Extended abstract submitted to ICALP'01
Track A (Algorithms, Automata, Complexity and Games)
Document available at `http://algo.inria.fr/flajolet/icalp01-sub.ps`
— January 23, 2001

HIDDEN PATTERN STATISTICS*

Philippe Flajolet†, Yves Guivarc'h‡, Wojciech Szpankowski§, and Brigitte Vallée¶

Abstract. Two fundamental problems in combinatorics on words and string manipulation are string matching and sequence comparison. In string matching one searches for all occurrences of a given *string*, understood as a sequence of consecutive symbols, in a text. In sequence comparison a *subsequence* rather than a string is searched in a text The string matching problem has been extensively studied in literature from algorithmic and probabilistic points of view. The sequence comparison problem, also known as *hidden pattern* problem, is harder and it has been much less investigated. In this paper we study the number of occurrences of a given pattern w of length m as a subsequence in a random text of length n generated by a memoryless source. In particular, we consider two versions of this problem, namely the *unconstrained* one in which the subsequence w can appear anywhere in the text, and the *constrained* one that puts bounds on the distances between symbols of the word w. We determine the mean and the variance of the number of occurrences, and establish a Gaussian limit law. These results are obtained via combinatorics on words, formal languages, and methods of analytic combinatorics based on generating functions and moment methods. The motivation to study this problem comes from an attempt at finding a reliable threshold for intrusion detections, from textual data processing applications, and from molecular biology.

1. INTRODUCTION

String matching and *sequence comparison* are two basic problems of pattern matching known informally as "stringology". Hereafter, by a string we mean a sequence of consecutive symbols. In string matching, given a pattern $w = w_1 w_2 \dots w_m$ (of length m) one searches for some/all occurrences of w (as a block of consecutive symbols) in a text T_n of length n. The algorithms by Knuth–Morris–Pratt and Boyer–Moore [3, 9] provide efficient ways of finding such occurrences. Accordingly, the number of string occurrences in a random text has been intensively studied over the last two decades, with significant progress in this area being reported [4, 14, 15, 24, 26, 27, 32]. For instance Guibas and Odlyzko [14, 15] have revealed the fundamental rôle played by autocorrelation vectors and their associated polynomials. Régnier and Szpankowski [26, 27] established that the number of occurrences of a string is asymptotically normal under a diversity of models that include Markov chains. Nicodème *et al.* [24] showed more generally that the number of places in a random text at which a 'motif' (i.e., a mildly restricted regular expression pattern) terminates is asymptotically normally distributed.

In sequence comparisons, we search for a given pattern $w = w_1 w_2 \dots w_m$ in the text $T_n = t_1 t_2 \dots t_n$ as a *subsequence*, that is, we look for indices $1 \le i_1 < i_2 < \cdots < i_m \le n$ such that $t_{i_1} = w_1$, $t_{i_2} = w_2$, $\cdots$, $t_{i_m} = w_m$. We also say that the word w is *"hidden"* in the text; thus we call this the *hidden pattern* problem. For example, '*baba*' occurs as a subsequence in the text 'a<u>b</u>r<u>a</u>cada<u>b</u>r<u>a</u>', in fact three times, but not even once as a string. We can impose additional set of constraints $\mathcal{D}$ on the indices $i_1, i_2, \dots, i_m$ to record a valid subsequence occurrence: for given a family of integers d_j ($d_j > 0$ and possibly $d_j = \infty$), we have $(i_{j+1} - i_j - 1) < d_j$. In other words, the allowed lengths of the "gaps" $(i_{j+1} - i_j - 1)$ are bounded from above strictly by d_j. With $\#$ representing a 'don't-care-symbol' (similar to the unix '$\star$'-convention) and the subscript denoting a strict upper bound on the length of the associated gap, a typical pattern may look like

$$\texttt{idd}\#_5\texttt{n}\#\texttt{pat}\#_6\texttt{n}\#\texttt{s}\#_4\texttt{ti};$$

there, $\#$ abbreviates $\#_\infty$ and $\#_1$ is omitted; the meaning is that '`idd`' should occur first contiguously, followed by '`n`' with a gap of < 5 symbols, followed anywhere later in the text by '`pat`', etc. The case when all the d_j's are infinite is called the *unconstrained problem*; when all the d_j's are finite, we speak of

*This research was supported in part by sponsors of CERIAS at Purdue under contract 1419991431A, by the ALCOM-FT Project (# IST-1999-14186) of the European Union, and by NSF Grant C-CR 9804760.

†INRIA-Roquencourt, BP 105, 78 153 Le Chesnay, France
‡IRMAR, Université de Rennes I, F-35042 Rennes Cedex, France
§Department for Computer Science, Purdue University, W. Lafayette, IN 47907, U.S.A.
¶GREYC, Université de Caen, F-14032 Caen Cedex, France.

the *constrained* problem. The case where all d_j reduce to 1 gives back classical string matching as a limit case.

Motivations. Our original motivation to study this problem came from *intrusion detection* in the area of computer security. The problem is important due to the rise of attacks on computer systems. There are several approaches to intrusion detections, but, recently the pattern matching approach has found many advocates, most notably in [2, 22]. The main idea of this approach is to search in an audit file (the text) for certain patterns (known also as signatures) representing suspicious activities that might be indicative of an intrusion by an outsider, or misuse of the system by an insider. The key to this approach is to recognize that these patterns are **subsequences** because an intrusion signature specification requires the possibility of a variable number of intervening events between successive events of the signature. In practice one often needs to put some additional restrictions on the distance between the symbols in the searched subsequence, which leads to the constrained version of subsequence pattern matching. The fundamental question is *how many occurrences of a signature (subsequence) constitute a real attack?* In other words, how to set a *threshold* so that we can detect only real intrusions and avoid false alarms? It is clear that *random* (unpredictable) events occur and setting the threshold too low will lead to an unrealistic number of false alarms. On the other hand, setting the threshold too high may result in missing some attacks, which is even more dangerous. This is a fundamental problem that motivated our studies of hidden pattern statistics. By knowing the most likely number of occurrences and the probability of deviating from it, we can set a threshold such that with a small probability we miss real attacks.

Molecular biology provides another important source of applications [25, 32]. As a rule, there, one searches for sequences, not strings. Examples are in abundance: split genes where *exons* are interrupted by *introns*, starting and stopping signal in genes, etc. In general, for gene searching, the constrained hidden pattern matching (perhaps with an exotic constraint set) is the right approach for finding meaningful information about genes. The hidden pattern problem can also be viewed as a close relative of the longest common subsequence (LCS) problem, itself of immediate relevance to computational biology and still surrounded by many unresolved questions [29].

