On the diversity of pattern distributions in rational language

Cyril Banderier, Olivier Bodini, Yann Ponty, Hanane Tafat

On the diversity of pattern distributions in rational language.

Cyril Banderier Olivier Bodini Yann Ponty Hanane Tafat Bouzid

Abstract

It is well known that, under some aperiodicity and irreducibility conditions, the number of occurrences of local patterns within a Markov chain (and, more generally, within the languages generated by weighted regular expressions/automata) follows a Gaussian distribution with both variance and mean in $\Theta(n)$. By contrast, when these conditions no longer hold, it has been observed that the limiting distribution may follow a whole diversity of distributions, including the uniform, power-law or even multimodal distribution, arising as tradeoffs between structural properties of the regular expression and the weight/probabilities associated with its transitions/letters. However these cases only partially cover the full diversity of behaviors induced within regular expressions, and a characterization of attainable distributions remained to be provided.

In this article, we constructively show that the limiting distribution of the simplest foreseeable motif (a single letter!) may already follow an arbitrarily complex continuous distribution (or *cadlag* process). We also give applications in random generation (Boltzmann sampling) and bioinformatics (parsimonious segmentation of DNA).

1 Introduction.

Numerous phenomena, in all fields of science, can be modelled as a graph giving transitions between finitely many states. A rigorous foundation of this idea is due to Markov in 1906, whose "Markov chains" were applied by Markov himself to study the frequency of groups of letters in Pushkin's *Eugene Onegin*. This idea led to many applications in information theory (compression and Shannon's entropy), artificial intelligence (Viterbi's algorithm, spam detection and creation), information retrieval (Google pagerank) chemistry (Michaelis-Menten kinetics), thermodynamics (Boltzmann equilibrium), language theory and compilation (regular grammars), computer science (automata theory), population ecology (Leslie matrix), biology (cell division, ion channels proteins), music (creation, or attribution), just to name a few.

For all these models, it makes sense to "mark" a specific transition, and study the distribution of the random variable X_n, giving the number of times that this marked transition was used, in a walk of length n. It turns out that, in an overwhelming majority of cases, X_n has linear mean and variance, and follows a Gaussian limit law. In fact, this property is automatically guaranteed whenever the underlying graph is aperiodic and "strongly connected" (or "irreducible"), and even holds when several transitions are marked [21]. This is yet another instance of what Philippe Flajolet was calling the Borges Theorem [17]: for non degenerated models, any pattern appears with non-trivial probability, and follows a Gaussian limit law (as proven for automata, grammar, trees, maps, ...).

From a mathematical point of view, several definitions lead to objects that are more or less isomorphic to Markov chains, such as regular expressions, rational language, linear grammars, recurrences with integer coefficients, and finite state automata. All of these objects can be seen as "word generating" processes, satisfying some fixed internal constraints which can be verified using bounded memory, irrespectively of the word length (e.g., as opposed to context-free grammars, whose recognition requires a stack structure of unbounded cardinality). The number of words of size n, let us call it a_n, and the associated generating functions $A(z) := \sum_{n\in\mathbb{N}} a_n z^n$ are the fundamental objects in our approach. This class of functions forms an important subset $\mathbb{N}^{\text{rat}}[[z]]$ (the so-called $\mathbb{N}$-rational functions) of the rational functions $\mathbb{Q}(z)$, an analytic characterisation of this class was given by Soittola's theorem, which can be seen as the reciprocal of Berstel's theorem [6]. Note that many problems related to this misleadingly simple world are expected to be computationally intractable (e.g. the Pisot problem, i.e. the presence of a zero in a linear recurrence, was proven to be NP-hard [7] and is even conjectured to be undecidable).

Faced with such difficulties, a fruitful approach resides in the constructive characterisation of a wide class of generating functions, whose limiting can be fully established. In a sense, the present work proposes a two-way multivariate gen-

eralisation of the Perron–Frobenius Theorem where the underlying Markov chain is not irreducible. This goal is also partly related to the Schur and Frobenius problems, nice unimodality/log-concavity questions, and knapsack-like problems or even polytope volume computations, where one is interested in counting the number of nonnegative integer solutions to equations like $\sum a_i x_i = n$, for some fixed integers a_i. Going back to the language theoretic perspective, in an important sequence of articles [1, 3, 18, 9], Goldwurm, Lonati, Choffrut & Bertoni have studied the distribution of occurrences of a given pattern in the language of a regular expression. They identified an important influence of the number of strongly connected components in the transition matrix, and of their relative intrication. In a sense, this article addresses the following reverse and complementary question: what kind of regular expression/automaton can lead to a given limit law?

Plan of the article

In Section 2, we constructively show that the distribution of a letter can be arbitrarily close to any continuous distribution (or *cadlag* process), as illustrated by Figure 1. We give several examples of expressions that give rise to non-Gaussian elementary distributions (uniform, polynomials, ...). Moreover, we show how any two regular expressions can be combined into a regular expression whose limit distribution is the (weighted and renormalised) sum of its individual components.

In Section 3, we show how a deeper insight into the distributions arising from regular expressions can be used to extend and/or delimit the scope of a multidimensional Boltzmann samplers introduced by two of the authors [11]. The Boltzmann approach is a strategy for the uniform random generation of combinatorial objects, developed by Flajolet et al. [15], which shares some aspects of statistical approaches in physics (e.g. for generating self-avoiding walks [5] and importance sampling [26]). We extend here the scope of an efficient multi-parameterised Boltzmann sampling by giving a precise characterisation of difficult distributions.

