Large deviation properties for patterns

Jérémie Bourdon, Mireille Regnier

Large deviation properties for patterns

Jérémie Bourdon[1,2] and Mireille Régnier[3], Jérémie Bourdon[1,2] and Mireille Régnier[3]

[1] *LINA, CNRS UMR 6241, Université de Nantes, France*
[2] *DYLISS-Inria team, INRIA Rennes-Bretagne-Atlantique, France*
[3] *AMIB-Inria team, LIX-Ecole Polytechnique, 91128 Palaiseau, France*
mireille.regnier@inria.fr,Jeremie.Bourdon@univ-nantes.fr

Abstract

Deciding whether a given pattern is overrepresented or under-represented according to a given background model is a key question in computational biology. Such a decision is usually made by computing some p-values reflecting the "exceptionality" of a pattern in a given sequence or set of sequences. In the simplest cases (short and simple patterns, simple background model, small number of sequences), an exact p-value can be computed with a tractable complexity. The realistic cases are in general too complicated to get such an exact p-value. Approximations are thus proposed (Gaussian, Poisson, Large deviation approximations). These approximations are applicable under some conditions: Gaussian approximations are valid in the central domain while Poisson and Large deviations approximations are valid for rare events. In the present paper, we prove a large deviation approximation to the double strands counting problem that refers to a counting of a given pattern in a set of sequences that arise from both strands of the genome. Here dependencies between a sequence and its complement plays a fundamental role. General combinatorial properties of the pattern allow to deal with such a dependency. A large deviation result is also provided for a set of small sequences.

Keywords: Pattern matching, statistics, Large deviation

1. Introduction

Counting random words [Szp01] and calculating probabilities is an old and extensively studied problem in theoretical computer science for various applications in bioinformatics including finding motifs [TLB$^+$05, NBM$^+$11] and calculating p-values [TV07]. Rare or overrepresented words are commonly assumed to be the hint in the genomes of some biological functions.

An exact derivation of the distribution of pattern occurrences is theoretically solved [Szp01] but its computation is expensive, not to mention accuracy and numerical issues. Classical approximations of the distribution (Gaussian, Poisson) are known not to be applicable [MRSKL04]. Indeed, distribution convergence are valid in the central domain, while rare events are studied in the *tail* domain. Simulations on biological data are presented in [MRSKL04]. This point is extensively discussed in [RV06].

Large deviation properties are the properties in the tail domain. One typical result in this case lies in mathematically proving a so-called *Large Deviation principle* that consists in providing a big-O bound on the tail distribution. In this paper, we prove that *combinatorial* properties of words allow to go further by providing an explicit and tractable formula for the tail distribution. A survey can be found in [Szp01] in the context of word occurrences. Still, very few results are known. One may also cite [RD04] and [Nue05]. The latter compares several methods for proving the exceptionality of a single pattern including Gaussian approximation, compound Poisson approximation and large deviation approximation.

Recently, new sequencing methods appear requiring to push away our knowledge on pattern occurrences statistics in order to obtain more precise results while keeping the complexity of the computations under control. Our goal is to obtain some probabilistic results adapted to datasets containing a huge number of sequences (typically reads from a high throughput sequencing) possibly coming from a pair end sequencing. In this case, patterns and their complementaries have to be considered at the same time (in the sequel, this case is referred as the "double strands" counting problem). The main problem here is to take properly into account the dependencies between a DNA sequence and its complement.

We consider here two cases. Large sequences are addressed in Section 3 when one counts occurrences of words from a finite set $\mathcal{H}$, under a Bernoulli model. Large deviation results were obtained for a single word in [RD04]. The case where $\mathcal{H}$ admits two overlap classes [RKFR09] is solved here. This case is fundamental, as it allows to address *double strand* counting. Short sequences are addressed in Section 4. Large deviation results have been known for long for random independent trials in the case of *identical* distributions [DZ98] and [Wat95]. Non-identical distributions are considered here.

It is shown that these formula are quite effective. Not only computation is easy with a low complexity, but tightness is ensured. We discuss a possible extension to a larger number of overlap classes.

