A symbolic method to compute the probability distribution of the number of pattern occurences in random texts generated by stochastic 0L-systems

Cedric Loi, Paul-Henry Cournède, Jean Françon

A symbolic method to compute the probability distribution of the number of pattern occurences in random texts generated by stochastic 0L-systems

Cédric Loi[1,2] Paul-Henry Cournède[1,2] Jean Françon[3]

[1] *Ecole Centrale Paris, Laboratory of Applied Mathematics and Systems, 92290 Châtenay Malabry, France*
[2] *INRIA Saclay - Île-de-France, DIGIPLANTE, 91400 Orsay, France*
[3] *Université de Strasbourg, France*

The analysis of pattern occurrences has numerous applications, in particular in biology. In this article, a symbolic method is proposed to compute the distribution associated to the number of occurences of a specific pattern in a random text generated by a stochastic 0L-system. To that purpose, a semiring structure is set for combinatorial classes composed of weighted words. This algebraic structure relies on new union and concatenation operators which, under some assumptions, are admissible constructions. Decomposing the combinatorial classes of interest by using these binary operators enables the direct translation of specifications into a set of functional equations relating generating functions thanks to transformation rules. The article ends with two examples. The first one deals with unary patterns and the connection with multitype branching process is established. The second one is about a pattern composed of two letters and underlines the importance of writing a proper specification.

Keywords: symbolic method, pattern occurrence, stochastic 0L-system, semiring, weighted words

1 Introduction

The problem of pattern occurrences in a random text has been widely studied for these last twenty years. Numerous applications exist in telecommunications, data compression and molecular biology. Concerning the latter, gene recognition and exceptionnal occurrence frequencies in DNA sequences are fundamental questions (see Pevzner et al. (1991) and Régnier and Szpankowski (1998)). Generally, DNA sequences appear as a random text generated by a probability model such as the Bernouilli model (see Guibas and Odlyzko (1981)) or the Markovian model (see Chrysaphinou and Papastavridis (1990)). In this article, we consider random texts that are generated by stochastic 0L-systems, which are generative parallel rewriting grammars. This type of probabilistic model is widely used in development biology for the modelling of cellular organism development or for plant growth (see Lindenmayer (1968) and Prusinkiewicz and Lindenmayer (1990)).

The aim of the article is to introduce a symbolic method that enables the computation of the probability distribution associated to the number of pattern occurrences in a random text generated by a stochastic 0L-system. As detailed in Flajolet and Sedgewick (2009), the symbolic method has numerous applications in

combinatorics. It is used to analyse characteristics of a specific type in a combinatorial class. The procedure consists first in writing a set of combinatorial equations based on admissible constructions and then in translating it into a system of equations relating generating functions. As far as random texts generated by a stochastic 0L-system are concerned, a semiring structure is established for finite or countable sets of weighted words. This structure relies on binary operators which guarantee admissible constructions.

Section 2 introduces the probabilistic framework. A generating function for sets composed of weighted words is defined. Then, Section 3 details the symbolic method. The first part of this section establishes an algebraic structure for sets of weighted words. A new union and a concatenation operator are created. Under some assumptions, they guarantee admissible constructions. The second part gives the different steps of the symbolic method. This article ends with two examples. The first one deals with unary patterns and the connection with Galton-Watson multitype branching processes is done. The second example deals with a more complex pattern.

2 Random words generated by stochastic 0L-systems

2.1 Description of the probabilistic model

Let $V = \{v_1, v_2, \ldots, v_m\}$ be an alphabet and W the set of all words built from V. The empty word is denoted by ϵ. Let $W^+ = W \setminus \{\epsilon\}$ be the set of all non-empty words built from V. Stochastic 0L-systems generate a Markov chain with infinite state space W. They are generative parallel rewriting grammars whose production rules are associated to a set of probability distributions (see Prusinkiewicz and Lindenmayer (1990) and Françon (1990)).

Definition 2.1 (Stochastic 0L-system) *A stochastic 0L-system is a construct $L = \langle a, \pi \rangle$ where :*

- *$a \in W^+$ is called an axiom,*
- *π is a transition matrix from V to W (i.e., $\forall (u,v) \in V \times W$, $0 \leq \pi_{u,v} \leq 1$ and $\sum_{w \in W} \pi_{u,w} = 1$) with a finite number of non-zero components.*

Remark : Definition 2.1 is equivalent to the one given by Prusinkiewicz and Lindenmayer (1990) : the set $P_r = \{(x,y) \in V \times W \mid \pi_{x,y} > 0\}$ defines the production rules. It represents the possible evolutions for all letters throughout the L-system.

A stochastic 0L-system $L = \langle a, \pi \rangle$ generates a random sequence of words $(w_n)_{n \in \mathbb{N}}$ built from the alphabet V. By definition, an axiom is the word which initiates the sequence generated by L. Then, $w_0 = a$. We get w_{n+1} by replacing randomly every letter x of w_n by a word y with a probability $\pi_{x,y}$ (note that the evolution of a letter is independent from the evolution of the other letters) . By doing, we create a Markov chain on W : $(w_n)_{n \in \mathbb{N}}$.

Proposition 2.1 *Let $L = \langle a, \pi \rangle$ be a stochastic 0L-system. Let $P = (P_{x,y})$ be the square matrix on W such that, for all $x = x_1 x_2 \cdots x_n \in W$ with $(x_1, x_2, \cdots, x_n) \in V^n$ and for all $y \in W$:*

$$P_{x,y} = \sum_{\substack{(y_1, y_2, \cdots, y_n) \in W^n, \\ y_1 y_2 \cdots y_n = y}} \prod_{i=1}^{n} \pi_{x_i, y_i}.$$

Then, P is a Markov kernel on W. By definition, P is called the Markov kernel associated to L.

