Pattern Matching Statistics on Correlated Sources

Jérémie Bourdon[1] and Brigitte Vallée[2]

[1] LINA, (CNRS and Université de Nantes, France)
Jeremie.Bourdon@univ-nantes.fr
[2] GREYC, (CNRS and Université de Caen, France)
Brigitte.Vallee@info.unicaen.fr

Abstract. In pattern matching algorithms, two characteristic parameters play an important rôle : the number of occurrences of a given pattern, and the number of positions where a pattern occurrence ends. Since there may exist many occurrences which end at the same position, these two parameters may differ in a significant way. Here, we consider a general framework where the text is produced by a probabilistic source, which can be built by a dynamical system. Such "dynamical sources" encompass the classical sources –memoryless sources, and Markov chains–, and may possess a high degree of correlations. We are mainly interested in two situations : the pattern is a general word of a regular expression, and we study the number of occurrence positions – the pattern is a finite set of strings, and we study the number of occurrences. In both cases, we determine the mean and the variance of the parameter, and prove that its distribution is asymptotically Gaussian. In this way, we extend methods and results which have been already obtained for classical sources [for instance in [9] and in [6]] to this general "dynamical" framework. Our methods use various techniques: formal languages, and generating functions, as in previous works. However, in this correlated model, it is not possible to use a direct transfer into generating functions, and we mainly deal with generating operators which generate... generating functions.

1 Introduction

The problem of searching for a particular pattern in a text is an important problem in information theory. It is crucial to study precisely the number of *occurrences* of a given pattern in a typical text. Here, "typical" essentially means that the text is a random text produced by a probabilistic model that follows as far as possible the real complexity of the studied sequences. It is also very interesting to consider *positions of occurrence*, i.e., positions (in a text) where an occurrence of the pattern can terminate.
The two parameters – the number of occurrences, denoted in the following by Ω, and the number of occurrence positions, denoted by C – may differ in a significant way, since the number of occurrence positions is always bounded by the text length, whereas this is not true for the number of occurrences. [There may exist many occurrences which end at the same occurrence position].

With a precise probabilistic study of these two parameters, one obtains sharp statistical heuristics (like Z-scores) which permit to describe the related algorithms, and perhaps improve them.

Various pattern matching problems. There are also different pattern matching problems, which differ according to the nature of the pattern.
String matching. This is the basic pattern matching problem. Here, a string w is a block of (consecutive) symbols $w = w_1 w_2 \dots w_s$ (of length s).
Set of strings. Previously, the string w should appear exactly in the text, while, in the approximate case, a few mismatches are considered acceptable. The *approximate string matching* is then expressed as a matching against *a set $\mathcal{L}$ of words* which contains all the valid approximations of the string.
Sequence of patterns. Here, the symbols no longer need to be consecutive in the text: we are interested in occurrences of the string w as a subsequence of the text T. The problem is different, and it is called the hidden word problem.
Regular expressions. Searching words from a regular language is surely the most general pattern matching problem, since all the three previous pattern matching problems all consist in finding words of a given regular language.

Motivations. *Molecular biology* [12,17,18] provides an important source of applications. As a rule, there, one searches for subsequences, not consecutive strings. There are plenty of examples: split genes where exons are interrupted by introns, starting and stopping signal in genes, etc.... In general, for gene searching [8], regular expressions are used as a general pattern model (such as the PROSITE format used to scan in protein databases).
In this general context, it is of obvious interest to discern what constitutes meaningful information from what is statistically unavoidable phenomenon. This leads to a probabilistic study. In information theory context, a source is a mechanism which emits symbols from an alphabet Σ. A text of length n is just an element of Σ^n, and the various models of sources are related to the choice of a probabilistic model on Σ^n. When the probabilistic model has been chosen, the main variables of interest — the number of occurrences Ω, and the number of occurrence positions C— become random variables, and it is crucial to study their distribution, in order to set *thresholds* from which appearance of a pattern becomes meaningful.

Previous results. The two classical models of sources are the memoryless sources (where each symbol m is always emitted with the same probability, and independently of the previous history) and Markov chains (where the probability of emitting m only depends on the unique symbol emitted before m). In both cases, these sources have a "bounded" memory and only provide idealized models, while real-life sources are often complex objects. Most of the results are obtained only for such idealized sources.
Number of occurrences Ω. The number of string occurrences in a random text has been intensively studied over the last two decades. Guibas and Odlyzko have revealed in 1981 the fundamental rôle played by autocorrelation. Régnier and Szpankowski [10,11] established that the number of occurrences of a string is asymptotically normal under a diversity of models that include Markov chains.

The number of occurrences of finite sets of (finite) strings also obeys the "Guibas and Odlyzko" principle, which now deals with correlation matrices.
In the case of the hidden word problems, Flajolet, Szpankowski and Vallée show that the distribution of Ω is asymptotically Gaussian for memoryless sources [6].
Number of occurrence positions C. Nicodème, Salvy, and Flajolet [9] showed that, for a simple[3] regular expression $\mathcal{E}$, the variable $C_n(\mathcal{E})$ is asymptotically normally distributed, both for memoryless sources and Markov chains.

Our results. We use here a general framework of sources related to dynamical systems theory which goes beyond the cases of memoryless and Markov sources [16,4]. 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, related to symbolic dynamics, associates an infinite word $M(x)$ to a real number $x \in [0,1]$, and generalizes numeration systems. Once the mechanism has been fixed, the density f on the $[0,1]$ interval can vary. This induces different probabilistic behaviors for sources of words.
In this context, string matching problems have been already considered: In [1], the authors study the parameter $\Omega(\mathcal{L})$ when $\mathcal{L}$ is a particular regular expression (namely, a generalized pattern), which provides a generalization for the hidden word problem. The mean and the variance of Ω_n are shown to be polynomial in n, and the exponent r depends on the number of freedom degrees of $\mathcal{L}$. However, the asymptotic distribution – expected to be Gaussian– is not obtained.

