Generalized Pattern Matching Statistics

Jérémie BOURDON, Brigitte VALLÉE

ABSTRACT: *In pattern matching algorithms, a characteristic parameter is the number of occurrences of a given pattern in a random text of length n generated by a source. We consider here a generalization of the pattern matching problem in two ways. First, we deal with a generalized notion of pattern that encompasses classical patterns as well as "hidden patterns". Second, we consider a quite general probabilistic model of sources that may possess a high degree of correlations. Such sources are built with dynamical systems and are called dynamical sources. We determine the mean and the variance of the number of occurrences in this generalized pattern matching problem, and establish a property of concentration of distribution. These results are obtained via combinatorics, formal language techniques, and methods of analytic combinatorics based on generating operators and generating functions. The generating operators come from the dynamical system framework and generate themselves generating functions. 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.*

Keywords: Average-case analysis of algorithms, Algorithms on words, Dynamical systems and dynamical analysis.

1 Introduction

Various pattern matching problems. String matching is the basic pattern matching problem. Here, a string w is a sequence of symbols $w = w_1 w_2 \dots w_s$ (of length s), and one searches for occurrences of w (as a block of consecutive symbols) in a text T. However, there are several useful generalizations of this basic problem:

Set of patterns. In the classical string matching problem, the pattern w should appear exactly (and consecutively) in the text, while, in the approximate case, a few mismatches are considered acceptable. The *approximate string matching* is then expressed as matching against *a set $\mathcal{L}$ of words* that contains all the valid approximations of the initial string.

Sequence of patterns. If we are interested in occurrences of the pattern w as a subsequence of the text T (now, the symbols no longer need to be consecutive), the problem is quite different, and it is called the subsequence matching problem. If the lengths of the gaps between successive symbols are not bounded, this matching problem becomes that of finding *a subsequence of symbols.*

Hidden pattern problem. If some of the gap lengths are bounded, while some others are not, one has the hidden pattern problem. A typical hidden pattern may look like $\mathtt{ab\#_2r\#ac\#a\#d\#_4a\#br\#a}$. Here, the notation $\#_i$ abbreviates a sequence of "don't-care-symbols" where the subscript denotes a strict upper bound on the length of the associated gap, $\#$ abbreviates $\#_\infty$ and $\#_1$ is omitted.

Here, we present a *common framework* for all the previously discussed types of pattern matching problems. We consider a generalized pattern which consists in a sequence $\mathcal{L}$ of languages, i.e., $\mathcal{L} := (\mathcal{L}_1, \mathcal{L}_2, \dots, \mathcal{L}_r)$. Each language $\mathcal{L}_i$ itself represents the i-th set of patterns to be used; it may be of infinite cardinality, but it is supposed to be "nondense" (in a formal sense that will be precised in the sequel). This generalized pattern then models a succession of r tasks, the i-th task corresponding to language $\mathcal{L}_i$. Between each task, any event (namely, any word of $\mathcal{A}^\star$) may occur. The corresponding pattern matching problem encompasses all the problems that we have described previously, as well as most of the classical problems of pattern matching, as described in [7] for instance. Surprisingly enough and to the best of our knowledge, there are no studies in the literature that address the question at this level of generality.

Motivations. We cite the introduction of [8] devoted to hidden patterns : "This general problem arises in two domains: intrusion detection and molecular biology. In the area of computer security, the *intrusion detection* [1, 12] searches 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 events between successive events of the signature.
Molecular biology provides another important source of applications [16, 22, 23]. 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, etc .... In general, for gene searching, the generalized pattern matching is the right approach for finding meaningful information.
We wish to study the number of occurrences of the generalized pattern in a random text of length n produced by a source of symbols. 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. An immediate consequence of our results is the possibility to set *thresholds* at which appearance of a generalized pattern starts being meaningful".

Probabilistic model. In information theory contexts, data items are (infinite) words that are produced by a common mechanism, called a source. While real-life sources are often complex objects, pattern matching analyses only deal with quite idealized sources, such as memoryless sources or Markov chains. We use here a general framework of sources related to dynamical systems theory which goes beyond the cases of memoryless and Markov sources [19]. This model can describe non-Markovian processes, where the dependency on past history is unbounded, and as such, they attain a high level of generality. A probabilistic dynamical source is defined by two objects: a symbolic mechanism and a density. The mechanism is related to symbolic dynamics and associates an infinite word $M(x)$ to a real number x of the $[0, 1]$ interval. It can be viewed as a generalization of numeration systems. Once the mechanism has been fixed, the density f on the $[0, 1]$ interval can vary. This induces then different probabilistic behaviors for source words.
In dynamical systems theory, an important tool is the *density transformer*; here, we use it in a non classical way, and, since we show that it can easily generate objects that are essential in the analysis, we give it the rôle of a "generating operator".

Results. To the best of our knowledge, all the results that are already obtained in this area deal with classical sources –memoryless sources, Markov chains–. For instance, 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 re-

ported [2, 9, 10, 13, 14, 15, 23]. Guibas and Odlyzko [9, 10] have revealed the fundamental rôle played by autocorrelation vectors and their associated polynomials. Régnier and Szpankowski [14, 15] 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 [13] 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.
We consider here the general problem –a general pattern in a general source–. Let $\Omega_n(\mathcal{L})$ be the number of occurrences of a given nondense generalized pattern $\mathcal{L} = (\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_r)$ in a random text of length n generated by a dynamical source. Here, the number r of components of the generalized pattern $\mathcal{L}$ plays a fundamental rôle, since it measures the degree of freedom of the pattern. We study the expectation and the variance of the random variable $\Omega_n(\mathcal{L})$, and show the following estimates

$$\mathbf{E}\,[\Omega_n(\mathcal{L})] \sim \pi(\mathcal{L})\ \frac{n^r}{r!} \qquad \mathbf{Var}[\Omega_n(\mathcal{L})] \sim \sigma^2(\mathcal{L})\, n^{2r-1}$$

where $\pi(\mathcal{L})$ is the weight of $\mathcal{L}$. Here, the variance coefficient $\sigma^2(\mathcal{L})$ depends on two kinds of correlations. The first kind of is due to the source, and disappears when the source is memoryless. The second kind of correlations is due to the structure of the pattern, and it may exist even in the memoryless case.
When there exists only one degree of freedom (i.e., $r = 1$), the mean and the variance become of linear growth. This situation arises as soon as one considers only one set of patterns: for instance, the basic string matching, the approximate string matching, the totally constrained hidden pattern, etc .... For $r = 1$, it is possible to deal with finite–state models and de Bruijn graphs (as it is announced in [8] for memoryless sources) and obtain a central limit law.

Methodology. The first step we approach the probabilistic analysis is through a formal description of situations of interest by means of regular languages. Basically, as in [8], such a description of contexts of one or two occurrences gives access to expectation and variance, respectively. In previous works, based on the "generating function methodology", as in the main books of the area [11, 18, 17], one operates a systematic translation into generating functions. Due to correlations of the source, such a direct approach is no longer possible. Instead, we perform what we call a "dynamical analysis" and we first operate a systematic translation into *generating operators*. Now, there are many instances of this methodology, that can be applied in two main areas: text algorithms as in [3, 6, 19], or arithmetical algorithms as in [20, 21]. Here, the structure of the implied generating operators at the pole $z = 1$ provides the necessary asymptotic information. Finally, we come back to (classical) generating functions, and obtain the asymptotic behavior of the main parameters.