Therefore, a stochastic 0L-system $L = \langle a, \pi \rangle$ creates a Markov chain $(w_n)_{n \in \mathbb{N}}$ with infinite state space W via the associated kernel P and the starting measure δ_a where δ_a is the Dirac measure concentrated on a.

Let us now define a more general class of L-systems called stochastic F0L-system, extending the classical definition of F0L-system (Rozenberg and Salomaa (1980), p. 89) to the stochastic case :

Definition 2.2 (Stochastic F0L-system) *A stochastic F0L-system is a construct $L = \langle A, \pi \rangle$ where :*

- *A is a nonempty subset of W^+ (called the set of axioms of L),*
- *for every $a \in A$, $L[a] = \langle a, \pi \rangle$ is a stochastic 0L-system (called component system of L).*

N.B. : The component systems $L[a]$ with $a \in A$ have the same Markov kernel P. Therefore, by definition, P is also called the Markov kernel associated to the stochastic F0L-system L.

In the sequel, $L = \langle A, \pi \rangle$ will denote a stochastic F0L-system. For all $n \in \mathbb{N}$ and $a \in A$, let $W_n^L[a]$ be the set of all words that can be generated by the component system $L[a]$ after n steps :

$$W_n^L[a] = \{w \in W \mid (P^n)_{a,w} > 0\}.$$

N.B. : Note that, in that case, P^n means P to the power of n. $(P^n)_{a,w}$ represents the component (a, w) of the matrix P^n.

2.2 Sets of weighted words

2.2.1 Some combinatorial concepts

In this section, some basic concepts of combinatorics are first recalled and then extended to a more general framework. Let us recall the definition of a combinatorial class (see Flajolet and Sedgewick (2009) for more details).

Definition 2.3 (Combinatorial Class) *A combinatorial class, or simply a class, is a finite or countable set on which a size function is defined, satisfying the following conditions :*

- *the size of an element is a non-negative integer,*
- *the number of elements of any given size is finite.*

When a non-negative real number can be associated to each element of a combinatorial class, the latter is said weighted :

Definition 2.4 (Weighted Combinatorial Class) *Let C be a combinatorial class. A weighted combinatorial class is a set $WC = \{(t, p_t) \mid t \in C\}$ such that :*

- $\forall t \in C$, $p_t \in \mathbb{R}_+$.
- $\sum_{t \in C} p_t < \infty$.

N.B. : if $WC = \{(t, p_t) \mid t \in C\}\}$ is a weighted combinatorial class, then C is said the combinatorial class associated to WC.

Definition 2.5 (Stochastic Combinatorial Class) *A stochastic combinatorial class is a weighted combinatorial class $SC = \{(t, p_t) \mid t \in C\}$ such that $\sum_{t \in C} p_t = 1$.*

The concept of generating functions associated to the size function in a combinatorial class can be easily extended to the weighted case :

Definition 2.6 (Generating Function (= GF)) *Let C be a combinatorial class and $WC = \{(t, p_t) \mid t \in C\}\}$ a weighted combinatorial class. Let s be a size function in C. The generating function Ψ associated to s in WC is the formal power series defined as follows :*

$$\Psi(z) = \sum_{t \in C} p_t z^{s(t)}.$$

2.2.2 Generating function associated to a pattern in a stochastic combinatorial class

The objective of this article is to compute the probability distribution associated to the number of occurrences of a specific pattern (*i.e.* a non-empty word) in a random text generated by a stochastic 0L-system. To that purpose, we introduce the counting function :

Definition 2.7 (Counting Function) *The counting function c is a map from $W^+ \times W^+$ to $\mathbb{N}$ such that, for all $(w, u) \in W^+ \times W^+$, $c(w, u)$ gives the number of patterns u in the word w.*

For all $u \in W^+$, $c(\bullet, u)$ is a size function in $W_n^L[a]$. Therefore, $W_n^L[a]$ is a combinatorial class with $c(\bullet, u)$ as size function. Let $SW_n^L[a] = \{(w, (P^n)_{a,w}) \mid w \in W_n^L[a]\}$. Then, $SW_n^L[a]$ is a stochastic combinatorial class.

Definition 2.8 (Generating function associated to a pattern) *Let u be a non-empty word. The GF associated to the size function $c(\bullet, u)$ in $SW_n^L[a]$ is called GF associated to the pattern u in $SW_n^L[a]$ and is denoted by $\Psi_n[a]$:*

$$\Psi_n[a](z) = \sum_{w \in W_n^L[a]} (P^n)_{a,w} \; z^{c(w,u)}.$$

In the sequel, W^{wgt} will denote the set of all weighted combinatorial classes whose associated combinatorial class is a subset of W (the set of all words built from the alphabet V). If $\mathcal{F}$ is an element of W^{wgt}, thus, there exists $F \subset W$ and a set of non-negative real weights $\{p_w \mid w \in F\}$ such that $\mathcal{F} = \{(w, p_w) \mid w \in F\}$.

3 A symbolic approach to analyse patterns in random texts

The symbolic method is a very effective method to analyse combinatorial structures and, as a consequence, plays an important role in analytic combinatorics (see Flajolet and Sedgewick (2009) for more details). In this section, we develop a symbolic method which enables the computation of the distribution associated to the number of occurrences of a given pattern $u \in W^+$ in the set $SW_n^L[a]$. Like all symbolic methods, the aim is to build a set of combinatorial equations using the stochastic combinatorial classes $SW_n^L[a]$ with $n \in \mathbb{N}$ and $a \in A$ and to translate it into a set of equations with the corresponding GFs.

3.1 An algebraic framework for sets of weighted words

To build the set of combinatorial equations, we need to set an algebraic structure for W^{wgt}. We define a union operator (different from the classical union operator $\cup$) and a concatenation operator ' . ' in W^{wgt}. These operators will be used to build admissible constructions (see Definition 6.1 in Appendix 6) for the symbolic method developed in Section 3.2.