Here, we obtain two new results in this correlated model of dynamical sources. We prove here that many variables R defined on some set $\mathcal{R}$ follow asymptotically a gaussian law. We first provide a precise definition:

Definition [Asymptotic gaussian law.] *Consider a cost R defined on a set $\mathcal{R}$ and its restriction R_n to the subset $\mathcal{R}_n$ of size n. The cost R asymptotically follows a gaussian law if there exist three sequences a_n, b_n, r_n, with $r_n \to 0$, for which*

$$\Pr\left[(u,v) \in \mathcal{R}_n \;\middle|\; \frac{R_n(u,v) - a_n}{\sqrt{b_n}} \le y\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-t^2/2}\, dt + O(r_n)\,.$$

The sequence r_n defines the speed of convergence, denoted also by $r[R_n]$. The expectation $\mathbb{E}[R_n]$ and the variance $\mathbb{V}[R_n]$ satisfy $\mathbb{E}[R_n] \sim a_n$, $\mathbb{V}[R_n] \sim b_n$. The triple $(\mathbb{E}[R_n], \mathbb{V}[R_n], r_n)$ is a characteristic triple for the gaussian law of R.

We now state our main result:[4]

Theorem. *Let $\mathcal{S}$ be a nice dynamical source.*
(i) Consider a simple regular expression $\mathcal{E}$ whose useful part of the automaton is primitive. The number of occurrence positions of $\mathcal{E}$ in a word of length n built by $\mathcal{S}$, denoted by $C_n(\mathcal{E})$, follows an asymptotic gaussian law with a characteristic triple given by $r[C_n(\mathcal{E})] = O(1/\sqrt{n})$,

$$\mathbb{E}[C_n(\mathcal{E})] = \gamma_{\mathcal{E}} \cdot n + \gamma'_{\mathcal{E}} + O(\mu_{\mathcal{E}}^n), \quad \mathbb{V}[C_n(\mathcal{E})] = \nu_{\mathcal{E}} \cdot n + \nu'_{\mathcal{E}} + O(\mu_{\mathcal{E}}^n),$$

[3] See Section 2.4 for a definition
[4] The word "nice" is defined in Def. 4, Section 3.3, the words "simple" and "useful" are defined in Def. 1, Section 2.4, the word "primitive" in Section 3.3

The constants $\gamma_{\mathcal{E}}$ and $\nu_{\mathcal{E}}$ are expressible with the pression $\Lambda(t)$ of the operator $\mathbb{R}(e^t)$ defined in (8), namely $\gamma_{\mathcal{E}} = \Lambda'(0)$, $\nu_{\mathcal{E}} = \Lambda''(0)$, while $\mu_{\mathcal{E}} < 1$ is any real number strictly larger than the subdominant eigenvalue of $\mathbb{R}$.
(ii) Consider a finite set of words $\mathcal{W} \subset \Sigma^\star$. The number of occurrences of $\mathcal{W}$ in a text of length n built by $\mathcal{S}$, denoted by $\Omega_n(\mathcal{W})$, follows an asymptotic gaussian law with a characteristic triple given by $r[\Omega_n(\mathcal{W})] = O(1/\sqrt{n})$,

$$\mathbb{E}[\Omega_n(\mathcal{W})] = \alpha_{\mathcal{W}} \cdot n + \alpha'_{\mathcal{W}} + O(\eta^n_{\mathcal{W}}), \quad \mathbb{V}[\Omega_n(\mathcal{W})] = \beta_{\mathcal{W}} \cdot n + \beta'_{\mathcal{W}} + O(\eta^n_{\mathcal{W}}).$$

The constants $\alpha_{\mathcal{W}}$ et $\beta_{\mathcal{W}}$ are expressible with the pression $\Lambda(t)$ of the operator $\mathbb{B}(e^t)$ defined in (9), namely $\alpha_{\mathcal{W}} = \Lambda'(0)$, $\beta_{\mathcal{W}} = \Lambda''(0)$, while $\eta_{\mathcal{W}} < 1$ is any real number strictly larger than the subdominant eigenvalue of $\mathbb{B}$.

Methodology. For studying the parameter $C(\mathcal{E})$, Nicodème, Salvy and Flajolet describe in [9] a general method which directly translates a regular expression into rational generating functions. They use, as a main tool, the transition matrix of the automaton which recognizes the regular language $\Sigma^\star \cdot \mathcal{E}$, and the occurrence positions are related to the final states of the automaton. In [6], the authors also use similar methods, namely the de Bruijn graph, to study the parameter $\Omega(\mathcal{W})$. These two previous works, based on the "generating function methodology", as in the main books of the area [14,13], operate a systematic translation of each language into its generating function. Due to correlations of a dynamical source, such a direct approach is no longer possible here. Instead, we perform what we call a "dynamical analysis" and we first operate a systematic translation into *generating operators*. In dynamical systems theory, an important tool is the *density transformer*; here, we give it the role of a "generating operator". Now, there are many instances of this methodology, applied in two main areas: text algorithms as in [2,5,16], or arithmetical algorithms as in [15]. Here, we deal with a mixed structure, where we insert generating operators inside the transition matrix of the automaton. We obtain an operator matrix which takes into account both the complexity of the source and the algebraic structure of the problem (namely an automaton).

2 Various tools.

We first introduce the languages and the related generating functions that intervene in the analysis of the characteristic parameters C and Ω. Next, we precise the probabilistic model. We define dynamical sources and introduce the generating operators that are a basic ingredient associated to our correlated sources.

2.1. Probabilistic model and generating functions. As regards the probabilistic model, we consider a source that creates the text by emitting symbols from a finite alphabet Σ. For a given length n, a random text, denoted by T_n is an element of Σ^n which is drawn according to the induced probability on Σ^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. A language $\mathcal{L}$ is then a set of words. For any language, we denote by $\mathcal{L}_n$ the language formed with all the words w of $\mathcal{L}$ with length n.