We, computer scientists and mathematicians, are certainly not the first who invented hidden words and hidden meaning [1]. Rabbi Akiva in the first century A.D. wrote a collection of documents called *Maaseh Merkava* on secret mysticism and meditations. In the eleventh century Spanish Solomon Ibn Gabirol called these secret teachings *Kabbalah*. Kabbalists organized themselves as a secret society dedicated to study of the ancient wisdom of Torah, looking for mysterious connections and hidden truth, meaning, and words in Kaballah and elsewhere (without computers!). Recent versions of this activity are *knowledge discovery and data mining, bibliographic search, lexicographic research, textual data processing*, or even *web site indexing*. Public domain utilities like `agrep`, `grappe`, `webglimpse`[1], etc, depend crucially on approximate pattern matching algorithms for subsequence detection. More generally, many interesting algorithms, based on regular expressions and automata, dynamic programming, directed acyclic word graphs, digital tries or suffix trees have been developed; see [6, 10, 21, 34] for a flavour of the diversity of approaches.

In all of the contexts mentioned above, it is of obvious interest to discern what constitutes a meaningful observation of pattern occurrences from what is merely a statistically unavoidable phenomenon (noise!). This is precisely the problem addressed here. We establish *hidden pattern statistics*—i.e., precise probabilistic information on number of occurrences of a given pattern w as a subsequence in a random text T_n generated by a memoryless source, this in the most general case (covering the constrained and unconstrained versions and many other situations). Surprisingly enough and to the best of our knowledge, there are no results in the literature that address the question at this level of generality. An immediate consequence of our results is the possibility to set *thresholds* at which appearance of a (subsequence) pattern starts being meaningful.

Results. Let Ω_n be the number of occurrences of a given pattern as a subsequence in a random text of lenght n generated by a memoryless source (i.e., symbols are drawn independently). We investigate the general case where we allow some of the gaps to be restricted, and others to be unbounded. Then the most important parameter is the quantity b defined as 1 plus the number of unbounded gaps (the number of indices j for which $d_j = \infty$); the product D of all the finite constraints d_j plays also a rôle. We obtain

[1] Developed by Manber and Wu [34], Kucherov [21], and others, see, e.g.:
`http://webglimpse.org/` and `http://www.loria.fr/˜ kucherov/SOFTWARE/grappe-3.0/`.

the mean, the variance, all moments, and finally a central limit law. Precisely, we prove that the number of occurrences has mean and variance given by

$$E[\Omega_n] \sim \frac{n^b}{b!} D\,\pi(w), \qquad \mathrm{Var}[\Omega_n] \sim \sigma^2(w) n^{2b-1}$$

where $\pi(w)$ is the probability of w, and $\sigma^2(w)$ is a computable constant that depends explicitly (though intricately) on the structure of the pattern w and the constraints (Theorem 1). Then we prove the central limit law by moment methods, that is, we show that all centred moments $(\Omega_n - \mathbf{E}[\Omega_n])/n^{b-\frac{1}{2}}$ converge to the appropriate moments of the Gaussian distribution (Theorem 2). We stress that, except in the constrained case, the difficulty of the analysis lies in a nonlinear growth of the mean and the variance so that many standard approaches to establishing the central limit law tend to fail.

For the unconstrained problem, one has $b = m$ and $D = 1$, and both the mean and the variance admit pleasantly simple closed forms (Corollary 1). For the constrained case, one has $b = 1$, while the mean and the variance become of linear growth. To visualize the dependency of $\sigma^2(w)$ of w, we observe that, when all the d_j equal 1, the problem reduces to the traditional *string* matching that was extensively studied in the past as witnessed by the (incomplete) list of references: [4, 14, 15, 24, 26, 27, 32]. It is well known that for string matching the variance coefficient is a function of the so called *autocorrelation* of the string. In the general case of hidden pattern matching, the autocorrelation must be replaced by a more complex quantity that depends on the way pairs of constrained occurrences may intersect (Theorem 1 and Corollary 2).

Methodology. The way we approach the probabilistic analysis is through a formal description of situations of interest by means of regular languages. Basically we describe *contexts* of one, two, or several occurrences by means of regular languages and this gives the needed informations relative to expectation, variance, and higher moments, respectively. A systematic translation into *generating functions* is available by the methods of analytic combinatorics and the original Chomsky-Schützenberger theorem. Then, the structure of the generating functions at the pole $z = 1$ provides the necessary asymptotic informations. In fact, there is an important phenomenon of *asymptotic simplification* where the essentials of combinatorial-probabilistic features are reflected by the singular forms of generating functions. For instance, variance coefficients come out naturally from this approach together with, for each case, a suitable notion of correlation; higher moments are seen to arise from a fundamental singular symmetry of the generating functions, and this fact eventually carries with it the existence the possibility of estimating moments. From there Gaussian laws eventually result by basic moment convergence theorems. Perhaps the originality of the present approach lies in a joint use of combinatorial-enumerative techniques and of analytic-probabilistic methods.

2. Framework

A *pattern* $w = w_1 \cdots w_m$ of length m is fixed once and for all. The number of occurrences is then defined as follows.

Definition 1. *An element $\mathcal{D} = (d_1, \ldots, d_{m-1}) \in (\mathbb{N} \cup \{\infty\})^{m-1}$ is called a* constraint. *An m-tuple $I = (i_1, i_2, \ldots, i_m)$ ($1 \le i_1 < i_2 < \cdots < i_m$) satisfies the constraint $\mathcal{D}$ if $(i_{j+1} - i_j - 1) < d_j$, in which case it is called a* position. *An* occurrence *of pattern w in the text $T_n = t_1 \ldots t_n$ of length n subject to the constraint $\mathcal{D}$ is a position $I = (i_1, i_2, \ldots, i_m)$ (satisfying $\mathcal{D}$) such that $t_{i_1} = w_1$, $t_{i_2} = w_2, \ldots, t_{i_m} = w_m$. For a text T_n of length n, we let $\Omega_n(\mathcal{D})$ represent the number of occurrences (of w) subject to the constraint $\mathcal{D}$.*

The case $\mathcal{D} = (\infty, \ldots, \infty)$ models the *unconstrained problem*; at the other extreme of the spectrum, there lies the case where all d_j are finite, which we name the *constrained problem*. The subset of indices j for which d_j is unbounded ($d_j = \infty$) is denoted by $\mathcal{U}$, and we set $b = 1 + \mathrm{card}(\mathcal{U})$; the subset of indices j for which d_j is finite ($d_j < \infty$) is denoted by $\mathcal{F}$, its cardinality equals $m - b$. The two extreme values of b, namely, $b = m$ and $b = 1$, thus describe the unconstrained and the constrained problem respectively. Let $\mathcal{P}_n(\mathcal{D})$ be the set of all positions subject to the separation constraint $\mathcal{D}$, satisfying furthermore $i_m \le n$. Let also $\mathcal{P}(\mathcal{D}) = \bigcup_n \mathcal{P}_n(\mathcal{D})$. For any element $I \in \mathcal{P}_n(\mathcal{D})$, we can define the characteristic variable

$$X_I := [\![w \text{ occurs at position } I \text{ in } T_n]\!], \tag{1}$$

where (following Iverson's notation) $[\![B]\!]$ is equal to 1 if the property B holds, and to 0 otherwise. Then the number of occurrences of w in T_n under the constraint $\mathcal{D}$ is a sum of characteristic variables