Plan of the paper. In Section 2, we describe the general pattern matching problem. We introduce the languages and the related generating functions that intervene in the analysis of the first two moments of the characteristic parameter, namely the number of occurrences of a given generalized pattern. Section 3 is devoted to the probabilistic model. Here, we define dynamical sources and introduce the generating operators that are a basic ingredient associated to the correlated sources considered. In Section 4, we come back to the average-case analysis of the characteristic parameter and prove our results. Finally, Section 5 presents important examples of pattern matching problems that fit in our general framework.

2 Collections of words and generating functions

We fix an alphabet $\mathcal{A}$, either finite or denumerable. The set of all possible texts is $\mathcal{A}^\star$, and a text of length n is an element $T = t_1 t_2 \cdots t_n$ of $\mathcal{A}^n$. We distinguish two notions: a language that is a set of words, and a collection (of words) that is a multi-set of words: in a collection, the same word may occur several times.

2.1. Generalized patterns. We give a higher degree of generalization to the pattern matching problem and consider *a sequence of sets of patterns.*

A generalized pattern is then specified by a finite sequence of languages $\mathcal{L} := (\mathcal{L}_1, \mathcal{L}_2, \dots, \mathcal{L}_r)$. Each language $\mathcal{L}_i$ represents the i-th set of patterns to be used; It is supposed to be *nondense*. A precise definition of this notion will be given in 3.4. When all the components $\mathcal{L}_i$ are nondense, the generalized pattern itself will be said to be nondense. We say that $\mathcal{L}$ *occurs* in the text T if the text T contains as a subsequence a sequence $\ell = (\ell_1, \ell_2, \dots, \ell_r)$ of $\mathcal{L}$. In this case, T is of the form

$$T = w_0 \ell_1 w_1 \ell_2 \dots w_i \ell_i w_{i+1} \dots w_r \ell_r w_{r+1} \qquad \text{with} \qquad w_i \in \mathcal{A}^\star \qquad \text{and} \qquad \ell_i \in \mathcal{L}_i.$$

The set of all valid *occurrences* of generalized pattern $\mathcal{L}$ is then the collection $\rho(\mathcal{L})$,

$$\rho(\mathcal{L}) = \mathcal{A}^\star \times \mathcal{L}_1 \times \mathcal{A}^\star \times \mathcal{L}_2 \times \cdots \times \mathcal{A}^\star \times \mathcal{L}_r \times \mathcal{A}^\star. \tag{1}$$

This operation ρ transforms a finite sequence of languages into a collection of words. This operation is called the completion, and $\rho(\mathcal{L})$ is the completion of the sequence $\mathcal{L}$.

An occurrence, that is an element of $\rho(\mathcal{L})$, defines a position I: this is the sequence of the r disjoint intervals $(I_1, I_2, \dots I_r)$, where interval $I_j := [a_j, b_j]$ represents the exact portion of text T where the word ℓ_j occurs. Then, for instance, $a_1 = |w_0| + 1, b_1 = a_1 + |\ell_1|$, etc .... We denote by $\mathcal{P}_n(\mathcal{L})$ the set of all valid positions relative to $\mathcal{L}$ satisfying $b_r \le n$. The number $\Omega_n(\mathcal{L})$ of occurrences of pattern $\mathcal{L}$ with size n is then a sum of characteristic variables

$$\Omega_n(\mathcal{L}) = \sum_{I \in \mathcal{P}_n(\mathcal{L})} X_I, \qquad \text{with} \qquad X_I(T) := [\![\mathcal{L} \text{ occurs at position } I \text{ in } T]\!], \tag{2}$$

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

2.2. An important particular case: hidden patterns. A hidden pattern matching problem is specified by a pair $(\mathcal{W}, \mathcal{D})$: the pattern $\mathcal{W} = w_1 \cdots w_s$ is a word of length s; the constraint $\mathcal{D} = (d_1, \dots, d_{s-1})$ is an element of $\bar{\mathbb{N}}^{s-1}$. An s-tuple $I = (i_1, i_2, \dots, i_s)$, $(1 \le i_1 < i_2 < \cdots < i_s)$, satisfies the constraint $\mathcal{D}$ if each gap $i_{j+1} - i_j$ is at most d_j, in which case it defines a *position*. An occurrence of pattern $\mathcal{W}$ subject to the constraint $\mathcal{D}$ is a pair (I, T) formed with a position $I = (i_1, i_2, \dots, i_s)$ and a text $T = t_1 t_2 \cdots t_n$ for which $t_{i_j} = w_j$ $(1 \le j \le s)$.

The case $\mathcal{D} = (\infty, \dots, \infty)$ models the *unconstrained problem*: in this case, the i-th language $\mathcal{L}_i$ reduces to the symbol $\{w_i\}$; at the other extreme of the spectrum, there lies the case where all d_j are finite, which we name the *constrained problem*: in this case, there exists only one finite-length language (i.e., $r = 1$) formed by all valid words that begin with the first symbol of $\mathcal{W}$, that end with the last symbol of $\mathcal{W}$, and fulfill all the constraints of $\mathcal{D}$. In the general case, the subset $\mathcal{U}$ of indices j for which d_j is unbounded ($d_j = \infty$) has cardinality $r-1$. It separates the pattern $(\mathcal{W}, \mathcal{D})$ into r independent sub-patterns that are called the blocks and are denoted by $(\mathcal{W}_1, \mathcal{D}_1), (\mathcal{W}_2, \mathcal{D}_2), \ldots, (\mathcal{W}_r, \mathcal{D}_r)$. Each block $(\mathcal{W}_k, \mathcal{D}_k)$ gives rise to a finite–length language $\mathcal{L}_k$ defined as the language of all valid words that begin with the first symbol of $\mathcal{W}_k$, that end with the last symbol of $\mathcal{W}_k$, and fulfill all the constraints of $\mathcal{D}_k$.

2.3. Probabilistic model and generating functions. As regards the probabilistic model, we consider a *random source* that emits symbols of the text from the fixed alphabet $\mathcal{A}$. For a given length n, a random *text*, denoted by T_n, is drawn according to the induced probability on $\mathcal{A}^n$, and, for any word w of length n, we denote by p_w the probability that the source emits a prefix equal to w.
We associate to any collection $\mathcal{M}$ its generating function $M(z)$, where the complex variable z marks the length of the word w

$$M(z) = \sum_{w \in \mathcal{M}} p_w \, z^{|w|} = \sum_{n \geq 0} p(\mathcal{M}_n) \, z^n .$$

The last expression involves the probability of the collection $\mathcal{M}_n$ of words of $\mathcal{M}$ of length n. If the series $M(1)$ and $M'(1)$ converge, they define two objects, the total weight $p(\mathcal{M})$, and the average–length $N(\mathcal{M})$,

$$p(\mathcal{M}) := \sum_{w \in \mathcal{M}} p_w = M(1), \qquad N(\mathcal{M}) := \frac{\sum_{w \in \mathcal{M}} |w| p_w}{\sum_{w \in \mathcal{M}} p_w} = \frac{M'(1)}{M(1)}. \tag{3}$$

For a generalized pattern $\mathcal{L}$, the weight and the average length are defined by

$$\pi(\mathcal{L}) = \prod_{i=1}^{r} p(\mathcal{L}_i), \qquad N(\mathcal{M}) := \sum_{i=1}^{r} N(\mathcal{L}_i), \tag{4}$$

two quantities that surface throughout the analysis.