We aim at studying the random variables $Y = C$ (the number of occurrence positions) and $Y = \Omega$ (the number of occurrences). In both cases, we consider

the restriction of Y to Σ^n, denoted by Y_n, and analyze its probabilistic behavior for $n \to \infty$. Our main tool is the moment generating function of Y_n, defined as

$$\mathbb{E}[\exp(tY_n)] := \sum_{w \in \Sigma^n} p_w \cdot \exp[tY(w)], \tag{1}$$

and the main challenge is to show that it behaves as a "quasi-power". Then, it will be possible to obtain an asymptotic Gaussian law:

Theorem 0. [Hwang] *Let Y_n be a sequence of variables whose moment generating functions satisfies $\mathbb{E}[\exp(tY_n)] = [\exp(nU(t)+V(t))]\cdot[1+O(W_n)]$, $W_n \to \infty$, with a uniform error term on the complex closed disk $|t| \le t_0$, $t_0 > 0$. Suppose that $U(t)$ and $V(t)$ are analytic in $|t| \le t_0$ and $U(t)$ satisfies $U''(0) \neq 0$. Then, Y_n follows an asymptotic gaussian law, with a characteristic triple given by*

$$\mathbb{E}[Y_n] = U'(0) \cdot n + V'(0) + O(W_n), \quad \mathbb{V}[Y_n] = U''(0) \cdot n + V''(0) + O(W_n),$$
$$r[Y_n] = O\left(\max(1/\sqrt{n}, W_n)\right).$$

2.2. Bivariate generating functions. The so–called probability generating function $F_Y(z,u)$ relative to parameter Y is defined as

$$F_Y(z,u) = \sum_{w \in \Sigma^\star} p_w \cdot u^{Y(w)} \cdot z^{|w|},$$

where $|w|$ denotes the length of w, the variables z and u respectively mark the length of the word and the parameter $Y(w)$. Remark that the moment generating function of parameter Y_n is closely related to $F_Y(z,u)$ via the relation

$$\mathbb{E}[\exp(tY_n)] = [z^n]F_Y(z,e^t) \tag{2}$$

where the notation $[z^n]G(z)$ denotes the coefficient of z^n in $G(z)$. Previous works, which deal with non correlated sources, directly work with the generating functions. Here, we cannot operate a direct translation from the problem into generating functions, and we mainly use generating operators.

2.3. Language vs automaton. Let us first recall that an automaton is defined by $(\Sigma, \mathcal{Q}, \mathcal{F}, s, \delta)$, where Σ is an alphabet, $\mathcal{Q}$ is the (finite) set of states, $\mathcal{F} \subset \mathcal{Q}$ corresponds to the final states, $s \in \mathcal{Q}$ is the initial state and $\delta : \mathcal{Q} \times \Sigma \to \mathcal{Q}$ is the transition function of the automaton. In the following, the set $\mathcal{Q}$ will be always $\{0, \dots, r-1\}$, and the state 0 will be the initial state.
The automaton recognizes a language $\mathcal{L}$ if, for all word $w := m_1 \dots m_n$ of $\mathcal{L}$, there exists a path $q_1, q_2, \dots, q_{n-1}$ of states and a final state f such that

$$\delta(s, m_1) = q_1, \quad \delta(q_i, m_{i+1}) = q_{i+1}, \quad [\text{for } 1 \le i \le n-2], \quad \delta(q_{n-1}, m_n) = f.$$

In this case, the language $\mathcal{L}$ is said to be a regular language. Every regular language can be described by a regular expression, composed of singletons and a finite number of unions, Cartesian products and star operations on those singletons. Conversely, it is possible to operate a direct translation from a regular expression to a deterministic finite automaton.
The transition matrix $\mathcal{T} := (\mathcal{T}_{i,j})$ is the $r \times r$ matrix whose element of index (i,j) is the set of symbols $m \in \Sigma$ for which there exists an edge from state i to state j labeled by m, namely $\mathcal{T}_{i,j} := \{m \in \Sigma; \quad \delta(i,m) = j\}$.

This matrix plays a fundamental rôle in the sequel. Thus, the component (i, j) of the matrix $\mathcal{T}^n$ is the language formed by all the words which allow to reach state j from state i in n steps. And, the component (i, j) of the matrix $\mathcal{T}^\star$ is the language formed by all the words which allow to reach state j from state i in an arbitrary number of steps. Finally, $\mathcal{L}_n = S \cdot \mathcal{T}^n \cdot F$, $\mathcal{L} = S \cdot \mathcal{T}^\star \cdot F$, where $F :=^t (f_1, \ldots, f_r)$ is a $\{0, 1\}$ column vector such that f_i equals 1 iff $i \in \mathcal{F}$, called the final vector and S is a row vector equal to $(1 \quad 0 \quad \cdots \quad 0)$.

2.4. Automata of interest. To each parameter [$C(\mathcal{E})$ or $\Omega(\mathcal{W})$], we associate an automaton which will be central in the study of this parameter.

Case $C(\mathcal{E})$- Automaton for the language $\mathcal{L} = \Sigma^\star \cdot \mathcal{E}$ associated to a regular expression $\mathcal{E}$. We consider the minimal automaton $\mathcal{A}$ which recognizes $\Sigma^\star \cdot \mathcal{E}$, and its decomposition into the acyclic graph of its strongly connected components (SCC):

Definition 1. *The expression $\mathcal{E}$ is simple if the minimal automaton $\mathcal{A}$ which recognizes $\Sigma^\star \cdot \mathcal{E}$ possesses a unique SCC which contains all the final states.*

Generally speaking, it is possible that all final states do not belong to the same SCC. Here, we mainly consider the case when $\mathcal{E}$ is simple[5]. However, we explain (in the conclusion) how our method extends to the general case.