Definition 3.1 (Union Operator ' + ') *Let F and G be two subsets of W. The union of the weighted combinatorial classes $\mathcal{F} = \{(w, p_w) \mid w \in F\}$ and $\mathcal{G} = \{(v, q_v) \mid v \in G\}$ is defined as follows :*

$$\mathcal{F} + \mathcal{G} = \left(\bigcup_{x \in F \setminus G} \{(x, p_x)\} \right) \cup \left(\bigcup_{x \in G \setminus F} \{(x, q_x)\} \right) \cup \left(\bigcup_{x \in F \cap G} \{(x, p_x + q_x)\} \right).$$

Example Let $V = \{c, d\}$ be an alphabet. Let $\mathcal{F} = \{(cd, p_1), (c, p_2)\}$ and $\mathcal{G} = \{(cd, p_3), (d, p_4)\}$ be two elements of W^{wgt}. Then :

$$\mathcal{F} + \mathcal{G} = \{(c, p_2), (d, p_4), (cd, p_1 + p_3)\}.$$

N.B.

1. Note that ' + ' is different from the classical union operator $\cup$. As a matter of fact, suppose that $\mathcal{F} \cap \mathcal{G} \neq \{\}$. Then, there exists $(w, p) \in \mathcal{F} \cap \mathcal{G}$. In that case, (w, p) will be an element of $\mathcal{F} \cup \mathcal{G}$ but not necessarily of $\mathcal{F} + \mathcal{G}$.
2. ' + ' is associative, commutative and has $\{\}$ (the empty set) as neutral element.

In the sequel, we will use the following notation :

$$\{(w_1, p_{w_1})\} + \{(w_2, p_{w_2})\} + \ldots + \{(w_n, p_{w_n})\} = \sum_{i=1}^{n} \{(w_i, p_{w_i})\}.$$

Definition 3.2 (Concatenation Operator ' . ') *Let F and G be two subsets of W. The concatenation of the weighted combinatorial classes $\mathcal{F} = \{(w, p_w) \mid w \in F\}$ and $\mathcal{G} = \{(v, q_v) \mid v \in G\}$ is defined as follows :*

$$\mathcal{F}.\mathcal{G} = \sum_{(w,v) \in F \times G} \{(w.v, p_w q_v)\}$$

where $w.v$ denotes the concatenation of the words w and v.

Example Let $V = \{c, d\}$ be an alphabet. Let $\mathcal{F} = \{(d, p_1), (dc, p_2)\}$ and $\mathcal{G} = \{(cc, p_3), (c, p_4)\}$ be two elements of W^{wgt}. Then :

$$\begin{aligned} \mathcal{F}.\mathcal{G} &= \{(dcc, p_1 p_3)\} + \{(dc, p_1 p_4)\} + \{(dccc, p_2 p_3)\} + \{(dcc, p_2 p_4)\} \\ &= \{(dc, p_1 p_4), (dccc, p_2 p_3), (dcc, p_1 p_3 + p_2 p_4)\}. \end{aligned}$$

By convention, we set $\mathcal{F}.\{\} = \{\}.\mathcal{F} = \{\}$. Thus, W^{wgt} has a semiring structure (see Duchamp et al. (2005) and Klima and Polak (2008) for examples of semirings in automata and language theory).

Definition 3.3 (Semiring) $(S, +, ., 0, 1)$ *is a semiring if :*

1. $(S, +, 0)$ *is a commutative monoid with neutral element* 0*,*
2. $(S, ., 1)$ *is a monoid with neutral element* 1*,*
3. *Multiplication distributes over addition,*
4. 0 *annihilates* S*, with respect to multiplication.*

Then, we have :

Theorem 3.1 $(W^{wgt}, \ + \ , \ . \ , \{\}, \{(\epsilon, 1)\})$ *is a semiring. The empty set* $\{\}$ *and* $\{(\epsilon, 1)\}$ *are respectively the neutral element for ' + ' and ' . '.*

Proof: The proof is immediate and relies on basic handling of the operators ' + ' and ' . '. Note that, for every $w \in W$, $w.\epsilon = \epsilon.w = w$. The fourth property of the semiring structure (see Definition 3.3) corresponds to the fact that, for all $\mathcal{F} \in W^{wgt}$, $\mathcal{F}.\{\} = \{\}.\mathcal{F} = \{\}$. □

The operators ' + ' and ' . ' are fundamental to build the set of combinatorial equations for the symbolic method. Under some assumptions, these operators are admissible constructions (whose definition is recalled in appendix, see Definition 6.1).

Definition 3.4 (Non-Generative Concatenation) *Let* $u \in W^+$ *be a pattern and* F *and* G *two subsets of* W*. The concatenation of* F *and* G *is said non-generative with respect to the pattern* u *if the following condition holds :*

$$\forall (w, v) \in F \times G, \quad c(w.v, u) = c(w, u) + c(v, u).$$

If not, the concatenation of F *and* G *is said generative with respect to* u*.*

By definition, the concatenation of $\mathcal{F} = \{(w, p_w) \mid w \in F\}$ and $\mathcal{G} = \{(v, q_v) \mid v \in G\}$ is said non-generative for the pattern u if the concatenation of F and G is non-generative with respect to u. The following theorem states that the union and concatenation operators are admissible constructions.