$$\Omega_n(\mathcal{D}) = \sum_{I \in \mathcal{P}_n(\mathcal{D})} X_I. \tag{2}$$

Probabilistic model. As regards the probabilistic model, we consider a source that emits symbols of the text independently from the fixed alphabet $A = \{a_1, a_2, \dots a_r\}$ and denote by p_α $(0 < p_\alpha < 1)$ the probability of the symbol $\alpha \in A$. For a given length n, a random *text*, denoted by T_n is drawn according to what is known as the *memoryless source* (or Bernoulli model) that is defined by the product probability on A^n,

$$\pi(t_1 \cdots t_n) = \prod_{i=1}^{n} p_{t_i}. \tag{3}$$

The pattern probability $\pi(w)$ is defined similarly by the product formula (3) and it surfaces throughout the analysis. Under the memoryless model of random text, the quantity $\Omega_n(\mathcal{D})$ becomes a *random variable* that is itself a sum (2) of correlated random variables X_I (defined in (1)) for all allowable $I \in \mathcal{P}_n(\mathcal{D})$.

Generating functions. We shall consider throughout this paper structures superimposed on words. For $\mathcal{V}$ a class of structures and given a weight function c (usually, c will be a weight induced by the probabilities of individual letters), we introduce the *generating function*

$$V(z) \equiv \sum_n V_n z^n := \sum_{v \in \mathcal{V}} c(v) z^{|v|},$$

where $|v|$ denotes size (usually the number of letters involved in the structure's construction). Then[2], $V_n = [z^n]V(z)$ is the total weight of all structures of size n. It is then known that disjoint unions and Cartesian products correspond respectively to sums and products of generating functions; see [12, 28, 30] for a general framework. Such correspondences make it possible to translate symbolically combinatorial descriptions into generating function equations and a great use is made of this in what follows.

Aggregates and blocks. Aggregates and blocks to be introduced now are essential in the analysis of variance and of higher moments. Given a constraint $\mathcal{D}$ and a position $I = (i_1, i_2, \dots, i_m)$ that satisfies it, we define the *aggregate* of I, denoted by $\alpha(I)$, as follows. First, associate to each index i_j of I an interval A_j of $\mathbb{N}$ by

$$\text{if } (d_j < \infty \text{ and } j < m), \text{ then } A_j := [i_j, i_{j+1}], \text{ else } A_j := [i_j].$$

In this way a system of intervals (some intersecting at their boundaries and some not) encodes I. Then the *aggregation* of the A_j (and also of I) is obtained by scanning the list of A_j and successively merging together all intersecting intervals. It is then seen that $\alpha(I)$ is composed of exactly b disjoint intervals and these are called *blocks*.

For instance, when $\mathcal{D} = (3, 2, \infty, 1, \infty, \infty, 4, \infty)$, taking $I = (5, 7, 9, 18, 19, 22, 30, 33, 50)$, the system of intervals and the resulting aggregate are

$$\begin{array}{rcl} (A_1, A_2, \dots A_m) & = & (\overbrace{[5,7],[7,9],[9]}, \overbrace{[18,19],[19]}, [22], \overbrace{[30,33],[33]}, [50]) \\ \alpha(I) & = & ([5,9], \quad [18,19], \quad [22], \quad [30,33], \quad [50]). \end{array}$$

3. Mean and Variance Estimates

3.1. **The mean number of occurrences.** The first moment analysis is easily obtained by describing the collection of all occurrences by means of words with some additional structure added; this description then involves extended regular expressions (with Cartesian products replacing catenation products). Let $\mathcal{O}$ be the collection of all occurrences of w as a hidden word. Each occurrence can be viewed as a "context" with an initial string, then the first letter of the pattern, then a separating string, then the second letter, etc. The collection $\mathcal{O}$ is then described by

$$\mathcal{O} = A^\star \times \{w_1\} \times A^{<d_1} \times \{w_2\} \times A^{<d_2} \times \dots \times \{w_{m-1}\} \times A^{<d_{m-1}} \times \{w_m\} \times A^\star. \tag{4}$$

---

[2]The notation (popularized by Graham, Knuth, and Patashnik) $[z^n]f(z)$ represents the coefficient of z^n in the series $f(z)$.

For $d < \infty$, $A^{<d}$ denotes the collection of all the words of length strictly less d, whereas, for $d = \infty$, $A^{<\infty}$ denotes the collection of all finite words,

$$A^{<d} := \sum_{i<d} A^i, \qquad A^{<\infty} := A^\star = \sum_{i<\infty} A^i.$$

The associated generating functions are

$$A_d(z) = 1 + z + z^2 + \cdots + z^{d-1} = \frac{1 - z^d}{1 - z}, \qquad A_\infty(z) = 1 + z + z^2 + \cdots + z^{d-1} + \cdots = \frac{1}{1 - z}.$$

Here, we weight each occurrence by a quantity equal to $E[Y_I] = \pi(w)$, so that the associated generating function $O(z)$ of $\mathcal{O}$ is

$$(5) \qquad \begin{aligned} O(z) &= \frac{1}{1-z} \times \left(\prod_{i=1}^{m} p_{w_i} z \right) \times \left(\prod_{i \in \mathcal{F}} \frac{1 - z^{d_i}}{1 - z} \right) \times \left(\frac{1}{1-z} \right)^{b-1} \times \frac{1}{1-z}, \\ &= \left(\frac{1}{1-z} \right)^{b+1} \times \left(\prod_{i=1}^{m} p_{w_i} z \right) \times \left(\prod_{i \in \mathcal{F}} \frac{1 - z^{d_i}}{1 - z} \right). \end{aligned}$$

This coincides with the generating function of the expectation $\mathbf{E}[\Omega_n]$, that is,

$$O(z) = \sum_{n \geq 1} \mathbf{E}[\Omega_n]\, z^n.$$

But Newton's binomial theorem,

$$[z^n] \frac{1}{(1-z)^{b+1}} = \binom{n+b}{b} = \frac{n^b}{b!} \left(1 + O(\frac{1}{n}) \right),$$

implies, with $\pi(w)$ the probability of the pattern w,

$$\mathbf{E}[\Omega_n] = \frac{n^b}{b!} \left(\prod_{i \in \mathcal{F}} d_i \right) \pi(w) \left(1 + O(\frac{1}{n}) \right).$$

3.2. **The variance.** For variance and higher moments analysis, it is essential to work with centred RV's defined as

$$(6) \qquad Y_I := X_I - \mathbf{E}[X_I] = X_I - \pi(w), \qquad \Xi_n(\mathcal{D}) := \Omega_n(\mathcal{D}) - \mathbf{E}[\Omega_n(\mathcal{D})] = \sum_{I \in \mathcal{P}_n(\mathcal{D})} Y_I.$$

The second moment of the centred variable $\Xi_n(\mathcal{D})$ equals the variance of $\Omega_n(\mathcal{D})$ and with the centred variables defined by (6), one has

$$\mathbf{E}[\Xi_n^2(\mathcal{D})] = \mathbf{E}\left[\left(\sum_{I \in \mathcal{P}_n(\mathcal{D})} Y_I \right)^2 \right] = \sum_{I,J \in \mathcal{P}_n(\mathcal{D})} \mathbf{E}[Y_I Y_J].$$

