strings.

Our linear time algorithm does *not* specify where the pattern occurs within the text. It outputs only the estimate of $M_{m,n}$, from which one could deduce whether the pattern (approximately) occurs in the text. There are several practical situations where such an information is useful, notably in molecular biology and pattern recognition. In particular, when searching for a homology between two biological strings (e.g., DNA, RNA or protein sequences) one needs to know how close is one strings from the other one [21]. Finally, our algorithm can also be used in the case when the position of occurrence of the pattern is important. In that case, our algorithm can act as a "cheap" test whether one should pay the $O(n\sqrt{m}\text{polylog}(m))$ time cost (Abrahamson's algorithm [1]) to find the position of occurrence in the text.

The paper is organized as follows. The next section makes a more precise statement of

our results. In particular, Theorem 2 contains our main probabilistic result. All proofs are delayed until Section 3, which is also of independent interest. It discusses a fairly general approach that can be used to analyze pattern matching in strings. In that section we apply extensively the saddle point method [14] to evaluate necessary asymptotic approximations. Finally, Section 4 contains some comparisons of our theoretical results with computer simulation results. Interestingly enough, our estimate is typically within 5% of the true value even for strings of length only in the hundreds. That section also suggests some future research problems associated with approximate pattern matching.

2. MAIN RESULTS

This section presents our main probabilistic and algorithmic results derived under the Bernoulli model discussed above. Note that, in such a model, $P = \sum_{i=1}^{V} p_i^2$ represents the probability of a match in a *given* position of the text string **a** and the pattern string **b**. It is easy to see that the distribution of C_i is binomial , that is, for any i

$$\Pr\{C_i = \ell\} = \binom{m}{\ell} P^{\ell}(1-P)^{m-\ell} . \tag{1}$$

Naturally, $\{C_i\}$ is a stationary sequence, and the average number of matches EC_1 is equal to $EC_1 = mP$. Furthermore, C_i tends *almost surely* to its mean mP (by the *Strong Law of Large Numbers* [8]).

The evaluation of $M_{m,n} = \max_{1 \le i \le n}\{C_i\}$ is more intricate, although the first order asymptotics are easy to obtain (cf. Theorem 1). For simplicity of the presentation, it helps to imagine that **a** is written on a cycle of size $n \ge 2m$, and that $\rho^i(\mathbf{b})$ is written on that same cycle, cyclically shifted by i positions relative to **a**. Then C_i can alternatively be thought of as the number of places on this cycle in which **a** and $\rho^i(\mathbf{b})$ agree.

Theorem 1. *If $m = O(n^{\alpha})$ for some $0 \le \alpha \le 1$, then for every $\epsilon > 0$ the following holds*

$$\lim_{n \to \infty} \Pr\{1 - \varepsilon < M_{m,n} - mP < 1 + \varepsilon\} = 1 , \tag{2}$$

that is, $M_{m,n} \sim EC_1 = mP$ in probability (pr.).

Proof. A lower bound on $M_{m,n}$ follows from the fact that the maximum $M_{m,n}$ over n values of C_i must be greater than C_1 which tends in probability to mP. So, now we concentrate on an upper bound. From Boole's inequality we have

$$\Pr\{M_{m,n} > r\} = \Pr\{C_1 > r \ \ or \ \ C_2 > r \ \ or \ \ \ldots C_n > r\} \le n\Pr\{C_1 > r\} . \tag{3}$$

It suffices to show that for $r \sim mP$ the above probability becomes $o(1)$, that is, $n\Pr\{C_1 > (1+\epsilon)mP\} = o(1)$. For this one needs an estimate of the tail for the binomial distribution (1). Such an estimate is computed in Section 2 by the saddle point method. A simpler (and more general) approach, however, is necessary for the purpose of this proof. We note that C_1 can be represented as a sum of m independent Bernoulli (spanned on two points) distributed random variables X_i, where X_i is equal to one when there is a match at the ith position, and zero otherwise. From the *Central Limit Theorem* we know that $(C_1 - EC_1)/(\sqrt{mP(1-P)}) \to \mathcal{N}(0,1)$, where $\mathcal{N}(0,1)$ is the standard normal distribution. Let $G_m(x)$ and $\Phi(x)$ be the distribution of $\sum_{i=1}^m X_i$ and the standard normal distribution respectively. Then, from Feller [8]

$$G_m(x) = \Phi(x) + \frac{e^{-x^2/2}}{\sqrt{2\pi}} o(\sqrt{m}) \ ,$$

where $\Phi(x) \sim \frac{e^{-x^2/2}}{x\sqrt{2\pi}}$. Define in (3) $r = mP + (1+\varepsilon)\sqrt{m2P(1-P)\log n}$. Then, the above directly implies our theorem. ∎

Theorem 1 does not provide much useful information, and an estimate of $M_{m,n}$ based on it would be a very poor one. From the proof of Theorem 1 we learn, however, that $M_{m,n} - EC_1 = O(\sqrt{m \log n})$, hence the next term in the asymptotics of $M_{m,n}$ can have a very significant value, and definitely cannot be omitted in any reasonable computation. The next theorem – our main result – provides an extension of Theorem 1, and shows how much the maximum $M_{m,n}$ differs from the average EC_1. In Section 3 we prove the following result.

Theorem 2. *Let* $\log m / \log n \to \alpha$, *that is,* $m = \Theta(n^\alpha)$ *for some* $\alpha > 0$.
(i) For every $\varepsilon > 0$ *we have*

$$\lim_{n\to\infty} \Pr\{\frac{M_n - mP}{\sqrt{2m(P-P^2)\log n}} < 1+\varepsilon\} = 1 \ , \tag{4}$$

where $P = \sum_{i=1}^V p_i^2$.