2.4. Generating function of the mean number of occurrences. Under this randomness model, all the quantities X_I defined in (2) become random variables whose expectations satisfy, for all allowable $I \in \mathcal{P}_n(\mathcal{L})$,

$$\mathbf{E}\,[X_I] = \sum_{T \in \mathcal{A}^n} p(T) \; [\![\mathcal{L} \text{ occurs in } T \text{ at position } I]\!].$$

Then, the quantity $\Omega_n(\mathcal{L})$ is itself a sum of correlated random variables, and the expectation $\mathbf{E}\,[\Omega_n(\mathcal{L})]$ of the number of occurrences of the generalized pattern $\mathcal{L}$ in a text of length n, equal to

$$\mathbf{E}\,[\Omega_n(\mathcal{L})] = \sum_{I \in \mathcal{P}_n(\mathcal{L})} \mathbf{E}\,[X_I] = \sum_{w \in \rho(\mathcal{L})_n} p_w = p\left(\rho(\mathcal{L})_n\right),$$

is exactly the probability of the sub-collection $\rho(\mathcal{L})_n$ defined in (1). Finally, the generating function of the expectations coincides exactly with the generating function $L(z)$ of the collection $\rho(\mathcal{L})$,

$$L(z) := \sum_{w \in \rho(\mathcal{L})} p_w \, z^{|w|} = \sum_{n \geq 1} \mathbf{E}\,[\Omega_n(\mathcal{L})] \; z^n .$$

2.5. Generating functions that intervene in the second moment analysis. The second moment of the variable $\Omega_n(\mathcal{L})$ equals

$$\mathbf{E}\,[\Omega_n^2(\mathcal{L})] = \sum_{I,J\in\mathcal{P}_n(\mathcal{L})} \mathbf{E}\,[X_I X_J].$$

As previously, the generating function $\sum_n \mathbf{E}\,[\Omega_n^2(\mathcal{L})]z^n$ coincides with the generating function $L^{[2]}(z)$ of the collection $\mathcal{L}^{[2]}$ of pairs of $\mathcal{L}$–occurrences.

We now describe the collection $\mathcal{L}^{[2]}$. The aggregate $\alpha(I,J)$ between two positions I and J is the sequence of system of intervals obtained by merging together all intersecting intervals of I and J. The number $\beta(I,J)$ of intervals of $\alpha(I,J)$ plays a fundamental rôle here, since it measures the *degree of freedom* of pairs. As an example, suppose that $I = ([2,6],[10,13])$, $J = ([5,11],[12,14])$ are two valid positions for a generalized pattern. Then $\alpha(I,J) = [2,14]$ and $\beta(I,J) = 1$.
Next, we group the pairs (I,J) according to the value of $\beta(I,J)$ and consider the sub–collections $\mathcal{L}_p^{[2]}$ formed by the pairs of $\mathcal{L}$–occurrences that arise at a pair (I,J) of positions for which $\beta(I,J)$ equals $2r-p$. However, it is sufficient to deal with the two sub–collections $\mathcal{L}_0^{[2]}$ and $\mathcal{L}_1^{[2]}$ relative to cases $p=0$ and $p=1$ since they give rise to the two main terms in the asymptotics of the second moment.

The first collection $\mathcal{L}_0^{[2]}$ (relative to non-intersecting positions) is just obtained from a completion of a generalized shuffle. Consider two sequences of languages $\mathcal{M}$ and $\mathcal{N}$. The shuffle of $\mathcal{M}$ and $\mathcal{N}$, denoted by $\mathcal{M} \amalg \mathcal{N}$, is a set of sequences obtained by shuffling the two sequences; if the sequence $\mathcal{M}$ has i components, and $\mathcal{N}$ has j components, the set $\mathcal{M} \amalg \mathcal{N}$ contains $\binom{i+j}{i}$ sequences with $i+j$ components. For instance, for a generalized pattern $\mathcal{L}$, the shuffle $\mathcal{L} \amalg \mathcal{L}$ contains $\binom{2r}{r}$ sequences, and

$$\mathcal{L}_0^{[2]} := \rho(\mathcal{L} \amalg \mathcal{L}). \tag{5}$$

The second collection $\mathcal{L}_1^{[2]}$ is relative to positions I and J for which only one pair (I_i, J_j) intersects. It is obtained from the operations of completion and generalized shuffle, but it also uses another operation that we denote by $\uparrow$ that transforms a pair $(\mathcal{B},\mathcal{C})$ of languages into a collection $\mathcal{B} \uparrow \mathcal{C}$. This collection gathers all the words that can be obtained by overlaps of words of $\mathcal{B}$ with words of $\mathcal{C}$. It is itself the union of three collections

$$\mathcal{B} \uparrow \mathcal{C} := <\mathcal{B},\mathcal{C}> \bigcup <\mathcal{C},\mathcal{B}> \bigcup (\mathcal{B} \cap \mathcal{C}) \tag{6}$$

where $<\mathcal{B},\mathcal{C}> := \{w = \beta u \gamma \mid \beta,\gamma,u \in \mathcal{A}^\star, u \neq \varepsilon, (\beta,\gamma) \neq (\varepsilon,\varepsilon), \beta u \in \mathcal{B}, u\gamma \in \mathcal{C}\}$.

Then, for each pair (i,j) of indices with $1 \leq i,j \leq r$, the collection $\mathcal{L}_{[i,j]}$ is obtained by mixing $\mathcal{L}_i \uparrow \mathcal{L}_j$ with two beginning sequences $b_i(\mathcal{L}), b_j(\mathcal{L})$ and two ending sequences $e_i(\mathcal{L}), e_j(\mathcal{L})$ of sequence $\mathcal{L}$,

$$\mathcal{L}_{[i,j]} := (b_i(\mathcal{L}) \amalg b_j(\mathcal{L}), \mathcal{L}_i \uparrow \mathcal{L}_j, e_i(\mathcal{L}) \amalg e_j(\mathcal{L}))\,, \tag{7}$$

$$\text{with} \quad b_i(\mathcal{L}) := (\mathcal{L}_1,\ldots,\mathcal{L}_{i-1}), \qquad e_j(\mathcal{L}) := (\mathcal{L}_{j+1},\ldots,\mathcal{L}_r),\ .$$

The collection $\mathcal{L}_{[i,j]}$ is then the union of $t(i,j)$ sequences of languages, with

$$t(i,j) := \binom{i+j-2}{i-1}\binom{2r-i-j}{r-i}. \tag{8}$$

Note that each sequence has the same components, but with a different order. Finally, the collection $\mathcal{L}_1^{[2]}$ is obtained with the completion operator,

$$\mathcal{L}_1^{[2]} := \sum_{1\le i,j\le r} \rho(\mathcal{L}_{[i,j]}). \tag{9}$$

2.6. Main differences with the memoryless case. In the memoryless case, we can obtain quite easily a direct translation of collections defined in (1, 5, 9) into generating functions. Moreover, it is possible to work directly with centered variables $Y_I := X_I - \mathbf{E}(X_I)$.
Here, this approach does not work (or at least we did not succeed in making it work). Due to correlations of the sources, we no longer use a direct translation in generating functions. However, in the framework of dynamical sources, it is possible to deal with "generating operators".