Theorem 3.2 (admissible constructions) *Let* $\mathcal{F}$*,* $\mathcal{G}$ *and* $\mathcal{H}$ *be three elements of* W^{wgt}*. Let* u *be an element of* W^+ *and* α*,*β *and* γ *be respectively the GFs associated to* u *in* $\mathcal{F}$*,* $\mathcal{G}$ *and* $\mathcal{H}$*. Let us assume that the concatenation of* $\mathcal{F}$ *and* $\mathcal{G}$ *is non-generative with respect to the pattern* u*. Therefore, the union operator ' + ' and the concatenation operator ' . ' are admissible constructions with the following transformation rules :*

$$\begin{array}{lll} \mathcal{H} = \mathcal{F} + \mathcal{G} & \Longrightarrow & \gamma(z) = \alpha(z) + \beta(z) \\ \mathcal{H} = \mathcal{F}.\mathcal{G} & \Longrightarrow & \gamma(z) = \alpha(z)\beta(z) \end{array}$$

Proof: Let u be in W^+. Suppose that $\mathcal{F} = \{(w, p_w) \mid w \in F\}$ and $\mathcal{G} = \{(v, q_v) \mid v \in G\}$.

– Suppose that $\mathcal{H} = \mathcal{F} + \mathcal{G}$. Then, we deduce from Definition 2.8 and Definition 3.1 :

$$\begin{aligned}\gamma(z) &= \sum_{x\in F\setminus G} p_x z^{c(x,u)} + \sum_{x\in G\setminus F} q_x z^{c(x,u)} + \sum_{x\in F\cap G} (p_x + q_x) z^{c(x,u)} \\ &= \sum_{w\in F} p_w z^{c(w,u)} + \sum_{v\in G} q_v z^{c(v,u)} = \alpha(z) + \beta(z).\end{aligned}$$

– Suppose that $\mathcal{H} = \mathcal{F}.\mathcal{G}$. Then, we deduce from Definition 2.8 and Definition 3.2 :

$$\gamma(z) = \sum_{(w,v)\in F\times G} (p_w q_v) z^{c(w.v,u)}.$$

Since the concatenation of $\mathcal{F}$ and $\mathcal{G}$ is non-generative with respect to the pattern u, then, for all $(w, v) \in F \times G$, $c(w.v, u) = c(w, u) + c(v, u)$. Therefore,

$$\gamma(z) = \sum_{(w,v)\in F\times G} \left(p_w z^{c(w,u)}\right)\left(q_v z^{c(v,u)}\right) = \left(\sum_{w\in F} p_w z^{c(w,u)}\right)\left(\sum_{v\in G} q_v z^{c(v,u)}\right) = \alpha(z).\beta(z).$$

□

N.B. : contrary to the standard union operator $\cup$, there is no need to impose $\mathcal{F} \cap \mathcal{G} = \{\}$ to make the operator ' + ' an admissible construction.

3.2 Description of the method

Let $L = \langle A, \pi\rangle$ be a stochastic F0L-system. For all $a \in A$, the component system $L[a]$ generates a Markov chain $(w_n)_{n\in\mathbb{N}}$ via the Markov kernel P associated to L and the starting measure δ_a (see Section 2.1). Let u be an element of W^+. The aim of this section is to set a symbolic method that enables the computation of the distribution associated to the number of patterns u in w_n for all $n \in \mathbb{N}$. To that purpose, we need to compute the GF associated to u in $SW_n^L[a] : \Psi_n[a]$. As a matter of fact, by reordering the terms of $\Psi_n[a]$, the GF can be expressed by the following power series :

$$\Psi_n[a](z) = \sum_{w\in W_n^L[a]} (P^n)_{a,w}\ z^{c(w,u)} = \sum_{l\in\mathbb{N}} p_l^{n,a}\ z^l$$

where $p_l^{n,a}$ represents the probability to get l patterns u in a word generated by the component system $L[a]$ after n steps. However, most of the time, it is impossible to compute directly $\Psi_n[a]$. In that case, $\Psi_n[a]$ is determined recursively by using a symbolic method.

Suppose we are interested in counting the number of characteristics of a given type c in a combinatorial class $\mathcal{A}$. Let s be the size function that counts the number of characteristics c and let Φ be the GF associated to s in $\mathcal{A}$. To get Φ, we need to find a specification for $\mathcal{A}$ (see Definition 6.2 in Appendix 6), *i.e.* a collection of combinatorial equations composed of admissible constructions. By using a set of transformation rules, we get a collection of functional equations (involving Φ) from the specification. Finally, the coefficients of Φ are determined from the functional equations.

When dealing with random words generated by stochastic 0L-systems, the combinatorial classes of interest are $SW_n^L[a]$ with $n \in \mathbb{N}$ and $a \in A$. Suppose we are interested in computing the distribution

$\{P_l^{N,a} \mid l \in \mathbb{N}\}$ for a given $N \in \mathbb{N}$. The first step of the symbolic method is to find an iterative specification for the set of stochastic combinatorial classes $\{SW_n^L[a] \mid n \in \{0,\dots,N\}\}$ which is a closed system of combinatorial equations using admissible constructions based on the operators ' + ' and ' . ' (see Section 3) and basic weighted combinatorial classes. A first idea to get these combinatorial equations is the use of Theorem 3.3 and Property 3.4 :

Theorem 3.3 (General Decomposition) *Suppose $A = W^+$. Then :*

$$\forall a \in W^+,\ \forall n \in \mathbb{N} : \quad SW_{n+1}^L[a] = \sum_{w \in W_1^L[a]} \{(\epsilon, \pi_{a,w})\}.SW_n^L[w].$$

Proof: Let $x \in SW_{n+1}^L[a]$. Then, there exists $s \in W_{n+1}^L[a]$ such that $x = (s, (P^{n+1})_{a,s})$ and $(P^{n+1})_{a,s} > 0$. By using the Chapman-Kolmogorov equation, we get :

$$(P^{n+1})_{a,s} = \sum_{w \in W_1^L[a]} P_{a,w}(P^n)_{w,s} = \sum_{w \in W_1^L[a]} \pi_{a,w}(P^n)_{w,s}. \tag{1}$$

