# Pattern Occurrences in Multicomponent Models

Massimiliano Goldwurm and Violetta Lonati

Dip. Scienze dell'Informazione, Università degli Studi di Milano
Via Comelico 39/41, 20135 Milano – Italy
{goldwurm, lonati}@dsi.unimi.it

**Abstract.** In this paper we determine some limit distributions of pattern statistics in rational stochastic models, defined by means of nondeterministic weighted finite automata. We present a general approach to analyse these statistics in rational models having an arbitrary number of connected components. We explicitly establish the limit distributions in the most significant cases; these ones are characterized by a family of unimodal density functions defined by polynomials over adjacent intervals.

**Keywords:** Automata and Formal Languages, Limit Distributions, Nonnegative Matrices, Pattern Statistics, Rational Formal Series.

## 1 Introduction

This work presents some results on the limit distribution of pattern statistics. The major problem in this context is to estimate the frequency of pattern occurrences in a random text. This is a classical problem that has applications in several research areas of computer science and biology: for instance, it is considered in connection with the search of motifs in DNA sequences [6, 14] while the earlier motivations are related to code synchronization [10] and approximated pattern-matching [12, 18, 5]. In the usual setting, established in the seminal paper [11] and developed in many subsequent works (see for instance [15, 13, 3]), one considers a finite alphabet $\Sigma$, a set of patterns $R \subseteq \Sigma^*$, a probabilistic source $P$ generating words at random in $\Sigma^*$, and studies the number $X_n$ of occurrences of elements of $R$ in a word of length $n$ generated by $P$. Typical goals are the asymptotic evaluation of the moments of $X_n$, its limit distribution (also in the local sense) and the corresponding large deviations. These results depend in particular on the stochastic model $P$, which is usually assumed to be a Bernoulli or a Markovian model.

A rather general result is presented in [13], where Gaussian limit distributions are obtained, for any regular set of patterns $R$ and any Markovian source $P$, under a primitivity hypothesis on the associated stochastic matrix. This result is extended in [2] to the so-called *rational stochastic model*, where the text is generated at random according to a probability distribution defined by means of a rational formal series in noncommutative variables. In particular cases, this is simply the uniform distribution over the set of words of given length in an

V. Diekert and B. Durand (Eds.): STACS 2005, LNCS 3404, pp. 680–692, 2005.

arbitrary regular language. For this reason, results for this model are also related to the analysis of additive functions over strings [9].

The rational stochastic model properly extends the Markovian models in the following sense: the frequency problem of regular patterns in a text generated in the Markovian model (as studied in [13]) is a special case of the frequency problem of a single symbol in a text over a binary alphabet generated in the rational stochastic model; it is also known that the two models are not equivalent [2]. We recall that extensions of the Markovian models have already been considered in the literature [3]. Furthermore, finding results under more general probabilistic assumptions is of interest since, for some applications, the Markovian models seem to be too restrictive.

Also in the rational stochastic models, Gaussian limit distributions are obtained under a primitive hypothesis, i.e. when the matrix associated with the rational formal series (counting the transitions between states) is primitive [2]. A complete study of the limit distributions is given in [4] in the bicomponent models, that is when the previous matrix has two primitive components.

In this paper we present a general approach to the analysis of multicomponent rational models, explicitly establishing the limit distribution in the most significant cases. The paper is organized as follows. In Section 2 we give the definition and the main properties of rational models. In Section 3 we show how this model can be decomposed and we introduce the notions of main chain and simple model. Under a special assumption on the main chain, in Section 4 we determine the limit distributions of pattern statistics for simple models. They are characterized by an interesting family of unimodal density functions defined by polynomials over adjacent intervals. Finally in Section 5 we extend the results to all simple models and also provide a natural method to determine the limit distribution in the general case.

## 2 Rational Models for Pattern Statistics

In this section we recall some basic notions on rational formal series [16, 1] and the corresponding stochastic models to study the number of symbol occurrences in words chosen at random.

Let $\mathbb{R}_+$ be the semiring of all nonnegative real numbers and consider a finite alphabet $\Sigma$. A formal series over $\Sigma$ with coefficients in $\mathbb{R}_+$ is a function $r : \Sigma^* \longrightarrow \mathbb{R}_+$, usually represented in the form $r = \sum_{\omega \in \Sigma^*} (r, \omega) \cdot \omega$, where $(r, \omega)$ denotes the value of $r$ at $\omega \in \Sigma^*$. Moreover, $r$ is called *rational* if it admits a *linear representation*, that is a triple $(\xi, \mu, \eta)$ where, for some integer $m > 0$, $\xi$ and $\eta$ are (column) vectors in $\mathbb{R}_+^m$ and $\mu : \Sigma^* \longrightarrow \mathbb{R}_+^{m \times m}$ is a monoid morphism, such that $(r, \omega) = \xi^T \mu(\omega)\ \eta$ holds for each $\omega \in \Sigma^*$. Observe that considering such a triple $(\xi, \mu, \eta)$ is equivalent to defining a (weighted) nondeterministic automaton, where the state set is given by $\{1, 2, \ldots, m\}$ and the transitions, the initial and the final states are assigned weights in $\mathbb{R}_+$ by $\mu$, $\xi$ and $\eta$, respectively. To avoid redundancy it is convenient to assume that $(\xi, \mu, \eta)$ is trim (meaning that all indices are used to define the series), i.e. for every index $i$ there are two indices $p, q$ and two words $x, y \in \Sigma^*$ such that $\xi_p \mu(x)_{pi} \neq 0$ and $\mu(y)_{iq}\eta_q \neq 0$. We say that $(\xi, \mu, \eta)$ is *primitive* if $M = \sum_{\sigma\in\Sigma} \mu(\sigma)$ is a primitive matrix, that is for some $n \in \mathbb{N}$ all entries of $M^n$ are strictly positive. We also recall that a matrix $M \in \mathbb{R}_+^{m\times m}$ is called *irreducible* if for every pair of indices $p, q$ there exists $n \in \mathbb{N}$ such that $M^n_{pq} > 0$.