Proposition 1. *Let $\mathcal{E}$ be a simple regular expression and $\mathcal{A}$ be the minimal automaton which recognizes $\Sigma^\star \cdot \mathcal{E}$. Then, there exists a partition of its set of states $\mathcal{Q}$ into two sets $\mathcal{X}$ and $\mathcal{Y}$ for which the transition matrix $\mathcal{T}$ of this automaton can be written as*

$$\mathcal{T} = \begin{pmatrix} \mathcal{M} & \mathcal{U} \\ 0 & \mathcal{R} \end{pmatrix};$$

Here, $\mathcal{M}$ is the matrix restricted to $\mathcal{X}$, $\mathcal{R}$ is the matrix restricted to $\mathcal{Y}$, and $\mathcal{U}$ is the matrix from $\mathcal{X}$ to $\mathcal{Y}$. If $\mathcal{X}$ is non empty, it contains the initial state, while the graph $(\mathcal{Y}, \mathcal{R})$ is the SCC of the automaton, which contains all the final states and is called the useful part of the automaton.

Remarks. Then, the language $\mathcal{L}$ decomposes as $\mathcal{L} = S_\mathcal{X} \cdot \mathcal{M}^\star \cdot \mathcal{U} \cdot \mathcal{R}^\star \cdot F_\mathcal{Y}$, where $S_\mathcal{X}$ is the initial row vector restricted to $\mathcal{X}$, and $F_\mathcal{Y}$ is the final column vector restricted to $\mathcal{Y}$. Note that the language $\mathcal{L}_+$ of the words which contain at least one occurrence of the regular expression $\mathcal{E}$ satisfies $\mathcal{L}_+ \subset S_\mathcal{X} \cdot \mathcal{M}^\star \cdot \mathcal{U} \cdot \mathcal{R}^\star \cdot \mathbf{1}_\mathcal{Y}$, where $\mathbf{1}_\mathcal{Y}$ is a column vector, indexed with $\mathcal{Y}$, whose all components equal 1.

Example. See Figure 1 (at the end) for $\mathcal{E} := (ba|c)^+a^+$.

Case $\Omega(\mathcal{W})$- The de Bruijn automaton relative to an alphabet Σ and a length ℓ. In the sequel, ℓ will be the maximum length of a word of $\mathcal{W}$, minus 1. We consider a "sliding window" of length ℓ that scans a text of $\Sigma^\star$ and, at each stage, keeps in its (finite) memory the last ℓ letters read from the text. Formally, the de Bruijn graph is a finite automaton with state space $\mathcal{Q} = \Sigma^\ell$; when the symbol m is read, in a state $b \in \Sigma^\ell$, one erases the left symbol of b, which provides a word denoted by $\tau(b)$, and m is added on the right of $\tau(b)$, so

[5] This is also the only case which is considered in [9]

that the new state is $\delta(b, m) = \tau(b) \cdot m$. A text of length $n \geq \ell$ is then associated to a path of length $n - \ell$ that begins at the state b formed with the first ℓ symbols of the text. This transition matrix is denoted by $\mathcal{B}$. Let us define the initial vector S as a row vector whose components are all the words of Σ^ℓ, and the final vector as a column vector whose components are all equal to 1. Then

$$\Sigma^n = S \cdot \mathcal{B}^{n-\ell} \cdot F, \qquad \Sigma^{\geq \ell} = S \cdot \mathcal{B}^\star \cdot F.$$

Example. See Fig. 2 (at the end) for the de Bruijn graph with $\Sigma := \{a, b\}, \ell = 2$.

We now present the probabilistic model for symbol generation. This model is based on dynamical systems. Here, probabilities are "generated" by operators, and the main generating functions of interest can be generated themselves by operators. Furthermore, unions and Cartesian products of sets translate into sums and compositions of the associated operators. This allows us to define a matrix generating operator related to a regular language.

2.5. Dynamical sources. We first recall the definition of a dynamical system (of the interval). We refer to [16,4] for more details. See Fig. 3 for an example.

Definition 2. *A* dynamical system $(\mathcal{I}, \mathcal{S})$ *is defined by four elements:*

(a) *a finite alphabet* Σ,

(b) *a topological partition of* $\mathcal{I} :=]0, 1[$ *with disjoint open intervals* $\mathcal{I}_m, m \in \Sigma$,

(c) *an encoding mapping* σ *which is constant and equal to* m *on each* $\mathcal{I}_m$,

(d) *a shift mapping* S *whose restriction to* $\mathcal{I}_m$ *is a bijection of class* $\mathcal{C}^2$ *from* $\mathcal{I}_m$ *to* $\mathcal{J}_m := S(\mathcal{I}_m)$. *The local inverse of* $S|_{\mathcal{I}_m}$ *is denoted by* h_m.

Such a dynamical system can be viewed as a "dynamical source", since, on an input x of $\mathcal{I}$, it outputs the word $M(x)$ formed with the sequence of symbols $\sigma S^j(x)$, i.e., $M(x) := (\sigma x, \sigma Sx, \sigma S^2 x, \ldots)$.
The branches of S^k, and also its inverse branches, are then indexed by Σ^k, and, for any $w = m_1 \ldots m_k \in \Sigma^k$, the mapping $h_w := h_{m_1} \circ h_{m_2} \circ \cdots \circ h_{m_k}$ is a $\mathcal{C}^2$ bijection from $\mathcal{J}_w$ onto $\mathcal{I}_w$. It is possible that the word w cannot be produced by the source: this means that $\mathcal{J}_w$ is empty, and the inverse branch h_w does not exist. All the words that begin with the same prefix w correspond to real numbers x that belong to the same interval $\mathcal{I}_w$.

Such sources may possess a high degree of correlations, due to the *geometry* of the branches [i.e., the respective positions of intervals $\mathcal{I}_m$ and $\mathcal{J}_\ell := S(\mathcal{I}_\ell)$] and also to the *shape* of branches. [See [4] for more details]. For instance, the classical sources correspond to dynamical systems with affine branches, for which the derivatives are constant. Generally speaking, the probability of emitting a symbol m is closely related to the shape of branches, as we now see.

