Hidden Pattern Statistics

Philippe Flajolet[1], Yves Guivarc'h[2],
Wojciech Szpankowski[3], and Brigitte Vallée[4]

[1] Algorithms Project, INRIA-Rocquencourt, 78153 Le Chesnay, France
[2] IRMAR, Université de Rennes I, F-35042 Rennes Cedex, France
[3] Dept. Computer Science, Purdue University, W. Lafayette, IN 47907, U.S.A
[4] GREYC, Université de Caen, F-14032 Caen Cedex, France

Abstract. We consider the sequence comparison problem, also known as "hidden pattern" problem, where one searches for a given *subsequence* in a text (rather than a string understood as a sequence of consecutive symbols). A characteristic parameter is 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. Spacings between letters of the pattern may either be constrained or not in order to define valid occurrences. 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 language techniques, and methods of analytic combinatorics based on generating functions and convergence of moments. 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 [7] 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 [3, 9, 10, 15–17, 24]. For instance Guibas and Odlyzko [9, 10] have revealed the fundamental rôle played by autocorrelation vectors and their associated polynomials. Régnier and Szpankowski [16, 17] established that the number of occurrences of a string is asymptotically normal under a diversity of models that include Markov chains. Nicodème, Salvy, and Flajolet [15] showed generally that the number of places in a random text at which a 'motif' (i.e., a general regular expression pattern) terminates is asymptotically normally distributed.

In sequence comparisons, we search for a given pattern $\mathcal{W} = w_1 w_2 \dots w_m$ 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, `date` occurs as a subsequence in the text `hidden pattern`, in fact four times, but not

even once as a string. We can impose an additional set of constraints $\mathcal{D}$ on the indices $i_1, i_2, \ldots, i_m$ to record a valid subsequence occurrence: for a given family of integers d_j ($d_j \geq 1$, possibly $d_j = \infty$), one should have $(i_{j+1} - i_j) \leq d_j$. In other words, the allowed lengths of the "gaps" $(i_{j+1} - i_j - 1)$ should be $< d_j$. 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

$$\texttt{ab}\#_2\texttt{r}\#\texttt{ac}\#\texttt{a}\#\texttt{d}\#_4\texttt{a}\#\texttt{br}\#\texttt{a};$$

there, $\#$ abbreviates $\#_\infty$ and $\#_1$ is omitted; the meaning is that '`ab`' should occur first contiguously, followed by '`r`' with a gap of < 2 symbols, followed anywhere later in the text by '`ac`', 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 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, 14, 25]. 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 then: *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 [18, 23, 24]. As a rule, there, one searches for subsequences, not strings. Examples are in abundance: split genes where *exons* are interrupted by *introns*, *starting* and *stopping* signal in genes, *tandem repeats* in DNA, 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. 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 mystery [20].

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 *Kab-*

balah. 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` (developed by Manber and Wu [26], Kucherov [13], and others) depend crucially on approximate pattern matching algorithms for subsequence detection. Many interesting algorithms, based on regular expressions and automata, dynamic programming, directed acyclic word graphs, digital tries or suffix trees have been developed; see [5, 8, 13, 26] 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 as well as mixed 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 $\mathcal{W}$ as a subsequence in a random text of length 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 the number of unbounded gaps (the number of indices j for which $d_j = \infty$) plus 1; the product D of all the finite constraints d_j plays also a rôle. We obtain the mean, the variance, all moments, and finally a central limit law. Precisely, we prove in Theorem 1 that the number of occurrences has mean and variance given by

$$\mathbf{E}[\Omega_n] \sim \frac{n^b}{b!}\, D\, \pi(\mathcal{W}), \quad \mathbf{Var}[\Omega_n] \sim \sigma^2(\mathcal{W})\, n^{2b-1}$$

where $\pi(\mathcal{W})$ is the probability of $\mathcal{W}$, and $\sigma^2(\mathcal{W})$ is a computable constant that depends explicitly (though intricately) on the structure of the pattern $\mathcal{W}$ and the constraints. Then we prove the central limit law by moment methods, that is, we show that all centered 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 both the mean and the variance admit pleasantly simple closed forms. For the constrained case, one has $b = 1$, while the mean and the variance become of linear growth. To visualize the dependency of $\sigma^2(\mathcal{W})$ of $\mathcal{W}$, we observe that, when all the d_j equal 1, the problem reduces to traditional *string matching* that was extensively studied in the past as witnessed by the (incomplete) list of references: [3, 9, 10, 15–17, 24]. 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 (cf. Theorem 1).