In Section 4, we give another application, more related to bioinformatics. Several people tried to modelled DNA as a Markov chain of low order (and it was shown in [24, 4] that order 7 was already enough for most of the DNA properties). Following this spirit, we tackle in this section the question of finding a kind of "minimal complexity regular expression" mimicking a given distribution. To this aim, we use the families of basic distributions introduced in Section 2 to address the construction of a regular expression that explains an observed distribution in a bioinformatics context. More precisely, one can take advantage of such a "modular decomposition" to devise algorithms that attempts to *guess* the specification/regular expression of a language, based on some indirect evidences (occurrences of patterns), a situation which routinely occurs in bioinformatics.

In Section 5, we conclude this presentation by outlining some perspectives.

2 Limit laws.

This section illustrates the wide variety of distributions followed by a parameter in a non strongly-connected regular grammar. We are going to define a family of rational languages whose respective distributions of the parameter will be used as building blocks to re-compose almost any distribution, as illustrated by the following picture.

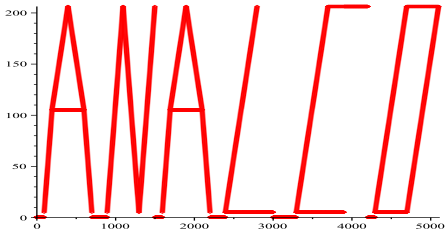


Figure 1: This figures gives the distribution of the letter "b" in words of a language $\mathcal{L} \in \{a, b\}^*$, generated by an ad-hoc regular expression of few lines (A Maple session is available here: http://lipn.univ-paris13.fr/~tafat/ANALCO/analco.mws). This distribution is converging towards a curve, "ANALCO". Note that this curve is, *at the limit*, a curve of a *multivalued* functional (as can be seen in the A, L, C, O letters), however we achieve it for *finite length words* via a *single valued* function, by interlacing two sequences mod 2. This figure illustrates the huge biodiversity of possible limit laws, even for the distribution of a single letter.

We propose two types of theorems. The first one proves that there is a class of rational languages for which some parameter may follow a large class of

distributions. The second one works on bivariate rational functions, and shows a very large domain of accessible distributions.

Theorem 1 *For any $P(n,k) = \sum_i \lambda_i \binom{k}{i}$ with $\lambda_i \in \mathbb{R}^+$, there is a rational language such that $\forall n \in \mathbb{N}$, the probability generating function of the distribution of the parameter u is exactly the probability generating function of n-th element X_n of the sequence of P-distribution $(X^{\langle P \rangle})_{n\in\mathbb{N}}$.*

The proof of these theorem is essentially based on the following lemma and its corollary:

In the first lemma, we explicitly give for every $\alpha > 0$ a rational language $\mathcal{L}_\alpha$ such that its dominant pole is of order α and for every $n > 0$, the distribution of parameter u is uniform on the class of words of size n in $\mathcal{L}_\alpha$.

Lemma 2 *For $\alpha > 1$, let $\mathcal{L}_\alpha$ be the language on the alphabet*

$$\{a_1, ..., a_{\alpha-1}, b_1, ..., b_{\alpha-1}, c\}$$

with $2\alpha - 1$ letters defined by the following regular expression:

$$\sum_{i+j=\alpha-2} \prod_{k=1}^{i+1} (c.a_k)^* \prod_{k=1}^{j+1} (b_k^2)^* .$$

The dominant pole of the generating function of $\mathcal{L}_\alpha$ is of order α and the distribution of the number of c in a word of size $2n$ is uniform.

Proof 1 *The generating function for the words of $\mathcal{L}_\alpha$ (In order to simplify the generating function, we consider here that the size of word of length $2n$ is n) is $C_\alpha(u,z) = \sum_{i+j=\alpha-1} \frac{1}{(1-uz)^{i+1}} \frac{1}{(1-z)^{j+1}}$. Now, let us recall that the distribution of the parameter u for a word of a given size n, is the distribution which the law follows the probability generating function $\frac{[z^n]C(z,u)}{[z^n]C(z,1)}$.*

A straightforward calculation shows that for $\alpha > 1$,

$$[z^n]C_\alpha(u,z) = \binom{n+\alpha-1}{\alpha-2} \sum_{k=0}^{n} u^k$$

and for $\alpha = 1$,

$$[z^n]C_1(u,z) = \sum_{k=0}^{n} u^k .$$

The lemma ensues immediately.

As a direct corollary of this lemma, we can build, for any $\beta \in \mathbb{N}$, languages with a distribution of the parameter u following a $\binom{n}{\beta}$-distribution sequence. More precisely:

Corollary 3 *For $\alpha > 1$ and $\beta > 1$, let $\mathcal{L}_{\alpha,\beta}$ be the language on an alphabet with $2\alpha - 1 + \beta$ letters defined by the following regular expression:*

$$\sum_{i+j=\alpha-2} \binom{j+\beta}{\beta} \prod_{k=1}^{(i+1+\beta)} (ca_k)^* \prod_{k=1}^{j+1} (b_k^2)^*,$$

the distribution of the number of c in a word of size $2n$ follows the law of X_n of the $\binom{k}{\beta}$-distribution sequence $(X^{\langle \binom{k}{\beta} \rangle})_{n\in\mathbb{N}}$. Moreover, the dominant pole of the generating function of $\mathcal{L}_\alpha$ is of order $\alpha + \beta$.