3 Dynamical sources and generating operators

We now present a quite general model of source –dynamical sources– that are associated to dynamical systems. Then, probabilities are "generated" by some generating operators, and the main generating functions to be studied can be generated themselves by operators. More precisely, we prove that an operator $\mathbf{L}(z)$ can be associated to each generalized pattern $\mathcal{L}$ from which the generating function $L(z)$ is easily deduced. Furthermore, unions and Cartesian products of sets translate into sums and compositions of the associated operators.

Moreover, such dynamical sources encompass and generalize the two classical models of sources, namely, the memoryless sources and Markovian sources. We refer to [19] for more details.

3.1. Dynamical sources. *A dynamical system $\mathcal{S}$ is defined by four elements:*
(a) an alphabet $\mathcal{A}$, either finite or denumerable,
(b) a topological partition of $\mathcal{I} :=]0,1[$ with disjoint open intervals $\mathcal{I}_a, a \in \mathcal{A}$,
(c) an encoding mapping σ which is constant and equal to a on each $\mathcal{I}_a$,
(d) a shift mapping T whose restriction to $\mathcal{I}_a$ is a bijection of class $\mathcal{C}^1$ from $\mathcal{I}_a$ to $\mathcal{I}$. The local inverse of T restricted to $\mathcal{I}_a$ is denoted by h_a.

Such a dynamical system can be viewed as a "dynamical source". since it produces infinite words of $\mathcal{A}^\infty$. With an input x of $\mathcal{I}$, it outputs the word $M(x)$ formed with the sequence of symbols $\sigma T^j(x)$, i.e., $M(x) := (\sigma x, \sigma T x, \sigma T^2 x, \ldots)$.

3.2. Generating operators. The mappings $h_w := h_{m_1} \circ h_{m_2} \circ \cdots \circ h_{m_k}$ relative to prefix words $w := m_1 \ldots m_k$ are then the inverse branches of T^k. All the infinite words that begin with the same prefix w correspond to real numbers x that belong to the same interval $\mathcal{I}_w =]h_w(0), h_w(1)[$. Then, if the unit interval is endowed with some density f, the probability p_w that a word begins with prefix w is the measure of the interval $\mathcal{I}_w$. Such a probability p_w is now easily generated by the operator $\mathbf{G}_{[w]}$, defined as

$$\mathbf{G}_{[w]}[f](t) = |h'_w(t)|\ f \circ h_w(t),$$

since one has

$$p_w = |\int_{h_w(0)}^{h_w(1)} f(t)dt| = \int_0^1 |h'_w(t)| f \circ h_w(t)dt = \int_0^1 \mathbf{G}_{[w]}[f](t)dt. \qquad (10)$$

The following composition property holds:

$$\text{for any prefixes } \ w, w', \qquad \text{one has:} \qquad \mathbf{G}_{[w.w']} = \mathbf{G}_{[w']} \circ \mathbf{G}_{[w]}, \qquad (11)$$

and "replaces" the relation $p_{w.w'} = p_w p_{w'}$ which is no longer true when the source has some memory.
The generating operator $\mathbf{B}(z)$ of a collection $\mathcal{B}$ of $\mathcal{A}^\star$ is then defined by

$$\mathbf{B}(z) := \sum_{w \in \mathcal{B}} z^{|w|}\ \mathbf{G}_{[w]}.$$

where the complex variable z marks the length of the word w. If the operator $\mathbf{B}(z)$ is well-defined at $z = 1$, the operator $\mathbf{B} := \mathbf{B}(1)$ is called the normalized operator of collection $\mathcal{B}$. For instance, the operator

$$\mathbf{G} := \sum_{a \in \mathcal{A}} \mathbf{G}_{[a]}, \qquad (12)$$

is the normalized operator of the alphabet $\mathcal{A}$ and it plays a fundamental rôle in the sequel. It is the density transformer of the dynamical system in the sense that if X is a random variable with density f, the density of TX is $\mathbf{G}[f]$.
Then, from (11), unions and Cartesian products of languages translates into sums and compositions of the associated operators. For instance, the operator associated to the set $\mathcal{A}^\star$ is the quasi-inverse

$$(I - z\mathbf{G})^{-1} := \sum_{i \geq 0} z^i \mathbf{G}^i.$$

If $\mathcal{L}$ is a generalized pattern, i.e., a sequence of languages $\mathcal{L} := (\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_r)$, the generating operator $\mathbf{L}(z)$ relative to collection $\rho(\mathcal{L})$ defined in (1) satisfies

$$\mathbf{L}(z) = (I - z\mathbf{G})^{-1} \circ \mathbf{L}_r(z) \circ (I - z\mathbf{G})^{-1} \circ \cdots \circ \mathbf{L}_1(z) \circ (I - z\mathbf{G})^{-1}, \qquad (13)$$

and involves $r + 1$ occurrences of the quasi-inverse $(I - z\mathbf{G})^{-1}$ "mixed" with the generating operators $\mathbf{L}_i(z)$ of languages $\mathcal{L}_i$.
Then, equation (10) provides a relation between the generating operator $\mathbf{B}(z)$ and the generating function $B(z)$ relative to the same collection $\mathcal{B}$, that is

$$B(z) := \sum_{w \in \mathcal{B}} z^{|w|} p_w = \sum_{w \in \mathcal{B}} z^{|w|} \int_0^1 \mathbf{G}_{[w]}[f](t)dt = \int_0^1 \mathbf{B}(z)[f](t)dt. \qquad (14)$$

3.3. Some nice dynamical sources. There exist dynamical sources for which "nice" properties of shift T can be transferred into "nice" properties of their density transformers. Under some "natural" properties of shift T, there exists a spectral decomposition of the density transformer $\mathbf{G}$ defined in (12) that induces a spectral decomposition of the quasi–inverse $(I - z\mathbf{G})^{-1}$. This property will be fundamental in the analysis. More precisely, we give the following definition, perhaps a little bit informal :

Definition. *A dynamical system is said to be decomposable if the density transformer* $\mathbf{G}$ *satisfies the following: It acts on a convenient Banach space* $\mathcal{F}$ *on which it is quasi–compact, and it has positivity properties that entail the existence of dominant spectral objects.*