Methodology. The way we approach the probabilistic analysis is through a formal description of situations of interest by means of regular languages. Basically such a description of *contexts* of one, two, or several occurrences gives access to expectation, variance, and higher moments, respectively. A systematic translation into *generating functions* is available by methods of analytic combinatorics deriving from the original Chomsky-Schützenberger theorem. Then, the structure of the implied generating functions at the pole $z = 1$ provides the necessary asymptotic information. 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 problem, a fact that eventually carries with it 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 such a joint use of combinatorial-enumerative techniques and of analytic-probabilistic methods.

2 Framework

We fix an alphabet $\mathcal{A} := \{a_1, a_2, \ldots, a_r\}$. The text is $T_n = t_1 t_2 \cdots t_n$. A particular matching problem is specified by a pair $(\mathcal{W}, \mathcal{D})$: the *pattern* $\mathcal{W} = w_1 \cdots w_m$ is a word of length m; the *constraint* $\mathcal{D} = (d_1, \ldots, d_{m-1})$ is an element of $(\mathbf{N}^+ \cup \{\infty\})^{m-1}$.

Positions and occurrences. 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 $i_{j+1} - i_j \le d_j$, in which case it is called a *position*. Let $\mathcal{P}_n(\mathcal{D})$ be the set of all positions subject to the separation constraint $\mathcal{D}$, satisfying furthermore $i_m \le n$. An *occurrence* of pattern $\mathcal{W}$ in the text T_n of length n subject to the constraint $\mathcal{D}$ is a position $I = (i_1, i_2, \ldots, i_m)$ of $\mathcal{P}_n(\mathcal{D})$ for which $t_{i_1} = w_1$, $t_{i_2} = w_2, \ldots, t_{i_m} = w_m$. For a text T_n of length n, the number of occurrences $\Omega_n(\mathcal{D})$ (of w) subject to the constraint $\mathcal{D}$ is then a sum of characteristic variables

$$\Omega_n(\mathcal{D}) = \sum_{I \in \mathcal{P}_n(\mathcal{D})} X_I, \quad \text{with} \quad X_I := [\![w \text{ occurs at position } I \text{ in } T_n]\!], \tag{1}$$

where $[\![B]\!] = 1$ if the property B holds, and $[\![B]\!] = 0$ otherwise (Iverson's notation).

Blocks and aggregates. In the general case, the subset $\mathcal{F}$ of indices j for which d_j is finite ($d_j < \infty$) has cardinality $m - b$ with $1 \le b \le m$. The two extreme values of b, namely, $b = m$ and $b = 1$, thus describe the (fully) unconstrained and the (fully) constrained problem respectively. The subset $\mathcal{U}$ of indices j for which d_j is unbounded ($d_j = \infty$) has cardinality $b - 1$. It separates the pattern $\mathcal{W}$ into b independent subpatterns that are called the blocks and are denoted by $\mathcal{W}_1, \mathcal{W}_2, \ldots \mathcal{W}_b$. All the possible d_j "inside" $\mathcal{W}_r$ are finite and form the subconstraint $\mathcal{D}_r$. In the example described in the introduction, one has $b = 6$ and the six blocks are

$\mathcal{W}_1$ =a$\#_1$b$\#_2$r, $\mathcal{W}_2$ = a$\#_1$c, $\mathcal{W}_3$= a, $\mathcal{W}_4$= d$\#_4$a, $\mathcal{W}_5$=b$\#_1$r, $\mathcal{W}_6$= a.