Any formal series can define a stochastic model for studying the frequency of occurrences of a letter in a word of given length. Consider the binary alphabet $\{a, b\}$ and, for any $n \in \mathbb{N}$, let $\{a, b\}^n$ denote the set of all words of length $n$ in $\{a, b\}^*$. Consider a formal series $r : \{a, b\}^* \longrightarrow \mathbb{R}_+$ and let $n$ be a positive integer such that $(r, x) \neq 0$ for some $x \in \{a, b\}^n$. A probability measure over $\{a, b\}^n$ can be defined by setting

$$\Pr\{\omega\} = \frac{(r, \omega)}{\sum_{x\in\{a,b\}^n}(r, x)} \qquad (\omega \in \{a, b\}^n). \tag{1}$$

In particular, if $r$ is the characteristic series $\chi_L$ of a language $L \subseteq \{a, b\}^*$, then Pr is just the uniform probability function over $L \cap \{a, b\}^n$. Then, we define the random variable (r.v. for short) $Y_n : \{a, b\}^n \to \{0, 1, \ldots, n\}$ such that $Y_n(\omega) = |\omega|_a$ for every $\omega \in \{a, b\}^n$. For every $j = 0, 1, \ldots, n$, we have

$$\Pr\{Y_n = j\} = \frac{\sum_{|\omega|=n, |\omega|_a=j}(r, \omega)}{\sum_{x\in\{a,b\}^n}(r, x)} .$$

If $r = \chi_L$ for some $L \subseteq \{a, b\}^*$, then $Y_n$ represents the number of occurrences of $a$ in a word chosen at random in $L \cap \{a, b\}^n$ under uniform distribution.

When $r$ is rational, the probability space given by (1) defines a stochastic model we call *rational* stochastic model. It is a generalization of the Markovian models in the sense that the r.v.'s $Y_n$ for rational $r$ represent, in special cases, the number of occurrences of patterns from an arbitrary regular language in words generated at random by Markovian processes [2–Section 2.1].

Let $(\xi, \mu, \eta)$ be a linear representation for the rational series $r$ and set $\mathcal{A} = \mu(a)$, $\mathcal{B} = \mu(b)$, $\mathcal{M} = \mathcal{A} + \mathcal{B}$. To study the behaviour of the random variables $Y_n$ and in particular their limit distribution, it is useful to introduce the sequence of functions $\{r_n(z)\}_n$ in one complex variable $z$ defined by

$$r_n(z) = \sum_{x\in\{a,b\}^n} (r, x) \cdot e^{z|x|_a} = \xi^T (\mathcal{A}e^z + \mathcal{B})^n \eta.$$

Indeed, it is immediate to see that the characteristic function of $Y_n$ satisfies the relation

$$\Psi_{Y_n}(t) = \mathbb{E}(e^{itY_n}) = \frac{r_n(it)}{r_n(0)} \tag{2}$$

for $t \in \mathbb{R}$. We recall that a sequence of random variables $X_n$ converges in distribution to a random variable $X$ if and only if the sequence of characteristic functions $\Psi_{X_n}(t)$ pointwise converges to $\Psi_X(t)$ [7].

Now consider the generating function of $\{r_n(z)\}_n$. Note that $\sum_{n=0}^{\infty} r_n(z)w^n = \xi^T H(z, w)\eta$, where $H(z, w)$ is the matrix function defined by

$$H(z, w) = \sum_{n=0}^{\infty}(\mathcal{A}e^z + \mathcal{B})^n w^n = (I - w(\mathcal{A}e^z + \mathcal{B}))^{-1} . \tag{3}$$

If $\mathcal{M}$ is irreducible, by the Perron–Frobenius Theorem (see [17–Theorem 1.5]) it has a nonnegative real eigenvalue $\lambda$ of maximum modulus. Moreover, if $\mathcal{M}$ is primitive, then all other eigenvalues have modulus strictly lower than $\lambda$. If further $\mathcal{A} \neq 0 \neq \mathcal{B}$, then there are two constants $\beta \in (0,1)$, $\gamma > 0$, both depending on the matrix $\mathcal{M}$ and its eigenvectors (see [2] for details), such that, as $n$ tends to infinity, the following relations hold:

$$\mathbb{E}(Y_n) = \beta n + \mathrm{O}(1) \ , \qquad \mathbb{V}ar(Y_n) = \gamma n + \mathrm{O}(1) \ . \tag{4}$$

For sake of brevity we say that $\beta$ and $\gamma$ are the *mean constant* and the *variance constant* of the primitive matrix $\mathcal{M}$, respectively. Under the same hypothesis, one can also prove [2] that the distribution of $\frac{Y_n - \beta n}{\sqrt{\gamma n}}$ converges to the normal distribution of mean value 0 and variance 1.

## 3 Decomposition of a Rational Model

Up to now, the properties of $Y_n$ have been studied only in the primitive models [2] and in the case of two primitive components [4]. Here we present a general approach to deal with an arbitrary rational model. To this aim, we describe the construction of the reduced graph of the strongly connected components of the corresponding linear representation. This is a usual approach in the analysis of counting problems on regular languages (see for instance [8] for an application concerning trace languages).