Then, there exist a unique dominant eigenvalue λ positive and a dominant eigenfunction denoted by φ. Under the normalization condition $\int_0^1 \varphi(t)dt = 1$, this last object is unique too. Then, quasi–compacity entails the existence of a spectral gap between the dominant eigenvalue and the remainder of the spectrum, that separates the operator $\mathbf{G}$ in two parts $\mathbf{G} = \lambda\mathbf{P} + \mathbf{N}$, where $\mathbf{P}$ is the projection of $\mathbf{G}$ onto the dominant eigenspace, and $\mathbf{N}$ is relative to the remainder of the spectrum. Since the operator $\mathbf{G}$ is a density transformer, its dominant eigenvalue satisfies $\lambda = 1$ and φ is also the (unique) stationary density. One then splits the quasi-inverse $(I - z\mathbf{G})^{-1}$ into two parts,

$$(I - z\mathbf{G})^{-1} = \frac{1}{1-z}\mathbf{P} + \mathbf{R}(z), \tag{15}$$

$$\text{with } \mathbf{R}(z) := (I - z\mathbf{N})^{-1} - \mathbf{P} = \sum_{k\geq 0} z^k(\mathbf{G}^k - \mathbf{P}), \qquad \mathbf{P}[f](t) = \varphi(t)\int_0^1 f(x)dx.$$

The first term of (15) defines an operator which has a pole at $z = 1$. Due to the existence of the spectral gap, the operator $\mathbf{N}$ has a spectral radius μ less than 1, and then the second term of (15) defines an operator $\mathbf{R}(z)$ that is analytic in $|z| < (1/\mu)$. At $z = 1$, the operator $\mathbf{R} := \mathbf{R}(1)$ describes the correlations of the source.

A decomposable dynamical source is proven to be ergodic and mixing with exponential rate. The main class of decomposable dynamical sources is provided by the analytic expanding sources, as defined in [19]; however, there are some other instances, as systems described in [5].

In the remainder of the paper, we deal with a decomposable dynamical source and the initial density will be always the stationary density φ. For any function g of $\mathcal{F}$, one has:

$$\mathbf{R}[\varphi] = 0, \qquad \int_0^1 \mathbf{P}[g](t)dt = \int_0^1 g(t)dt, \qquad \int_0^1 \mathbf{R}[g](t)dt = 0. \tag{16}$$

3.4. Nondense collections and correlations. We are now ready to give the precise definition of a nondense collection.

Definition. *The collection* $\mathcal{B}$ *is said to be nondense if the associated generating operator* $\mathbf{B}(z) : \mathcal{F} \to \mathcal{F}$ *is analytic on the disk* $|z| < (1/\rho)$ *for some* $\rho < 1$. *A generalized pattern* $\mathcal{L}$ *whose all components* $\mathcal{L}_i$ *are nondense is said to be nondense.*

Denote by $\mathcal{B}_n$ the subcollection of $\mathcal{B}$ formed with words of length n, and by $\mathbf{B}_n$ its normalized generating operator. The collection $\mathcal{B}$ is nondense if and only if there exists $\rho < 1$ for which one has $||\mathbf{B}_n||_{\mathcal{F}} = O(\rho^n)$. Remark that, if $\mathcal{B}$ is nondense, one has $p(\mathcal{B}_n) = O(\rho^n)$. Then the generating function $B(z)$ is analytic at $z = 1$, so that both quantities $p(\mathcal{B})$ and $N(\mathcal{B})$ defined in (3) are well-defined.

Here are some instances of nondense languages: finite languages, finite–length languages. For memoryless sources, languages where a symbol is forbidden, etc ...

For a nondense collection $\mathcal{B}$, the normalized generating operator $\mathbf{B}$ satisfies

$$\int_0^1 \mathbf{P} \circ \mathbf{B} \circ \mathbf{P}[g](t) = p(\mathcal{B}) \left(\int_0^1 g(t)dt \right). \tag{17}$$

In the analysis, two kinds of correlations may occur between two nondense collections $\mathcal{B}, \mathcal{C}$, according to the relative position of $\mathcal{B}$ and $\mathcal{C}$. If $\mathcal{B}$ and $\mathcal{C}$ do not overlap, there are two cases: $\mathcal{B}$ before $\mathcal{C}$, or $\mathcal{C}$ before $\mathcal{B}$. In these cases, the correlation coefficients are $c(\mathcal{B}, \mathcal{C})$ or $c(\mathcal{C}, \mathcal{B})$ where $c(\mathcal{B}, \mathcal{C})$ is defined as

$$p(\mathcal{B})p(\mathcal{C})c(\mathcal{B}, \mathcal{C}) := \sum_{k \geq 0} \left[p(\mathcal{B} \times \mathcal{A}^k \times \mathcal{C}) - p(\mathcal{B})p(\mathcal{C}) \right] = \int_0^1 \mathbf{C} \circ \mathbf{R} \circ \mathbf{B}[\varphi](t). \tag{18}$$

If $\mathcal{B}$ and $\mathcal{C}$ overlap, the correlation coefficient

$$d(\mathcal{B}, \mathcal{C}) := \frac{p(\mathcal{B} \uparrow \mathcal{C})}{p(\mathcal{B})p(\mathcal{C})} \tag{19}$$

involves the nondense collection $\mathcal{B} \uparrow \mathcal{C}$ of words w defined in (6), so that coefficient $d(\mathcal{B}, \mathcal{C})$ is well-defined. Finally, the total correlation coefficient $m(\mathcal{B}, \mathcal{C})$ between $\mathcal{B}$ and $\mathcal{C}$ gathers the three possible cases

$$m(\mathcal{B}, \mathcal{C}) = c(\mathcal{B}, \mathcal{C}) + c(\mathcal{C}, \mathcal{B}) + d(\mathcal{B}, \mathcal{C}), \tag{20}$$

so that

$$p(\mathcal{B})p(\mathcal{C})m(\mathcal{B}, \mathcal{C}) = p(\mathcal{B} \uparrow \mathcal{C}) + \sum_{k \geq 0} \left[p(\mathcal{B} \times \mathcal{A}^k \times \mathcal{C}) + p(\mathcal{C} \times \mathcal{A}^k \times \mathcal{B}) - 2p(\mathcal{B})p(\mathcal{C}) \right].$$

4 Analysis of the number of pattern occurrences

We are now able to come back to the analysis of the characteristic parameter and we prove a similar asymptotic behavior for the number of occurrences as in the classical case [8]. We first study the mean, then the second moment. We conclude with an expression of the variance, and a proof of the concentration of distributions.

The main steps of our analysis which can be called a "dynamical analysis" are as follows:

(a) We first describe the generating operators relative to each collection, namely $\mathbf{L}(z)$ relative to collection $\rho(\mathcal{L})$ for the mean, and $\mathbf{L}_0^{[2]}, \mathbf{L}_1^{[2]}$ relative to collections $\mathcal{L}_0^{[2]}, \mathcal{L}_1^{[2]}$ for the second moment. Each generating operator will contain some occurrences of the quasi-inverse $(I - z\mathbf{G})^{-1}$, and some occurrences of generating operators $\mathbf{L}_i(z)$.