Proof 2 *The idea is to obtain a new rational language by pointing the parameter u. Now, as for Lemma 2, the lemma follows a straightforward, but fastidious calculation. The generating function for the words of $\mathcal{L}_{\alpha,\beta}$ is $C_{\alpha,\beta}(u,z) = \sum_{i+j=\alpha-2} \binom{i+\beta}{\beta} \frac{1}{(1-uz)^{i+1+\beta}} \frac{1}{(1-z)^{j+1}}$. Now, we can show that for $\alpha > 1$,*

$$[z^n]C_{\alpha,\beta}(u,z) = \binom{n+\beta+\alpha-1}{\alpha-2} \sum_{k=0}^{n} \binom{k+\beta}{\beta} \cdot u^k$$

The lemma ensues immediately.

Now, one can gives the proof of Theorem 1:

Proof 3 *The theorem is an easy consequence of Corollary 3, just by taking a non negative combination of the language's $\mathcal{L}_{2,\beta}$.*

Lemma 4 *The power series $\sum_{n\geq 0} n^i z^n$ belongs to a subset of $\mathbb{Q}^{rat}[[z]]$ and more precisely*

$$\sum_{n\geq 0} n^i z^n = \sum_{j=0}^{i} j! \left\{ {i \atop j} \right\} \frac{z^j}{(1-z)^{j+1}} = \frac{\sum_{j=0}^{i} \left[{i \atop j} \right] z^{j+1}}{(1-z)^{i+1}}.$$

Proof 4 *The proof is done by induction: first, apply zdz to the equality and then use the recurrence defining the Stirling numbers of the second kind $\left\{ {i \atop j} \right\} = \left\{ {i-1 \atop j} \right\} + j \left\{ {i-1 \atop j} \right\}$ and the Eulerian numbers $\left[{i \atop j} \right] = (j+1) \left[{i-1 \atop j} \right] + (i-j) \left[{i-1 \atop j-1} \right]$. Note that there is also a bijective proof (combinatorial explanation) of these relations.*

Theorem 5 *One can reach any polynomial distribution, i.e.* $\forall P(n,k) \in \mathbb{R}^+[[n,k]]$, $\exists F(z,u) \in \mathbb{Q}^{rat}[[z]]$ *such that*

$$\Pr(X_n^{\langle P\rangle} = k) = \frac{P(n,k)}{\sum_{i=0}^{n} P(n,i)} = \frac{[u^k] f_n(u)}{f_n(1)}.$$

Proof 5 *Let* $\mathcal{C}_1$ *the class of rational languages obtained by substituting* z *by* zu *in Lemma 4 and* $\mathcal{C}_2$ *the class of languages having as rational generating function* $\sum_{n\geq 0}\sum_{k\geq 0} k^\alpha n^\beta u^k z^n = \sum_{j=0}^{\alpha} j! \left\{ \alpha \atop j \right\} \frac{u^j}{(1-u)^{j+1}} \sum_{j=0}^{\beta} j! \left\{ \beta \atop j \right\} \frac{z^j}{(1-z)^{j+1}}$.

It is easy to see that the set of rational languages having $P(n,k)$ *as distribution limit is described by a linear combination of rational languages defined by* $\mathcal{C}_1$ *and* $\mathcal{C}_2$.

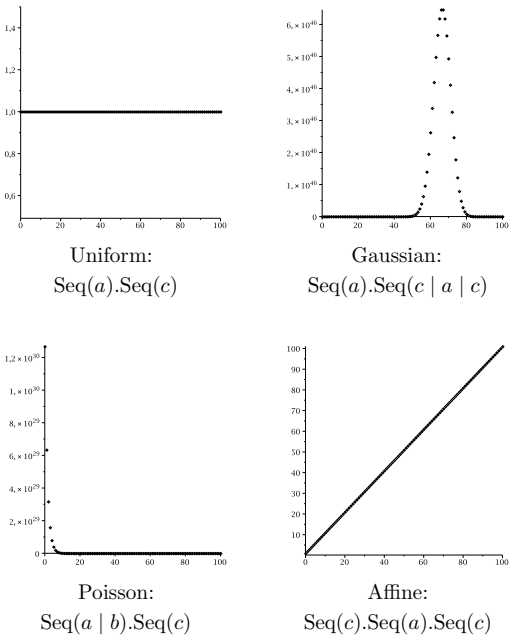
The second part of this section is organized around two propositions which allow us to extend the field of reachable distributions (the proof being simple, we omit them). The first proposition shows which generating function would give the "translation" of a given distribution.

Proposition 6 *Let* $A(z,u)$ *the generating function of the combinatorial structure* $\mathcal{A}$ *which admits a distribution of parameter* u *equal to* $D_{\mathcal{A}}$ *with support* $[0,n[$, *then the generating function* $B(z,u) = A(uz,u)$ *admits a distribution of parameter* u *such that for* $k \in [0,n[$, $D_{\mathcal{B}}(X = n+k) = D_{\mathcal{A}}(X = k)$. *So, the support of* $D_{\mathcal{B}}$ *is* $[n,2n[$.