2. Preliminaries

In this section, we present the basic definitions and framework necessary to understand our work.

. Given two words w and h, we note $w \preceq h$ (resp. $w \sqsubseteq h$) when w is a prefix (resp. a suffix) of h. When $w \preceq h$, the word $e(w,h)$ that satisfies $h = w \cdot e(w,h)$ is called the extension of w into h. Prefix and suffix relations are order relations. An equivalence relation on $\mathcal{H}$ has been defined in [RKFR09] that is *stable* for prefix and suffix relations.

Definition 1. *Given a set $\mathcal{H}$, the* overlap set *is the set of words that are prefix and suffix of two (possibly equal) words in $\mathcal{H}$. It is denoted $\mathcal{OV}(\mathcal{H})$.*

Two words f and g are said overlap equivalent *iff*

$$max_{\{w \in \mathcal{OV}(\mathcal{H})\}}\{w \preceq f\} = max_{\{w \in \mathcal{OV}(\mathcal{H})\}}\{w \preceq g\} \quad (1)$$

$$max_{\{w \in \mathcal{OV}(\mathcal{H})\}}\{w \sqsubseteq f\} = max_{\{w \in \mathcal{OV}(\mathcal{H})\}}\{w \sqsubseteq g\} \quad (2)$$

Definition 2. *A set $\mathcal{H}$ is called a k-set if the quotient set admits exactly k classes.*

Let G denote an overlap class. Given a prefix w of a member of G that belongs to $\mathcal{OV}(\mathcal{H})$, this word is a prefix of any word in G and one notes

$$e(w, \mathrm{G}) = \cup_{g \in \mathrm{G}} \{e(w, g)\} \ .$$

. A probability model on words steadily extends to overlap classes by a summation of members probabilities. This allows an extension to such patterns of classical generating functions for words [Szp01].

Definition 3. *Given an overlap class* F*, one denotes $H_{\mathrm{F}}(z)$ the* probability generating series

$$H_{\mathrm{F}}(z) = \sum_{f \in \mathrm{F}} Prob(f) z^{|f|} \ . \tag{3}$$

Given two overlap classes F *and* G*, the* probability matrix $\mathbb{H}(z)$ *is defined as*

$$\mathbb{H}(z) = \begin{pmatrix} H_{\mathrm{F}}(z) & H_{\mathrm{G}}(z) \\ H_{\mathrm{F}}(z) & H_{\mathrm{G}}(z) \end{pmatrix} \ . \tag{4}$$

Given two overlap classes F *and* G*, the* correlation polynom $A_{\mathrm{F,G}}(z)$ *is defined as*

$$A_{\mathrm{F,G}}(z) = \sum_{f \in \mathrm{F}, g \in \mathrm{G}} \sum_{w \sqsubseteq f, w \preceq g} Prob(e(w, g)) z^{|e(w,g)|} \ , \tag{5}$$

and the correlation matrix $\mathbb{A}(z)$ *is defined as*

$$\mathbb{A}(z) = \begin{pmatrix} A_{\mathrm{F,F}}(z) & A_{\mathrm{F,G}}(z) \\ A_{\mathrm{G,F}}(z) & A_{\mathrm{G,G}}(z) \end{pmatrix} \ . \tag{6}$$

Definition 4. *Given an integer n, one denotes X_n the random variable that counts the number of $\mathcal{H}$-occurrences in a random text of size n. Given an integer k, the* generating function *for k-occurrences is defined as*

$$L_k(z) = \sum_{n \geq 0} Prob(X_n = k) z^n \ . \tag{7}$$

The computation of $L_k(z)$ for a finite set of words was addressed in [Rég00, RD04], and derived formula depend on matrices of dimensions $|\mathcal{H}|$. Overlap classes allow for a reduction of these formula that makes use of 2-dimensional matrices.