There are two kinds of pairs (I, J) according as they intersect or not. When I and J do not intersect, the corresponding RV's Y_I and Y_J are independent, and the corresponding covariance $E[Y_I Y_J]$ reduces to 0. It is thus sufficient to consider intersecting subsets I and J, that is,

$$(7) \qquad \mathbf{E}[\Xi_n^2(\mathcal{D})] = \sum_{\substack{I,J \in \mathcal{P}_n(\mathcal{D}), \\ I \cap J \neq \emptyset}} \mathbf{E}[Y_I Y_J].$$

When I and J intersect at ℓ distinct places, the k-th intersection point being the r_k-th in the natural ordering of I and the s_k-th in the natural ordering of J, the expectation $\mathbf{E}[Y_I Y_J]$ involves a correlation number $e_w(I, J)$,

$$\mathbf{E}[Y_I Y_J] = \pi^2(w)\, e_w(I, J),$$

which only depends on the pairs $(r_1, s_1), (r_2, s_2), \ldots, (r_\ell, s_\ell)$ under the form

$$(8) \qquad e_w(I, J) = \left(\prod_{k=1}^{\ell} \frac{[\![w_{r_k} = w_{s_k}]\!]}{p_{w_{r_k}}} \right) - 1.$$

In this case, we take the pair of occurrences relative to (I, J) as weighted by $E[Y_I Y_J] = e_w(I, J)$, and consider the collection $\mathcal{O}_2$ of pairs of intersecting occurrences. The associated generating function $O_2(z)$ coincides with the generating function of the expectations $E[Y_I Y_J]$, that is,

$$O_2(z) = \sum_{n \geq 1} z^n \sum_{\substack{I, J \in \mathcal{P}_n(\mathcal{D}), \\ I \cap J \neq \emptyset}} E[Y_I Y_J].$$

We define the *aggregate* $\alpha(I, J)$ to be the system of intervals obtained by aggregation of the collection $\alpha(I) \cup \alpha(J)$ according to the process defined at the end of Section 2; the number of blocks $\beta(I, J)$ of $\alpha(I, J)$ plays a fundamental rôle here. Since I and J intersect, there exists at least one block of $\alpha(I)$ that intersects a block of $\alpha(J)$, so that $\beta(I, J)$ is at most equal to $2b - 1$. Realizing that the relative rather than absolute values of I and J play the prime role, we say that a (I, J) of $\mathcal{P}_q(\mathcal{D}) \times \mathcal{P}_q(\mathcal{D})$ is *full* if it is intersecting and the aggregate $\alpha(I, J)$ completely covers the interval $[1, q]$. (Clearly, the possible values of q are finite.) We denote by $\mathcal{B}_2^{[p]}$ (with $p \geq 1$) the following collection:

(9) $$\mathcal{B}_2^{[p]} := \{(I, J) \mid (I, J) \text{ is full and } \beta(I, J) = 2b - p\}.$$

Next, we group the sets I, J according to the value of $\beta(I, J)$ and write $\mathcal{O}_2^{[p]}$ for the collection of pairs of occurrences relative to intersecting pairs (I, J) of positions for which $\beta(I, J)$ equals $2b - p$. Then the collection $\mathcal{O}_2^{[p]}$ can be described as ($\cong$ represents combinatorial isomorphism)

$$\mathcal{O}_2^{[p]} \cong (A^\star)^{2b-p+1} \times \mathcal{B}_2^{[p]}.$$

The generating function of $\mathcal{O}_2^{[p]}$ is accordingly

$$O_2^{[p]}(z) = \left(\frac{1}{1-z}\right)^{2b-p+1} \times B_2^{[p]}(z).$$

Here, $B_2^{[p]}(z)$ is the generating function of all full pairs of occurrences whose aggregate has a number of blocks equal to $2b - p$, and from our earlier discussion, it is a *polynomial*—in fact of degree at most $2d_{\max}(m-1)$, where $d_{\max} = \max_{i \in \mathcal{F}} d_i$. Now, an easy dominant pole analysis entails that

$$[z^n]O_2^{[p]} = \frac{n^{2b-p}}{(2b-p)!}\, \beta_2^{[p]} \left(1 + O(\frac{1}{n})\right),$$

where $\beta_2^{[p]} = B_2^{[p]}(1)$ is the total weight of the collection $\mathcal{B}_2^{[p]}$. Since the generating function $O_2^{[p]}$ coincides with the generating function of the expectations,

$$O_2^{[p]}(z) = \sum_{n \geq 0} z^n \sum_{\substack{I, J \in \mathcal{P}_n, I \cap J \neq \emptyset, \\ \beta(I,J) = 2b-p}} \mathbf{E}[Y_I Y_J],$$

it is the term $[z^n]O_2^{[1]}$ that gives the dominant contribution to the variance. Indeed, its contribution is $O(n^{2b-1})$. The polynomial $B_2^{[1]}(z)$ is conceptually an extension of Guibas and Odlyzko's autocorrelation polynomial. It is the generating function of all full pairs of occurrences whose aggregate has a number of blocks equal to $2b - 1$. The constant $\beta_2^{[1]} = B_2^{[1]}(z)$ is in particular the total weight of the collection $\mathcal{B}_2^{[1]}$,

$$\beta_2^{[1]} = \pi(w)^2 \sum_{(I,J) \in \mathcal{B}_2^{[1]}} e_w(I, J).$$

In summary, we have found

(10) $$\mathbf{E}[\Xi_n^2] \sim [z^n]O_2^{[1]} \sim \frac{n^{2b-1}}{(2b-1)!}\, \pi(w)^2 \sum_{(I,J) \in \mathcal{B}_2^{[1]}} e_w(I, J)$$

with the correlation coefficients defined in (8). The relative error in the estimate is clearly $O(1/n)$. Also, the standard deviation is of an order, $O(n^{b-1/2})$, that is smaller than the mean, $O(n^b)$, a fact that entails concentration of distribution (by a well-known argument based on Chebyshev's inequalities). In summary:

Theorem 1. *Consider a general constraint $\mathcal{D}$ and the number of occurrences $\Omega_n \equiv \Omega_n(\mathcal{D})$. The mean and variance of Ω_n satisfy*

$$\mathbf{E}[\Omega_n] = \frac{\pi(w)}{b!}\left(\prod_{j\in\mathcal{F}} d_j\right) n^b \left(1 + O(\frac{1}{n})\right), \qquad \mathbf{Var}[\Omega_n] = \sigma^2 n^{2b-1}\left(1 + O(\frac{1}{n})\right),$$

where $\mathcal{F}$ is the set of j such that $d_j < \infty$, and the "variance coefficient" σ^2 is given by

$$\sigma^2 := \frac{\pi(w)^2}{(2b-1)!}\left(\sum_{(I,J)\in\mathcal{B}_2^{[1]}} e_w(I,J)\right), \tag{11}$$

with the correlation coefficients $e_w(I,J)$ defined in (8) *and the full pairs $\mathcal{B}_2^{[1]}$ defined by* (9)*. Consequently, the distribution of the random variable Ω_n converges in probability:*

$$\textit{for any } \epsilon > 0, \qquad \lim_{n\to\infty} \Pr\left\{\left|\frac{\Omega_n}{\mathbf{E}[\Omega_n]} - 1\right| < \epsilon\right\} = 1. \tag{12}$$

(Note that dynamic programming makes it possible to evaluate easily the variance coefficient σ^2 for any given pattern.)