Note that, since $(P^{n+1})_{a,s} > 0$, $(P^n)_{w,s} > 0$ for all $w \in W_1^L[a]$. Therefore :

$$\{x\} = \{(s, \sum_{w \in W_1^L[a]} \pi_{a,w}(P^n)_{w,s})\} = \sum_{w \in W_1^L[a]} \{(s, \pi_{a,w}(P^n)_{w,s})\} = \sum_{w \in W_1^L[a]} \{(\epsilon, \pi_{a,w})\}.\{(s, (P^n)_{w,s})\}.$$

Since $(P^n)_{w,s} > 0$, then $(s, (P^n)_{w,s}) \in SW_n^L[w]$ for all $w \in W_1^L[a]$. We deduce that, for all $x \in SW_{n+1}^L[a]$, $x \in \sum_{w \in W_1^L[a]} \{(\epsilon, \pi_{a,w})\}.SW_n^L[w]$ and therefore :

$$SW_{n+1}^L[a] \subset \sum_{w \in W_1^L[a]} \{(\epsilon, \pi_{a,w})\}.SW_n^L[w].$$

Conversely, let $x = (s, p_s)$ be in $\sum_{w \in W_1^L[a]} \{(\epsilon, \pi_{a,w})\}.SW_n^L[w]$. Then, there exists a sequence

$$\{(r_w, (P^n)_{w,r_w}) \mid w \in SW_1^L[w],\ (r_w, (P^n)_{w,r_w}) \in SW_n^L[w]\}$$

such that :

$$\{x\} = \{(s, p_s)\} = \sum_{w \in W_1^L[a]} \{(\epsilon, \pi_{a,w})\}.\{(r_w, (P^n)_{w,r_w})\} = \sum_{w \in W_1^L[a]} \{(r_w, \pi_{a,w}(P^n)_{w,r_w})\}.$$

The previous equation imposes $s = r_w$ for all $w \in W_1^L[a]$. Therefore, by using the Chapman-Kolmogorov equation (1), we have :

$$\{x\} = \sum_{w \in W_1^L[a]} \{(s, \pi_{a,w}(P^n)_{w,s})\} = \{(s, \sum_{w \in W_1^L[a]} \pi_{a,w}(P^n)_{w,s})\} = \{(s, (P^{n+1})_{a,s})\}.$$

with $(P^{n+1})_{a,s} > 0$. We deduce that, for all $x \in \sum_{w \in W_1^L[a]} \{(\epsilon, \pi_{a,w})\}.SW_n^L[w]$, $x \in SW_{n+1}^L[a]$. Therefore :

$$\sum_{w \in W_1^L[a]} \{(\epsilon, \pi_{a,w})\}.SW_n^L[w] \subset SW_{n+1}^L[a].$$

Finally,

$$SW_{n+1}^L[a] = \sum_{w \in W_1^L[a]} \{(\epsilon, \pi_{a,w})\}.SW_n^L[w].$$

□

Property 3.4 (Decomposition Property) *Suppose $A = W^+$. Let w_1, w_2,..., w_k be k words of W^+. Then :*

$$\forall n \in \mathbb{N}, \quad SW_n^L[w_1.w_2.\cdots.w_k] = SW_n^L[w_1].SW_n^L[w_2].\cdots.SW_n^L[w_k].$$

Proof: Let $k = 2$. Let m_1 and m_2 be respectively the number of letters in w_1 and w_2. Then, there exists $(v_1^1, \ldots, v_{m_1}^1) \in V^{m_1}$ and $(v_1^2, \ldots, v_{m_2}^2) \in V^{m_2}$ such that $w_1 = v_1^1.\cdots.v_{m_1}^1$ and $w_2 = v_1^2.\cdots.v_{m_2}^2$. Since each letter is acting independently (see Definition 2.1), we get :

$$SW_n^L[w_1.w_2] = SW_n^L[v_1^1.\cdots.v_{m_1}^1.v_1^2.\cdots.v_{m_2}^2] = SW_n^L[v_1^1].\cdots.SW_n^L[v_{m_1}^1].SW_n^L[v_1^2].\cdots.SW_n^L[v_{m_2}^2].$$

In the same way, for $j \in \{1, 2\}$:

$$SW_n^L[w_j] = SW_n^L[v_1^j.\cdots.v_{m_j}^j] = SW_n^L[v_1^j].\cdots.SW_n^L[v_{m_j}^j].$$

Therefore,

$$SW_n^L[w_1.w_2] = SW_n^L[w_1].SW_n^L[w_2].$$

The induction for $k > 2$ is straightforward. □

N.B. : Theorem 3.3 and Property 3.4 provide a combinatorial equation which appears as a natural decomposition for the combinatorial classes of interest. However, they do not always give admissible constructions (see Section 4.2).

Then, once the specification is determined, we use the transformation rules of Theorem 3.2 and we get a set of recursive functional equations involving $\Psi_n[a]$ with $n \in \{0, \ldots, N\}$ and $a \in A$. By doing so, we are able to compute recursively the coefficients of the power series $\Psi_N[a]$.

Suppose you have an alphabet V and a stochastic F0L-system $L = \langle A, \pi \rangle$. The symbolic method can be decomposed into the following steps :

1. Determine the objective : computing the distribution associated to the number of occurrences of a pattern $u \in W^+$ in a word generated randomly by the component system $L[a]$ after N steps with $N \in \mathbb{N}$ and $a \in A$.
2. Write the GF associated to u in $SW_N^L[a]$: $\Psi_N[a]$. The coefficients of $\Psi_N[a]$ seen as power series give the distribution of interest.
3. Write an iterative specification for the set of stochastic combinatorial classes $\{SW_n^L[a] \mid n \in \{0, \ldots, N\}\}$ using admissible constructions based on the union operator ' + ' and the concatenation operator ' . '.