In the same way, an occurrence $I = (i_1, i_2, \ldots, i_m)$ of $\mathcal{W}$ subject to constraint $\mathcal{D}$ gives rise to b subpositions $I^{[1]}, I^{[2]}, \ldots I^{[b]}$, the rth term $I^{[r]}$ being an occurrence of $\mathcal{W}_r$ subject to constraint $\mathcal{D}_r$. The rth *block* $B^{[r]}$ is the closed segment whose end points are the extremal elements of $\mathcal{I}^{[r]}$, and the *aggregate* of position I, denoted by $\alpha(I)$, is the collection of these b blocks. In the example of the introduction, the position

$$I = (6, 7, 9, 18, 19, 22, 30, 33, 50, 51, 60)$$

satisfies the constraint $\mathcal{D}$ and gives rise to six subpositions,

$$I^{[1]} = (6, 7, 9), \quad I^{[2]} = (18, 19), \; I^{[3]} = 22, \; I^{[4]} = (30, 33), \; I^{[5]} = (50, 51), \; I^{[1]} = 60;$$

accordingly, the resulting aggregate $\alpha(I)$ is formed with six blocks,

$$B^{[1]} = [6, 9], \; B^{[2]} = [18, 19], \; B^{[3]} = [22], \; B^{[4]} = [30, 33], \; B^{[5]} = [50, 51], \; B^{[6]} = [60].$$

Probabilistic model. We consider a *memoryless source* that emits symbols of the text independently and denote by p_α $(0 < p_\alpha < 1)$ the probability of the symbol $\alpha \in \mathcal{A}$ being emitted. For a given length n, a random *text*, denoted by T_n is drawn according to the product probability on $\mathcal{A}^n$. For instance, the pattern probability $\pi(\mathcal{W})$ is defined by $\pi(\mathcal{W}) = \prod_{i=1}^{n} p_{w_i}$, a quantity that surfaces throughout the analysis. Under this randomness model, the quantity $\Omega_n(\mathcal{D})$ becomes a *random variable* that is itself a sum 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 a class $\mathcal{V}$ of structures and given a weight function c (induced by the probabilities of individual letters), we introduce the *generating function*

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

where the size $|\upsilon|$ is the number of letters involved in the structure. Then[1], $V_n = [z^n]V(z)$ is the total weight of all structures of size n. The collection of occurrences is described by means of regular expressions extended with disjoint unions, and Cartesian products. It is then known that disjoint unions and Cartesian products correspond respectively to sums and products of generating functions; see [19, 21] 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. All the resulting generating functions turn out to be *rational*, of the form $V(z) = (1 - z)^{-(k+1)}P(z)$ for some integer $k \geq 0$ and polynomial P, so that

$$V_n := [z^n]\frac{1}{(1-z)^{k+1}}P(z) = \frac{n^k}{k!}\, P(1) \; \left(1 + O(\frac{1}{n})\right). \tag{2}$$

3 Mean and Variance Estimates of the Number of Occurrences

Mean value analysis. The first moment analysis is easily obtained by describing the collection of all occurrences in terms of formal languages. Let $\mathcal{O}$ be the collection of all occurrences of $\mathcal{W}$ as a hidden word. Each occurrence can be viewed as a "context"

[1] The notation $[z^n]f(z)$ represents the coefficient of z^n in the series $f(z)$.

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

$$\mathcal{O} = \mathcal{A}^\star \times \{w_1\} \times \mathcal{A}^{<d_1} \times \{w_2\} \times \mathcal{A}^{<d_2} \times \ldots \times \{w_{m-1}\} \times \mathcal{A}^{<d_{m-1}} \times \{w_m\} \times \mathcal{A}^\star. \quad (3)$$

There, for $d < \infty$, $\mathcal{A}^{<d}$ denotes the collection of all words of length strictly less d, i.e., $\mathcal{A}^{<d} := \bigcup_{i<d} \mathcal{A}^i$, whereas, for $d = \infty$, $\mathcal{A}^{<\infty}$ denotes the collection of all finite words, i.e., $\mathcal{A}^{<\infty} := \mathcal{A}^\star = \bigcup_{i<\infty} \mathcal{A}^i$. The associated generating functions are