In particular, this proposition and Theorem 1 allows us to find a rational language such that the distribution of one of its letters simulates the ANALCO spelling (see figure in the introduction).

Another very simple proposition consists to observe that if a distribution D is realizable by a rational language, then its mirror is also realizable:

Proposition 7 *Let* $A(z,u)$ *the generating function of the combinatorial structure* $\mathcal{A}$ *which admits a distribution of parameter* u *equals to* $D_{\mathcal{A}}$, *then the generating function* $B(z,u) = A(uz,u^{-1})$ *admits a distribution of parameter* u *such that for* $k \in [0,n[$, $D_{\mathcal{B}}(X = n-k) = D_{\mathcal{A}}(X = k)$.



Uniform: Seq(a).Seq(c)	Gaussian: Seq(a).Seq($c \mid a \mid c$)
Poisson: Seq($a \mid b$).Seq(c)	Affine: Seq(c).Seq(a).Seq(c)

Table 1: Basic regular expressions and their associated limit distributions for the number of occurrences of the distinguished letter c.

3 On the robustness of multi-parameterized Boltzmann sampling around a phase change.

Knowledge of the limit laws for parameters is crucial in multi-parameterised random generation. Indeed, it plays a central role in assessing the complexity of the multi-parameterised generators under Boltzmann model.

Let us remind briefly the context of random generation under Boltzmann model: In 2004, Duchon, Flajolet, Louchard, Schaeffer [15] introduced the Boltzmann samplers for the uniform random generation of decomposable objects. Contrasting with the so-called recursive method, the key idea here was to draw objects of any size within a Boltzmann-induced distribution of parameter x, and reject those of unsuitable sizes. A careful fine-tuning of the parameter x allowed for $\Theta(n^2)$ exact-size and $\Theta(n)$ approximate-size samplers for a large number of operators [15], later extended by subsequent efforts [16, 14, 12, 13]. In particular, two of the authors generalized this idea to multi-

dimensional objects [11] and proved the effectiveness of the Boltzmann generator for the regular and context-free languages when the limit law of the parameters is a multidimensional Gaussian law. Nevertheless, in many situations (as illustrated in Section 2) the limiting law of the parameter is not Gaussian. A detailed analysis of the complexity of the Boltzmann generators requires a general understanding of the effect of the limit laws on it.

Let $\mathcal{A}(\mathcal{Z},\mathcal{U})$ be a parameterised combinatorial class (usually $\mathcal{Z}$ represents each atom of the structure, e.g. the length of our words, and $\mathcal{U}$ marks some auxiliary parameter, e.g. the letter c in the previous section of this article) then a probabilistic algorithm is called a *free multi-parameterised Boltzmann sampler* if it depends on two tuning parameters that returns an object γ of $\mathcal{A}(\mathcal{Z},\mathcal{U})$ of size n and parameter k with probability $\mathbb{P}_{x,u}(\gamma) = \frac{x^n u^k}{A(x,u)}$, where $A(x,u)$ is the bivariate generating function of $\mathcal{A}(\mathcal{Z},\mathcal{U})$. Such a sampler is denoted by $\Gamma_{x,u}\mathcal{A}(\mathcal{Z},\mathcal{U})$. The constructions introduced in [15] extend easily to the multi-parameterised framework. However, the analysis of the complexity of approximate-size Boltzmann sampler, denoted by $\Gamma_{x,u,\varepsilon,\delta_u}\mathcal{A}(\mathcal{Z},\mathcal{U})$, where there is two rejection phases, one to select the size to be in a window of type $[(1-\varepsilon)n,(1+\varepsilon)n]$ and one to enforce the frequency of the parameter $\mathcal{U}$ to belong to a window of type $[(1-\delta_u)k,(1+\delta_u)k]$ is much more complicated.

In this section, we first prove in the framework of rational languages that the size-rejection phase of Boltzmann generators is extremely robust in terms of efficiency to the change of nature of the limits laws (see Theorem 8). On the other side, Theorem 9 shows that the frequency-rejection phase can be completely inefficient if the choice of the frequency tolerance is unrealizable. This is coherent with the wide variety of distribution that can take the parameter $\mathcal{U}$. More precisely, we prove the following results:

Theorem 8 *Let $\mathcal{A}(\mathcal{Z},\mathcal{U})$ be a parameterised regular language and $C(z,u) = \frac{P(z,u)}{Q(z,u)}$ the rational fraction corresponding to its bivariate generating function. Let $\mathcal{P}$ be a compact set of values of the parameter u. Let us denote by $\rho(u)$ the modulus of the smallest pole of $C(z,u)$ (which is a real number according to Pringsheim's theorem). For any fixed tolerance $\varepsilon > 0$, the approximate-size Boltzmann sampler $\Gamma_{\rho(u_0)(1-1/n),u_0,\varepsilon}\mathcal{A}(\mathcal{Z},\mathcal{U})$ generates, in average, in $O(n)$ independently in $u_0 \in \mathcal{P}$, an object of size N in the interval $[(1-\varepsilon),(1+\varepsilon)n]$.*