(ii) Let $T = \sum_{i=1}^V p_i^3$. *Then for every* $\varepsilon > 0$ *the following holds*

$$\lim_{n\to\infty} \Pr\{1-\varepsilon < \frac{M_n - mP}{\sqrt{2m(P-P^2)\log n}}\} = 1 \tag{5}$$

provided $0 < \alpha \le \beta = 1 - 4(T-P^2)/(P-3P^2+2T)$. *In other words,* $M_{m,n} = EC_1 + \sqrt{m\tau}$ *where* $EC_1 = mP$ *and* $\tau/\log n$ *converges in probability to* $2(P-P^2)$ *when* α *satisfies the above constraint.*

(iii) Let $\delta = \sqrt{\frac{(1-\alpha)(P-3P^2+2T)}{4(T-P^2)}}$. *Then for every* $\varepsilon > 0$ *we have*

$$\lim_{n\to\infty} \Pr\{(1-\varepsilon)\delta < \frac{M_n - mP}{\sqrt{2m(P-P^2)\log n}}\} = 1 \tag{6}$$

provided $1 \geq \alpha > \beta = 1 - 4(T-P^2)/(P-3P^2+2T)$. ∎

Remark 1. If all the p_i's are equal to $1/V$ (the so called uniform case), the condition on α becomes $\alpha \leq 1$, which is always true since $m \leq n$. In all other cases $\alpha < 1$. The proof of Theorem 2 is considerably simpler for the uniform case, but that case is unlikely to arise in practice. Much of the difficulty lies in establishing Theorem 2 in the general (nonuniform) case.

Remark 2. Theorem 2 holds in a much more general probabilistic framework provided the probability of a match P (and also T) is appropriately interpreted. For example, if the probability of occurrence of a symbol k at any position of **a** (resp., **b**) is p_k (resp., p'_k), then Theorem 2(ii) holds if $\alpha < 1 - 4(T-P^2)/(P-3P^2+T+T')$ where $P = \sum_{k=1}^{k=Q} p_k p'_k$, and $T = \sum_{k=1}^{k=Q} p_k^2 p'_k$ and $T' = \sum_{k=1}^{k=Q} p_k (p'_k)^2$. ∎

Theorem 2 suggests a simple algorithm that can approximately determine $M_{m,n}$ provided that the strings are independent (see [5] for justification of the independence assumption for some DNA sequence searches). We also predict that our theorem – and hence the algorithm discussed below – holds under much more general probabilistic model (cf. [15], [20]). We first note that in many practical applications the probabilities $\{p_i\}_{i=1}^{V}$ are *unknown*. In such situations, Theorem 2 provides an algorithmic tool to obtain an estimate of $M_{m,n}$:

Algorithm .

1. Compute $C_1 + C_2 + ... + C_n$. This is straightforward to do in $O(m+n)$ time, but we nevertheless sketch how it is done, for the sake of completeness: first, in $O(m)$ time, we compute the number of occurrences (call it $\text{count}_1(i)$) of each symbol $i \in \Sigma$ in the pattern. This is done by scanning the pattern and, if the current symbol being scanned is (say) i, incrementing $\text{count}_1(i)$ by one. We do the same for the text, obtaining in $O(n)$ time the number of times (call it $\text{count}_2(i)$) that each $i \in \Sigma$ occurs in the text. Now, observe that

$$C_1 + C_2 + ... + C_n = \sum_{i=1}^{V} (\text{count}_1(i) * \text{count}_2(i)),$$

and hence we can compute the quantity we seek with an extra $O(V)$ time.

2. Evaluate $\tilde{C}$ as follows

$$\tilde{C} = \frac{1}{n}\sum_{i=1}^{n} C_i \ . \tag{7}$$

(This takes constant time, in view of the previous step.)

3. From the *Strong Law of Large Numbers* [8] one concludes that $\tilde{C} \rightarrow EC_1$ *almost surely* for large n. Hence, we may estimate the probability P as $\tilde{P} = \tilde{C}/m$, and finally the estimate of $M_{m,n}$ is evaluated (in constant time) from Theorem 2 as

$$\tilde{M}_{m,n} = \tilde{C} + \sqrt{2\tilde{C}(1 - \tilde{C}/m)\log n} \ . \tag{8}$$

¿From Theorem 2 we know that $M_{m,n} = \tilde{M}_{m,n}$ *with high probability* (whp). Formally, (8) is true only for $\alpha \leq \beta$. The parameter β can be estimated in the algorithm by using the fact that $P^2 \leq T \leq P^{3/2}$ (cf. [16], [20]). We also note that even in the case $\alpha > \beta$ the estimate (8) gives a good approximation (see Section 4).

The above algorithm in $O(m+n)$ time provides the right answer (whp) to our problem even in the presence of unknown probabilities of symbol occurrences. The assumption of a very large n that we make in the analysis is not overly restrictive, since extremely large values of n can arise in many of the application areas of this problem, notably, in text processing, speech recognition, machine vision and, last but not least, molecular sequence comparison. For example, in the human genome project one estimates that n can be as large as 10^9 [7]. Surprisingly enough, our computer experiments indicate that our estimate works well even for moderate values of n, namely n close to 100 (see Section 4 for more details).

3. ANALYSIS

In this Section we prove our main theorem (Theorem 2). In the course of proving it we establish some interesting combinatorial properties of pattern matching that have some similarities with the work of Guibas and Odlyzko [12, 13] (see also [15]). The proof itself consists of two parts: upper bound (easy) and lower bound (difficult).

We start with the upper bound. Although it is easy to derive, in order to illustrate the technique that we adopt for the (much harder) lower bound, we present one result that directly implies the upper bound for Theorem 2. The following lemma suffices.