3.3. **Two particular cases.** The previous estimates can be made somewhat explicit in the two extreme situations: the unconstrained case, and the constrained case.

In the *unconstrained case*, the generating function of the mean simplifies considerably. Combinatorially, there are exactly m blocks, each of them being a singleton. The dominant term of the variance is then given by intersecting pairs (I,J) that intersect at exactly one place. We abbreviate as $I \cap J \rightsquigarrow (r,s)$ the fact that I and J intersect at exactly one point and that this point is of rank r in the natural ordering of I and of rank s in J. In that case, we shall say that I and J "*join*" at (r,s) and the pair (r,s) will be referred to as the "joining place" of I and J. The value of the coefficient $e_w(I,J)$ depends only on the joining place (r,s),

$$e_w(I,J) = e_w(r,s) = \frac{[\![w_r = w_s]\!]}{p_{w_r}} - 1, \tag{13}$$

so that the generating function $B_2^{[1]}(z)$ is of the form

$$B_2^{[1]}(z) = z^{2m-1}\,\pi(w)^2 \sum_{1\le r,s\le m} M_{m,r,s}\, e_w(r,s) \tag{14}$$

where $M_{m,r,s}$ is the number of pairs of occurrences of $\mathcal{B}_2^{[1]}$ that join at (r,s). It is clear that

$$M_{m,r,s} = \binom{r+s-2}{r-1}\binom{2m-r-s}{m-r}. \tag{15}$$

In words: since the pivot (i.e., the intersection of the occurrences) has a fixed rank equal to $r+s-1$, then, amongst the $r+s-2$ elements smaller than the pivot, assign freely $r-1$ to the first occurrence and the remaining $s-1$ to the second; proceed similarly for the $2m-r-s$ elements larger than the pivot.

We have thus found that

$$\mathbf{E}[\Xi_n^2] \sim O_{2,n}^{[1]} \sim \frac{n^{2m-1}}{(2m-1)!}\,\pi(w)^2 \sum_{1\le r,s\le m} M_{m,r,s}\, e_w(r,s), \tag{16}$$

with the correlation coefficients defined in (13). Hence:

Corollary 1. *Consider the unconstrained problem $\mathcal{D} = (\infty,\ldots,\infty)$ and the number of occurrences $\Omega_n \equiv \Omega_n(\mathcal{D})$. The mean and variance of Ω_n satisfy*

$$\mathbf{E}[\Omega_n] = \binom{n}{m}\pi(w) \sim \frac{\pi(w)}{m!}n^m, \qquad \mathbf{Var}[\Omega_n] = \sigma^2 n^{2m-1}\left(1 + O(\frac{1}{n})\right),$$

where the "variance coefficient" σ^2 is given by

$$\sigma^2 := \frac{\pi(w)^2}{(2m-1)!}\left(\sum_{1\le i,j\le m}\binom{i+j-2}{i-1}\binom{2m-i-j}{m-i}\left(\frac{[\![w_i=w_j]\!]}{p_{w_i}}-1\right)\right). \tag{17}$$

Regarding the *constrained case*, it is possible to give complicated expressions from the mean in the form of binomial convolutions that result from Eq. (5). The main parameter b now equals 1, so that the mean and the variance are both of order $O(n)$. Furthermore, the number of blocks $\beta(I,J)$ of intersecting pairs (I,J) always equals 1. We have the generalized "autocorrelation polynomial"

$$B_2^{[1]}(z) = \pi(w)^2\sum_{k=m}^{2d(m-1)} z^k\left(\sum_{\substack{(I,J)\text{ full}\\ |I\cup J|=k}} e_w(I,J)\right). \tag{18}$$

Thus the "variance coefficient" $\sigma^2 := B_2^{[1]}(1)$ is

$$\sigma^2 = \pi(w)^2\sum_{k=m}^{2d(m-1)}\left(\sum_{\substack{(I,J)\text{ full}\\ |I\cup J|=k}} e_w(I,J)\right) = \pi(w)^2\left(\sum_{\ell=1}^{m}\sum_{\substack{(I,J)\text{ full}\\ |I\cap J|=\ell}} e_w(I,J)\right), \tag{19}$$

the correlation number $e_w(I,J)$ being given by (8).

Corollary 2. *Consider the constrained problem $\mathcal{D}=(d_1,\dots,d_m)$ and the number of occurrences $\Omega_n \equiv \Omega_n(\mathcal{D})$. The mean and variance of Ω_n satisfy, with σ^2 defined in (19):*

$$\mathbf{E}[\Omega_n] = \pi(w)\left(\prod_{j=1}^{m-1} d_j\right)n + O(1), \qquad \mathbf{Var}[\Omega_n] = \sigma^2 n + O(1).$$

4. Central Limit Laws

In this section we establish the central limit laws for the hidden pattern matching by symbolic methods in conjunction with the moment convergence theorem that we briefly review below.

4.1. **The general method.** Our goal is to prove that Ω_n appropriately normalized tends to the standard normal distribution. We consider the following normalized random variable

$$\widetilde{\Xi}_n := \frac{\Xi_n}{n^{b-1/2}} = \frac{\Omega_n - \mathbf{E}[\Omega_n]}{n^{b-1/2}},$$

where b is the number of blocks of the constraint $\mathcal{D}$. We shall show that $\widetilde{\Xi}_n$ behaves asymptotically as a normal variable with mean 0 and standard deviation σ. By the classical *moment convergence theorem* (Theorem 30.2 of [5]), this is established once we show that all moments of $\widetilde{\Xi}_n$ converge to the appropriate moments of the standard normal distribution. We remind the reader that if G is a standard normal variable (i.e., a Gaussian distributed variable with mean 0 and standard deviation 1), then for any integral $s\ge 0$

$$\mathbf{E}[G^{2s}] = 1\cdot 3\cdots(2s-1), \qquad \mathbf{E}[G^{2s+1}] = 0. \tag{20}$$

We shall accordingly distinguish two cases based on the parity of r, $r=2s$ and $r=2s+1$, and prove that

$$\lim_{n\to+\infty}\mathbf{E}[\widetilde{\Xi}_n^{2s+1}] = 0, \qquad \lim_{n\to+\infty}\mathbf{E}[\widetilde{\Xi}_n^{2s}] = \sigma^{2s}\left(1\cdot 3\cdots(2s-1)\right), \tag{21}$$

which establishes Gaussian convergence (Theorem 2 below) for the general case.