(b) We then decompose the three main operators with the help of (15). We will consider only the two transformations which give rise to the main asymptotic terms: The transformation called AllP, where we replace all the occurrences of the quasi-inverse by the first term $(1-z)^{-1}\mathbf{P}$, and the transformation AllP-1R, where we replace all occurrences except one of the quasi inverse $(I-z\mathbf{G})^{-1}$ by the first term $(1-z)^{-1}\mathbf{P}$, the last occurrence being replaced by the operator $\mathbf{R}(z)$. Then, formulae (16,17) are useful in the computations of constants involved.

(c) We thus come back to generating functions thanks to (14).

(d) We finally extract the asymptotic behavior of coefficients of generating functions.

4.1. Mean number of occurrences. We state our first result:

Theorem 1. *Consider a decomposable dynamical source endowed with its stationary density φ and a generalized nondense pattern $\mathcal{L} = (\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_r)$. The expectation $\mathbf{E}[\Omega_n(\mathcal{L})]$ of the number of occurrences of the generalized pattern $\mathcal{L}$ in a text of length n satisfies*

$$\mathbf{E}[\Omega_n(\mathcal{L})] = \binom{n+r}{r}\pi(\mathcal{L}) + \binom{n+r-1}{r-1}\pi(\mathcal{L})\ [C(\mathcal{L}) - N(\mathcal{L})]\ + O(n^{r-2}),$$

where $\pi(\mathcal{L})$ is the total weight and $N(\mathcal{L})$ is the average length. The coefficient $C(\mathcal{L})$ equals the sum of the correlations $c(\mathcal{L}_{i-1}, \mathcal{L}_i)$ between two consecutive languages, where the correlation coefficient $c(\mathcal{B}, \mathcal{C})$ is defined in (18).

Proof. The collection relative to the mean is $\rho(\mathcal{L})$ defined in (1), and the generating function is $\mathbf{L}(z)$ defined in (13). When using the transformation AllP in $\mathbf{L}(z)$, we obtain an operator $\mathbf{M}_1(z)$ which has a pole of order $r+1$ at $z=1$,

$$\mathbf{M}_1(z) = \left(\frac{1}{1-z}\right)^{r+1} \mathbf{P} \circ \mathbf{L}_r(z) \circ \mathbf{P} \circ \cdots \circ \mathbf{P} \circ \mathbf{L}_1(z) \circ \mathbf{P}.$$

Near $z=1$, each operator $\mathbf{L}_i(z)$ is analytic and admits the expansion $\mathbf{L}_i(z) = \mathbf{L}_i + (z-1)\mathbf{L}_i'(1) + O(z-1)^2$, so that the main term of the expansion is

$$\left(\frac{1}{1-z}\right)^{r+1} \mathbf{P} \circ \mathbf{L}_r \circ \mathbf{P} \circ \cdots \circ \mathbf{P} \circ \mathbf{L}_1 \circ \mathbf{P} \tag{21}$$

while the second main term is obtained as a sum of r terms, each of them obtained by replacing the operator $\mathbf{L}_i(z)$ by its derivative $\mathbf{L}_i'(1)$ at $z=1$. The corresponding generating function $M_1(z)$ satisfies near $z=1$

$$M_1(z) = \left(\frac{1}{1-z}\right)^{r+1} \pi(\mathcal{L}) - \left(\frac{1}{1-z}\right)^{r} \pi(\mathcal{L})N(\mathcal{L}) + O\left(\frac{1}{1-z}\right)^{r-1}. \tag{22}$$

The apparition of the weight $\pi(\mathcal{L})$ and the average length $N(\mathcal{L})$ is due to formulae (16,17), together with definitions (4).

When using now AllP-1R in $\mathbf{L}(z)$, we obtain an operator $\mathbf{M}_2(z)$ which has a pole of order r at $z=1$. This is a sum of $r+1$ terms, each of the term containing an occurrence of the operator $\mathbf{R}(z)$ between two generating operators of consecutive languages $\mathcal{L}_{i-1}, \mathcal{L}_i$. The relative generating function $M_2(z)$ has also a pole of order r at $z=1$ and satisfies near $z=1$

$$M_2(z) = \left(\frac{1}{1-z}\right)^{r} \pi(\mathcal{L}) \sum_{i=2}^{r} c(\mathcal{L}_{i-1}, \mathcal{L}_i) + O\left(\frac{1}{1-z}\right)^{r-1}.$$

Here, the correlation number $c(\mathcal{B}, \mathcal{C})$ between $\mathcal{B}$ and $\mathcal{C}$ is defined in (18). ◻

4.2. The second moment. We prove here that the analysis of the second moment is very similar to the analysis of the expectation, even if it is more intricate.

Theorem 2. *The second moment* $\mathbf{E}\,[\Omega_n^2(\mathcal{L})]$ *of the number of occurrences of the nondense generalized pattern* $\mathcal{L} = (\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_r)$ *in a text of length* n *produced by a decomposable dynamical source endowed with stationary density* φ *satisfies*

$$\mathbf{E}\,[\Omega_n^2(\mathcal{L})] = \binom{n+2r}{2r}\binom{2r}{r}\pi^2(\mathcal{L}) + \binom{n+2r-1}{2r-1}\pi^2(\mathcal{L})\widetilde{C}(\mathcal{L}) + O(n^{r-2}),$$

$$\text{with } \widetilde{C}(\mathcal{L}) := 2\binom{2r-1}{r-1}C(\mathcal{L}) - 2\binom{2r}{r}N(\mathcal{L}) + \sum_{1\le i,j\le r} t(i,j)\, m(\mathcal{L}_i, \mathcal{L}_j).$$

Here, $\pi(\mathcal{L})$ *is the total weight,* $N(\mathcal{L})$ *is the average length,* $C(\mathcal{L})$ *is the coefficient of Theorem 1 and the correlation–coefficient* $m(\mathcal{B}, \mathcal{C})$ *is defined in (20). The "choice" coefficients* $t(i,j)$ *are defined in (8).*

Proof. We deal now with collections $\mathcal{L}_0^{[2]}, \mathcal{L}_1^{[2]}$ defined in (5, 9), and the generating functions $\mathbf{L}_0^{[2]}, \mathbf{L}_1^{[2]}$ relative to these collections.
We begin with $\mathbf{L}_0^{[2]}$. When using the transformation AllP in $\mathbf{L}_0^{[2]}$, we obtain an operator $\mathbf{M}_3(z)$ which has a pole of order $2r+1$ at $z=1$. The corresponding generating function $M_3(z)$ has also a pole of order $2r+1$ at $z=1$ and satisfies near $z=1$

$$M_3(z) = \binom{2r}{r}\left(\frac{1}{1-z}\right)^{2r+1}\pi^2(\mathcal{L}) - 2\binom{2r}{r}\left(\frac{1}{1-z}\right)^{2r}\pi^2(\mathcal{L})N(\mathcal{L}) + O\left(\frac{1}{1-z}\right)^{2r-1}.$$

Here, the weight $\pi(\mathcal{L})$ and the average length $N(\mathcal{L})$ occur thanks to the same arguments as in Theorem 1.