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

We now weight each occurrence by the quantity $\pi(w) = E[X_I]$, so that the generating function $O(z)$ of $\mathcal{O}$ coincides with the generating function of the expectations $\mathbf{E}[\Omega_n]$,

$$O(z) = \sum_{n \geq 1} \mathbf{E}[\Omega_n]\, z^n = \left(\frac{1}{1-z}\right)^{b+1} \times \left(\prod_{i=1}^{m} p_{w_i} z\right) \times \left(\prod_{i \in \mathcal{F}} \frac{1-z^{d_i}}{1-z}\right), \quad (4)$$

and, with $\pi(\mathcal{W})$ the probability of the pattern $\mathcal{W}$, one finds from (2) and (4):

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

Variance analysis. For variance and higher moment analysis, it is essential to work with centred random variables defined as

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

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 above one has

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

There are two kinds of pairs (I, J) according as they intersect or not. When I and J do not intersect, the corresponding random variables 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. Suppose that there exist two occurrences of pattern $\mathcal{W}$ at positions I and J which 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. (This is only possible if, for all $k, 1 \leq k \leq \ell$, one has $w_{r_k} = w_{s_k}$.) We then denote by $\mathcal{W}_{I \cap J}$ the subpattern of $\mathcal{W}$ that occurs at position $I \cap J$, and by $\pi(\mathcal{W}_{I \cap J})$ the probability of this subpattern. Since the expectation $\mathbf{E}[X_I X_J]$ equals $\pi(\mathcal{W})^2/\pi(\mathcal{W}_{I \cap J})$, the expectation $\mathbf{E}[Y_I Y_J] = \mathbf{E}[X_I X_J] - \pi(w)^2$ involves a correlation number $e(I, J)$

$$\mathbf{E}[Y_I Y_J] = \pi^2(\mathcal{W})\, e(I, J), \quad \text{with} \quad e(I, J) = \frac{1}{\pi(\mathcal{W}_{I \cap J})} - 1. \quad (5)$$

In this case, we take the pair of occurrences relative to (I, J) as weighted by $E[Y_I Y_J]$, 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,

$$O_2(z) = \sum_{n\geq 1} z^n \sum_{\substack{I,J\in\mathcal{P}_n(\mathcal{D}),\\ I\cap J\neq\emptyset}} E[Y_I Y_J] = \sum_{n\geq 1} z^n \mathbf{E}[\Xi_n^2(\mathcal{D})].$$

We now need to estimate $O_2(z)$ as $z \to 1$. First, define the *aggregate* $\alpha(I,J)$ to be the system of blocks obtained by merging together all intersecting blocks of the two aggregates $\alpha(I)$ and $\alpha(J)$. The number of blocks $\beta(I,J)$ of $\alpha(I,J)$ plays a fundamental rôle here, since it measures the *degree of freedom* of pairs. 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 $2b-1$. 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 intersecting pairs (I,J) of occurrences for which $\beta(I,J)$ equals $2b-p$. Since there is a fundamental translation invariance, we introduce a notion of *full pairs*: a pair (I,J) of $\mathcal{P}_q(\mathcal{D}) \times \mathcal{P}_q(\mathcal{D})$ is *full* if the aggregate $\alpha(I,J)$ completely covers the interval $[1,q]$. (Clearly, the possible values of q are finite.) Then the collection $\mathcal{O}_2^{[p]}$ is isomorphic to $(\mathcal{A}^\star)^{2b-p+1} \times \mathcal{B}_2^{[p]}$, where $\mathcal{B}_2^{[p]}$ is the subset of full pairs such that $\beta(I,J)$ equals $2b-p$. 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 the collection $\mathcal{B}_2^{[p]}$ and from our earlier discussion, it is a *polynomial* of degree at most $2d(m-1)+1$, with $d = \max_{i\in\mathcal{F}} d_i$. Now, an easy dominant pole analysis entails that $[z^n]O_2^{[p]} = O(n^{2b-p})$. This proves that the dominant contribution to the variance is given by $[z^n]O_2^{[1]}$, which is of order $O(n^{2b-1})$. Then, the variance $\mathbf{E}[\Xi_n^2]$ involves the constant $B_2^{[1]}(1)$ that is the total weight of the collection $\mathcal{B}_2^{[1]}$; the polynomial $B_2^{[1]}(z)$ is itself the generating function of the collection $\mathcal{B}_2^{[1]}$, conceptually an extension of Guibas and Odlyzko's autocorrelation polynomial.