The proof below is combinatorial. It basically reduces to grouping and enumerating adequately the various combinations of indices in the sum that expresses $\mathbf{E}[\Xi_n^r]$. Once more, $\mathcal{P}_n(\mathcal{D})$ is formed of all the positions of $[1,n]$ subject to the constraint $\mathcal{D}$ and $\mathcal{P}(\mathcal{D})=\bigcup_n \mathcal{P}_n(\mathcal{D})$ Then totally distributing the terms in $\Xi_n^r(\mathcal{D}) = (\sum_I Y_I)^r$ yields $\mathbf{E}[\Xi_n^r] = \sum \mathbf{E}[Y_{I_1}\cdots Y_{I_r}]$, where the sum is taken over all $I_1,\dots,I_r\in\mathcal{P}_n(\mathcal{D})$. A collection of sets $(I_1,\dots,I_r)$ in $\mathcal{P}^{(r)}(\mathcal{D}) := \mathcal{P}\times\cdots\times\mathcal{P}$ is said to be *friendly* if each I_k intersects at least one other I_ℓ, with $\ell\ne k$ and we let $\mathcal{Q}^{(r)}(\mathcal{D})$ be the set of all friendly collections in $\mathcal{P}^{(r)}(\mathcal{D})$. For $\mathcal{P}^{(r)}$, $\mathcal{Q}^{(r)}$, and their derivatives below, we add the subscript n each time the situation is particularized to texts of length n. If $(I_1,\dots,I_r)$ does not lie in $\mathcal{Q}^{(r)}(\mathcal{D})$, then $\mathbf{E}[Y_{I_1}\cdots Y_{I_r}]=0$, since at least one of the Y_I's is independent of the other factors in the product and the Y_I's have been centred, $\mathbf{E}[Y_I] = 0$. One can thus restrict attention to friendly families and get the basic formula

$$\mathbf{E}[\Xi_n^r] = \sum_{(I_1,\dots,I_r)\in\mathcal{Q}_n^{(r)}(\mathcal{D})} \mathbf{E}[Y_{I_1}\cdots Y_{I_r}], \tag{22}$$

where the expression involves fewer terms than before. From there, we proceed roughly in two stages: first, restrict attention to friendly families that give rise to the dominant contribution by introducing a suitable subfamily $\mathcal{Q}_\star^{(r)} \subset \mathcal{Q}^{(r)}$: in so doing, we prove that moments of odd order appear to be negligible. For even order r, the family $\mathcal{Q}_\star^{(r)}$ involves a symmetry; we reduce the analysis to another subfamily $\mathcal{Q}_{\star\star}^{(r)} \subset \mathcal{Q}_\star^{(r)}$ that corresponds to a "standard" form of occurrence intersections, and this reduction gives precisely rise to the Gaussian moments.

4.2. **Proof of the general result.** We operate with a general constraint $\mathcal{D}$, $\Omega_n = \Omega_n(\mathcal{D})$, and $\widetilde{\Xi}_n = \Xi_n/n^{b-1/2}$.

Theorem 2. *The random variable Ω_n satisfies a Central Limit Law:*

$$\lim_{n\to\infty} \Pr\left\{\frac{\Omega_n - \mathbf{E}[\Omega_n]}{\sqrt{\mathbf{Var}[\Omega_n]}} \le x\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}\,dt. \tag{23}$$

We start from the basic formula (22). Given $(I_1,\dots,I_r)\in\mathcal{Q}^{(r)}$, one defines the aggregate $\alpha(I_1,I_2,\dots,I_r)$ as the aggregation (in the sense of Section 2 and the variance calculation above) of $\alpha(I_1)\cup\cdots\cup\alpha(I_r)$. Next, the *block number* of $(I_1,\dots,I_r)$ is the number of blocks of the aggregate $\alpha(I_1,\dots,I_r)$. If p is the total number of intersecting blocks of the aggregate $\alpha(I_1,\dots,I_r)$, the aggregate $\alpha(I_1,I_2,\dots I_r)$ has $rb-p$ blocks. Like previously, we say that the family $(I_1,\dots,I_r)$ of $\mathcal{Q}_q^{(r)}$ is full if the aggregate $\alpha(I_1,I_2,\dots I_r)$ completely covers the interval $[1..q]$. Since the length of the aggregate is at most $rd(m-1)$, the generating function of full families is a polynomial $P_r(z)$ of degree at most $rd(m-1)$ with $d=\max_{j\in\mathcal{F}} d_j$. Then, the generating function of families of $\mathcal{Q}_n^{(r)}$ whose block number equals k is of the form

$$\left(\frac{1}{1-z}\right)^{k+1} \times P_r(z),$$

so that the number of families of $\mathcal{Q}^{(r)}$ whose block number equals k is $O(n^k)$. This observation proves that the dominant contribution to (22) arises from friendly families with a maximal block number. It is clear that the minimum number of intersecting blocks of any element of $\mathcal{Q}^{(r)}$ equals equals $\lceil r/2\rceil$, since it coincides exactly with the minimum number of edges of a graph with r vertices which contains no isolated vertex. Then the maximum block number of a friendly family equals $rb-\lceil r/2\rceil$.

In view of this fact and the remarks above regarding cardinalities, we immediately have

$$\mathbf{E}\left[\widetilde{\Xi}_n^{2s+1}\right] = O\left(\frac{n^{(2s+1)b-s-1}}{n^{(2s+1)(b-1/2)}}\right) = O(\frac{1}{\sqrt{n}})$$

which establishes the limit form of odd moments in the form of the first relation in (21).

We are thus left with estimating the even moments. The dominant term is relative to friendly families of $\mathcal{Q}^{(2s)}$ with an intersecting block number equal to s, that we denote by $\mathcal{Q}_\star^{(2s)}$. In such a family, each subset I_k intersects one and only one another subset I_k. Furthermore, if the blocks of $\alpha(I_h)$ are denoted by $B_h^{[t]}, 1\le t\le b$, there exists only one block $B_k^{[t_k]}$ of $\alpha(I_k)$ and only one block $B_\ell^{[t_\ell]}$ that contains the points of $I_k\cap I_\ell$. This defines an involution τ such that $\tau(k)=\ell$ and $\tau(\ell)=k$ for all pairs of indices (ℓ,k) for which I_k and I_ℓ intersect. Furthermore, the symmetry relation

$$\mathbf{E}[Y_{I_1}\cdots Y_{I_{2s}}] = \mathbf{E}[Y_{I_{\rho(1)}}\cdots Y_{I_{\rho(2s)}}].$$

shows that one can consider only friendly families of $\mathcal{Q}_\star^{(2s)}$ for which the involution τ is the standard one with cycles $(1,2)$, $(3,4)$, etc. For such "standard" families whose set is denoted by $\mathcal{Q}_{\star\star}^{(2s)}$, the pairs that intersect are (I_1,I_2), $\dots$, (I_{2s-1},I_{2s}). Since the set $\mathcal{K}_{2s}$ of involutions of $2s$ elements has cardinality

$K_{2s} = 1 \cdot 3 \cdot 5 \cdots (2s-1)$, the equality

$$\sum_{\mathcal{Q}^{(2s)}_{\star n}} \mathbf{E}[Y_{I_1} \cdots Y_{I_{2s}}] = K_{2s} \sum_{\mathcal{Q}^{(2s)}_{\star\star n}} \mathbf{E}[Y_{I_1} \cdots Y_{I_{2s}}], \tag{24}$$

entails that we can work now solely with standard families.