When using AllP-1R in $\mathbf{L}_0^{[2]}$, we obtain an operator $\mathbf{M}_4(z)$ which has a pole of order $2r$ at $z=1$. This is a sum of terms, where each term contains an occurrence of the operator $\mathbf{R}(z)$ between two generating operators of languages $\mathcal{L}_i, \mathcal{L}_j$. Now, there are two cases, according as the two languages $\mathcal{L}_i, \mathcal{L}_j$ "come from" the same occurrence or not. In the second case, these languages are not necessary consecutive, and all pairs (i,j) may intervene. More precisely, each pair (i,j), for $1 \le i,j \le r$, intervenes in a number of terms equal to $2t(i,j)$ –here, the "choice coefficient" $t(i,j)$ is defined in (8)–. In the first case, the pair (i,j) is necessary of the form $(i-1,i)$, with $2 \le i \le r$, and such a pair occurs in exactly $2\binom{2r-1}{r-1}$ terms. Finally, the associated generating function $M_4(z)$ has also a pole of order $2r$ at $z=1$ and satisfies

$$M_4(z) = \left(\frac{1}{1-z}\right)^{2r}\pi^2(\mathcal{L})C_2(\mathcal{L}) + O\left(\frac{1}{1-z}\right)^{2r-1},$$

$$\text{where } C_2(\mathcal{L}) := 2\binom{2r-1}{r-1}C(\mathcal{L}) + \sum_{1\le i,j\le r} t(i,j)\,[c(\mathcal{L}_i,\mathcal{L}_j) + c(\mathcal{L}_j,\mathcal{L}_i)]$$

involves the coefficient $C(\mathcal{L})$ that intervenes in Theorem 1 together with supplementary correlations $c(\mathcal{B},\mathcal{C})$ defined in (18).

The second generating operator $\mathbf{L}_1^{[2]}(z)$ is the sum of generating operators $\mathbf{L}_{[i,j]}$ relative to collections $\rho(\mathcal{L}_{[i,j]})$ defined in (7,9). Each operator $\mathbf{L}_{[i,j]}$ is itself the sum of $t(i,j)$ terms. If, we use AllP in each term, we obtain an operator $\mathbf{M}_5(z)$ which has a pole of order $2r$ at $z=1$. Near $z=1$, the associated generating function

$$M_5(z) = \left(\frac{1}{1-z}\right)^{2r} \pi^2(\mathcal{L}) \sum_{1\le i,j\le r} t(i,j)\, d(\mathcal{L}_i,\mathcal{L}_j) + O\left(\frac{1}{1-z}\right)^{2r-1},$$

involves correlation–coefficients $d(\mathcal{B},\mathcal{C})$ defined in (19). □

4.3. Concentration of distributions for the number of occurrences. In the computation of the variance, the two main terms of order n^{2r} in $\mathbf{E}[\Omega_n^2]$ and in $\mathbf{E}[\Omega_n]^2$ cancel, and thus, the standard deviation is of an order, $O(n^{r-1/2})$, that is smaller than the mean, $O(n^r)$. This fact entails concentration of distribution, via a well-known argument based on Chebyshev's inequalities. In summary:

Theorem 3. *The mean and the variance of the number of occurrences $\Omega_n(\mathcal{L})$ relative to a nondense generalized pattern $\mathcal{L} = (\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_r)$ in a text of length n produced by a decomposable dynamical source endowed with stationary density φ satisfy*

$$\mathbf{E}[\Omega_n(\mathcal{L})] = \frac{\pi(\mathcal{L})}{r!}\, n^r \left(1 + O(\frac{1}{n})\right), \qquad \mathbf{Var}[\Omega_n(\mathcal{L})] = \sigma^2(\mathcal{L})\; n^{2r-1}\left(1+O(\frac{1}{n})\right),$$

where the variance coefficient

$$\sigma^2(\mathcal{L}) = \pi^2(\mathcal{L})\left[\frac{r-2N(\mathcal{L})}{r!(r-1)!} + \frac{m(\mathcal{L})}{(2r-1)!}\right]$$

involves the weight $\pi(\mathcal{L})$, the average length $N(\mathcal{L})$ and the total correlation–coefficient

$$m(\mathcal{L}) := \sum_{1\le i,j\le r} \binom{i+j-2}{i-1}\binom{2r-i-j}{r-i} m(\mathcal{L}_i,\mathcal{L}_j)$$

that deals with the correlation–coefficients $m(\mathcal{B},\mathcal{C})$, defined as

$$p(\mathcal{B})p(\mathcal{C})m(\mathcal{B},\mathcal{C}) := p(\mathcal{B}\uparrow\mathcal{C}) + \sum_{k\ge 0}\left[p(\mathcal{B}\times\mathcal{A}^k\times\mathcal{C}) + p(\mathcal{C}\times\mathcal{A}^k\times\mathcal{B}) - 2p(\mathcal{B})p(\mathcal{C})\right].$$

Consequently, the distribution of the random variable $\Omega_n(\mathcal{L})$ converges in probability:

$$\text{for any } \epsilon > 0, \qquad \lim_{n\to\infty} \Pr\left\{\left|\frac{\Omega_n(\mathcal{L})}{\mathbf{E}[\Omega_n(\mathcal{L})]} - 1\right| < \epsilon\right\} = 1.$$

4.4. A Gaussian law ? In the memoryless case, we can adapt the method used in [8], based on the study of higher centered moments. We then obtain a limit Gaussian law.

However, in the case of a general decomposable source, we did not succeed in analysing directly centered moments. We can obtain several main terms of each moment $\mathbf{E}[\Omega_n^k]$ of order k, but, in order to compute the main term of the moment of order k of the centered variable $X_n := \Omega_n - \mathbf{E}[\Omega_n]$, we would have to get a very precise expansion of each $\mathbf{E}[\Omega_n^k]$. We do not succeed to obtain these expansions. However, we conjecture that the asymptotic Gaussian law holds, even for a general decomposable dynamical source.

5 Examples of various pattern matching problems

For memoryless sources, all the correlation–coefficients $c(\mathcal{L}_i, \mathcal{L}_j)$ disappear. We now study some particular cases where the coefficients $d(\mathcal{L}_i, \mathcal{L}_j)$ admit some expressions that are related to classical objects that appear in pattern matching problems, such as correlation polynomials, for instance, and we find again some classical results that are described in the book [18].