Lemma 3. *When m and τ both tend to infinity with $\tau = O(\log m)$, then*

$$\Pr\{C_1 \geq mP + \sqrt{m\tau}\} \sim \frac{1}{\sqrt{2\pi(P - P^2)\tau}} \exp[\frac{\tau}{2(P - P^2)}] \ . \tag{9}$$

Proof: According to (1) C_1 is binomially distributed, that is, $\Pr\{C_1 = r\} = \binom{m}{r}P^r(1-P)^{m-r}$. Introducing the generating function $C(u) = \sum_r \Pr\{C_1 = r\}u^r$ for u complex, we easily get the formula $C(u) = (1 + P(u-1))^m$. Then, by the Cauchy's celebrated formula [14]

$$\Pr\{C_1 \geq r\} = \frac{1}{2i\pi}\oint (1 + P(u-1))^m \frac{1}{u^r(u-1)}du \ , \tag{10}$$

where the integration is along a path encircling the unit disk for u complex. The problem is how to evaluate this integral for large m. In this case the best suited method seems to be a simplified saddle point method (see also the Laplace's method) [14, 9]. This method applies to integrals of the following form

$$I(m) = \int_\Gamma \phi(x)e^{-mh(x)}dx \ , \tag{11}$$

where Γ is a closed curve, and $\phi(x)$ and $h(x)$ are analytical functions inside Γ, and we evaluate $I(m)$ for large m. It is noticed that the main contribution to this integral comes from the point where $h(x)$ is minimum, that is $h'(x) = 0$ (some additional assumptions are necessary; for details see [14]). To apply this idea to our integral (10) we represent it in the form of (11) and find the minimum of the exponent. Define $u = 1 + h$, and then the exponent in our integral can be expended as

$$\begin{aligned}
\log((1 + P(u-1))^m/u^r) = & \quad m\log(1+hP) - r\log(1+h) \\
= & \quad (Pm - r)h - 1/2(P^2m - r)h^2 + O(m+r)h^3) \\
= & \quad -1/2\tfrac{(r-mP)^2}{r-mP^2} + (r - mP^2)\tfrac{(h-h_0)^2}{2} + O(m+r)h^3)
\end{aligned}$$

with $h_0 = (r - mP)/(r - mP^2)$. Let $r = mP + \sqrt{m}x$, with $x > 0$. Changing the scale of the variable under the integrand: $h = h_0 + it/\sqrt{m(P - P^2)}$, we obtain

$$\Pr\{C_1 \geq mP + \sqrt{m}x\} = \exp[-1/2\frac{x^2}{P - P^2 + x/\sqrt{m}}]\frac{1}{2\pi}\int \frac{\exp[-t^2/2]}{\frac{x}{\sqrt{P-P^2}} + it}dt\,(1 + O(1/\sqrt{m})) \ .$$

Therefore, when $x \to \infty$ (like $\sqrt{\log m}$) we get

$$\int \frac{\exp[-t^2/2]}{\frac{x}{\sqrt{P-P^2}} + it}dt \sim \frac{\sqrt{P - P^2}}{x}\int_{-\infty}^{\infty} \exp(-t^2/2)dt = \frac{\sqrt{2\pi(P - P^2)}}{x}$$

where the last integral can be computed from the error function [2]. This completes the proof of our lemma. ■

The rest of this section is devoted to the lower bound. We attack the problem through the *second moment method.* We will use a form due to Chung and Erdös [6], which states that for events $\{C_i > r\}$, the following holds

$$\Pr\{M_{m,n} > r\} = \Pr\{\bigcup_{i=1}^{n}(C_i > r)\} \geq \frac{(\sum_i \Pr\{C_i > r\})^2}{\sum_i \Pr\{C_i > r\} + \sum_{(i \neq j)} \Pr\{C_i > r \ \& \ C_j > r\}} \ . \quad (12)$$

Thus, one needs to estimate, unfortunately, the joint distribution $\Pr\{C_i > r \ \& \ C_j > r\}$, and prove that the right-hand side of (12) goes to one when $r = mP + \sqrt{m2(P - P^2)\log n}(1-\varepsilon)$ for any $\varepsilon > 0$. Define $F_{m,n}(r) = \sum_{i=2}^{i=n} \Pr\{C_1 > r \ \& \ C_i > r\}$. Then, the following lemma summarizes what we have already said.

Lemma 4. *For every $\varepsilon > 0$ we have $\lim_{n\to\infty} \Pr\{M_{m,n} \geq mP + \sqrt{ma_n}(1-\varepsilon)\} = 1$ provided that the following is fulfilled*

$$\lim_{n\to\infty} \frac{m}{n} \cdot \left(\frac{F_{m,n}(mP + \sqrt{ma_n}(1-\varepsilon))}{(\Pr\{C_1 > mP + \sqrt{ma_n}(1-\varepsilon)\})^2 m} - 1\right) = 0 \ . \quad (13)$$

where a_n is a solution of $n\Pr\{C_1 > mP + \sqrt{ma_n}\} = 1$, that is, $a_n \sim (P - P^2)\log n$ for large values of n.

Proof. Note that $\rho^i(\mathbf{b})$ and $\rho^j(\mathbf{b})$ do not overlap when $|i - j| > m$, and therefore corresponding C_i and C_j are mutually independent. Applying this to (12) one immediately obtains

$$\Pr\{M_{m,n} > r\} \geq \frac{n^2(\Pr\{C_1 > r\})^2}{(n^2 - 2n(m + 1/2))(\Pr\{C_1 > r\})^2 + n\Pr\{C_1 > r\} + 2nF_{m,n}(r)} \ .$$

Thus

$$\Pr\{M_{m,n} \geq mP + \sqrt{ma_n}(1-\varepsilon)\} \geq$$

$$[1 + 1/n + 1/g_n(-\varepsilon) + 2m/n(\frac{F_{m,n}(mP + \sqrt{ma_n}(1-\varepsilon))}{m(\Pr\{C_1 > mP + \sqrt{ma_n}(1-\varepsilon)\})^2} - 1)]^{-1} \ .$$

where $g_n(\varepsilon) = n\Pr\{C_1 > mP + \sqrt{ma_n}(1+\varepsilon)\}$ and $a_n \sim 2(P - P^2)\log n$. From Lemma 3 we easily see that $a_n \sim (P - P^2)\log n$, and $g(-\varepsilon) \to \infty$ as $n \to \infty$, hence (13) follows. ∎