Theorem 9 *For any fixed frequency-tolerance $0 \leq \delta_u < 1$, there is a parameterised regular language $\mathcal{A}(\mathcal{Z},\mathcal{U})$ with bivariate generating function $C(z,u) = \frac{P(z,u)}{Q(z,u)}$ such that the approximate-frequency Boltzmann sampler $\Gamma_{\rho(u_0)(1-1/n),u_0,\varepsilon,\delta_u}\mathcal{A}(\mathcal{Z},\mathcal{U})$ is able to generate an object with a frequency of $\mathcal{U}$ in the interval $[(f_u-\delta_u),(f_u+\delta_u)]$.*

Before proving these theorems, let us consider the following two examples:

- We are interested in generating a word in $Seq(\mathcal{Z}) \times Seq(\mathcal{U}\mathcal{Z})$. In this case, the bivariate generating function is $\frac{1}{(1-z)(1-uz)}$. The question of what happens for the complexity of a Boltzmann sampler when the parameter u becomes close to 1 which is a phase change for the distribution of the parameter. Indeed, in this case, the nature of the limit distribution of the parameter $\mathcal{U}$ changes from a Gaussian law to a uniform law. The previous Theorem 8 entails that up to a constant factor independent of u, the Boltzmann complexity for the size-rejection stage is not changed. The following graphic shows for various choice of u, the mean number of rejections effectuated by the approximate-size Boltzmann sampler $\Gamma_{\rho(u_0)(1-1/n),u_0,\varepsilon}\mathcal{A}(\mathcal{Z},\mathcal{U})$. We can observe that, the worst situation is when $u_0 = 1$. But, in every case, the number of rejections is bounded by the common limit when n tends to the infinity. Our proof will follow these observations.

- Let us consider the rational language $\mathcal{L}$ of specification

  $$Seq(\mathcal{Z}^3)\times Seq(\mathcal{U}\mathcal{Z}^3)+Seq(\mathcal{U}^2\mathcal{Z}^3)\times Seq(\mathcal{U}^3\mathcal{Z}^3).$$

  The previous theorems show that there is no word of size n having a frequency of $\mathcal{U}$ in the range $]1/3,2/3[$. For instance, the distribution of the parameter $\mathcal{U}$ for a word of $\mathcal{L}$ of size 300 is given in the following graphic. So, this example illustrates what Theorem 9 explains: a generator of Boltzmann cannot generate an object which does not exist. Such a phenomena can be anticipated, for instance by looking

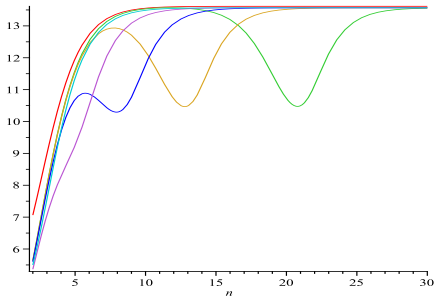


Figure 2: The number of rejection in Boltzmann sampling when the control parameter u_0 is close to the change of phases $u = 1$. The red curve corresponds to the parameter $u_0 = 1$, the green one to the farthest value of u_0. The x-axis is in logarithmic scale.

at the variance of the parameter $\mathcal{U}$. In a previous work, Bodini and Ponty showed that under concentration conditions of the distribution $\mathcal{U}$ that Boltzmann generators are totally efficient to generate objects (see Theorem 3 in [11]).

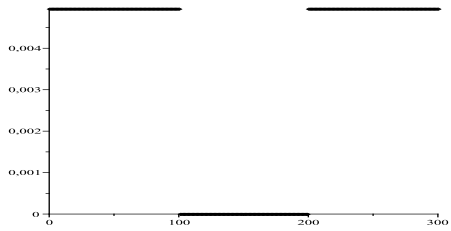


Figure 3: Distribution of the number of $\mathcal{U}$ in a word of size 300 in $Seq(\mathcal{Z}^3) \times Seq(\mathcal{U}\mathcal{Z}^3) + Seq(\mathcal{U}^2\mathcal{Z}^3) \times Seq(\mathcal{U}^3\mathcal{Z}^3)$.

Proof 6 (Sketch of the proof of Theorem 8) *The proof follows closely that of Theorem 6.3 in the seminal paper [15]. The proof is in two steps: First, we prove that asymptotically, the mean number of rejections during the size-rejection of the approximate-size sampler $\Gamma_{\rho(u_0)(1-1/n),u_0,\varepsilon}\mathcal{A}(\mathcal{Z},\mathcal{U})$ is bounded by a constant which does not depend on u. Secondly, we show using the compactness of the parameter that the mean number of rejections is bounded by a constant independent of u.*

So, for a fixed u_0, we have $C(z,u_0) \sim c(u_0)(1 - z/\rho(u_0))^{-\alpha(u_0)}$ with $\alpha(u_0)$ an integer smaller than $\deg_x(Q(x,u))$. Taking $x = \rho(u_0)(1-1/n)$ and u_0 as parameter, the probability to draw an object of size $\lfloor \beta n \rfloor$ is

$$\mathbb{P}(N = \lfloor \beta n \rfloor) \sim \frac{c(u_0)}{\Gamma(\alpha)} \rho(u_0)^{-\beta n} (\beta n)^{\alpha-1} \frac{(x^{\beta n})}{C(x,u_0)}.$$