5.1. Basic pattern matching. Here, we consider a pattern α that is a finite string of length s. This fits in our general framework with $r = 1$ and $\mathcal{L} = \{\alpha\}$. The set $\mathcal{L} \uparrow \mathcal{L}$ is here a collection that contains all the words of the form $\beta u \gamma$ where u satisfies $\beta u = u\gamma = \alpha$. Classically, as in [18], the set of all such possible suffixes γ is called the autocorrelation set and is denoted by $\mathcal{D}_\alpha$. Then $\mathcal{L} \uparrow \mathcal{L}$ coincides with the collection of the words $\alpha\gamma$, where the suffix γ belongs to $\mathcal{D}_\alpha$. For $\gamma = \varepsilon$, the word has multiplicity 1, while, for $\gamma \neq \varepsilon$, the word has multiplicity 2. Then, the autocorrelation-coefficient $d(\alpha, \alpha)$ is related to the auto-correlation polynomial $D_\alpha(z)$ via the relation $p^2(\alpha) d(\alpha, \alpha) = 2D_\alpha(1) - p(\alpha)$. The following theorem holds:

Theorem 4. *Let α be a single pattern of length s. The number Ω_n of occurrences of α in a text of length n produced by a decomposable dynamical source satisfies*

$$\mathbf{E}[\Omega_n] = [n - s + 1]\, p(\alpha) + O(\frac{1}{n}), \quad \mathbf{Var}[\Omega_n] = n\, p^2(\alpha)\, [1 - 2s + m(\alpha)] + O(1),$$

$$\textit{with} \quad m(\alpha) := \frac{2}{p^2(\alpha)} \left(\sum_{k \geq 0} [p(\alpha \times \mathcal{A}^k \times \alpha) - p^2(\alpha)] + \sum_{\gamma \in \mathcal{D}_\alpha} p(\alpha\gamma) \right) - \frac{1}{p(\alpha)}.$$

5.2. Matching a set of patterns. Approximate pattern matching as well as multiple pattern matching are very similar to the previous problem. In these cases, the generalized pattern $\mathcal{L}$ still satisfies $r = 1$ and there is a unique finite-length language that we denote by $\mathcal{L}$. The pattern–correlation coefficient is related to a correlation polynomial $A_{(\alpha,\beta)}(z)$ that extends the autocorrelation polynomial for two words α and β of $\mathcal{L}$. Totally constrained hidden patterns, as defined in paragraph 2.2, are also a particular case of the general setting where $r = 1$.

5.3. The gaussian law for a single regular language. In this case, our analysis, described in [4], deals with a generating matrix operator that is a mixing between generating operators relative to the source and the automaton A relative to the regular language $\mathcal{L}$. We associate to $\mathcal{L}$ subsets $\mathcal{B}^{[i,j]}$ of the alphabet $\mathcal{A}$ formed with all the symbols a that allow automaton transitions from the i-th state to the j-th state,

$$\mathcal{B}^{[i,j]} := \{a \in \mathcal{A} \mid A(a, i) = j\}.$$

We denote by $\mathbf{B}^{[i,j]}$ the normalized operator relative to language $\mathcal{B}^{[i,j]}$, and, we mark with variable u all the operators that lead to an accepting state. Finally, we consider the matrix operator $\mathbf{B}_u$ whose general coefficient is

$$\mathbf{B}_u^{[i,j]} := \begin{cases} \mathbf{B}^{[i,j]} & \text{if } j \text{ is not accepting,} \\ u\,\mathbf{B}^{[i,j]} & \text{if } j \text{ is accepting.} \end{cases}$$

This mixed operator $\mathbf{B}_u$, that depends on parameter u, possesses dominant spectral properties when u is near the real axis. We then obtain the following result:

Theorem 5. *The number of occurrences of a regular language in a random text of length n produced by a decomposable dynamical source, once normalized by the mean and standard deviation, obeys in the asymptotic limit a Gaussian law.*

5.4. Examples of problems with $r > 1$. We recall that a hidden pattern $(\mathcal{W},\ D)$ gives rise to a generalized pattern with r components where $r+1$ equals the number of infinite constraints. Then, as soon as the hidden pattern contains an infinite constraint, one has $r > 1$.

In the totally unconstrained hidden pattern problem, one considers a pattern $\alpha := \alpha_1 \dots \alpha_s$ of $\mathcal{A}^s$ and we are interested in counting all the occurrences of α as a subsequence (commonly called *hidden word*) of a text T. The associated generalized pattern corresponds exactly to $r = s$ and $\mathcal{L}_i := \{\alpha_i\}$. The pattern–correlation coefficient $d(\alpha_i, \alpha_j)$ between two symbols α_i and α_j satisfies

$$d(\alpha_i, \alpha_j) = \frac{[\![\alpha_i = \alpha_j]\!]}{p_{\alpha_i}},$$

and the constant $N(\mathcal{L})$ equals the length s of the pattern.

The problem of searching for a sequence of words (i.e. $\mathcal{L}_i := \{w_i\}$) is a slight generalization of this problem. In this case, the correlation coefficient d is expressed with the correlation polynomial $A_{(w_i,w_j)}(z)$ between two words w_i and w_j, and the constant $N(\mathcal{L})$ is the sum $N(\mathcal{L}) = \sum_i |w_i|$.

References

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

[2] E. Bender and F. Kochman, The Distribution of Subword Counts is Usually Normal, *European Journal of Combinatorics*, 14, 265-275, 1993.

[3] J. Bourdon, Size and Path-length of Patricia Tries: Dynamical Sources Context, *Random Structures and Algorithms*, Vol 19 (3–4), 289–315, 2001.

[4] J. Bourdon, B. Vallée, Dynamical Sources in Information Theory : Motif Statistics, not published.

[5] J. Bourdon, B. Daireaux, B. Vallée, Dynamical Analysis of α-Euclidean Algorithms, submitted.

[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] Ph Flajolet, Y. Guivarch, W. Szpankowski and B. Vallée, Hidden Pattern Statistics,*Proc. of ICALP'2001*, LNCS 2076, 152–165, 2001.

[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] D. E. Knuth, *The Art of Computer Programming, Fundamental Algorithms*, Vol. 1, Third Edition, Addison-Wesley, Reading, MA, 1997.

[12] S. Kumar and E.H. Spafford, A Pattern-Matching Model for Intrusion Detection, *Proceedings of the National Computer Security Conference*, 11–21, 1994.

[13] P. Nicodème, B. Salvy, and P. Flajolet, Motif Statistics, *European Symposium on Algorithms, Lecture Notes in Computer Science*, No. 1643, 194–211, 1999.

[14] 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.

[15] M. Régnier and W. Szpankowski, On Pattern Frequency Occurrences in a Markovian Sequence, *Algorithmica*, 22, 631–649, 1998.

[16] 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.

[17] R. Sedgewick and P. Flajolet, *An Introduction to the Analysis of Algorithms*, Addison-Wesley, Reading, MA, 1995.

[18] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, John Wiley & Sons, New York, 2001.

[19] B. Vallée, Dynamical Sources in Information Theory: Fundamental Intervals and Word Prefixes, *Algorithmica*, 29, 262–306, 2001.

[20] B. Vallée, Dynamical Analysis of a Class of Euclidean Algorithms, to appear in *Theoretical Computer Science* (2002), also available in Les cahiers du GREYC 2000.

[21] B. Vallée, Digits and Continuants in Euclidean Algorithms. Ergodic Versus Tauberian Theorems, JTNB 12, 531–570, 2000.

[22] A. Vanet, L. Marsan, and M.-F. Sagot, Promoter sequences and algorithmical methods for identifying them, *Res. Microbiol.*, 150, 779-799, 1999.

[23] M. Waterman, *Introduction to Computational Biology*, Chapman and Hall, London, 1995.

Jérémie BOURDON, **Brigitte** VALLÉE
GREYC–UMR 6072, Université de Caen, F-14032 Caen (France)
e-mail: {jeremie.bourdon, brigitte.vallee}@info.unicaen.fr