4. Use the transformation rules of Theorem 3.2 and write a closed system of functional equations involving $\Psi_n[a]$ for $n \in \{0, \dots, N\}$.
5. Either solve directly the previous system or find a recursive set of equations satisfied by the coefficients of $\Psi_n[a]$ for $n \in \{0, \dots, N\}$.
6. Deduce the coefficients of $\Psi_N[a]$.

4 Examples

This section illustrates the symbolic method throughout two examples. The first one (Section 4.1) deals with unary patterns and the connection between the symbolic method and the multitype branching process approach is established. The second one (Section 4.2) deals with a pattern composed of two letters. In that example, the main issue is to find the best specification. Both examples are inspired by botanical issues (as a matter of fact, the structure of a plant can be coded by a Dyck word, see Loi et al. (2010)).

4.1 Example of unary pattern

Let $V = \{s, m\}$ be an alphabet. Let $L = \langle W^+, \pi \rangle$ be a stochastic F0L-system such that the components of π are all equal to zero except for :

$$\pi_{s,ms} = p_1 \qquad \pi_{s,ss} = 1 - p_1 \qquad \pi_{m,mm} = p_2 \qquad \pi_{m,s} = 1 - p_2$$

with $(p_1, p_2) \in]0, 1[^2$. We are interested in computing the distribution associated to the number of patterns m in a word generated randomly by the component system $L[s]$ after N steps. To solve the problem, we need to compute the GF associated to m in $SW_N^L[s]$:

$$\Psi_N[s](z) = \sum_{w \in W_N^L[s]} (P^N)_{s,w} \; z^{c(w,m)} = \sum_{l \in \mathbb{N}} p_l^{N,s} \; z^l$$

where $p_l^{N,s}$ represents the probability to get l patterns m in a word generated randomly by the component system $L[s]$ after N steps. Let us find now an adequate iterative specification for $\{SW_n^L[s] \mid n \in \{0, \dots, N\}\}$. A first idea is to use Theorem 3.3 :

$$\forall n \in \{0, \dots, N-1\}, \quad SW_{n+1}^L[s] = \{(\epsilon, 1-p_1)\}.SW_n^L[ss] + \{(\epsilon, p_1)\}.SW_n^L[ms].$$

Then, by using the decomposition property (see Property 3.4), we get :

$$\forall n \in \{0, \dots, N-1\}, \quad SW_{n+1}^L[s] = \{(\epsilon, 1-p_1)\}.SW_n^L[s].SW_n^L[s] + \{(\epsilon, p_1)\}.SW_n^L[m].SW_n^L[s]. \quad (2)$$

In the same way, we get a combinatorial equation for $SW_{n+1}^L[m]$:

$$\forall n \in \{0, \dots, N-1\}, \quad SW_{n+1}^L[m] = \{(\epsilon, 1-p_2)\}.SW_n^L[s] + \{(\epsilon, p_2)\}.SW_n^L[m].SW_n^L[m]. \quad (3)$$

Equation (2) and (3) form an iterative specification. Note that $SW_0^L[s] = \{(s, 1)\}$ and $SW_0^L[m] = \{(m, 1)\}$. Therefore, by using the transformation rules (see Theorem 3.2), we get :

$$\forall n \in \{0, \dots, N-1\}, \quad \begin{cases} \Psi_{n+1}[s](z) = (1-p_1)\left(\Psi_n[s](z)\right)^2 + p_1 \Psi_n[m](z) \Psi_n[s](z) \\ \Psi_{n+1}[m](z) = (1-p_2)\Psi_n[s](z) + p_2 \left(\Psi_n[m](z)\right)^2 \end{cases} \quad (4)$$

with $\Psi_0[s](z) = 1$ and $\Psi_0[m](z) = z$. The coefficients of $\Psi_N[s](z)$ can be easily extracted from the system (4) by identifying the coefficients of the corresponding power series. By doing so, the coefficients of $\Psi_N[s](z)$ can be computed recursively.

N.B. : as detailed in Loi and Cournède (2008), the underlying stochastic process in this section is that of a Galton-Watson multitype branching process (see Mode (1971) and Athreya and Ney (2004)). As a matter of fact, let S_n and M_n be two random variables on some probability space $(\Omega, \mathcal{F}, \mathbf{P})$ (where $\mathbf{P}$ is a probability measure) such that S_n and M_n gives respectively the number of letters s and m in a word generated randomly by either the component system $L[s]$ or the component system $L[m]$ after n steps. Then, the sequence of random vectors $\Big((S_n, M_n)\Big)_{n\in\mathbb{N}}$ is a Galton-Watson multitype branching process. Let Φ_n be the probability generating function associated to (S_n, M_n) for all $n \in \mathbb{N}$. In that case, the system of equations (4) is simply the translation of the composition formula for branching processes : $\Phi_{n+1} = \Phi_1(\Phi_n)$ (Harris (1963)).

4.2 Complex example

This example corresponds to a concrete botanical issue : determining the probability distribution associated to the number of Y-structures in a random tree. Let $V = \{s, m, d\}$ be an alphabet. Let $L = \langle W^+, \pi\rangle$ be a stochastic F0L-system such that the components of π are all equal to zero except for :

$$\pi_{s,mssd} = p \qquad \pi_{s,\epsilon} = 1 - p \qquad \pi_{m,m} = 1 \qquad \pi_{d,d} = 1$$

with $p \in]0, 1[$. We are interested in computing the distribution associated to the number of patterns dm in a word generated randomly by the component system $L[s]$ after N steps. To solve the problem, we need to compute the GF associated to dm in $SW_N^L[s]$:

$$\Psi_N[s](z) = \sum_{w\in W_N^L[s]} (P^N)_{s,w}\ z^{c(w,dm)} = \sum_{l\in\mathbb{N}} p_l^{N,s}\ z^l$$

where $p_l^{N,s}$ represents the probability to get l patterns dm in a word generated randomly by the component system $L[s]$ after N steps. Let us find now a good iterative specification for $\{SW_n^L[s] \mid n \in \{0, \ldots, N\}\}$. A first approach consists in using Theorem 3.3 and Property 3.4. By doing so, we get :

$$\forall n \in \{0, \ldots, N-1\}, \quad SW_{n+1}^L[s] = \{(\epsilon, 1-p)\} + \{(\epsilon, p)\}.SW_n^L[m].SW_n^L[s].SW_n^L[s].SW_n^L[d].$$