Proposition 1. *The* generating function *for k-occurrences is*

$$L_k(z) = (H_{\mathrm{F}}(z), H_{\mathrm{G}}(z)) \cdot \mathbb{D}^{-1}(z) \cdot \mathbb{M}(z)^{k-1} \cdot \mathbb{D}^{-1}(z) \begin{pmatrix} 1 \\ 1 \end{pmatrix} \ , \tag{8}$$

where

$$\mathbb{D}(z) = (1 - z)\mathbb{A}(z) + \mathbb{H}(z) \tag{9}$$

$$\mathbb{M}(z) = \mathbb{I} + (z - 1)\mathbb{D}^{-1}(z) \ . \tag{10}$$

Our proofs rely on the *saddle point method* and provide an approximate expression for probabilities $Prob(X_n = k)$. Section 4 makes use of *Large powers* Theorem that is proved in [FS09] [chapter 8]. In Section 3, problem is reduced to a variant. This theorem rewrites, according to our notations

Theorem 1. Large powers: *Let $A(z)$ and $B(z)$ be two analytic functions with non-negative coefficients. Let ρ_A and ρ_B denote their radius of convergence. Assume that B is a-periodic and that $\rho_B < \rho_A$. For any real number a,* saddle point Equation $az_a \frac{B'(z_a)}{B(z_a)} = 1$ *has a unique positive root. Moreover*

$$[z^n]A(z)B(z)^{na} = e^{-nI(a)} \frac{A(z_a)}{\sqrt{2\pi n z_a}}(1 + o(1)) \tag{11}$$

where $I(a) = a \log B(z_a) - \log z_a$.

Detailed examples for Bernoulli, Poisson or normal distributions can be found in [Van04], where the classical Large Deviation results [DZ98] are derived by a generating function approach. The saddle point method was used in [RD04] for a single word and extended to deal with conditional probabilities. An extension to several words or overlap classes is addressed in the next section.

3. Large sequences

The main result of this section is Theorem 2 that addresses the case of 2-sets. It is presented in 3.1 and the proof is given in 3.2, 3.3 and Appendix.

Example 1. *A typical example is the so-called RY-element*

$$\mathcal{H} = \cup_{X \in \{AT,AG,TT,GT,TG,TA\}} \{X \cdot GCATGCA\} \ .$$

It is a 2-set, where $\mathcal{OV}(\mathcal{H})$ reduces to $\{A\}$. The two overlap equivalence classes are

$$\begin{aligned} \mathrm{F} &= \cup_{X \in \{TT,GT,TG,TA\}} \{X \cdot GCATGCA\} \ , \\ \mathrm{G} &= \cup_{X \in \{AT,AG\}} \{X \cdot GCATGCA\} \ . \end{aligned}$$

Example 2. *Two strands counting provides an other important example. A Transcription Factor Binding Site,* h*, is usually searched on two DNA strands, that are not independent. Nevertheless, this search is equivalent to the search of $\mathcal{H} = \{\mathrm{h}, \tilde{\mathrm{h}}\}$, where $\tilde{\mathrm{h}}$ is the complementary word of* h*.*

3.1. Large deviation property for a 2-set

Results will be expressed as a function of the following parameters

Notation:

$$K(z) = det(\mathbb{A}(z))(1 - z + Trace(\mathbb{H}\mathbb{A}^{-1}(z)) \tag{12}$$

$$\theta(z) = (1 - z)Trace(\mathbb{A}(z)) + Trace(\mathbb{H}(z)) \ . \tag{13}$$

Remark 1. *Generating function K occurs whenever a* fixed *finite number of occurrences k is counted. When the set $\mathcal{H}$ reduces to one word* h*, $K(z)$ reduces to $(1-z)A(z)+P(\mathrm{h})z^{|\mathrm{h}|}$ that can be found in numerous works [GO81, Szp01]. Generalization of function K for a set of words was introduced in [BCRV05] .*

Definition 5. *Given a real a, the* fundamental equation *is the equation:*

$$K(z)\psi^2(z)+\theta(z)\psi(z,a)\phi(z,a)+(1-z)\phi^2(z)=0 \tag{14}$$

where

$$\begin{aligned}\psi(z,a) &= az[K(z)+K'(z)(1-z)]+K(z)[2(1-z)-\theta(z)] &(15)\\ \phi(z,a) &= az[K'(z)\theta(z)-K(z)\theta'(z)]-K(z)[2K(z)-\theta(z)]\ . &(16)\end{aligned}$$