Since the standard deviation is of an order, $O(n^{b-1/2})$, that is smaller than the mean, $O(n^b)$, concentration of distribution holds, via 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*

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

where $\mathcal{F}$ is the set of j such that $d_j < \infty$, and the "variance coefficient" $\sigma^2(\mathcal{W})$ involves the autocorrelation $\kappa(\mathcal{W})$

$$\sigma^2(\mathcal{W}) = \frac{\pi^2(\mathcal{W})}{(2b-1)!}\,\kappa^2(\mathcal{W}) \quad \textit{with} \quad \kappa^2(\mathcal{W}) := \sum_{(I,J)\in\mathcal{B}_2^{[1]}} \left(\frac{1}{\pi(\mathcal{W}_{I\cap J})} - 1\right). \quad (6)$$

*The set $\mathcal{B}_2^{[1]}$ is the collection of all pairs of occurrences (I, J) that satisfy three conditions: (i) they are full; (ii) they are intersecting; (iii) there is a single pair (r, s) with $1 \leq r, s \leq b$ for which the r*th *block $B^{[r]}$ of $\alpha(I)$ and the s*th *block $C^{[s]}$ of $\alpha(J)$ intersect.*

Computation of the variance. The computation of the autocorrelation $\kappa(\mathcal{W})$ reduces to b^2 computations of correlations $\kappa(\mathcal{W}_r, \mathcal{W}_s)$, relative to pairs $(\mathcal{W}_r, \mathcal{W}_s)$ of blocks. Note that each correlation of the form $\kappa(\mathcal{W}_r, \mathcal{W}_s)$ involves a totally constrained problem and can be evaluated by dynamic programming. Precisely, one has

$$\kappa^2(\mathcal{W}) = D^2 \sum_{1 \leq r,s \leq b} \frac{1}{D_r D_s} \binom{r+s-2}{r-1} \binom{2b-r-s}{b-r} \kappa(\mathcal{W}_r, \mathcal{W}_s), \qquad (7)$$

where $\kappa(\mathcal{W}_r, \mathcal{W}_s)$ is the sum of the $e(I, J)$ taken over all full intersecting pairs (I, J) formed with an occurrence I of $\mathcal{W}_r$ subject to constraint $\mathcal{D}_r$ and an occurrence J of $\mathcal{W}_s$ subject to constraint $\mathcal{D}_s$. Let us explain the formula (7) in words: for a pair (I, J) of the set $\mathcal{B}_2^{[1]}$, there is a single pair (r, s) of indices with $1 \leq r, s \leq b$ for which the rth block $B^{[r]}$ of $\alpha(I)$ and the sth block $C^{[s]}$ of $\alpha(J)$ intersect. Then, there exist $r + s - 2$ blocks before the block $\alpha(B^{[r]}, C^{[s]})$ and $2b-r-s$ blocks after it. We then have three different degrees of freedom: (i) the relative order of blocks $B^{[i]}(i < r)$ and blocks $C^{[j]}(j < s)$, and similarly the relative order of blocks $B^{[i]}(i > r)$ and blocks $C^{[j]}(j > s)$; (ii) the lengths of the blocks (there are D_j possible lengths for the jth block); (iii) finally the relative positions of the blocks $B^{[r]}$ and $C^{[s]}$.