Let $(\xi, \mu, \eta)$ be a linear representation over the alphabet $\{a, b\}$ with coefficients in $\mathbb{R}_+$. As in the previous section, set $\mathcal{A} = \mu(a)$, $\mathcal{B} = \mu(b)$, $\mathcal{M} = \mathcal{A} + \mathcal{B}$ and consider the directed graph defined by $\mathcal{M}$, where the set of nodes is $\{1, 2, \ldots, m\}$ and $(p, q)$ is an (oriented) edge if and only if $\mathcal{M}_{pq} \neq 0$. Then, let $C_1, C_2, \ldots, C_s$ be the strongly connected components of the graph and define $C_i$ *initial* (resp. *final*) if $\xi_p \neq 0$ (resp. $\eta_p \neq 0$) for some $p \in C_i$. The *reduced graph* of $(\xi, \mu, \eta)$ is then defined as the directed acyclic graph $G$ where $C_1, C_2, \ldots, C_s$ are the vertices and any pair $(C_i, C_j)$ is an edge if and only if $i \neq j$ and $\mathcal{M}_{pq} \neq 0$ for some $p \in C_i$ and some $q \in C_j$.

Up to a permutation of indices, the matrix $\mathcal{M}$ can be represented as a triangular block matrix of the form

$$\mathcal{M} = \begin{pmatrix} M_1 & M_{12} & M_{13} & \cdots & M_{1s} \\ 0 & M_2 & M_{23} & \cdots & M_{2s} \\ & & \cdots & & \\ 0 & 0 & 0 & \cdots & M_s \end{pmatrix}$$

where each $M_i$ corresponds to the strongly connected component $C_i$ and every $M_{ij}$ corresponds to the transitions from vertices of $C_i$ to vertices of $C_j$ in the original graph of $\mathcal{M}$. Also $\mathcal{A}$, $\mathcal{B}$, $\xi$ and $\eta$ admit similar decompositions: we define the matrices $A_i, A_{ij}, B_i, B_{ij}$ and the vectors $\xi_i, \eta_i$ in the corresponding way and we say that the component $C_i$ *degenerates* if $A_i = 0$ or $B_i = 0$. Since each $M_i$ is either irreducible or null, by the Perron–Frobenius Theorem it has a nonnegative real eigenvalue $\lambda_i$ of maximum modulus. We call *main eigenvalue* of $\mathcal{M}$ the value

$\lambda = \max\{\lambda_i \mid i = 1, 2, \ldots, s\}$ and we say that $C_i$ is a *dominant component* if $\lambda_i = \lambda$. Observe that $\lambda_i = 0$ only if $C_i$ reduces to a loopless single node and hence from now on we assume $\lambda > 0$. If further $M_i$ is primitive, we say that $C_i$ is a *primitive component*.

The block decomposition of $\mathcal{M}$ also induces a decomposition of the matrix $H(z, w)$ defined in (3). More precisely, the blocks under the diagonal are all null, while the upper triangular part is composed by a family of matrices, say $H_{ij}(z, w)$, $1 \leq i \leq j \leq s$. Note that the bivariate generating function $\xi^T H(z, w)\eta$, which is the main tool of our investigation, is now given by

$$\xi^T H(z,w)\eta = \sum_{n=0}^{\infty} \xi^T (\mathcal{A}e^z + \mathcal{B})^n \eta \cdot w^n = \sum_{1\leq i \leq j \leq s} \xi_i^T H_{ij}(z,w)\eta_j \ . \tag{5}$$

Setting $M_{ij}(z) = A_{ij}e^z + B_{ij}$ and reasoning by induction on $j - i$, one can prove that, for each $1 \leq j \leq s$, the following equality holds

$$H_{jj}(z,w) = (I - w(A_j e^z + B_j))^{-1} = \frac{\mathrm{Adj}(I - w(A_j e^z + B_j))}{\det(I - w(A_j e^z + B_j))} \ , \tag{6}$$

while for each $1 \leq i \leq j \leq s$ we have

$$H_{ij}(z,w) = \sum_{*} H_{i_1 i_1}(z,w) M_{i_1 i_2}(z) H_{i_2 i_2}(z,w) \cdots M_{i_{\ell-1} i_\ell}(z) H_{i_\ell i_\ell}(z,w) \cdot w^{\ell-1}, \tag{7}$$

where the sum $(*)$ is extended over all sequences of integers $(i_1, i_2, \ldots, i_\ell)$, $\ell \geq 2$ such that $i_1 = i$, $i_t < i_{t+1}$ for each $t = 1, \ldots, \ell - 1$ and $i_\ell = j$.

The previous equation suggests us to introduce the notion of chain of the reduced graph $G$ associated with $(\xi, \mu, \eta)$. A *chain* is a simple path in $G$, i.e. any sequence of distinct components $\kappa = (C_{i_1}, C_{i_2}, \ldots, C_{i_\ell})$, $\ell \geq 1$, such that $M_{i_j i_{j+1}} \neq 0$ for every $j = 1, 2, \ldots, \ell - 1$. We say that $\ell$ is the *length* of $\kappa$ while the *order* of $\kappa$ is the number of its dominant components. Let $\Gamma$ denote the family of all chains in $G$ starting with an initial component and ending with a final component. We say that a chain $\kappa$ is a *main chain* if $\kappa \in \Gamma$ and its order is maximal in $\Gamma$. We denote by $\Gamma_m$ the set of all main chains in $G$.

In Section 3.1 we illustrate the role of main chains, which leads us to study the simple but representative case when the model has just one main chain, say $\kappa$. We first determine the limit distribution of $Y_n$ when all dominant components of $\kappa$ are primitive, non-degenerate and have distinct mean constants. A similar approach can be developed when the above mean constants are partially or totally coincident.