According to Lemma 4 our problem reduces to a sharp estimate of the joint distribution of C_1 and C_ℓ. We achieve it by first evaluating the generating function $H_{m,\ell}(u,v) = \sum_{r_1,r_2} \Pr\{C_1 = r_1 \ \& \ C_\ell = r_2\}u^{r_1}v^{r_2}$, and then computing the probability $\Pr\{C_1 = r_1 \ \& \ C_\ell = r_2\}$ through the Cauchy integral as it was done in Lemma 3. Let $\mathbf{x}$ and $\mathbf{y}$ be column vectors of dimension V, that is, $\mathbf{x} = \{x_i\}_{i=1}^{V}$ and $\mathbf{y} = \{y_i\}_{i=1}^{V}$. We define the scalar product $\langle \mathbf{x}, \mathbf{y} \rangle$

by $x_1y_1 + \cdots + x_Vy_V$. Then, the next crucial theorem captures some important combinatorial properties of $\{C_1, C_\ell\}$ that allow to estimate the generating function $H_{m,2}(u,v)$ for $\ell = 2$, and finally $H_{m,\ell}(u,v)$ for any ℓ (cf. Theorem 6).

Now we are ready to establish a closed-form formula for the generating function $H_{m,2}(u,v)$. This is achieved by showing a recurrence relationship between the distributions of $\{C_1(\mathbf{b}), C_2(\mathbf{b})\}$ and $\{C_2(\mathbf{b}'), C_3(\mathbf{b}')\}$ where $\mathbf{b}'$ is the suffix of $\mathbf{b}$ of length $m-1$. In the above we write $C_i(\mathbf{b})$ instead of C_i in order to show explicitly a dependency of C_i on the string $\mathbf{ba}$. With this in mind, we can proceed to the following key theorem.

Theorem 5. *We have the identity* $H_{m,2}(u,v) = \langle \mathbf{x}(u), \mathbf{A}^{m-1}(u,v)\mathbf{y}(v)\rangle$, *where* $\mathbf{A}(u,v)$ *is a* $V \times V$ *square matrix whose generic element* $a_{ij}(u,v)$ *satisfies the following*

$$a_{ij}(u,v) = p_i(1 + p_i(v-1) + p_j(u-1))$$

when $i \neq j$, *and*

$$a_{ii}(u,v) = p_i(1 + p_i(uv-1))$$

for $i = j$. *The row vectors* $\mathbf{x}(u)$ *and* $\mathbf{y}(v)$ *are defined as* $\mathbf{x}(u) = \{1 + p_i(u-1)\}_{i=1}^V$ *and* $y(v) = \{p_i(1 + p_i(v-1))\}_{i=1}^V$.

Proof: Let us define a random variable Γ_i as the number of matches between string $\mathbf{a}$ and $\rho^i(\mathbf{b})$ *without* counting the eventual first matching at position i (recall that $\rho^i(\mathbf{b})$ is the shifted version of $\mathbf{b}$ by i positions on the cycle). For example, $\Gamma_1 = C_1$ if there is no matching at position 1, and $\Gamma_1 = C_1 - 1$ otherwise. Define next the generating function $P_{i,m}(u,v)$ as

$$P_{i,m}(u,v) = \sum_{r_1,r_2} \Pr\{\Gamma_1 = r_1 \ \&\ C_2 = r_2 \ \&\ \text{string } b \text{ starts with symbol } i\} u^{r_1} v^{r_2} \ ,$$

and $\mathbf{P}_m(u,v)$ denote the row vector $\{P_{i,m}(u,v)\}_{i=1}^V$. Note that $\mathbf{P}_1(u,v) = \mathbf{y}(v)$ and $H_{m,2}(u,v) = \sum_{i=1}^{i=V}(1 + p_i(u-1))P_{i,m}(u,v)$, thus $H_{m,2}(u,v) = \langle \mathbf{x}(u), \mathbf{P}_m(u,v)\rangle$.

The most interesting fact that we prove next is the following relationship $\mathbf{P}_m(u,v) = \mathbf{A}(u,v)\mathbf{P}_{m-1}(u,v)$ when $m > 1$. A proof of this relies on building a recurrence relationship between $\{C_1(\mathbf{b}), C_2(\mathbf{b})\}$ and $\{C_2(\mathbf{b}'), C_3(\mathbf{b}')\}$, as explained above. Let i and j be the two first symbols of string $\mathbf{b}$ and let k be the second symbol of string $\mathbf{a}$. When $i \neq j$ we have $\Gamma_1(\mathbf{b}) = \Gamma_2(\mathbf{b}') + 1$ and $C_2(\mathbf{b}) = C_3(\mathbf{b}')$ if $k = j$, $\Gamma_1(\mathbf{b}) = \Gamma_2(\mathbf{b}')$ and $C_2(\mathbf{b}) = C_3(\mathbf{b}') + 1$ if $k = i$, and $\Gamma_1(\mathbf{b}) = \Gamma_2(\mathbf{b}')$ and $C_2(\mathbf{b}) = C_3(\mathbf{b}')$ otherwise. When $i = j$, we have $\Gamma_1(\mathbf{b}) = \Gamma_2(\mathbf{b}') + 1$ and $C_2(\mathbf{b}) = C_3(\mathbf{b}') + 1$ if $k = i$, and $\Gamma_1(\mathbf{b}) = \Gamma_2(\mathbf{b}')$ and $C_2(\mathbf{b}) =$

$C_3(\mathbf{b}')$ otherwise. Since the $\Gamma_\ell(\mathbf{b})$'s and $C_\ell(\mathbf{b})$'s are stationary random variables, hence the following identity follows

$$\frac{P_{i,m}(u,v)}{p_i} = (1+p_i(uv-1))P_{i,m-1}(u,v) + \sum_{j\neq i}(1+p_i(v-1)+p_j(u-1))P_{j,m-1}(u,v)\ ,$$

which proves our theorem. ■

The next theorem extends Theorem 5 and give an ultimate formula for the generating function $H_{m,\ell}(u,v)$, which is of its own interest.