In the unconstrained problem, the parameter b equals m, and each block $\mathcal{W}_r$ is reduced to the symbol w_r. Then the " correlation coefficient" $\kappa^2(\mathcal{W})$ simplifies to

$$\kappa^2(\mathcal{W}) := \sum_{1 \leq r,s \leq m} \binom{r+s-2}{r-1} \binom{2m-r-s}{m-r} \Gamma(r, s) \left(\frac{1}{p_{w_r}} - 1 \right), \qquad (8)$$

where the "autocorrelation matrix" Γ of pattern $\mathcal{W}$ is defined by $\Gamma(r, s) := [\![w_r = w_s]\!]$.

4 Central Limit Laws

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 [4]) this is established once all moments of $\widetilde{\Xi}_n$ are known to 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 $s \geq 0$

$$\mathbf{E}[G^{2s}] = 1 \cdot 3 \cdots (2s-1), \quad \mathbf{E}[G^{2s+1}] = 0. \qquad (9)$$

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

$$\mathbf{E}[\Xi_n^{2s+1}] = o(n^{(2s+1)(b-1/2)}), \quad \mathbf{E}[\Xi_n^{2s}] \sim \sigma^{2s}\,(1\cdot 3\cdots(2s-1))\,n^{2sb-s}, \quad (10)$$

which implies Gaussian convergence of $\widetilde{\Xi}_n$.

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

$$\lim_{n\to\infty} \Pr\left\{ \frac{\Omega_n - \mathbf{E}[\Omega_n]}{\sqrt{\mathbf{Var}[\Omega_n]}} \le x \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\,dt \quad (11)$$

Proof. 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 $\Xi_n^r(\mathcal{D})$ yields

$$\mathbf{E}[\Xi_n^r] = \sum_{(I_1,\ldots,I_r)\in\mathcal{P}_n^r(\mathcal{D})} \mathbf{E}[Y_{I_1}\cdots Y_{I_r}]. \quad (12)$$

An r-tuple of sets $(I_1,\ldots,I_r)$ in $\mathcal{P}^r(\mathcal{D})$ is said to be *friendly* if each I_k intersects at least one other I_ℓ, with $\ell \neq k$ 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 $(I_1,\ldots,I_r)$ 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

$$\mathbf{E}[\Xi_n^r] = \sum_{(I_1,\ldots,I_r)\in\mathcal{Q}_n^{(r)}(\mathcal{D})} \mathbf{E}[Y_{I_1}\cdots Y_{I_r}], \quad (13)$$

where the expression involves fewer terms than in (12). From there, we proceed in two stages. First, restrict attention to friendly families that give rise to the dominant contribution and introduce a suitable subfamily $\mathcal{Q}_\star^{(r)} \subset \mathcal{Q}^{(r)}$; in so doing, moments of odd order appear to be negligible. Next, for even order r, the family $\mathcal{Q}_\star^{(r)}$ involves a symmetry and it suffices to consider another smaller subfamily $\mathcal{Q}_{\star\star}^{(r)} \subset \mathcal{Q}_\star^{(r)}$ that corresponds to a "standard" form of occurrence intersection; this last reduction precisely gives rise to the even Gaussian moments.

Odd moments. Given $(I_1,\ldots,I_r) \in \mathcal{Q}^{(r)}$, one defines the aggregate $\alpha(I_1, I_2, \ldots, I_r)$ as the aggregation (in the sense of the variance calculation above) of $\alpha(I_1) \cup \cdots \cup \alpha(I_r)$. Next, the *number of blocks* of $(I_1,\ldots,I_r)$ is the number of blocks of the aggregate $\alpha(I_1,\ldots,I_r)$; if p is the total number of intersecting blocks of the aggregate $\alpha(I_1,\ldots,I_r)$, the aggregate $\alpha(I_1, I_2, \ldots I_r)$ has $rb - p$ blocks. Like previously, we say that the family $(I_1,\ldots,I_r)$ of $\mathcal{Q}_q^{(r)}$ is *full* if the aggregate $\alpha(I_1, I_2, \ldots I_r)$ completely covers the interval $[1,q]$. In this case, the length of the aggregate is at most $rd(m-1)+1$, and the generating function of full families is a polynomial $P_r(z)$ of

degree at most $rd(m-1)+1$ with $d=\max_{j\in\mathcal{F}} d_j$. Then, the generating function of families of $\mathcal{Q}^{(r)}$ 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 $\mathcal{Q}_n^{(r)}$ whose block number equals k is $O(n^k)$. This observation proves that the dominant contribution to (13) 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

$$\mathbf{E}\left[\Xi_n^{2s+1}\right] = O\left(n^{(2s+1)b-s-1}\right) = o\left(n^{(2s+1)(b-1/2)}\right)$$

which establishes the limit form of odd moments in (10).