2.6. Probabilities and generating operators. When the interval $\mathcal{I}$ is endowed with some density g, this induces a probabilistic model on $\Sigma^{\mathbb{N}}$, and 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 easily generated by an operator $\mathbf{G}_{[w]}$, defined as

$$\mathbf{G}_{[w]}[f](x) = |h'_w(x)| \; f \circ h_w(x) \mathbb{1}_{\mathcal{J}_w}(x), \tag{3}$$

since one has $\quad p_w = \int_{\mathcal{I}_w} g(x)dx = \int_{\mathcal{J}_w} |h'_w(x)| g \circ h_w(x)dx = \int_0^1 \mathbf{G}_{[w]}[g](x)dx.$

Then, the operator $\mathbf{G}_{[w]}$ is called the generating operator of the prefix w. The generating operator $\mathbf{L}$ relative to a collection $\mathcal{L}$ of words is defined as the sum of all the generating operators relative to the words of $\mathcal{L}$, namely $\mathbf{L} := \sum_{w\in\mathcal{L}} \mathbf{G}_{[w]}$, and the generating operator $\mathbf{G}$ of the alphabet Σ

$$\mathbf{G} := \sum_{m\in\Sigma} \mathbf{G}_{[m]}. \tag{4}$$

plays a fundamental rôle here, since it is the density transformer of the dynamical system; it describes the evolution of densities on $\mathcal{I}$ under iterations of S: if X is a random variable with density g, then SX has density $\mathbf{G}[g]$.

For two prefixes w, w', the relation $p_{w.w'} = p_w p_{w'}$ is no longer true when the source has some memory, and is replaced by the following composition property

$$\mathbf{G}_{[w.w']} = \mathbf{G}_{[w']} \circ \mathbf{G}_{[w]}, \tag{5}$$

so that unions and Cartesian products of collections of words translate into sums and compositions of the associated generating operators. Remark just that, due to (5), the generating operator of $\mathcal{L} \times \mathcal{M}$ is $\mathbf{M} \circ \mathbf{L}$.

2.7. Matrix generating operators. Here, we transform the transition matrix of an automaton into a matrix generating operator that combines both information from the dynamical source and the automaton. We associate to each element $\mathcal{T}_{j,i}$ of the matrix T, its generating operator $\mathbb{T}_{i,j}$

$$\mathbb{T}_{i,j} := \sum_{w\in\mathcal{T}_{j,i}} \mathbf{G}_{[w]}. \tag{6}$$

Then, $\mathbb{T}$ is a matrix generating operator which is related to ${}^t\mathcal{T}$, due to (5).

Examples. In the case when $\mathcal{L}$ is $\Sigma^\star\cdot\mathcal{E}$, there are three matrix operators, $\mathbb{M}, \mathbb{U}, \mathbb{R}$, respectively associated to matrices $\mathcal{M}, \mathcal{U}, \mathcal{R}$ [See Prop. 1]. For the de Bruijn graph, the generating operator is denoted by $\mathbb{B}$. See Figures 1 and 2 for examples.

2.8. The mixed source. We now build a source $\mathcal{S}_\mathcal{T}$ that combines both a transition matrix $\mathcal{T}$ of an automaton $\mathcal{A}$, and the original source $\mathcal{S}$. The set of states of $\mathcal{A}$ is $\mathcal{Q}$ and the matrix $\mathcal{T}$ has order r. The initial source $\mathcal{S}$ is defined by an interval $\mathcal{I}$, an alphabet Σ, a topological partition $(\mathcal{I}_m)_{m\in\Sigma}$ and a shift S whose each local inverse $h_m := (S\big|_{\mathcal{I}_m})^{-1}$ maps $\mathcal{J}_m :=]c_m, d_m[$ on $\mathcal{I}_m :=]a_m, b_m[$. The source $\mathcal{S}_\mathcal{T}$ [see Fig. 3 (at the end) for an example] is defined with the interval $\mathcal{I}^{[r]} = [0, r]$, the alphabet $\Gamma := \Sigma \times \mathcal{Q}$, a topological partition $(\mathcal{I}_{m,i})_{(m,i)\in\Gamma}$ and a shift function that maps $\mathcal{I}^{[r]}$ on $\mathcal{I}^{[r]}$. Each local inverse $h_{m,i}$ maps $\mathcal{J}_{m,i}$ on $\mathcal{I}_{m,i}$. More precisely, $\mathcal{I}_{m,i} = \mathcal{I}_m + i :=]a_m + i, b_m + i[$, $\quad \mathcal{J}_{m,i} = \mathcal{J}_m + \delta(i, m) :=]c_m + \delta(i, m), d_m + \delta(i, m)[$, and $h_{m,i}(x) = h_m(x - \delta(i, m)) + i$. The density transformer $\mathfrak{G}$ of the source $\mathcal{S}_\mathcal{T}$ defined, as in (4), by

$$\mathfrak{G}[f](x) := \sum_{(m,i)\in\Sigma\times\mathcal{Q}} |h'_{m,i}(x)| \cdot f \circ h_{m,i}(x) \cdot \mathbb{1}_{\mathcal{J}_{m,i}}(x), \tag{7}$$

is conjugated to the matrix operator $\mathbb{T}$ defined in (6) via a mapping Ψ [namely $\mathfrak{G} = \Psi^{-1} \circ \mathbb{T} \circ \Psi$] which associates to g (defined on $\mathcal{I}^{[r]}$) the vector ${}^t[g_1, \ldots, g_r]$ where each g_i is defined on $\mathcal{I}$ by $g_i(x) := g\big|_{[i-1,1]}(x + i)$.

3 Probabilistic behavior of parameters C and Ω.

Now, we come back to the two situations of interest. The next step consists in weighting operator matrices $\mathbb{T}$ in order to study our parameters $C(\mathcal{E})$ and $\Omega(\mathcal{W})$.