By replacing x by $\rho(u_0)(1-1/n)$ and $C(z,u_0)$ by $c(u_0)(1-z/\rho(u_0))^{-\alpha(u_0)}$, we get

$$\mathbb{P}(N = \lfloor \beta n \rfloor) \sim \frac{e^{-\beta}\beta^{\alpha(u_0)-1}}{\Gamma(\alpha(u_0))} \frac{1}{n},$$

uniformly for β in a compact subinterval of $[0,\infty[$. Cumulating the estimates in the formula above, we find by Euler–MacLaurin summation:

$$\mathbb{P}((1-\varepsilon)n < N < (1+\varepsilon)n) \sim \frac{2\varepsilon}{\Gamma(\alpha(u_0))} \int_{1-\varepsilon}^{1+\varepsilon} e^{-x} x^{\alpha(u_0)-1} dx.$$

So, as $\alpha(u_0)$ is bounded by the degree in x of $Q(x,u)$. Asymptotically, we proved that $\mathbb{P}((1-\varepsilon)n < N < (1+\varepsilon)n)$ is greater than a strictly positive real number μ which does not depend on u but only the degree in x of $Q(x,u)$.

So, $\forall u_0 \in \mathcal{P}$, and $\forall \nu > 0$ there is a n_{ν,u_0} such that for $n > n_{\nu,u_0}$, $\mathbb{P}((1-\varepsilon)n < N < (1+\varepsilon)n)$ is ν-closed to μ, and by compactness of the domain $\mathcal{P}$, $\sup_{u_0 \in \mathcal{P}}(n_{\nu,u_0}) = N_\nu < \infty$. So, $\forall \nu > 0$ there is a N_ν such that $\forall u_0 \in \mathcal{P}$ and $\forall n > N_\nu$, $\mathbb{P}((1-\varepsilon)n < N < (1+\varepsilon)n)$ is ν-closed to μ. Now, again by compactness of $\mathcal{P}$, for $n \leq N_\nu$, $\mathbb{P}((1-\varepsilon)n < N < (1+\varepsilon)n)$ reaches its lower bounds m. Thus, $\mathbb{P}((1-\varepsilon)n < N < (1+\varepsilon)n)$ is greater than a strictly positive real number $\mu_1 = \max(\mu+\nu, m)$ which does not depend on u. Consequently, the theorem is proved.

Proof 7 (Sketch of the proof of Theorem 9) *We only have to build a specification with an arbitrary large gap in the distribution of the parameter $\mathcal{U}$. The following specification $Seq(\mathcal{Z}^k) \times Seq(\mathcal{U}\mathcal{Z}^k) + Seq(\mathcal{U}^{k-1}\mathcal{Z}^k) \times Seq(\mathcal{U}^k\mathcal{Z}^k)$ meets our expectations. The reachable frequencies for the parameter $\mathcal{U}$ are in $[0, 1/k[$ and $]\frac{k-1}{k}, 1]$.*

4 Parsimonious structural segmentation of DNA sequences.

In Section 2, we presented a family of regular expressions, giving rise to families of elementary

distributions. Here we propose a simple algorithm that combines these expressions in order to approximate an observed distribution in an additive model, motivated by bioinformatics applications.

4.1 Motivation: A computational genomics perspective

One of the goals of computational genomics is to reverse-engineer the mechanisms of life, or to relate constraints and motifs observed at the DNA level to functional mechanisms. A fruitful framework to study this relationship postulates a principle of parsimony, which dictates that, among the set of models whose induced behavior matches the observables, the simplest one should be preferred. This principle, usually referred to as the *Occam's razor* in a broader context, is one of the pillar of the scientific method.

Regular expressions and their probabilistic counterpart, the hidden Markov models, have long been used to model the modular architecture of genomes [24]. For instance, the PROSITE database [19] uses simple regular expressions, called patterns, to encode sequential signatures associated with functional domains of proteins. Such patterns can then be used for multiple tasks such as the search, within sequenced genomes, of new occurrences of functionally similar genes. One of the underlying computational challenge is that of grammatical induction/inference, namely the construction of a formal grammar which recovers all existing examples, while being general enough to describe novel instances [2]. In the context of bioinformatics, such efforts will ideally unravel some structural property weighing on a set of sequences, and will offer in any case testable hypotheses for a putative common mode of action.

Here we address a natural probabilistic variant of the grammatical inference problem, which we call the grammatical segmentation problem. Namely, given an observed discrete distribution of occurrences for a (local) motif, our goal is to construct a weighted rational expression whose (asymptotic) distribution has minimal distance to the observed distribution. Furthermore, the parsimony principle will be implemented as a limit on the complexity of the regular expression.

4.2 Statement of the problem

Let $p = [(x_i, y_i)]_{i=1}^n$ be a discrete distribution, i.e. a sequence of n points indexed by increasing x-ordinate. Let us define a **segmentation** of p as a pair $(\mathbf{s}, \mathbf{f})$, where $\mathbf{s} = (s_1, \ldots, s_k)$ is a partition of $[1, n]$ into k integral intervals, and $\mathbf{f} = (f_1, \ldots, f_k)$ is a k-tuple of functions. A segmentation can also be seen as a piecewise function, whose restriction to the x-ordinates of p **approximates** p. Let us use the usual **squared Euclidean distance**, defined as

$$\Delta(p, \mathbf{s}, \mathbf{f}) = \sum_{i=1}^{k} \sum_{j \in s_i} (y_i - f_i(x_j))^2, \tag{1}$$

as a quality measure for the approximation of p induced by $(\mathbf{s}, \mathbf{f})$.