Lemma 1. *Fundamental equation admits a real positive root. The real positive root of smallest modulus is called the* fundamental root *and denoted z_a.*

Theorem 2. *Given a real number $a \neq Prob(\mathcal{H})$, let z_a be the fundamental root of the fundamental equation. The number of occurrences of words from a given 2-set $\mathcal{H}$ satisfies a large deviation property*

$$\lim_{n\to\infty} Prob(X_n \geq na) = I(a) \tag{17}$$

where

$$I(a) = -a\log(1-\frac{\psi(z_a,a)}{\phi(z_a,a)})+\log z_a \tag{18}$$

Function I is called the rate function*.*

Remark 2. *When $a=\mathrm{H_F}(1)+\mathrm{H_G}(1)=Prob(\mathcal{H})$, $z_a=1$ is the fundamental root and $\lambda(1)$ is 1. Therefore, $I(a)=0$. According to the central limit theorem [Szp01], this is the expected value in the central domain.*

3.2. Algebraic properties and eigenvalues derivation

Equation (8) allows for a reduction of our problem to an application of Large Powers Theorem. Intuition is as follows. $\mathbb{M}(z)$ is similar to matrix $\left(\begin{smallmatrix}\lambda(z) & 0\\ 0 & \mu(z)\end{smallmatrix}\right)$ and $\mathbb{M}^k(z)$ is similar to $\left(\begin{smallmatrix}\lambda(z)^k & 0\\ 0 & \mu(z)^k\end{smallmatrix}\right)$. Therefore, $L_k(z)$ rewrites as $A_1(z)\lambda(z)^k+A_2(z)\mu(z)^k$. When $|\mu(z)|<\alpha|\lambda(z)|$, with $\alpha<1$, the dominating part of the integrand reduces to $A_1(z)\lambda(z)^k$.

In a first step, $\mathbb{M}$ is rewritten as a function of K and θ in Proposition 2. The technical proof is deferred to the Appendix.

Proposition 2. *Matrix $\mathbb{M}(z)$ rewrites*

$$\mathbb{M}(z)=\mathbb{I}-\mathbb{A}^{-1}(z)(\mathbb{I}-\frac{\mathbb{H}\mathbb{A}^{-1}(z)}{1-z+Trace(\mathbb{H}\mathbb{A}^{-1}(z))}) \tag{19}$$

When its dimension is 2, it satisfies

$$Trace(\mathbb{M}(z)) = 2-\frac{\theta(z)}{K(z)} \tag{20}$$

$$det(\mathbb{M}(z)) = \frac{(1-z)+K(z)-\theta(z)}{K(z)}\ . \tag{21}$$

In a second step, we proceed to the search of the eigenvalues of matrix $\mathbb{M}(z)$. According to spectral Theorem, eigenvalues are the roots of the characteristic polynom

$$X^2 - Trace(\mathbb{M}(z))X + det(\mathbb{M}(z)) \ .$$

Let $T(z)$ denote $K(z)Trace(\mathbb{M}(z))$ and $D(z)$ denote $K(z)det(\mathbb{M}(z))$. When $K(z) \neq 0$, eigenvalues are the roots of the polynom P defined as

$$P(X) = K(z)X^2 - T(z)X + D(z) \ . \tag{22}$$

Let $\lambda(z)$ and $\mu(z)$ denote the two eigenvalues. Deriving $P(\lambda(z))$ with respect to z yields a functional equation to be satisfied by $\lambda(z)$ and $\lambda'(z)$,

$$K'(z)\lambda^2(z) + 2K(z)\lambda(z)\lambda'(z) - T(z)\lambda'(z) - T'(z)\lambda(z) + D'(z) = 0 \ . \tag{23}$$

The same equation is satisfied by $\mu(z)$ and $\mu'(z)$.

In a third step, we show that one of the two eigenvalues admits a saddle point that is given by the fundamental equation.