Since $\pi_{m,m} = 1$ and $\pi_{d,d} = 1$, we have :

$$\forall n \in \mathbb{N}, \qquad SW_n^L[m] = \{(m, 1)\} \qquad SW_n^L[d] = \{(d, 1)\}. \tag{5}$$

Therefore :

$$\forall n \in \{0, \ldots, N-1\}, \quad SW_{n+1}^L[s] = \{(\epsilon, 1-p)\} + \{(\epsilon, p)\}.\{(m, 1)\}.SW_n^L[s].SW_n^L[s].\{(d, 1)\}. \tag{6}$$

However, the combinatorial equation (6) is not an admissible construction because the concatenation of $SW_n^L[s]$ and $SW_n^L[s]$ is generative with respect to the pattern dm (see Definition 3.4). As a matter of fact, $mssd$ is an element of $W_1^L[s]$. The concatenation $mssd.mssd = mssdmssd$ is an element of the

combinatorial class associated to $SW_1^L[s].SW_1^L[s]$ and we have $c(mssdmssd, dm) \neq c(mssd, dm) + c(mssd, dm)$. As a consequence, we need to find another decomposition. Since the pattern dm can only be created from the concatenation of two letters s, the idea is to find a decomposition involving $SW_n^L[ss]$. We get it from Equation (6) and Property 3.4 :

$$\forall n \in \{0, \ldots, N-1\}, \quad SW_{n+1}^L[s] = \{(\epsilon, 1-p)\} + \{(\epsilon, p)\}.\{(m, 1)\}.SW_n^L[ss].\{(d, 1)\}. \tag{7}$$

However, Equation (7) is not sufficient to have a good iterative specification. We need a recursive equation for $SW_n^L[ss]$. Once again, according to the description of the transition matrix π, we get :

$$\begin{aligned} SW_{n+1}^L[ss] =& \{(\epsilon, (1-p)^2)\} + \{(\epsilon, p(1-p))\}.SW_n^L[mssd].\{(\epsilon, 1)\} \\ &+ \{(\epsilon, p(1-p))\}.\{(\epsilon, 1)\}.SW_n^L[mssd] + \{(\epsilon, p^2)\}.SW_n^L[mssd].SW_n^L[mssd]. \end{aligned}$$

By using succesively the decomposition property (Property 3.4), Equation (5) and the concatenation operator, we get :

$$\begin{aligned} SW_{n+1}^L[ss] =& \{(\epsilon, (1-p)^2)\} + \{(m, 2p(1-p))\}.SW_n^L[ss].\{(d, 1)\} \\ &+ \{(m, p^2)\}.SW_n^L[ss].\{(dm, 1)\}.SW_n^L[ss].\{(d, 1)\}. \end{aligned} \tag{8}$$

Equations (7) and (8) form an adequate specification for $\{SW_n^L[s] \mid n \in \{0, \ldots, N\}\}$. As a matter of fact, we can prove by an immediate recursion that all words of $W_n^L[ss]$ begin either by s or m and end either by s or d. Therefore, in Equations (7) and (8), the concatenation of two consecutive sets of weighted words is always non-generative with respect to the pattern dm. Finally, by using the transformation rules (see Theorem 3.2), we get :

$$\forall n \in \{0, \ldots, N-1\}, \quad \begin{cases} \Psi_{n+1}[s](z) = 1 - p + p\Psi_n[ss](z) \\ \Psi_{n+1}[ss](z) = (1-p)^2 + 2p(1-p)\Psi_n[ss](z) + p^2 z \left(\Psi_n[ss](z)\right)^2 \end{cases} \tag{9}$$

with $\Psi_0[s](z) = 1$ and $\Psi_0[ss](z) = 1$. The coefficients of $\Psi_N[s](z)$ can be easily extracted from the system (9) by identifying the coefficients of the corresponding power series. By doing so, the coefficients of $\Psi_N[s](z)$ can be computed recursively.

5 Conclusion

In this article, we developed a symbolic method to compute the distribution associated to the number of pattern occurrences in random texts generated by stochastic 0L-systems. In order to get good specifications, we set an algebraic structure for combinatorial classes composed of weighted words. We defined new union and concatenation operators which enable to build admissible constructions. A set of transformation rules was given to turn specifications into a set of functional equations involving the generating functions associated to the problem. The crucial point of the method is to find a good specification. However, finding the adequate decomposition is not always systematic and, as a consequence, is the most difficult step of the method.

It is straightforward to extend the method to compute the probability distributions associated to the number of occurences for sets of patterns, by considering multivariate generating functions and extending Theorem 3.2 on admissible constructions. The multivariate approach is interesting in order to compute the covariances between the occurences of different patterns.

This symbolic method has applications in botany. As a matter of fact, the structure of a plant can be coded by a Dyck word (see Loi et al. (2010)). Therefore, we can use the method developed in this article to compute the distribution associated to the number of structures of a particular type in the plant. In that sense, the symbolic method appears as a generalization of the multitype branching process approach introduced by Kang et al. (2007). We could also imagine other applications in molecular biology with, for instance, mutations in the DNA sequence.