Let us consider regular expressions $r \in \mathcal{R}$ having asymptotic distribution f_r, and denote by s_r the minimal interval of the indices of p that contains all of the non-null values in the distribution. Given a discrete distribution p, the **optimal grammatical segmentation** of p consists in computing the regular expression, made of k non-overlapping parts, that best approximates p. In other words, one needs to compute $\{r_1, \ldots, r_k\} \subset \mathcal{R}$ such that:

1. Cardinality: $|r| = k$
2. Non-overlap: $\forall r, r' \in \mathcal{R}, s_r \cap s_{r'} = \emptyset$
3. Optimality: $\forall \{r'_1, \ldots, r'_k\} \subset \mathcal{R}$ such that (1) and (2) hold, one has $\Delta(p, (s_{r_1}, \ldots, s_{r_k}), (f_{r_1}, \ldots, f_{r_k})) \leq \Delta(p, (s_{r'_1}, \ldots, s_{r'_k}), (f_{r'_1}, \ldots, f_{r'_k}))$.

4.3 A dynamic programming algorithm for the optimal grammatical segmentation problem

The minimal distance $\alpha_{k,p}$ between a subset of $\mathcal{R}$ of cardinality k and a sequence p follows

$$\alpha_{0,p} = \begin{cases} +\infty & (p \neq \varepsilon) \\ 0 & (p = \varepsilon) \end{cases} \tag{2}$$

and $\alpha_{k,p} = \min_{p'.p''=p \atop p' \neq \varepsilon} (\beta_{p'} + \alpha_{k-1,p''}), \forall k > 1$

$\beta_p := \min_{r \in \mathcal{R} \atop \text{s.t. } s_r = p} (\Delta(p, [1, |p|], f_r))$.

These equations can be computed using dynamic programming in $\mathcal{O}(k \cdot |p|^2)$ time and memory, since

the subsets involved in the subsequent calls are always intervals.

This algorithm can be easily extended to the case of parameterised regular expressions/distributions. By substituting a least-square fitting procedure to Equation 4.3, one can then optimize over infinite enumerable families of regular expressions/distributions.

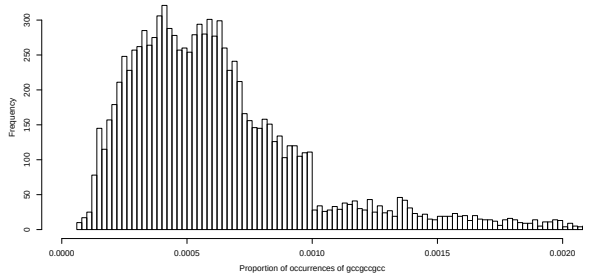


Figure 4: Distribution for the normalized number of occurrences of the GCCGCCGCC motif.

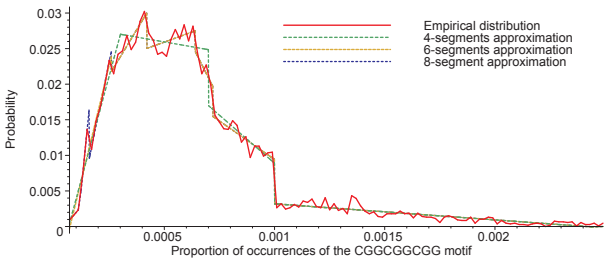


Figure 5: The best segmentation of the GCCGCCGCC motif using 4, 6, and 8 affine segments respectively.

4.4 Grammatical model for mini-satellites

To illustrate our algorithm, let use consider the distribution of occurrences of the CGGCGGCGG motif DNA sequences, as retrieved from the GenBank database [8]. The resulting distribution, shown in Figure 5 (Left) does not strictly follow a normal law, as an outcome of both the highly autocorrelating nature of the motif, and of the existence of duplicated essential motifs called mini-satellites [25]. Remarking that Section 2 implies the existence of a regular expressions for any affine positive targeted distribution, we can replace Equation 4.3 with a least-square fitting procedure, and obtain the best segmentations shown in Figure 5 (Right).

5 Conclusion and perspectives

In this article, we gave examples of distributions which can be reached as a pattern occurrence distribution. In the full version of this article, we plan to give a tighter characterization of the class of functions which can be reached (exactly, and not in a ε-approximation sense).

In a forthcoming work, we will tackle the question of limit laws in non strongly connected context free-grammars, and foresee an implementation (in SageMath) of the following algorithmic questions: Taking as input a language and a pattern (e.g. described by a regular expression), output the associated limit law and a Boltzmann (multivariate) sampler. This also incidentally forces us to investigate several intriguing phenomena related to multivariate Boltzmann sampling (coalescence of singularities [10], computational impact, etc).

Acknowledgements:
$$$$: All of the authors wish to acknowledge support from the ANR Magnum project funded by the French *Agence Nationale de la Recherche*.
♡♡♡♡: All the authors had the pleasure to interact, in different ways, with the late Philippe Flajolet. It is with great sadness that we will always remember his death last March, leaving orphan a whole community, but it is also with great pleasure that we will remember him, not only through his many wonderful writings or for the marvelous world he opened for us, but also for what he transmitted to us, this priceless community spirit and scientific impetus.