For this reason we introduce the notion of simple model. Formally, we say that $(\xi, \mu, \eta)$ is a *simple* linear representation, or just a *simple model*, if $\Gamma_m$ contains only one chain $\kappa$ and, for every dominant component $C_i$ in $\kappa$, $M_i$ primitive and $A_i \neq 0 \neq B_i$. Note that, for such a matrix $M_i$, the mean constant $\beta_i$ and the variance constants $\gamma_i$ can be defined as in (4), $0 < \beta_i < 1$ and $\gamma_i > 0$.

In simple models the limit distribution of $Y_n$ first depends on the order $k$ of $\kappa$, i.e. the number of its dominant components. If $k \leq 2$ the limit distribution is known and derives from the analysis of the bicomponent models given in [4]:

- If $\kappa$ has only one dominant component $C_i$ then the limit distribution of $\frac{Y_n - \beta_i n}{\sqrt{\gamma_i n}}$ is a Gaussian distribution of mean value 0 and variance 1;
- If $\kappa$ has two dominant components $C_i$, $C_j$ then we have the following three subcases:
  1. If $\beta_i \neq \beta_j$ then $Y_n/n$ converges in law to a random variable uniformly distributed in the interval $[b_1, b_2]$, where $b_1 = \min\{\beta_i, \beta_j\}$ and $b_2 = \max\{\beta_i, \beta_j\}$;
  2. If $\beta_i = \beta_j = \beta$ but $\gamma_i \neq \gamma_j$ then the limit distribution of $\frac{Y_n - \beta n}{\sqrt{n}}$ is a mixture of normal distributions of mean value 0 and variance uniformly distributed in the interval $[c_1, c_2]$, where $c_1 = \min\{\gamma_i, \gamma_j\}$ and $c_2 = \max\{\gamma_i, \gamma_j\}$. In other words, $\frac{Y_n - \beta n}{\sqrt{n}}$ converges in law to a random variable with density function

$$f(x) = \frac{1}{c_2 - c_1} \int_{c_1}^{c_2} \frac{e^{-x^2/(2v)}}{\sqrt{2\pi v}} dv \quad ;$$

  3. If $\beta_i = \beta_j = \beta$ and $\gamma_i = \gamma_j = \gamma$ then the distribution of $\frac{Y_n - \beta n}{\sqrt{\gamma n}}$ again converges to a Gaussian distribution of mean value 0 and variance 1.

In Section 4 we determine the limit distribution for simple models having main chain (of arbitrary order) with distinct mean constants of the dominant components; this result generalizes point 1) above. In Section 5, we extend these results to all simple models (with partially or totally coincident mean constants of dominant components) and also to all models whose main chains are simple (i.e. with primitive, non-degenerate dominant components).

We observe that the only cases not covered by our analysis concern the rational models where some dominant component of main chain is either non-primitive or degenerate. In the first case periodicity phenomena occur while in the second one a large variety of possible behaviours can be obtained even in bicomponent models [4].

### 3.1 The Role of Main Chains

In this section we show how the main chains determine the limit distribution of the sequence $\{Y_n\}$ associated with the linear representation $(\xi, \mu, \eta)$. Intuitively, this is a consequence of two facts. First, by equation (2) the characteristic function of (a normalization of) $Y_n$ depends on the sequences $\{r_n(z)\}$ for $z$ near 0, and hence on the generating function $\xi^T H(z, w)\eta$. Second, by (5), this function is a sum of products of the form given in (7), each of which is identified by a chain.

Thus, let us examine such terms. First consider the case $i = j$ and hence the terms of the form $\xi_j^T H_{jj}(z, w)\eta_j$. Relation (6) implies that, as $z$ tends to 0, the singularities of each of its entries approach the inverses of eigenvalues of $M_j$. Then, we can distinguish three cases according whether $M_j$ is dominant and primitive, dominant but non-primitive, or non-dominant. In each of these cases, the Perron–Frobenius theory gives us the necessary information on the eigenvalues of $M_j$, that allows us to analyse the singularities of $\xi_j^T H_{jj}(z,w)\eta_j$ in a neighbourhood of $z=0$.

The results of this analysis can be applied to functions $\xi_i^T H_{ij}(z,w)\eta_j$ where $i \neq j$. Recalling (7), we consider an arbitrary chain $\kappa = (C_{i_1}, C_{i_2}, \ldots, C_{i_\ell})$ with $\ell \geq 2$ and we define the sequence $\{{r_n}^{(\kappa)}(z)\}$ by setting

$$\sum_{n=0}^{\infty} r_n^{(\kappa)}(z) w^n = \xi_{i_1}^T H_{i_1 i_1}(z,w) M_{i_1 i_2}(z) H_{i_2 i_2}(z,w) \cdots M_{i_{\ell-1} i_\ell}(z) H_{i_\ell i_\ell}(z,w) \eta_{i_\ell} \cdot w^{\ell-1}. \qquad (8)$$

Then one can prove that for $z = c/n$, $c \in \mathbb{C}$, the terms corresponding to the main chains have singularities of smallest modulus with the largest degree, and hence they yield the main asymptotic contribution to the associated sequence $\{r_n(c/n)\}$. Formalizing the previous intuitive argument, one gets the following result.

**Theorem 1.** *If all dominant components of the main chains are primitive and non-degenerate then, for every constant $c \in \mathbb{C}$, we have*

$$r_n(c/n) = \sum_{\kappa \in \Gamma_m} r_n^{(\kappa)}(c/n) \ \ (1 + O(1/n)) = \Theta(\lambda^n n^{k-1})$$

*where $k$ is the order of the main chains.*[1]

We observe that Theorem 1 may not hold if the main chains admit non-primitive dominant components.