Even moments. 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, whose set we denote by $\mathcal{Q}_\star^{(2s)}$. In such a family, each subset I_k intersects one and only one other subset I_ℓ. Furthermore, if the blocks of $\alpha(I_h)$ are denoted by $B_h^{[u]}, 1\le u\le b$, there exists only one block $B_k^{[u_k]}$ of $\alpha(I_k)$ and only one block $B_\ell^{[u_\ell]}$ 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, given the symmetry relation $\mathbf{E}[Y_{I_1}\cdots Y_{I_{2s}}]=\mathbf{E}[Y_{I_{\rho(1)}}\cdots Y_{I_{\rho(2s)}}]$ it suffices to restrict attention to 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 thus $(I_1,I_2),\ldots,(I_{2s-1},I_{2s})$. 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

$$\sum_{\mathcal{Q}_{\star n}^{(2s)}} \mathbf{E}[Y_{I_1}\cdots Y_{I_{2s}}] = K_{2s} \sum_{\mathcal{Q}_{\star\star n}^{(2s)}} \mathbf{E}[Y_{I_1}\cdots Y_{I_{2s}}], \tag{14}$$

entails that we can work now solely with standard families.

The class of occurrences relative to standard families is $\mathcal{A}^\star\times(\mathcal{A}^\star)^{2sb-s-1}\times\mathcal{B}_{2s}^{[s]}\times\mathcal{A}^\star$, and involves the collection $\mathcal{B}_{2s}^{[s]}$ of all full friendly $2s$-tuples of occurrences with a number of blocks equal to s. Since $\mathcal{B}_{2s}^{[s]}$ is exactly a shuffle of s copies of $\mathcal{B}_2^{[1]}$ (as introduced in the study of the variance), the associated generating function is

$$\left(\frac{1}{1-z}\right)^{2sb-s+1} (2sb-s)!\left(\frac{B_2^{[1]}(z)}{(2b-1)!}\right)^s,$$

where $B_2^{[1]}(z)$ is the already introduced autocorrelation polynomial. Upon taking coefficients, we obtain the estimate

$$\sum_{\mathcal{Q}_{\star\star n}^{(2s)}} \mathbf{E}[Y_{I_1}\cdots Y_{I_{2s}}] \sim n^{(2b-1)s}\sigma^{2s}. \tag{15}$$

In view of the formulæ (12), (13), (14), and (15) above, this yields the estimate of even moments and leads to the second relation of (10). (Note that the even Gaussian moments eventually come out of the number of involutions, which corresponds to a fundamental symmetry present in the problem.) This completes the proof of Theorem 2.

5 Conclusion

As a test case, we took the full text of Hamlet where all nonalphabetic characters are suppressed. This gives us a (rather unpoetical looking) text that has one long line with 30,316 words and $n = 120,057$ alphabetical characters: "*who s there nay answer me stand and unfold yourself long live the king bernardo he you come most carefully upon your hour [...]*". 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 any two letters 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
20	5.886E+04	124,499	2.11	41,066	0.69
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

This (together with many other experiments) shows a fair fit between the theoretical model and the observed data even though the text chosen is far from being "random".

Extensions. For the constrained case where all the distances are finite, based on finite state models and the de Bruijn graph, it is possible to obtain local limit laws (i.e., a direct estimation of probability densities), a characterization of the speed of convergence to the asymptotic limit (it is $n^{-1/2}$), as well as large deviation estimates (that are exponentially small); see the full paper. For the unconstrained case, the corresponding problems appear to be related to products of random matrices and to the difficult case of random walks on nilpotent Lie groups; see Guivarc'h's paper [11] for context and references. Finally, preliminary investigations indicate that the methods developed here apply to Markovian sources and more generally to all dynamical sources in the sense of Vallée [6, 22].