Saddle point equation rewrites $az\lambda'(z) - \lambda(z) = 0$. Plugging this equation into Equation 23 yields $Q(\lambda(z)) = 0$ where

$$Q(X) = (2K(z) + azK'(z))X^2 - (T(z) + azT'(z))X + azD'(z) \ . \tag{24}$$

Eigenvalue $\lambda(z)$ being a common root of P and Q, it is a root of $GCD(P, Q)$. Computing this GCD and substituting $T(z)$ and $D(z)$ expressions yields:

$$GCD(P, Q) = -X\phi(z, a) - \psi(z, a) + \phi(z, a) \ .$$

This GCD has a single root $1 - \frac{\psi(z,a)}{\phi(z,a)}$. It is a common root of (22) and (24) iff it is a root of either one. Substituting $1 - \frac{\psi(z,a)}{\phi(z,a)}$ to X in P leads to the *fundamental equation* (14).

3.3. Saddle point

Finally, we show that the saddle point provides the main contribution to the integrand. Property below is shown in Appendix 2. It generalizes a result established for a single word [Szp01]. Lemma 1 and 2 are shown in Appendix.

Proposition 3. *Equation $K(z) = 0$ admits a root ρ that is real positive. Moreover, $\rho > 1$ and ρ is the root of smallest modulus.*

Exists $h > 0$ such that any other root σ satisfies $|\sigma| > |\rho| + h$.

Lemma 2. *Fundamental root z_a satisfies*

$$0 < z_a < \rho \ .$$

Roots λ and μ satisfy

$$\mu(z_a) < \lambda(z_a) < 1 \ .$$

Proposition 3 and Lemma 2 allow for an application of Large Power Theorem. One computes $[z^n]A_2(z)\mu^k(z)$ by integration on the contour $z = z_a$. As $|\mu(z)| < |\mu(|z|)|$, this is upper bounded by $e^{-na\mu(z_a)}$ that is exponentially smaller than $e^{-na\lambda(z_a)} = e^{-nI(a)}$. This yields Equations 17 and 18.

Remark 3. *This scheme should extend to several words. The degree of the characteristic polynom, determinant*$(\mathbb{M}(z) - \lambda(z)\mathbb{I})$ *can be reduced to the number p of overlap classes [RKFR09]. The fundamental equation becomes Resultant*$(P, Q) = 0$*, that is a polynom of degree* $p - 1$*. Although no close formula can be given here, a numerical approximation should be computable and* $I(a)$ *follows whenever* $GCD(P, Q)$ *can be efficiently computed.*

4. Short sequences

Studying short sequences statistics is of great interest too. Here, "short" means that sequences do not contain enough symbols to enter in an asymptotic behavior (the sequence length is in the order of hundreds of symbols). In computational biology, this is the case when one searches for cis-acting elements in regulatory sequences [NBM+11] that may be known, for example from ChIP-chip or ChipSeq experiments, as being under a similar regulatory control. Formally, this can be viewed as the search of common similar motifs in short random sequences. The particularity of this case is that first, the probability that a sequence contains an occurrence can be precisely computed but second, the number of sequences is huge allowing asymptotic statistics to be derived.

Let $\mathcal{H}$ denote the set of admissible motifs. In the *sequence number model*, a sequence is considered as *positive* if it contains at least one $\mathcal{H}$-occurrence.

One assumes below that M random sequences are given. Their lengths may be different but are usually similar, due to experimental constraints. The probability to find one (or k) motifs from a set $\mathcal{H}$ in sequence i actually depends on (the symbol composition of) sequence i. Let us be more formal.

Definition 6. *Given a set of* M *random sequences, let* p_i *be the probability to find one* $\mathcal{H}$*-occurrence in a sequence with number* i*. One notes*

$$\mu(u) = \prod_{i=1}^{M} [p_i(u-1) + 1] \ . \tag{25}$$

Remark 4. *When all probabilities* p_i *are equal, say* $p_i = p$*, and* M *is large,* $\phi(t) = \mu(e^t) = (p(e^t - 1) + 1)^M$ *converges to the probability generating function of the Gaussian law, and large deviation results are known [DZ98] or, in the context of computational biology [Wat95, RV06]. It is worth noticing that Large Power Theorem steadily applies. Our goal is to focus on the case when the* p_i*'s are different, meaning, in a biological sense, that each sequence may comes from distinct region of the genome with different nucleotide compositions.*