3.1. Case of $C(\mathcal{E})$. We consider here the language $\mathcal{L} := \Sigma^\star\mathcal{E}$ and the three matrix operators $\mathbb{M}, \mathbb{U}, \mathbb{R}$. We now mark the transitions which arrive at final states and define three new operators $\mathbb{R}(u)$, $\mathbb{U}(u), \mathbb{X}(u)$ by the relations.

$$\mathbb{R}(u)_{j,i} = u^{[\![j\in\mathcal{F}]\!]} \cdot \mathbb{R}_{j,i}, \quad \mathbb{U}(u)_{j,i} = u^{[\![j\in\mathcal{F}]\!]} \cdot \mathbb{U}_{j,i}, \quad ([\![\cdot]\!]\text{is Iverson's bracket}) \tag{8}$$

$$\mathbb{X}(z,u) := z \cdot \mathbb{U}(u) \circ (I - z\mathbb{M})^{-1}.S_{\mathcal{X}},$$

where the vector $S_{\mathcal{X}}$ is a column vector (of length $|\mathcal{X}|$) equal to ${}^t(1,0,\ldots,0)$.

Example. Figure 1 describes the marked matrix operators for $\mathcal{E} = (ba|c)^+a^+$.

3.2. Case of $\Omega(\mathcal{W})$. We consider here a set of finite words $\mathcal{W}$, and we choose the length ℓ of the de Bruijn graph to be equal to the maximal length of a word of $\mathcal{W}$, minus 1. this de Bruijn automaton is weighted with a counter that gets incremented each time a transition is effected, so that the value of the counter will contain at the end of the text the number $\Omega(\mathcal{W})$. A transition of the automaton, of the form $c = \delta(b,m)$ requires $b\cdot m \in \Sigma\cdot c$. When this transition is effected, one can "cash in" all the "new" occurrences of $\mathcal{W}$ which arise when reading the last letter m, i.e., all the occurrences of the pattern that *end* at the letter m. Precisely, for a transition $c = \delta(b,m)$ of the automaton, the number of occurrences of the pattern $\mathcal{W}$ contained in $b\cdot m$ and *ending* at the letter m is determined by either the pair (b,m) or the pair (b,c); we denote this number by $\phi(b,m)$ or $\psi(b,c)$, depending on context, so that $\phi(b,m) = \psi(b,c)$ whenever $c = \delta(b,m)$. Since the length of word $b\cdot m$ exactly equals $\ell+1$ that is the maximum length of a word of $\mathcal{W}$, all the occurrences of $\mathcal{W}$ that end at m are contained in a text of the form $b\cdot m$ with $b\in\Sigma^\ell$ so that the relation $\phi(b,m) = \Omega(b\cdot m) - \Omega(b)$ holds. We build a operator matrix $\mathbb{B}(u)$ indexed by $\mathcal{Q}\times\mathcal{Q}$ as follows

$$\mathbb{B}(u)_{c,b} := u^{\phi(b,m)} \cdot [\![\, bm \in \Sigma c \,]\!] \cdot \mathbf{G}_{[m]} = u^{\Omega(bm)-\Omega(b)} \cdot [\![\, bm \in \Sigma c \,]\!] \cdot \mathbf{G}_{[m]}, \tag{9}$$

and the initial vector $\mathbb{X}(z,u)$ is a column vector defined by

$$\left(\mathbb{X}(z,u)\right)_b = z^\ell \cdot u^{\Omega(b)} \cdot \mathbf{G}_{[b]}\,. \tag{10}$$

Example. Figure 2 describes the matrix $\mathbb{B}(u)$ relative to $\mathcal{W} := \{ab, aab, aba\}$

In both cases, the operator $\mathbb{F}_Y(z,u) := (I - z\mathbb{T}(u))^{-1}\circ\mathbb{X}(z,u)$, with $\mathbb{T}(u) = \mathbb{R}(u)$ or $\mathbb{T}(u) = \mathbb{B}(u)$ itself generates, with (3), the generating function $F_Y(z,u)$, and we obtain, with (1):

Proposition 2. *The probability generating functions of parameters $Y = C$ and $Y = \Omega$ are expressible with the quasi-inverse of a matrix operator $\mathbb{T}(u)$,*

$$\mathbb{E}[u^{Y_n}] = [z^n]\cdot\left(\int_0^1 \left(\mathbf{1}_{\mathcal{X}}\cdot(I - z\mathbb{T}(u))^{-1}\circ\mathbb{X}(z,u)\right)[g](t)dt\right).$$

In the case $Y = C$, the operator $\mathbb{T}(u)$ equals $\mathbb{R}(u)$, and $\mathbb{R}(u), \mathbb{X}(z,u)$ involve the decomposition of Proposition 1 [see (8)]. In case $Y = \Omega$, the operator $\mathbb{T}(u)$ equals $\mathbb{B}(u)$ and $\mathbb{B}(u), \mathbb{X}(z,u)$ involve the de Bruijn graph [see (9,10)].

In the sequel, we prove that, provided that the source $\mathcal{S}$ and the transition matrix $\mathcal{T}$ possesses good properties, it is the same for the source $\mathcal{S}_{\mathcal{T}}$.

3.3. Nice sources and convenient sources. Under quite general hypotheses, and on a convenient functional space, the density transformer admits $\lambda = 1$ as an eigenvalue of largest modulus. But, generally speaking, this is not a unique dominant eigenvalue isolated from the remainder of the spectrum.