Theorem 6. *For all $q < m$ the following holds $H_{m,1+q}(u,v) = (H_{h,2}(u,v))^{q-\ell}(H_{h+1,2}(u,v))^{\ell}$, where $h = \lfloor \frac{m}{q} \rfloor$ and $\ell = m - hq$.*

Proof. For $i \leq q$ define $\mathbf{b}_{(i)}$ as a subsequence of string $\mathbf{b}$ obtained by selecting the ith symbol of $\mathbf{b}$, then the $i+q$th, then $i+2q$th, and so forth. For $1 \leq i \leq \ell$, strings $\mathbf{b}_{(i)}$ are of length $h+1$, for $\ell+1 \leq i \leq q$, strings $\mathbf{b}_{(i)}$ are of length h. We can do the same with string $\mathbf{a}$ and obtain subsequences $\mathbf{a}_{(1)}, \ldots, \mathbf{a}_{(q)}$.

Let $\langle \mathbf{a}, \rho^i(\mathbf{b}), \rho^j(\mathbf{b}) \rangle$ be a new notation for the two dimensional row vector $[C_i, C_j]$, that is, it represents the number of matches between $\mathbf{a}$ and simultaneously $\rho^i(\mathbf{b})$ and $\rho^j(\mathbf{b})$. It is easy to see that $[C_1, C_{1+q}] = \langle \mathbf{a}_{(1)}, \rho^1(\mathbf{b}_{(1)}), \rho^2(\mathbf{b}_{(1)}) \rangle + \cdots + \langle \mathbf{a}_{(q)}, \rho^1(\mathbf{b}_{(q)}), \rho^2(\mathbf{b}_{(q)}) \rangle$, where the $\langle \mathbf{a}_{(i)}, \rho^1(\mathbf{b}_{(i)}), \rho^2(\mathbf{b}_{(i)}) \rangle$ are absolutely independent. Note that $\langle \mathbf{a}_{(i)}, \rho^1(\mathbf{b}_{(i)}), \rho^2(\mathbf{b}_{(i)}) \rangle$'s has the same distribution as $[C_1, C_2]$ when $\mathbf{b}_{(i)}$ is of length $h+1$ for $i \leq \ell$, and the same distribution as $[C_1, C_2]$ when $\mathbf{b}_{(i)}$ is of length h for $\ell < i \leq q$. This finally establishes the theorem. ■

Theorem 6 establishes a closed form formula for the generating function of the joint distribution $\Pr\{C_1 = r_1, C_\ell = r_2\}$. Therefore, in principle we can recover the probabilities $\Pr\{C_1 = r_1, C_\ell = r_2\}$ from $H_{m,\ell}(u,v)$ by the Cauchy's formula, as we did in Lemma 3. The difficulty is that the generating function $H_{m,\ell}(u,v)$ is expressed in terms of matrix $\mathbf{A}(u,v)$, so we need some tools from the linear algebra to apply the saddle point method. Before, however, we plunge into this, we should treat the uniform case (i.e., $p_k = 1/V$) separately since , as the next lemma shows, it possesses very special property.

Lemma 7. *In the uniform case, for all $i \neq j$, the random variables C_i and C_j are mutually independent.*

Proof: It suffices to prove that C_1 is independent of C_{1+q} for all $1 \leq q \leq m$. In the uniform case the $a_{ij}(u,v)$'s are all identical and equal to $\frac{1}{V}(1+\frac{1}{V}(u-1+v-1))$ except when $i=j$

where $a_{ii}(u,v) = \frac{1}{V}(1 + \frac{1}{V}(uv - 1))$. Note that $\mathbf{y}(v)$ coincides with an eigenvector of the matrix $\mathbf{A}$ and $\mathbf{A}(u,v)\mathbf{y}(v) = (1 + \frac{1}{V}(u-1))(1 + \frac{1}{V}(v-1))\mathbf{y}(v)$ and therefore $H_{m,2}(u,v) = (1 + \frac{1}{V}(u-1))^m(1 + \frac{1}{V}(v-1))^m$. This last formula shows that C_1 and C_2 are mutually independent. Applying Theorem 6 one concludes that also $H_{m,1+q}(u,v) = (1 + \frac{1}{V}(u - 1))^m(1 + \frac{1}{V}(v-1))^m$. Therefore C_1 and C_{1+q} are also mutually independent. ∎

This lemma completes the proof of the lower bound in the uniform case by the second moment method. Therefore, in the rest of this section we concentrate on the non-uniform case.

Let $\lambda(u,v)$ be the principal eigenvalue of the matrix $\mathbf{A}(u,v)$. Let $\xi(u,v)$ (resp. $\Pi(u,v)$) be the corresponding right (resp. left) eigenvector of $\mathbf{A}(u,v)$ such that $\langle \Pi(u,v), \xi(u,v)\rangle = 1$, that is, $A(u,v)\xi(u,v) = \lambda(u,v)\xi(u,v)$ and $\mathbf{A}^T(u,v)\Pi(u,v) = \lambda(u,v)\Pi(u,v)$ (cf. [19, 18]. We note the following three cases.

1. When $u = v = 1$, $\lambda(1,1) = 1$, $\xi(1,1) = \mathbf{y}(1)$ and $\Pi(1,1) = \mathbf{x}(1)$, the other eigenvalues are null.

2. When $v = 1$, $\lambda(u,1) = 1 + P(u-1)$, $\xi(u,1) = \xi(1,1) = \mathbf{y}(1)$ and $\Pi(u,1) = 1/\lambda(u,1)\,\mathbf{x}(u)$, the other eigenvalues are null.