# 4 Main Results

In this section we determine the limit distribution of $Y_n$ in the simple models that satisfy the following additional property: the dominant components of the main chain have (pairwise) distinct mean constants. This is related to a special family of distribution functions we call *polynomial*.

Consider a tuple $b = (b_1, b_2, \ldots, b_k)$ of $k \geq 2$ real numbers such that $0 < b_1 < b_2 < \cdots < b_k < 1$ and let $f_b : \mathbb{R} \longrightarrow \mathbb{R}$ be the function defined by

$$f_b(x) = \begin{cases} 0 & \text{if } x < b_1 \\ (k-1)\sum_{j=r}^{k} c_j (b_j - x)^{k-2} & \text{if } b_{r-1} \leq x < b_r \text{ for some } 1 < r \leq k \\ 0 & \text{if } x \geq b_k \end{cases} \qquad (9)$$

where $c_j = \prod_{i \neq j} (b_j - b_i)^{-1}$ for every $j = 1, 2, \cdots, k$. In the following we say that a random variable $X$ is a *polynomial* r.v. of parameters $b$ if $f_b$ is its density function. Note that if $k = 2$ then $f_b$ is the uniform density function over the interval $(b_1, b_2)$.

[1] In this work, for any pair of sequences $\{f_n\}$, $\{g_n\} \subseteq \mathbb{C}$, the expression $f_n = \Theta(g_n)$ means that there exist two positive constants $a, b$ such that $a|g_n| \leq |f_n| \leq b|g_n|$ holds for every $n$ large enough.

**Theorem 2.** *Let $Y_n$ be defined in a simple model of main chain $\kappa$ having order $k$ and let $\beta = (\beta_1, \ldots, \beta_k)$ be the tuple of mean constants of dominant components in $\kappa$ in non-decreasing order. If $k \geq 2$ and all $\beta_j$'s are distinct then $Y_n/n$ converges in law to a polynomial random variable of parameters $\beta$.*

*Sketch of the proof.* First consider $r_n(it/n)$. Theorem 1 allows us to focus on the contribution of ${r_n}^{(\kappa)}(it/n)$ corresponding to the main chain $\kappa$. Then, by the singularity analysis of its generating function (the right handside of equation (8)), one can show that for every $t \in \mathbb{R}$

$$r_n\left(\frac{it}{n}\right) = \sum_{h=0}^{k-1} S_h\left(\frac{it}{n}\right)\lambda^{n-h} D_h\left(\frac{it}{n}\right) \cdot (1 + \mathrm{O}(1/n)) \quad \text{as } n \to +\infty,$$

where, for each $h$, the function $S_h(z)$ is analytic at $z = 0$ and $D_h$ is defined by

$$D_h(it/n) = \sum_{j=1}^{k} \frac{(1 + it\beta_j/n)^{n-h+k-1}}{\prod_{\ell \neq j}(it\beta_j/n - it\beta_\ell/n)} \;\; \text{if } t \neq 0, \quad D_h(0) = \binom{n-h+k-1}{k-1}.$$

Recalling that the characteristic function $\Psi_{Y_n/n}(t)$ equals $r_n(it/n)/r_n(0)$, one can show that, as $n$ tends to infinity, $\Psi_{Y_n/n}(t)$ converges to

$$\Phi_\beta(t) = \frac{(k-1)!}{(it)^{k-1}} \sum_{j=1}^{k} \frac{e^{i\beta_j t}}{\prod_{\ell \neq j}(\beta_j - \beta_\ell)}.$$

Finally, one can prove that $f_\beta(x)$ is a density function such that $\Phi_\beta(t)$ is its characteristic function (for details see Proposition 7). □

The properties of the family of polynomial distributions, together with the most relevant parts of the proof of Theorem 2, are all based on the convolutions of sequences defined by powers of complex numbers. In the following section we illustrate such properties and give some details of the proof sketched above.

### 4.1 Polynomial Distributions

Let us first consider the function $G_a(w) = w^{k-1} \cdot \prod_{i=1}^{k}(1 - a_i w)^{-1}$ where the tuple $a = (a_1, a_2, \ldots, a_k)$ has $k \geq 2$ nonnull complex components. Then $G_a$ is the generating function of the convolution of the sequences $\{a_1^n\}_n, \{a_2^n\}_n, \ldots, \{a_k^n\}_n$ shifted of $k-1$ indices. More precisely, at the point $w = 0$ such a function admits the power series expansion $G_a(w) = \sum_{n \geq 0} g_a(n) w^n$ such that

$$g_a(n) = \begin{cases} 0 & \text{if} \quad 0 \leq n \leq k-2 \\ \sum_{*} a_1^{i_1} a_2^{i_2} \cdots a_k^{i_k} & \text{if} \quad n \geq k-1 \end{cases} \tag{10}$$

where the sum $(*)$ is extended over all $k$-tuples $(i_1, \ldots, i_k) \in \mathbb{N}^k$ such that $i_1 + \cdots + i_k = n - k + 1$. When all $a_j$'s are distinct, the following proposition allows us to express the terms of the sequence $\{g_a(n)\}_{n \geq 0}$ in a useful form and provides us an important relationship among the $a_j$'s.

**Proposition 3.** *Let $a = (a_1, a_2, \dots, a_k)$ be a tuple of $k \geq 2$ distinct nonnull complex numbers and let the sequence $\{g_a(n)\}_n$ be defined by (10). Then, for every $n \in \mathbb{N}$, we have*

$$g_a(n) = \sum_{j=1}^{k} c_j \; a_j^n$$

*where $c_j = \prod_{i \neq j}(a_j - a_i)^{-1}$ for every $j = 1, 2, \cdots, k$. Moreover, the polynomial $\sum_j c_j (a_j - x)^s$ is identically null for each $0 \leq s \leq k-2$ and in particular $\sum_j c_j a_j^s = 0$. Finally we have $\sum_j c_j a_j^{k-1} = 1$ .*

The application of the previous proposition yields the following results on $f_b$.