Definition 3. *A dynamical source is said to be decomposable if, when acting on a convenient Banach space $\mathcal{F}$, the density transformer $\mathbf{G}$ [defined in (4)] possesses a unique dominant eigenvalue (equal to 1) separated from the remainder of the spectrum by a spectral gap, i.e., $\rho := \sup\{|\lambda| \ ; \quad \lambda \in \operatorname{Sp} \mathbf{G}, \lambda \neq 1\} < 1$.*

Remarks. Let us explain the terminology: Consider the dominant eigenfunction φ which is an invariant function for $\mathbf{G}$. Under the normalization condition $\int_0^1 \varphi(t)dt = 1$, this last object is unique too, and it is also the (unique) stationary density. Due to the existence of the spectral gap, the operator $\mathbf{G}$ decomposes into two parts, namely $\mathbf{G} = \lambda\mathbf{P} + \mathbf{N}$, where $\mathbf{P}$ is the projection of $\mathbf{G}$ onto the dominant eigenspace generated by φ, and $\mathbf{N}$, relative to the remainder of the spectrum, has a spectral radius equal to ρ, which is strictly less than 1. The operator $\mathbf{N}$ describes the correlations of the source. A decomposable dynamical source is ergodic and mixing with an exponential rate equal to ρ.

Most of the classical sources –memoryless sources, or primitive Markov chains– are easily proven to be decomposable. We now present sufficient conditions under which a general dynamical source will be proven to be decomposable, together with all its associated mixed sources $\mathcal{S}_{\mathcal{T}}$ [the proofs are omitted here].

Definition 4. *A dynamical source (on a finite alphabet) is said to be "nice" if it satisfies the two conditions*
(i) [Expansiveness] *There exist two constants C, D with $D > 1$ for which one has, for any $m \in \Sigma$, for any $x \in \mathcal{I}_m$, $D < |S'(x)] < C$.*
(ii) [Topologically mixing] *For any pair of two nonempty open sets (V, W), there exists $n_0 \geq 1$ such that $S^{-n}V \cap W \neq \emptyset$ for all $n \geq n_0$.*

Proposition 3. *A nice dynamical system is decomposable, with respect to the space $BV(\mathcal{I})$ of functions with bounded variation, endowed with the norm $||f|| := \sup|f| + V(f)$ [Here, $V(f)$ is the total variation of f on $\mathcal{I}$].*

We consider now the mixed source $\mathcal{S}_{\mathcal{T}}$. Recall that a transition matrix $\mathcal{T}$ is primitive if there exists a power of the matrix $\mathcal{T}$ whose coefficients are never the empty language. A strongly connected graph gives rise to a matrix $\mathcal{T}$ which is primitive if and only if the gcd of the lengths of its cycles equals 1. If it is not primitive, the gcd d of its cycle lengths is called the period, and $\mathcal{T}^d$ is primitive.

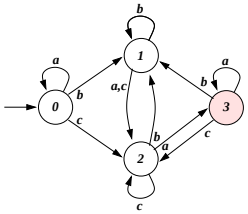
Proposition 4. *If $\mathcal{S}$ is a nice dynamical source, then the following holds:*
(i) the mixed source $\mathcal{S}_{\mathcal{T}}$ relative to any primitive graph $\mathcal{T}$ is nice too.
(ii) The mixed source $\mathcal{S}_{\mathcal{B}}$ relative to a de Bruijn graph $\mathcal{B}$ is always nice.
(iii) Define the period of a regular expression $\mathcal{E}$ to be equal to the period of the useful part of $\mathcal{R}$ of its automaton. Then, for any regular language $\mathcal{E}$ of period d, the source $\mathcal{S}_{\mathcal{R}}^d$ (whose shift equals $T_{\mathcal{R}}^d$) is nice.

3.4. Our main result. We are now ready for the proof of our main result.

Proof. We consider two graphs of interest: (i) in the case when we study $Y = C(\mathcal{E})$, the useful part $\mathcal{R}$ of the automaton $\mathcal{A}$ which recognizes the language $\Sigma^\star \cdot \mathcal{E}$ – (ii) in the case when we study $Y = \Omega(\mathcal{W})$, where $\ell + 1$ is the maximal length of the words of $\mathcal{W}$, the de Bruijn graph $\mathcal{B}$ of length ℓ.
With hypotheses of the present theorem, Propositions 3 and 4, and Definition 3, the density transformer $\mathfrak{G}$ has dominant spectral properties, and, by conjugation and perturbation theory, this transmits to the quasi-inverses of marked operators $\mathbb{R}(u)$ or $\mathbb{B}(u)$, when u is near 1, which admit a spectral decomposition too. Then, with Proposition 2, the moment generating functions of cost Y_n behave as approximate n-th powers. We end with Theorem 0 [7] (See 2.1). □

Conclusions. In this paper, as in [9], we restrict ourselves to the case when the expression $\mathcal{E}$ is simple. In the case when there does not exist a unique FSCC [see Section 2.4], all these FSCC's may play a rôle in the asymptotics, via their dominant eigenvalues. Our theorem extends to the general case by dealing with the super-dominant eigenvalues (which dominate the others).

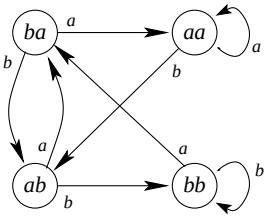
Acknowledgements. We wish to thank Julien Clément, Maxime Crochemore, Philippe Flajolet, and Véronique Terrier for their help.



$$ {}^t\mathcal{R} := \begin{pmatrix} \{b\} & \{b\} & \{b\} \\ \{a,c\} & \{c\} & \{c\} \\ \oslash & \{a\} & \{a\} \end{pmatrix}, \qquad {}^t\mathcal{M} := (\,\{a\}\,), \qquad {}^t\mathcal{U} := \begin{pmatrix} \{b\} \\ \{c\} \\ \oslash \end{pmatrix}. $$

$$ \mathbb{M} := (\,\mathbf{G}[a]\,), \quad \mathbb{R}(u) := \begin{pmatrix} \mathbf{G}_{[b]} & \mathbf{G}_{[b]} & \mathbf{G}_{[b]} \\ \mathbf{G}_{[a]} + \mathbf{G}_{[c]} & \mathbf{G}_{[c]} & \mathbf{G}_{[c]} \\ 0 & u \cdot \mathbf{G}_{[a]} & u \cdot \mathbf{G}_{[a]} \end{pmatrix}, \quad \mathbb{U}(u) := \begin{pmatrix} \mathbf{G}_{[b]} \\ \mathbf{G}_{[c]} \\ 0 \end{pmatrix} $$