3. When $u = 1$, $\lambda(1,v) = \lambda(v,1) = 1 + P(v-1)$, $\xi(1,v) = \mathbf{y}(v)$ and $\Pi(1,v) = 1/\lambda(1,v)\,\mathbf{x}(1)$, the other eigenvalues are null.

It follows that the other eigenvalues are $O((u-1)(v-1))$, and therefore we immediately prove the following fact.

Corollary 8. *We have* $H_{m,2}(u,v)/\lambda^{m-1}(u,v) = \langle \mathbf{x}(u), \xi(u,v)\rangle \langle \Pi(u,v), \mathbf{y}(v)\rangle + O((u-1)^m(v-1)^m)$.

Proof: This is a classical property of the principal eigenvalue and follows from the Perron-Frobenius theorem (the interested reader is referred to [3, 18, 19] for details). ∎

As a consequence of Corollary 8 we have the following important expansion of the generating function $H_{m,\ell}(u,v)$.

Corollary 9. *Let* $F_m(u,v) = \sum_{i=2}^{i=m} H_{m,i}(u,v)$. *We have*

$$\frac{F_m(u,v)}{(\lambda(u,v))^m} = \frac{a(u,v) - a^m(u,v)}{1 - a(u,v)} + O((u-1)(v-1))\,. \tag{14}$$

with $a(u,v) = \langle \mathbf{x}(u), \xi(u,v)\rangle \langle \Pi(u,v), \mathbf{y}(v)\rangle/\lambda(u,v)$

Proof: From Corollary 8 and Theorem 6 we have the estimate

$$H_{m,1+q}(u,v) = (\lambda(u,v))^m [a(u,v))^q + O((u-1)^{h+1}(v-1)^{h+1})] .$$

Therefore

$$\frac{F_m(u,v)}{(\lambda(u,v))^m} = \sum_{q=1}^{m-1} (a(u,v))^q + \sum_{h=2}^{m} O((u-1)^{h+1}(v-1)^{h+1}) ,$$

which completes the proof by summing the geometric series in the last expression. ■

The following two lemmas present more detailed Taylor's expansions of the principal eigenvalue of $\mathbf{A}(u,v)$ defined in Corollary 9. These expansions are next used (cf. Theorem 12) in the saddle point method to obtain a sharp estimate of $F_{m,n}(r)$ around $r = mP + \sqrt{2mP(1-P)\log n}$ necessary to prove our lower bound (see Lemma 4 for details).

Lemma 10. *The Taylor expansion of* $\lambda(u,v)$ *to the second order is* $1 + (u-1)P + (v-1)P + (u-1)(v-1)(2T - P^2)$, *with* $T = p_1^3 + \cdots + p_V^3$.

Proof. We know that $\lambda(u,v) = 1 + (u-1)P + (v-1)P + O((u-1)(v-1))$. We adopt the following notation. If $f(u,v)$ is a function of two variables u and v, then we denote by $f_u(u,v)$ (resp. $f_v(u,v)$) the partial derivative of $f(u,v)$ with respect to u (resp. to v). We have $\lambda = \langle \Pi, \mathbf{A}\xi \rangle$, where the variables (u,v) have been dropped in the last expression for some simplifications of the presentation. Thus, $\lambda_u = \langle \Pi_u, \mathbf{A}\xi \rangle + \langle \Pi, \mathbf{A}_u\xi \rangle + \langle \Pi, \mathbf{A}\xi_u \rangle$. Since $\mathbf{A}\xi = \lambda\xi$, $\mathbf{A}^T\Pi = \lambda\Pi$ and the fact that $\langle \Pi_u, \xi \rangle + \langle \Pi, \xi_u \rangle = 0$ (because we assume $\langle \Pi, \xi \rangle = 1$), we get $\lambda_u = \langle \Pi, \mathbf{A}_u\xi \rangle$. Substituting $u = 1$ in the last expression we obtain the identity $\lambda(1,v) = (P + 2T(v-1) + \sum_{i=1}^{i=V} p_i^4(v-1)^2)/(1 + P(v-1))$, and after some simple algebra this completes the proof. ■

Lemma 11. *The Taylor expansion of* $a(u,v)$ *defined in Corollary 9 to the second order is* $1 - (u-1)(v-1)(T - P^2)$.

Proof. Easy computations give $a(u,1) = a(1,v) = 1$, therefore $a(u,v) - 1$ is $O((u-1)(v-1))$. Differentiating twice $a(u,v)$ and setting $u = v = 1$ leads to a formula beginning with $\langle \Pi_{uv}, \xi \rangle + \langle \Pi, \xi_{uv} \rangle$ and ending with a linear combination of scalar products involving first partial derivatives of Π, ξ, $\mathbf{x}$ and $\mathbf{y}$. These first derivatives are already known since Π and ξ are completely determined when $u = 1$ or $v = 1$. For $\langle \Pi_{uv}, \xi \rangle + \langle \Pi, \xi_{uv} \rangle$, we differentiate twice $\langle \Pi, \xi \rangle = 1$ in order to get $\langle \Pi_{uv}, \xi \rangle + \langle \Pi, \xi_{uv} \rangle + \langle \Pi_u, \xi_v \rangle + \langle \Pi_v, \xi_u \rangle = 0$,which leads to a complete determination of $a_{uv}(1,1)$. ■

Lemma 10 and 11 are crucial to apply the Cauchy's formula in order to estimate $F_{m,n}(r)$ for $r > mP$ and complete the proof of our main result (Theorem 2), by the virtue of Lemma

4. To do that we can use double Cauchy formula (see also Section 4 for more details regarding this method)

$$F_{m,n}(r) = \frac{1}{(2i\pi)^2} \oint \oint F_m(u,v) \frac{dudv}{u^r(u-1)v^r(v-1)} .$$

This kind of integration is rather unusual. Since $\Pr\{C_i > r, C_j > r\} \leq \Pr\{C_i + C_j > 2r\}$ we can estimate $F^*_{m,n}(r) = \sum_{i=2}^m \Pr\{C_1 + C_i > 2r\}$ which leads to a single integration

$$F^*_{m,n}(r) = \frac{1}{2i\pi} \oint F_m(u,u) \frac{1}{u^{2r}(u-1)} du .$$

So finally we can prove the following asymptotics for the tail of $F_{m,n}(r)$.