**Proposition 4.** *If $k \geq 3$ then $f_b$ is continuously differentiable all over $\mathbb{R}$ up to the order $k-3$. Moreover the $(k-2)$-th derivative of $f_b$ is well defined in $\mathbb{R} \setminus \{b_1, \dots, b_k\}$ and is constant in each of the intervals $(b_i, b_{i+1})$, $i = 1, \cdots, k-1$.*

**Lemma 5.** *Let $f : \mathbb{R} \to \mathbb{R}$ be a function admitting $j$-th derivative all over $\mathbb{R}$ for some $j \geq 1$. Also assume that, for some reals $a < b$, $f$ has $m$ zeros in $(a, b)$ and $f(x) = 0$ for each $x \leq a$ or $x \geq b$. Then, for every $i = 1, \dots, j$, the $i$-th derivative of $f$ admits at least $m + i$ zeros in $(a, b)$.*

*Proof.* We reason by induction on $i = 1, \dots, j$. If $i = 1$, then consider the $m+1$ intervals determined by the zeros of $f$ in $[a, b]$. For each of them, say $(x_1, x_2)$, Rolle's Theorem guarantees that $f'(x) = 0$ for some $x \in (x_1, x_2)$.

Now assume $1 < i < j$ and consider the $i$-th derivative of $f$, that is $g = f^{(i)}$. By the properties of $f$, we have $g(a) = g(b) = 0$ and by the inductive hypotheses $g$ admits $m + i$ zeros in $(a, b)$. Therefore, by applying the previous argument to $g$, one proves that $g' = f^{(i+1)}$ admits $m + i + 1$ zeros in $(a, b)$. □

**Proposition 6.** *For every $k \geq 3$, the function $f_b$ is nonnegative and admits a unique maximum all over $\mathbb{R}$.*

*Proof.* If $k = 3$ the property follows by a direct inspection of the function, which is linear and nonnull in the intervals $(b_1, b_2)$ and $(b_2, b_3)$. If $k \geq 4$, let us consider the $(k-3)$-th derivative ${f_b}^{(k-3)}(x)$ of $f_b(x)$. It is immediate to see that ${f_b}^{(k-3)}(x)$ is linear with respect to $x$ in each of the $k-1$ intervals $(b_i, b_{i+1})$, $i = 1, \dots k-1$. Moreover, by Proposition 3, it does not vanish in $(b_1, b_2) \cup (b_{k-1}, b_k)$. Thus, ${f_b}^{(k-3)}$ has at most $k-3$ many zeros in $(b_1, b_k)$.

Now, assume by contradiction that $f_b$ is not unimodal. Then its derivative $f_b'$ vanishes in at least 3 points in the interval $(b_1, b_k)$ and hence $f_b'$ satisfies the hypotheses of Lemma 5 with $i = k-4$ and $m = 3$. As a consequence, ${f_b}^{(k-3)}$ admits at least $k-1$ zeros in $(b_1, b_k)$, which contradicts the previous property. □

Fig.1 and Fig.2 show the graphics of the functions $f_b$ having parameters $b = (0.1,\ 0.3, 0.4, 0.8)$ and $b = (0.008, 0.95, 0.96, 0.97, 0.98, 0.99)$, respectively. In each figure the first picture represents the entire curve, while the others show the details of the function in some subintervals. The vertical bars indicate the values

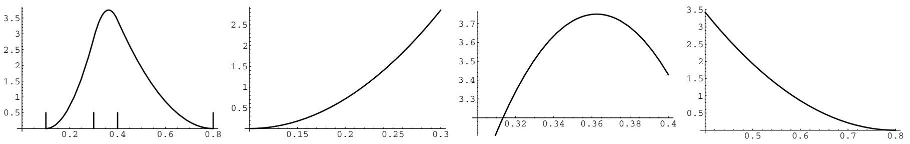

**Fig. 1.** Graphics of the function $f_b(x)$, where $b_1 = 0.1$, $b_2 = 0.3$, $b_3 = 0.4$, $b_4 = 0.8$. The vertical bars indicate the values of $b_j$'s

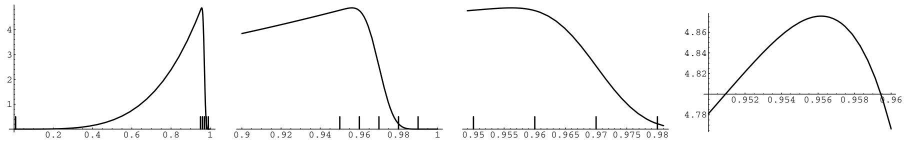

**Fig. 2.** Graphics of the function $f_b(x)$, where $b_1 = 0.008$, $b_2 = 0.95$, $b_3 = 0.96$, $b_4 = 0.97$, $b_5 = 0.98$, $b_6 = 0.99$. The vertical bars indicate the values of $b_j$'s

of $b_j$'s. Note that if $k = 4$ the maximum necessarily lays in the intermediate interval $(b_2, b_3)$. On the contrary, if $k > 4$ the maximum can lay in any interval between $b_2$ and $b_{k-1}$. For instance in Fig. 2, due to the asymmetric position of the points $b_j$'s, it lays in the second interval $(b_2, b_3)$.