References

[1] Alberto Bertoni and Christian Choffrut and Massimiliano Goldwurm and Violetta Lonati, *Local Limit Distributions in Pattern Statistics: Beyond the Markovian Models*, STACS, (2004), pp. 117–128.

[2] François Coste and Goulven Kerbellec, *A Similar Fragments Merging Approach to*

Learn Automata on Proteins, ECML, (2005), pp. 522–529.

[3] Massimiliano Goldwurm and Violetta Lonati, *Pattern Occurrences in Multicomponent Models*, STACS, (2005), pp. 680–692.

[4] David Abergel, *Caractérisation bioinformatique des régions inter-ORF chez la levure*, Université Paris Sud, École doctorale Gènes, Génomes et Cellules, (2004).

[5] Neil Madras and Alan D. Sokal, *The Pivot algorithm - A highly efficient Monte-Carlo method for the self-avoiding walk*, J. Stat. Phys 50, no. 1–2, (1988), pp. 109–186.

[6] Jean Berstel, *Another proof of Soittola's theorem*, Theoretical Computer Science 393, no. 1–3, (2008), pp. 196–203.

[7] Vincent Blondel and Natacha Portier, *The presence of a zero in an integer linear recurrent sequence is NP-hard to decide*, Linear Algebra and its Applications 351–352, (2002), pp. 91–98,

[8] Dennis A. Benson and Ilene Karsch-Mizrachi and David J. Lipman and James Ostell and Eric W. Sayers, *GenBank*, Nucleic Acids Res Database issue 39, (2011), pp. D32–D37.

[9] Alberto Bertoni and Christian Choffrut and Massimiliano Goldwurm and Violetta Lonati, *Local Limit Properties for Pattern Statistics and Rational Models*, Theory Comput. Syst.39 no 1, (2006), pp. 209–235.

[10] Cyril Banderier and Philippe Flajolet and Gilles Schaeffer and Michèle Soria, *Random Maps, Coalescing Saddles, Singularity Analysis, and Airy Phenomena*, Random Structures and Algorithms 19 no. 3–4, (2001), pp. 194–246.

[11] Bodini Olivier and Ponty Yann, *Multi-dimensional Boltzmann Sampling of Languages*, DMTCS Proceedings 0 no. 01, (2010), pp. 49–64.

[12] Olivier Bodini and Alice Jacquot, *Boltzmann samplers for colored combinatorial objects*, Proceedings of Gascom'08, (2008).

[13] Olivier Bodini and Olivier Roussel and Michèle Soria, *Boltzmann samplers for first order combinatorial differential equations*, to appear in Discrete Applied Mathematics, (2011).

[14] Manuel Bodirsky and Éric Fusy and Mihyun Kang and Stefan Vigerske, *An unbiased pointing operator for unlabeled structures, with applications to counting and sampling*, SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, (2007), pp. 356–365.

[15] Philippe Duchon and Philippe Flajolet and Guy Louchard and Gilles Schaeffer, *Boltzmann Samplers for the Random Generation of Combinatorial Structures*, Combinatorics, Probability, and Computing 13, (2004), pp. 577–625.

[16] Philippe Flajolet and Éric Fusy and Carine Pivoteau, *Boltzmann Sampling of Unlabelled Structures*, SIAM Proceedings in Applied Mathematics 126, (2007), pp. 201–211.

[17] Flajolet Philippe and Sedgewick Robert *Analytic combinatorics*, (2009).

[18] Massimiliano Goldwurm and Violetta Lonati, *Pattern statistics and Vandermonde matrices*, Theor. Comput. Sci 356 no. 1–2, (2006), pp. 153–169.

[19] Nicolas Hulo and Amos Bairoch and Virginie Bulliard and Lorenzo Cerutti and Edouard De Castro and Petra S Langendijk-Genevaux and Marco Pagni and Christian J A Sigrist, *The PROSITE database*, Nucleic Acids Res 34 no. Database issue, (2006), pp. D227–D230.

[20] Hsien-Kuei Hwang, *On convergence rates in the central limit theorems for Combinatorial structures*, European Journal of Combinatorics 19 no. 1, (1998), pp. 329–343.

[21] Nicodème Pierre and Salvy Bruno and Flajolet Philippe, *Motif statistics*, Theor. Comput. Sci. 287 no. 2, (2002), pp. 593-617.

[22] Thomas Crombie Schelling , *Dynamic Models of Segregation*, Journal of Mathematical Sociology 1, (1971), pp. 143–186.

[23] Stefan Gerhold and Lev Glebsky and Carsten Schneider and Howard Weiss and Bukhard Zimmermann, *Computing the Complexity for Schelling Segregation Models*, Communications in Nonlinear Science and Numerical Simulation 13, (2008), pp. 2236–2245.

[24] Michel Termier and Angelos Kalogeropoulos, *Discrimination between fortuitous and biologically constrained open reading frames in DNA sequences of Saccharomyces cerevisiae*, Yeast 12 no. 4 , (1996), pp. 369–384.

[25] Kevin J. Verstrepen and An Jansen and Fran Lewitter and Gerald R. Fink, *Intragenic tandem repeats generate functional variability* Nat Genet 37 no. 9, (2005), pp. 986–990.

[26] G. Peter Lepage, *A new algorithm for adaptive multidimensional integration*, Journal of Computational Physics 27 no. 2, (1978), pp. 192 – 203.