Theorem 12. *When m and τ both tend to infinity with $\tau = O(\log m)$, then*

$$F^*_{m,n}(mP + \sqrt{m\tau}) \sim \frac{m(P - 3P^2 + 2T)^{5/2}}{2(T - P^2)\sqrt{\pi\tau^3}} \exp[-\frac{\tau}{P - 3P^2 + 2T}] . \tag{15}$$

Proof. We parallel the proof of Lemma 4. By Cauchy and (12) we have

$$F^*_{m,n}(r) \sim \frac{1}{2i\pi} \oint \frac{\lambda^m(u,u)(a(u,u) - a^m(u,u))}{u^{2r}(1 - a(u,u))} \frac{du}{u-1} , \tag{16}$$

the integration path encircling the unit disk. Let $1 + h = u$, then using Lemmas 10 and 11 we can expand as follows

$$\begin{aligned} \log(\lambda^m(u,u)/u^{2r}) = & \quad m\log(1 + 2hP + h^2(2T - P^2) + O(h^3)) - 2r\log(1+h) \\ = & \quad -2(r - mP)h + (r + 2mT - 3mP^2)h^2 + O((m+r)h^3) \\ = & \quad -\frac{(r-mP)^2}{r - m(3P^2 - 2T)} + (r - 3mP^2 + 2mT)(h - h_0)^2 + O((m+r)h^3) , \end{aligned}$$

with $h_0 = (r - mP)/(r - 3mP^2 + 2mT)$. Let $r = mP + \sqrt{m}x$. Substituting $h = h_0 + it/\sqrt{m}$, and using $1 - a(1 + h, 1 + h) = h^2(T - P^2) + O(h^3)$ we get the first estimate

$$\frac{1}{2i\pi} \oint \frac{\lambda^m(u,u)a(u,u)du}{u^{2r}(u-1)(1 - a(u,u))} =$$

$$= \exp[-\frac{x^2}{P - 3P^2 + 2T}] \frac{m}{2(T - P^2)\pi} \int \frac{\exp[-(P - 3P^2 + 2T)t^2]}{(\frac{x}{P - 3P^2 + 2T} + it)^3} dt(1 + O(1/\sqrt{m})) .$$

Since $x = O(\sqrt{\log n}) \to \infty$, we obtain

$$\frac{1}{2i\pi} \oint \frac{\lambda^m(u,u)a(u,u)du}{u^{2r}(u-1)(1 - a(u,u))} \sim \frac{m(P - 3P^2 + 2T)^{5/2}}{2(T - P^2)\sqrt{\pi}x^3} \exp[-\frac{x^2}{P - 3P^2 + 2T}] . \tag{17}$$

It remains to evaluate the second term in (16), that is,

$$\frac{1}{2i\pi}\oint \frac{\lambda^m(u,u)a^m(u,u)\mathrm{d}u}{u^{2r}(u-1)(1-a(u,u))} \,. \tag{18}$$

Using the estimates from Lemmas 10 and 11 we find

$$\log(\lambda^m(u,u)a^m(u,u)/u^{2r}) = -(r-mP)h + (r+mT-2mP^2)h^2 + O((m+r)h^3) \,,$$

hence (18) becomes

$$\frac{1}{2i\pi}\oint \frac{\lambda^m(u,u)a^m(u,u)\mathrm{d}u}{u^{2r}(u-1)(1-a(u,u))} \sim \frac{m(P-2P^2+T)^{5/2}}{2\sqrt{\pi}x^3}\exp[-\frac{x^2}{P-2P^2+T}] \,. \tag{19}$$

Since $P-2P^2+T < P-3P^2+2T$ in the non-uniform case, the exponent in (19) is larger than in (17), so the latter is the leading term in the asymptotic expansion of $F^*_{m,n}(r)$. This concludes the proof. ∎

The rest is easy. For an ultimate proof of Theorem 2 we need only to put together all the estimates we have obtained so far. In particular, Lemma 4 requires to verify the following (cf. (13))

$$\lim_{n\to\infty}\frac{m}{n}\nu_n(\varepsilon) = 0 \tag{20}$$

where $\nu_n(\varepsilon)$ is defined as

$$\nu_n(\varepsilon) = \frac{F_{m,n}(mP+\sqrt{ma_n}(1-\varepsilon))}{m(\Pr\{C_1 > mP+\sqrt{ma_n}(1-\varepsilon)\})^2} \,.$$

Our previous estimates provided in Lemma 3 and Theorem 12 suggest the following bound

$$\nu_n(\varepsilon) \le \frac{(P-3P^2+2T)^{5/2}(P-P^2)}{(1-\varepsilon)(T-P^2)}\sqrt{\frac{\pi}{a_n}}\exp[(1-\varepsilon)^2\frac{2a_n(T-P^2)}{(P-P^2)(P-3P^2+2T)}] \,.$$

with $a_n \sim 2\log n(P-P^2)$. After some additional algebra one shows

$$\nu_n(\varepsilon) \le \frac{(P-3P^2+2T)^{3/2}(P-P^2)^{1/2}}{(T-P^2)(1-\varepsilon)^2}\sqrt{\frac{\pi}{2\log n}}n^{4(1-\varepsilon)\frac{T-P^2}{P-3P^2+2T}} \,. \tag{21}$$

Therefore, for $m = O(n^\alpha)$ (20) and (21) imply

$$\frac{m}{n}\nu_n(\varepsilon) = O(1/\log n)n^{\alpha-1+4(1-\varepsilon)\frac{T-P^2}{P-3P^2+2T}}$$

and this tends to zero provided $\alpha < 1-4(T-P^2)/(P-3P^2+2T)$ as required in Theorem 2(ii). Part (iii) of Theorem 2 follows from the above too. This completes the proof of our main result.