**Proposition 7.** *For every $b = (b_1, b_2, \ldots, b_k) \in \mathbb{R}^k$ such that $0 < b_1 < b_2 < \cdots < b_k < 1$ and $k \geq 2$, $f_b(x)$ is a density function and $\Phi_b(t)$ is its characteristic function.*

*Proof.* Using Proposition 3, one can show that $\lim_{t\to 0} \Phi_b(t) = 1$ by a direct computation. Therefore, it suffices to show that $\int_{-\infty}^{+\infty} f_b(x) e^{itx} dx = \Phi_b(t)$ for every $t \in \mathbb{R}$. We prove this equality by using Proposition 3 again. Set $I(t) = \int_{-\infty}^{\infty} f_b(x) e^{itx} dx$ and $c_j = \prod_{i \neq j} (b_j - b_i)^{-1}$ for every $j = 1, \ldots, k$. Observe that

$$I(t) = (k-1) \sum_{r=2}^{k} \sum_{j=r}^{k} c_j \int_{b_{r-1}}^{b_r} (b_j - x)^{k-2} e^{itx} dx \, .$$

Integrating by parts one can verify that for $t \neq 0$ the function $e^{itx}(c-x)^p$ admits the antiderivative

$$\frac{e^{itx}}{it} \sum_{s=0}^{p} \frac{p! \, (c-x)^{p-s}}{(p-s)! \, (it)^s} .$$

Hence we can write $I(t) = \sum_{r=2}^{k} \sum_{j=r}^{k} c_j (A_{r,j} - A_{r-1,j})$ where

$$A_{r,j} = e^{itb_r} \sum_{s=0}^{k-2} \frac{(k-1)! \, (b_j - b_r)^{k-2-s}}{(k-2-s)! \, (it)^{s+1}} \quad \text{and in particular} \quad A_{r,r} = \frac{(k-1)!}{(it)^{k-1}} \, e^{itb_r} .$$

Now set $B_r = \sum_{j=r}^{k} c_j A_{r,j}$ and $C_r = \sum_{j=r}^{k} c_j A_{r-1,j}$. For each $2 \leq r \leq k-1$ we have $B_r - C_{r+1} = c_r A_{r,r}$ and moreover $B_k = c_k A_{k,k}$. Finally, by Proposition 3 we have $C_2 = \sum_{j=1}^{k} c_j A_{1,j} - c_1 A_{1,1} = -c_1 A_{1,1}$. As a consequence we get the result, since the integral can be computed as follows

$$I(t) = \sum_{r=2}^{k} (B_r - C_r) = \sum_{j=1}^{k} c_j A_{j,j} = \frac{(k-1)!}{(it)^{k-1}} \sum_{j=1}^{k} c_j e^{itb_j} = \Phi_b(t) \ . \qquad \square$$

## 5 Further Developments

The analysis presented in the previous section can be extended to all simple models, also when the mean constants $\beta_j$'s (associated with the dominant components of the main chain) are partially or totally coincident. The limit distributions of our statistics in this more general case are defined extending the notion of polynomial density function given in (9) by allowing multiplicities in the associated tuple $b$ and proving an analogue of Proposition 3 for convolutions with multiplicities.

To state these results precisely we only have to introduce the following characteristic function. Let $b = (b_1, b_2, \ldots, b_r)$ be a tuple of $r \geq 2$ distinct real numbers lying in the interval $(0,1)$ and let $m = (m_1, m_2, \ldots, m_r) \in \mathbb{N}^r$ be a tuple of multiplicities, where $m_j \geq 1$ for each $j$ and $m_1 + \ldots + m_r = k$. Then define the function

$$\Phi_{b,m}(t) = (k-1)! \sum_{j=1}^{r} \sum_{s=1}^{m_j} c_{j,s} \cdot \frac{e^{itb_j}}{(it)^{k-s}(s-1!)}$$

$$\text{where} \quad c_{j,s} = (-1)^{m_j - s} \sum_{\sum_{\ell \neq j} h_\ell = m_j - s} \ \prod_{\ell \neq j} \binom{m_\ell + h_\ell - 1}{m_\ell - 1} \cdot \frac{1}{(b_j - b_\ell)^{m_\ell + h_\ell}} \ .$$

One can prove that this is a characteristic function and the corresponding density function can be obtained from (9) by a continuity argument. The main difference is that the new density may be non-continuous at the points $x = b_j$ such that $m_j > 1$, $j = 1, \ldots, k$.

Now, let $Y_n$ be defined in a simple model having main chain $\kappa$ of order $k$. Let $\beta_1, \ldots, \beta_k$ and $\gamma_1, \ldots, \gamma_k$ be, respectively, the mean and variance constants of the dominant components in $\kappa$. We also denote by $\beta$ and $\gamma$ the tuples of distinct $\beta_j$'s and $\gamma_j$'s in increasing order and by $u$ and $v$ the tuples of the corresponding multiplicities. Clearly, if $\beta_1, \ldots, \beta_k$ are pairwise distinct then Theorem 2 applies. Otherwise we have the following cases:

- If $\beta_1, \ldots, \beta_k$ are partially but not totally coincident (i.e. $\beta_i = \beta_j$ and $\beta_s \neq \beta_t$ for some indices $i, j, s, t$, $i \neq j$), then $Y_n/n$ converges in distribution to a random variable of characteristic function $\Phi_{\beta,u}(t)$;
- If $\beta_j = \beta_1$ for all $j = 2, \ldots, k$ and all $\gamma_j$'s are pairwise distinct, then $\frac{Y_n - \beta_1 n}{\sqrt{n}}$ converges in distribution to a random variable of characteristic function $\Phi_\gamma(-t^2/(2i))$;

- If $\beta_j = \beta_1$ for all $j = 2, \ldots, k$ and $\gamma_1, \ldots, \gamma_k$ are partially but not totally coincident, then $\frac{Y_n - \beta_1 n}{\sqrt{n}}$ converges in distribution to a random variable of characteristic function $\Phi_{\gamma,v}(-t^2/(2i))$;
- If $\beta_j = \beta_1$ and $\gamma_j = \gamma_1$ for all $j = 2, \ldots, k$, then $\frac{Y_n - \beta_1 n}{\sqrt{\gamma_1 n}}$ converges in distribution to a normal random variable of mean 0 and variance 1.