The class of occurrences relative to standard families is

$$A^\star \times (A^\star)^{2sb-s-1} \times \mathcal{B}^{[s]}_{2s} \times A^\star,$$

and involves the collection $\mathcal{B}^{[s]}_{2s}$ of all full friendly $2s$-tuples of occurrences with a number of blocks equal to s. It is clear that $\mathcal{B}^{[s]}_{2s}$ is exactly a shuffle product of s copies of $\mathcal{B}^{[1]}_{2}$ introduced in the study of the variance. The associated generating function is

$$\left(\frac{1}{1-z}\right)^{2sb-s+1} (2sb-s)! \left(\frac{B^{[1]}_2(z)}{(2b-1)!}\right)^s,$$

where $B^{[1]}_2(z)$ is the autocorrelation polynomial introduced in the study of the variance. Upon taking coefficients, we obtain the final estimates that proves Theorem 2:

$$\sum_{\mathcal{Q}^{(2s)}_{n,\star\star}} \mathbf{E}[Y_{I_1} \cdots Y_{I_{2s}}] \sim n^{(2b-1)s} \sigma^{2s}.$$

In view of the above this yields the estimate of even moments and leads to the second relation of (21). We have thus obtained a proof of the moment convergence estimates (21), the error term (for moments) being of order $O(1/n)$. This completes the proof of Theorem 2.

5. Refined Estimates and Further Results

Probability estimates as obtained before raise a number of interesting questions like: *How far is the finite–n regime from the asymptotic regime? Can* local *(instead of central) probability estimates be derived? What is the status of large deviations from expected values?*

At the moment, we do not have definite answers to offer regarding the most general case. We only note that the unconstrained problem can be easily rephrased as one concerning products of random matrices. However, all corresponding eigenvalues are equal to one, so that the standard method developed for random product of matrices (cf. [7, 18]) do not work. Then, another line of research in random matrices, that of random walks on nilpotent Lie groups [16, 31], is likely to be applicable, and should lead to answers to the questions above (work in progress).

The constrained problem is easier to deal with, given the closeness with classical string enumeration, finite automata, transfer matrices, Markov chains, and Perron-Frobenius theory. For lack of space, we only mention some of the key steps in a simplified case. Let $d < \infty$ be fixed and assume that each $d_j = d$; also, we take a binary alphabet with a uniform distribution of letters. The de Bruijn matrix $\mathbf{B}$ is the $2^{md} \times 2^{md}$ matrix that is the adjacency matrix of the de Bruijn graph/automaton that records the last factor of length md seen in the text; see for instance [11] and [19, Ex. 2.3.4.2.23]. Then, it can be seen that there is a diagonal matrix $\Delta(u)$ with diagonal entries each a monomial in u such that $\mathbf{C}(u) = \mathbf{B} \cdot \Delta(u)$ "generates" the constrained occurrence counts via the powers $\mathbf{C}(u)^n$. Perron-Frobenius theory applies to the matrix $\mathbf{C}(u)$ for $u > 0$. Then analytic perturbation theory applied to the dominant eigenvalue $\lambda(u)$ (a piecewise algebraic function) yields fairly complete answers to the questions above.

Theorem 3. *In the constrained case, the random variable Ω satisfies a Local Limit Law,*

$$\Pr\left\{\Omega_n = \left\lfloor \mathbf{E}[\Omega_n] + x\sqrt{\mathbf{Var}[\Omega_n]}\right\rfloor\right\} \sim \frac{1}{\sqrt{n}} \left(\frac{e^{-x^2/2}}{\sqrt{2\pi}}\right).$$

(The proof involves additionally the saddle point method; see, e.g., [8, 12, 30].)

Theorem 4. *In the constrained case, the speed of convergence to the limit law in either the central or local limit laws (Theorems 2 and 3) is $1/\sqrt{n}$. For instance:*

$$\Pr\left\{\frac{\Omega_n - \mathbf{E}[\Omega_n]}{\sqrt{\mathbf{Var}[\Omega_n]}} \le x\right\} = \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt\right) + O\left(\frac{1}{\sqrt{n}}\right).$$

(The proof involves additionally the Berry-Esseen inequalities; see, e.g., [12, 17, 30].)

Large deviations results can also be derived in this analytic framework. The bounds obtained are exponentially small but not very constructive since a determination of the constants involves diagonalizing a huge parameterized matrix. Weaker but fully constructive bounds can be based on the Azuma inequality [23, 30]. Let Ω_n' be the number of w occurrences in the text T_n' that is the variant of T_n in which *one* symbol is replaced by an independent copy. Since $|\Omega_n - \Omega_n')| \le md^m$, then by Theorem 2 and Azuma's inequality:

Theorem 5. *In the constrained case, large deviations from the mean have exponentially small probability:*

$$\Pr\{|\Omega_n - \mathbf{E}[\Omega_n]| > \delta\mathbf{E}[\Omega_n]\} \le 2\exp\left(-K\frac{n\delta^2}{2}\right), \qquad K = \frac{\pi^2(w)}{m^2d^2}.$$

Note finally that the symbolic approach adopted here seems flexible enough so as to be applicable to Vallée's general model of dynamical sources [33]. Preliminary investigations suggest that the results would naturally adapt to such sources, which includes all the Markovian ones and mixing sources.

6. CONCLUSION

We have provided here precise estimates of the probability of occurrence of a "hidden" pattern in a random text. Since, mean value estimates are supplemented by variance, moment, and full distribution analysis, the probabilistic phenomena are quite well quantified. The analytic results can then be used to set up thresholds below which observations are likely to be meaningful and beyond which they are merely statistically unavoidable.

The reader may be curious to know whether the randomness model adopted is meaningful or not in practical circumstances. Space limitations in this extended abstract prohibit us from elaborating much and we shall content ourselves with a unique experiment based on the detection of a hidden pattern in an English text of some 120,000 characters. The experiment were conducted with our own dynamic programming implementation of (constrained and unconstrained) sequence comparison.