Table 1: Simulation and theoretical results for uniform alphabet.

V	n=1000, m=300		n=2000, m=1000		n=4000, m=2000		n=8000, m=4000	
	Exact	Estimate	Exact	Estimate	Exact	Estimate	Exact	Estimate
10	49	49.3	132	136.9	243	254.6	477	480.4
20	27	29.0	73	76.8	141	139.7	245	258.4
30	25	21.5	57	55.1	100	98.7	191	181.1
40	17	17.1	42	44.2	78	78.4	139	141.8
50	15	15.0	35	37.2	64	65.5	118	117.5
55	13	13.2	32	34.4	58	60.3	108	107.8

4. FURTHER REMARKS

Our analysis assumes a large n, however we were pleasantly surprised, when experimenting with the method, that the algorithm gives a good estimate of $M_{m,n}$ *even for moderate values of* n. This is illustrated in the table below which compares our theoretical estimate from Theorem 2 with results obtained in a computer simulation for uniform alphabet. We note that for n in the hundreds, the estimate is typically within 5% of the true value. Very large values of n do arise in many of the application areas of this problem, notably, in text processing, speech recognition, machine vision and, last but not least, molecular sequence comparison. Furthermore, our results indicate that the estimate improves with higher values of V which is a particularly desired property (cf. [1]). Finally, the reader may conclude from the table that **b** is very unlikely to "almost occur" in **a** in the case of uniform alphabet, and this is in fact apparent if we set $p_i = 1/V$ in Theorem 2. However, in the nonuniform case one can draw no such conclusion (in fact if the alphabet is English and $p_\sigma = 1$ if $\sigma = z$ and zero otherwise, then **b** occurs everywhere in **a** !).

The answer returned by the algorithm can be interpreted as a rough measure of the extent to which the pattern occurs in the text, but it does not tell us *where* it occurs. To find out where, one would have to use the $O(n\sqrt{m}\text{polylog}(m))$ time (worst-case) algorithm of Abrahamson [1]. There are also practical situations where one only needs to know whether the pattern occurs, not where it occurs. Finally, when a pattern matching is performed on a huge database, then our algorithm can quickly rule out places in the database where the pattern is unlikely to occur.

Regarding future work, one may investigate when Theorem 2 can be extended to the

case where α is in the region of the interval $[0,1]$ not covered by the current statement of Theorem 2 (cf. [5]). In particular, one might be interested in finding the matching upper bound in Theorem 2(iii). Another problem worth investigating is how much our result relies on the assumption $m = O(n^\alpha)$. In particular, one should consider $m = f(n)$ for some function $f(\cdot)$ and see how this modifies our result (see [5] and [4] for examples of such functions f).

References

[1] K. Abrahamson, Generalized String Matching, *SIAM J. Comput.*, 16, 1039-1051, 1987.

[2] Abramowitz, M. and Stegun, I., *Handbook of Mathematical Functions,* Dover, New York (1964).

[3] Aldous, D., *Probability Approximations via the Poisson Clumping Heuristic*, Springer Verlag, New York 1989.

[4] Arratia, R., Gordon, L., and Waterman, M., An Extreme Value Theory for Sequence Matching, *Annals of Statistics*, 14, 971-993, 1986.

[5] Arratia, R., Gordon, L., and Waterman, M., The Erdös-Rényi Law in Distribution, for Coin Tossing and Sequence Matching, *Annals of Statistics*, 18, 539-570, 1990.

[6] Chung, K.L. and Erdös, P., On the Application of the Borel-Cantelli Lemma, *Trans. of the American Math. Soc.*, 72, 179-186, 1952.

[7] DeLisi, C., The Human Genome Project, *American Scientist*, 76, 488-493, 1988.

[8] Feller, W., *An Introduction to Probability Theory and its Applications,* Vol. II, John Wiley & Sons, New York (1971).

[9] Flajolet, P., Analysis of Algorithms, in *Trends in Theoretical Computer Science* (ed. E. Börger), Computer Science Press, 1988.

[10] Galambos, J., *The Asymptotic Theory of Extreme Order Statistics,* John Wiley & Sons, New York (1978).

[11] Z. Galil, Open Problems in Stringology, *Combinatorial Algorithms on Words* (Eds. A. Apostolico and Z. Galil), 1-8 (1984).

[12] L. Guibas and A. Odlyzko, Periods in Strings *Journal of Combinatorial Theory*, Series A, 30, 19-43 (1981).

[13] L. Guibas and A. W. Odlyzko, String Overlaps, Pattern Matching, and Nontransitive Games, *Journal of Combinatorial Theory,* Series A, 30, 183-208 (1981).

[14] Henrici, P., *Applied and Computational Complex Analysis*, vol. I., John Wiley& Sons, New York 1974.

[15] Jacquet, P. and Szpankowski, W., Autocorrelation on Words and Its Applications. Analysis of Suffix Trees by String-Ruler Approach, INRIA Technical report No. 1106, October 1989; submitted to a journal.

[16] Karlin, S. and Ost, F., Counts of Long Aligned Matches Among Random Letter Sequences, *Adv. Appl. Probab.*, 19, 293-351, 1987.

[17] Knuth, D.E., J. Morris and V. Pratt, Fast Pattern Matching in Strings, *SIAM J. Computing*, 6, 323-350, 1977.

[18] Noble, B. and Daniel, J., *Applied Linear Algebra*, Prentice-Hall, New Jersey 1988

[19] Seneta, E., *Non-Negative Matrices and Markov Chains*, Springer-Verlag, New York 1981.

[20] Szpankowski, W., On the Height of Digital Trees and Related Problems, *Algorithmica*, 5, 1991 (in press).

[21] M. Zuker, Computer Prediction of RNA Structure, *Methods in Enzymology*, 180, 262-288, 1989.

ISSN 0249-6399