Acknowledgments. We thank M. Atallah (Purdue U.) for introducing us to the intrusion detection problem that motivated this study. 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.

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. E. Bender and F. Kochman, The Distribution of Subword Counts is Usually Normal, *European Journal of Combinatorics*, 14, 265-275, 1993.
4. P. Billingsley, *Probability and Measure*, Second Edition, John Wiley & Sons, New York, 1986.
5. 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.
6. J. Clément, P. Flajolet, and B. Vallée, Dynamical Sources in Information Theory: A General Analysis of Trie Structures, *Algorithmica*, 29, 307–369, 2001.
7. M. Crochemore and W. Rytter, *Text Algorithms*, Oxford University Press, New York, 1994.
8. G. Das, R. Fleischer, L. G asieniec, D. Gunopulos, and J. Kärkkäinen, Episode Matching, In *Combinatorial Pattern Matching, 8th Annual Symposium*, *Lecture Notes in Computer Science* vol. 1264, 12–27, 1997.
9. L. Guibas and A. M. Odlyzko, Periods in Strings, *J. Combinatorial Theory Ser. A*, 30, 19–43, 1981.
10. L. Guibas and A. M. Odlyzko, String Overlaps, Pattern Matching, and Nontransitive Games, *J. Combinatorial Theory Ser. A*, 30, 183–208, 1981.
11. Y. Guivarc'h, Marches aléatoires sur les groupes, *Fascicule de probabilités*, Publ. Inst. Rech. Math. Rennes, 2000.
12. D. E. Knuth, *The Art of Computer Programming, Fundamental Algorithms*, Vol. 1, Third Edition, Addison-Wesley, Reading, MA, 1997.
13. G. Kucherov and M. Rusinowitch, Matching a Set of Strings with Variable Length Don't Cares, *Theoretical Computer Science* 178, 129–154, 1997.
14. S. Kumar and E.H. Spafford, A Pattern-Matching Model for Intrusion Detection, *Proceedings of the National Computer Security Conference*, 11–21, 1994.
15. P. Nicodème, B. Salvy, and P. Flajolet, Motif Statistics, *European Symposium on Algorithms*, *Lecture Notes in Computer Science*, No. 1643, 194–211, 1999.
16. 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.
17. M. Régnier and W. Szpankowski, On Pattern Frequency Occurrences in a Markovian Sequence, *Algorithmica*, 22, 631-649, 1998.
18. I. Rigoutsos, A. Floratos, L. Parida, Y. Gao and D. Platt, The Emergence of Pattern Discovery Techniques in Computational Biology, *Metabolic Engineering*, 2, 159-177, 2000.
19. R. Sedgewick and P. Flajolet, *An Introduction to the Analysis of Algorithms*, Addison-Wesley, Reading, MA, 1995.
20. J. M. Steele, *Probability Theory and Combinatorial Optimization*, SIAM, Philadelphia, 1997.
21. W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, John Wiley & Sons, New York, 2001.
22. B. Vallée, Dynamical Sources in Information Theory: Fundamental Intervals and Word Prefixes, *Algorithmica*, 29, 262–306, 2001.
23. A. Vanet, L. Marsan, and M.-F. Sagot, Promoter sequences and algorithmical methods for identifying them, *Res. Microbiol.*, 150, 779-799, 1999.
24. M. Waterman, *Introduction to Computational Biology*, Chapman and Hall, London, 1995.
25. A. Wespi, H. Debar, M. Dacier, and M. Nassehi, Fixed vs. Variable-Length Patterns For Detecting Suspicious Process Behavior, *J. Computer Security*, 8, 159-181, 2000.
26. S. Wu and U. Manber, Fast Text Searching Allowing Errors, *Comm. ACM*, 35:10, 83–991, 1995.