The complete works of Shakespeare are found under `http://the-tech.mit.edu/Shakespeare/`. We took the full text of Hamlet where all non alphabetic characters are collapsed to spaces and sequences of spaces are replaced by a single space. This gives us a (rather unpoetical looking) text that has one long line with 30,316 words and 150,373 characters: "*who s there nay answer me stand and unfold yourself long live the king bernardo he you come most carefully upon your hour [. . .]*". Stripped of its spaces (' '), the text has $n = 120,057$ characters. The pattern is "*The law is Gaussian*" [$w =$ thelawisgaussian] and its mirror image $\widetilde{w}$, corresponding to $m = 16$. Based on the empirical distribution of letter frequencies in the text, we anticipate the pattern to appear $1.330\,10^{48}$ times as a subsequence, while the observed counts are $1.365\,10^{48}$ and $1.388\,10^{48}$, a deviation of less than 4% from what is expected. Similarly, if we bound the separation distance between letters uniformly by d, analysis predicts that the pattern might start occurring near $d = 10$, while its presence is unlikely for smaller values, $d < 10$. In fact, w starts occurring at $d = 14$ while $\widetilde{w}$ starts at $d = 13$—a deviation of some 30–40% from what the model predicts. Here is a table of observed versus predicted values when d varies:

		$w =$ thelawisgaussian		$\widetilde{w} =$ naissuagsiwaleht	
d	Expected (E)	Occurred (Ω)	Ω/E	Occurred (Ω)	Ω/E
13	9.195E+01	0	0.00	18	0.19
14	2.794E+02	693	2.47	371	1.32
15	7.866E+02	1,526	5.46	2,379	3.02
18	1.211E+04	31,385	2.58	14,123	1.16
20	5.886E+04	124,499	2.11	41,066	0.69
30	2.577E+07	40,001,940	1.55	25,631,589	0.99
50	5.482E+10	76,146,232,395	1.38	48,386,404,680	0.88
∞	1.330E+48	1.36554E+48	1.03	1.38807E+48	1.04

The main conclusion is a fair fit between the theoretical model and the observed data, which vindicates our previous analyses—this even though the text chosen is quite far from being "random".

REFERENCES

[1] A. Aczel, *The Mystery of the Aleph. Mathematics, the Kabbalah, and the Search for Infinity*, Four Walls Eight Windows, New York, 2000.

[2] A. Apostolico and M. Atallah, Compact Recognizers of Episode Sequences, Submitted to *Information and Computation*.

[3] A. Apostolico and Z. Galil (Eds.), *Pattern Matching Algorithms*, Oxford University Press, New York, 1997.

[4] E. Bender and F. Kochman, The Distribution of Subword Counts is Usually Normal, *European Journal of Combinatorics*, 14, 265-275, 1993.
[5] P. Billingsley, *Probability and Measure*, Second Edition, John Wiley & Sons, New York, 1986.
[6] L. Boasson, P. Cegielski, I. Guessarian, and Yuri Matiyasevich, Window-Accumulated Subsequence Matching Problem is Linear, In *Proceedings of the Eighteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems: PODS 1999*, ACM Press, 327–336, 1999.
[7] P. Bougerol and J. Lacroix, *Products of Random Matrices with Applications to Schrödinger Operators*, Birhäuser, Boston, 1985.
[8] N. G. de Bruijn, *Asymptotic Methods in Analysis*, Dover, 1981.
[9] M. Crochemore and W. Rytter, *Text Algorithms*, Oxford University Press, New York, 1994.
[10] G. Das, R. Fleischer, L. G asieniec, D. Gunopulos, J. Kärkkäinen, Episode Matching, In *Combinatorial Pattern Matching, 8th Annual Symposium*, *Lecture Notes in Computer Science* vol. 1264, 12–27, 1997.
[11] P. Flajolet, P. Kirschenhofer, and R. F. Tichy, Deviations from uniformity in random strings, *Probability Theory and Related Fields*, 80, 139–150, 1988.
[12] P. Flajolet, and R. Sedgewick, *Analytic Combinatorics*, In prep., 2001. (Available electronically at `http:algo.inria.fr/flajolet/Publications`.)
[13] U. Grenander, *Probabilities on Algebraic Structures*, John Wiley & Sons, New York, 1963.
[14] L. Guibas, and A. M. Odlyzko, Periods in Strings, *J. Combinatorial Theory Ser. A*, 30, 19–43, 1981.
[15] L. Guibas and A. M. Odlyzko, String Overlaps, Pattern Matching, and Nontransitive Games, *J. Combinatorial Theory Ser. A*, 30, 183–208, 1981.
[16] Y. Guivarc'h, Marches Aléatoires sur les Groupes, *Fascicule de probabilités*, Publ. Inst. Rech. Math. Rennes, 2000.
[17] Hsien-Kuei Hwang, On convergence rates in the central limit theorems for combinatorial structures, *European Journal of Combinatorics*, 19, 329–343, 1998.
[18] N. Katz and P. Sarnak, *Random Matrices, Frobenius Eigenvalues, and Monodromy*, AMS, Providence, 1999.
[19] D. E. Knuth, *The Art of Computer Programming, Fundamental Algorithms*, Vol. 1, Third Edition, Addison-Wesley, Reading, MA, 1997.
[20] D. E. Knuth, *The Art of Computer Programming. Sorting and Searching*, Vol. 3, Second Edition, Addison-Wesley, Reading, MA, 1998.
[21] G. Kucherov and M. Rusinowitch, Matching a Set of Strings with Variable Length Don't Cares, *Theoretical Computer Science* 178, 129–154, 1997.
[22] S. Kumar and E.H. Spafford, A Pattern-Matching Model for Intrusion Detection, *Proceedings of the National Computer Security Conference*, 11–21, 1994.
[23] C. McDiarmid, On the Method of Bounded Differences, in *Surveys in Combinatorics* (Ed. J. Siemons), vol 141, 148–188, London Mathematical Society Lecture Notes Series, Cambridge University Press, Cambridge, 1989.
[24] P. Nicodème, B. Salvy, and P. Flajolet, Motif Statistics, *European Symposium on Algorithms*, Lecture Notes in Computer Science, No. 1643, 194–211, 1999.
[25] P. Pevzner, *Computational Molecular Biology: An Algorithmic Approach*, MIT Press, 2000.
[26] M. Régnier, and W. Szpankowski, On the Approximate Pattern Occurrences in a Text, *Proc. Compression and Complexity of SEQUENCE'97*, IEEE Computer Society, 253–264, Positano, 1997.
[27] M. Régnier and W. Szpankowski, On pattern frequency occurrences in a Markovian sequence *Algorithmica*, 22, 631-649, 1998.
[28] R. Sedgewick, and P. Flajolet, *An Introduction to the Analysis of Algorithms*, Addison-Wesley, Reading, MA, 1995.
[29] J. M. Steele, *Probability theory and combinatorial optimization*, SIAM, Philadelphia, 1997.
[30] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, John Wiley & Sons, New York, 2001.
[31] A. D. Virtser, Limit Theorems for Compositions of Distributions on Certain Nilpotent Lie Groups, *Theory Probab. and Its Appl.*, 19, 86-105, 1974.
[32] M. Waterman, *Introduction to Computational Biology*, Chapman and Hall, London, 1995.
[33] B. Vallée, Dynamical Sources in Information Theory: Fundamental Intervals and Word Prefixes, *Algorithmica*, 29, 262–306, 2001.
[34] S. Wu and U. Manber, Fast Text Searching Allowing Errors, *Comm. ACM*, 35:10, 83–991, 1995.