The previous results can be further extended by a standard conditioning argument (already used in [4]) to all rational models $(\xi, \mu, \eta)$ whose main chains are "simple", i.e. for every $\kappa \in \Gamma_m$ all dominant components in $\kappa$ are primitive and non-degenerate. In this case, by equation (8), for every $\kappa \in \Gamma_m$ one can easily see that

$$r_n^{(\kappa)}(z) = s_\kappa(z)\lambda^n n^{k-1} + \mathrm{O}(\lambda^n n^{k-2})$$

where $k$ is the degree of $\kappa$ and $s_\kappa(z)$ is a nonnull analytic function at $z = 0$. Then, by Theorem 1, we have

$$r_n(0) = R\lambda^n n^{k-1} + \mathrm{O}(\lambda^n n^{k-2})$$

where $R = \sum_{\kappa \in \Gamma_m} s_\kappa(0)$. We can also associate each $\kappa \in \Gamma_m$ with the probability value $p_\kappa$, given by $p_\kappa = s_\kappa(0)/R$. Note that the values $\{p_\kappa\}_{\kappa \in \Gamma_m}$ define a discrete probability measure and they can be explicitly computed from the triple $(\xi, \mu, \eta)$.

Moreover, each $\kappa \in \Gamma_m$ defines a simple rational model in its own right, with an associate sequence of random variables $\{{Y_n}^{(\kappa)}\}$ having its own limit distribution according to Theorem 2 and list items above. In particular, ${Y_n}^{(\kappa)}/n$ always converges in distribution to a random variable of distribution function $F_\kappa(x)$ defined according to the previous results. Note that if all constants $\beta_j$'s are here equal, then $F_\kappa(x)$ reduces to the degenerate distribution of mass point $\beta_1$. Now it is not difficult to see that the overall statistics $Y_n/n$ converges in distribution to a r.v. of distribution function $F(x)$ defined by $F(x) = \sum_{\kappa \in \Gamma_m} F_\kappa(x) p_\kappa$.

## References

1. J. Berstel and C. Reutenauer. *Rational series and their languages*, Springer-Verlag, New York - Heidelberg - Berlin, 1988.
2. A. Bertoni, C. Choffrut, M. Goldwurm, and V. Lonati. On the number of occurrences of a symbol in words of regular languages. *Theoret. Comput. Sci.*, 302(1-3):431–456, 2003.
3. J. Bourdon and B. Vallée. Generalized pattern matching statistics. *Mathematics and computer science II: algorithms, trees, combinatorics and probabilities.* Proc. of Versailles Colloquium, Birkhauser, 249–265, 2002.
4. D. de Falco, M. Goldwurm, V. Lonati, Frequency of symbol occurrences in bicomponent stochastic models. *Theoret. Comput. Sci.*, 327 (3):269–300, 2004.
5. I. Fudos, E. Pitoura and W. Szpankowski. On pattern occurrences in a random text. *Inform. Process. Lett.*, 57:307–312, 1996.
6. M. S. Gelfand. Prediction of function in DNA sequence analysis. *J. Comput. Biol.*, 2:87–117, 1995.
7. B.V. Gnedenko. *The theory of probability* (translated by G. Yankovsky). Mir Publishers - Moscow, 1976.

8. M. Goldwurm, Probabilistic estimation of the number of prefixes of a trace, *Theoret. Comp. Sci.*, 92:249–268, 1992.
9. P. Grabner and M. Rigo. Additive functions with respect to numeration systems on regular languages. *Monatshefte für Mathematik*, 139: 205–219, 2003.
10. L. J. Guibas and A. M. Odlyzko. Maximal prefix-synchronized codes. *SIAM J. Appl. Math.*, 35:401–418, 1978.
11. L. J. Guibas and A. M. Odlyzko. String overlaps, pattern matching, and nontransitive games. *Journal of Combinatorial Theory. Series A*, 30(2):183–208, 1981.
12. P. Jokinen and E. Ukkonen. Two algorithms for approximate string matching in static texts. *Proceedings MFCS'91*, Lect. Notes in Comput. Sci., vol. n. 520, Springer, 1991, 248–248.
13. P. Nicodeme, B. Salvy, and P. Flajolet. Motif statistics. *Theoret. Comput. Sci.*, 287(2):593–617, 2002.
14. B. Prum, F. Rudolphe and E. Turckheim. Finding words with unexpected frequencies in deoxyribonucleic acid sequence. *J. Roy. Statist. Soc. Ser. B*, 57:205–220, 1995.
15. M. Régnier and W. Szpankowski. On pattern frequency occurrences in a Markovian sequence. *Algorithmica*, 22 (4):621–649, 1998.
16. A. Salomaa and M. Soittola. *Automata-Theoretic Aspects of Formal Power Series*, Springer-Verlag, 1978.
17. E. Seneta. *Non-negative matrices and Markov chains*, Springer–Verlag, New York Heidelberg Berlin, 1981.
18. E. Ukkonen. Approximate string-matching with $q$-grams and maximal matchings. *Theoret. Comput. Sci.*, 92: 191–211, 1992.