Fig. 1. The automaton relative to $\Sigma^\star.\mathcal{E}$ with $\mathcal{E} = (ba|c)^+a^+$, the transition matrices, and the marked operators.



$$ {}^t\mathcal{B} = \begin{array}{c} \\ aa \\ ab \\ ba \\ bb \end{array} \begin{array}{c} \begin{array}{cccc} aa & ab & ba & bb \end{array} \\ \begin{pmatrix} \{a\} & \oslash & \{a\} & \oslash \\ \{b\} & \oslash & \{b\} & \oslash \\ \oslash & \{a\} & \oslash & \{a\} \\ \oslash & \{b\} & \oslash & \{b\} \end{pmatrix} \end{array}. $$

$$ \mathbb{B} = \begin{pmatrix} \mathbf{G}_{[a]} & 0 & \mathbf{G}_{[a]} & 0 \\ \mathbf{G}_{[b]} & 0 & \mathbf{G}_{[b]} & 0 \\ 0 & \mathbf{G}_{[a]} & 0 & \mathbf{G}_{[a]} \\ 0 & \mathbf{G}_{[b]} & 0 & \mathbf{G}_{[b]} \end{pmatrix}. \qquad \mathbb{B}(u) = \begin{pmatrix} \mathbf{G}_{[a]} & 0 & \mathbf{G}_{[a]} & 0 \\ u^2 \cdot \mathbf{G}_{[b]} & 0 & u \cdot \mathbf{G}_{[b]} & 0 \\ 0 & \mathbf{G}_{[a]} & 0 & \mathbf{G}_{[a]} \\ 0 & u \cdot \mathbf{G}_{[b]} & 0 & \mathbf{G}_{[b]} \end{pmatrix}. $$

Fig. 2. The De Bruijn automaton, with its transition matrix, the operator, with its use for $\mathcal{W} = \{ab, aab, aba\}$ and the marked operator.

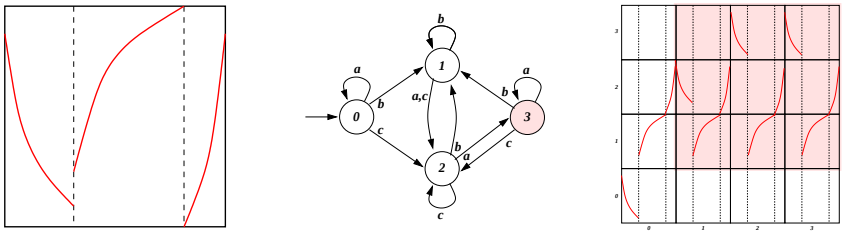


Fig. 3. The original source $\mathcal{S}$, the automaton $\mathcal{A}$ relative to $\Sigma^\star\mathcal{E}$, and the mixed sources $\mathcal{S}_\mathcal{A}$, $\mathcal{S}_\mathcal{R}$ when the states are restricted to be in $\mathcal{Y} := \{1, 2, 3\}$.

References

1. J. Bourdon, B. Vallée, Generalized pattern matching statistics. In Birkhauser, T.i.M., ed.: Mathematics and Computer Science II. (2002) 249–265
2. J. Bourdon, Size and path length of Patricia tries: dynamical sources context. Random Structures Algorithms **19** (2001) 289–315
3. E. Bender, F. Kochman, The distribution of subword counts is usually normal. European Journal of Combinatorics **14** (1993) 265–275
4. F. Chazal, V. Maume-Deschamps, B. Vallée, Systèmes dynamiques et algorithmique. In INRIA Research Report 5003. (2003) 121–150
5. J. Clément, P. Flajolet, B. Vallée, Dynamical sources in information theory: a general analysis of trie structures. Algorithmica **29** (2001) 307–369
6. P. Flajolet, W. Szpankowski, B. Vallée, Hidden word statistics. to appear in Journal de l'ACM (2005)
7. H.K. Hwang, Théorèmes limites pour les structures combinatoires et les fonctions arithmétiques. PhD thesis, Ecole Polytechnique, Palaiseau, France (1994)
8. P. Nicodème, T. Doerks, M. Vingron, Proteome analysis based on motif statistics. Bioinformatics **18** (2002) 161–171
9. P. Nicodème, B. Salvy, P. Flajolet, Motif statistics. Theoretical Computer Science **287** (2002) 593–617
10. M. Régnier, W. Szpankowski, On the approximate pattern occurrences in a text. In: Proc. SEQUENCE'97, IEEE Computer Society. (1997) 253–264
11. M. Régnier, W. Szpankowski, On pattern frequency occurrences in a Markovian sequence. Algorithmica **22** (1998) 631–649
12. I. Rigoutsos, A. Floratos, L. Parida, Y. Gao, D. Platt, The emergence of pattern discovery techniques in computational biology. J. of Met. Eng. **2** (2000) 159–177
13. R. Sedgewick, P. Flajolet, An introduction to the analysis of algorithms. Foreword by D. E. Knuth. Amsterdam: Addison-Wesley. xv, 492 p. (1996)
14. W. Szpankowski, Average case analysis of algorithms on sequences. Wiley-Interscience Series in Discrete Mathematics and Optimization (2001)
15. B. Vallée, Euclidean Dynamics to appear in Discrete and Continuous Dynamical Systems, 2005, web page: `www.info.unicaen.fr/~brigitte`
16. B. Vallée, Dynamical sources in information theory: fundamental intervals and word prefixes. Algorithmica **29** (2001) 262–306
17. A. Vanet, L. Marsan, M.F. Sagot, Promoter sequences and algorithmical methods for identifying them. Research in Microbiology **150** (1999) 779–799
18. M. S. Waterman, Introduction to Computational Biology. Chapman & Hall (1995)