6 Appendix

The concepts developed in this section can all be found in detail in Flajolet and Sedgewick (2009). Let $\mathcal{A}$ be a combinatorial class and s a size function counting the number of characteristics of a given type, say c. Let A be the generating function associated to the size function s in $\mathcal{A}$:

$$A(z) = \sum_{t \in \mathcal{A}} z^{s(t)} = \sum_{n \in \mathbb{N}} A_n z^n$$

where A_n gives the number of elements of $\mathcal{A}$ having n times the characteristic c. The sequence $(A_n)_{n\in\mathbb{N}}$ is called the counting sequence of $\mathcal{A}$.

Definition 6.1 (Admissible Construction) *Let $\mathcal{B}^{(1)}, \ldots, \mathcal{B}^{(m)}$ be a collection of combinatorial classes with s as size function and $B^{(1)}, \ldots, B^{(m)}$ be the corresponding GFs. Let $(B_n^{(1)})_{n\in\mathbb{N}}$, ..., $(B_n^{(m)})_{n\in\mathbb{N}}$ be the associated counting sequences. Let Φ be an m-ary construction that associates to the collection of classes $\mathcal{B}^{(1)}, \ldots, \mathcal{B}^{(m)}$ a new class*

$$\mathcal{A} = \Phi[\mathcal{B}^{(1)}, \ldots, \mathcal{B}^{(m)}].$$

The construction Φ is admissible if the counting sequence $(A_n)_{n\in\mathbb{N}}$ of $\mathcal{A}$ only depends on the counting sequences $(B_n^{(1)})_{n\in\mathbb{N}}$,...,$(B_n^{(m)})_{n\in\mathbb{N}}$ of $\mathcal{B}^{(1)}, \ldots, \mathcal{B}^{(m)}$. For such an admissible construction, there exists a well-defined operator Ψ acting on the corresponding generating functions :

$$A(z) = \Psi[B^{(1)}(z), \ldots, B^{(m)}(z)].$$

Definition 6.2 (Specification) *A specification for an r-tuple $(\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(r)})$ of combinatorial classes is a collection of r equations,*

$$\begin{cases} \mathcal{A}^{(1)} & = \Phi_1(\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(r)}) \\ \mathcal{A}^{(2)} & = \Phi_2(\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(r)}) \\ & \cdots \\ \mathcal{A}^{(r)} & = \Phi_r(\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(r)}) \end{cases} \tag{10}$$

where Φ_1, Φ_2, ..., Φ_r are admissible constructions.

Formally, the system (10) is an iterative or non-recursive specification if it is strictly lower-triangular, that is $\mathcal{A}^{(1)}$ can be expressed by only using basic combinatorial classes and, for all $k \in \{1, \ldots, r-1\}$, the construction of $\mathcal{A}^{(k+1)}$ depends only on $\mathcal{A}^{(1)}$, ..., $\mathcal{A}^{(k)}$ and on basic combinatorial classes.

Références

K. Athreya and P. Ney. *Branching Processes*. Dover Publications, 2004.

C. Chrysaphinou and S. Papastavridis. The Occurence of Sequence of Patterns in Repeated Dependent Experiments. *Theory of Probability and Applications*, pages 167–173, 1990.

G. Duchamp, H. Kacem, and E. Laugerotte. Algebraic elimination of epsilon-transitions. *Discrete Mathematics and Theoretical Computer Science*, 7 :51–70, 2005.

P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.

J. Françon. Sur la Modélisation Informatique de l'Architecture et du Développement des Végétaux. In *2ème Colloque International : L'Arbre. Institut de Botanique, Montpellier, France*, 1990.

L. Guibas and A. Odlyzko. String Overlaps, Pattern Matching and Nontransitive Games. *Journal of Combinatorial Theory Serie A*, 30 :183–208, 1981.

T. Harris. *The theory of branching processes*. Springer, Berlin, 1963.

M. Kang, P.-H. Cournède, J.-P. Quadrat, and P. de Reffye. A stochastic language for plant topology. In T. Fourcaud and X. Zhang, editors, *Plant growth Modeling, simulation, visualization and their Applications.* IEEE Computer Society (Los Alamitos, California), 2007.

O. Klima and L. Polak. On varieties of meet automata. *Theoretical Computer Science*, 407 :278–289, 2008.

A. Lindenmayer. Mathematical models for cellular interactions in development. i. filaments with one-sided inputs. *Journal of Theoretical Biology*, 18 :280–289, 1968.

C. Loi and P.-H. Cournède. Generating Functions of Stochastic L-Systems and Application to Models of Plant Development. *Discrete Mathematics and Theoretical Computer Science Proceedings*, AI : 325–338, 2008.

C. Loi, P.-H. Cournède, and J. Francon. Plants as Combinatorial Structures and Applications. In *Plant growth Modeling, simulation, visualization and their Applications (PMA09).*, In Press. IEEE Computer Society (Los Alamitos, California), 2010.

C. Mode. *Multitype branching processes : Theory and applications*. American Elsevier Publishing Co. Inc, New York, 1971.

P. Pevzner, M. Borodovski, and A. Mironov. Linguistic of Nucleotide Sequences : the Significance of Deviations from Mean : Statistical Characteristics and Prediction of the Frequency of Occurence of Words. *J. Biomol. Struct. Dyn.*, 6 :1013–1026, 1991.

P. Prusinkiewicz and A. Lindenmayer. *The Algorithmic Beauty of Plants*. Springer-Verlag, New-York, 1990.

M. Régnier and W. Szpankowski. On Pattern Frequency Occurrences in a Markovian Sequence. *Algorithmica*, 22(4) :631–649, 1998. URL `http://algo.inria.fr/papers/other/ReSz97b.ps.gz`.

G. Rozenberg and A. Salomaa. *The Mathematical Theory of L-systems*. Academic Press, New York, 1980.