Remark 5. *Equation* (25) *can easily be turned into a recurrence formula that permits to design a dynamical programming algorithm for computing* $\mu(u)$ *in an efficient way (with a complexity of order* $O(M^2)$*), In fact, our goal is to obtain the k-th coefficient*

of polynom $\mu(u)$ that equals the probability that exactly k sequences out of M sequences contains one $\mathcal{H}$-occurrence. Indeed, denoting $\mu_n(u) = \prod_{i=1}^{M}[p_i(u-1)+1]$, there exists a trivial recurrence relation relating the coefficients of $\mu_n(u)$, that is $[u^k]\mu_n(u) = (1-p_n)[u^k]\mu_{n-1}(u) + p_n[u^{k-1}]\mu_{n-1}(u)$ with $[u^0]\mu_1(u) = (1-p_1)$ and $[u^1]\mu_1(u) = p_1$. Its computation is then straightforward.

In the following section, we give a novel approximation formula for computing $[u^k]\mu_n(u)$ in a constant time.

4.1. Large deviations formulae

Our aim here is to provide an approximate computation that relies on large deviations [DZ98]. Our aim is the derivation of the probability to have at least k positive sequences out of M, where $a = \frac{k}{M}$ is (significantly) greater, or smaller, than the expectation.

Notation: One denotes, for each integer j,

$$\sigma_j = \sum_{i=1}^{i=M} p_i^j,\ \tau_j = \frac{\sigma_j}{M}\ . \tag{26}$$

Remark 6. *It follows from Equation (25) above, generating functions properties and general probability theory that the expected size of a positive cluster is σ_1 and its variance is $\sigma_1 - \sigma_2$.*

Theorem 3. *The distribution of the number of positive sequences in a set of sequences satisfies a large deviation property. Let K be the number of positive sequences and* a *some real number satisfying $Ma \neq \sigma_1$ and $0 < a < 1$.*

$$\lim_{M\to\infty} \frac{1}{M} \log(Prob(K \geq Ma) = -I(a) \tag{27}$$

where

$$I(a) = at_a + \sum_{j=1}^{\infty}(-1)^j \tau_j (e^{t_a} - 1)^j\ . \tag{28}$$

$I(a)$ can be approximated as

$$a\tilde{t_a} - \tau_1 \frac{a-\tau_1}{\tau_1 - \tau_2}\ . \tag{29}$$

Proof. Large Power Theorem extends for $\phi(t) = \mu(e^t)$. Let $h_a(t)$ be the function $\frac{1}{M}\log\phi(t) - at$. One searches for the roots of smallest modulus of Equation $h'_a(t) = 0$. They are described in Proposition 4,

Proposition 4. *The root of smallest modulus of Equation $h'_a(t) = 0$ is* real. *It is called the* fundamental *root. It can be approximated by $\tilde{t}_a$ where*

$$\tilde{t}_a = \log[1 + \frac{a-\tau_1}{\tau_1 - \tau_2} + \frac{(a-\tau_1)^2(\tau_2 - \tau_3)}{(\tau_1 - \tau_2)^3}]\ . \tag{30}$$

PROOF.

$$h_a(t) = \frac{1}{M}\log\phi(t) - at = \frac{1}{M}\sum_{i=1}^{M}\log[p_i(e^t-1)+1] \ . \tag{31}$$

and a, $0 < a < 1$ is real. Namely, $h'_a(t) = 0$ leads to

$$1 - a = \frac{1}{M}\sum_{i=1}^{M}\frac{1-p_i}{p_i(e^t-1)+1} \ . \tag{32}$$

A Taylor expansion of (32) yields

$$\begin{aligned} M(a-1) &= -\sum_i (1-p_i)[1+\sum_{j\geq 1}(-1)^j p_i^j (e^t-1)^j] \\ &= -M(1-\tau_1) + M\sum_j (-1)^{j+1}(\tau_j - \tau_{j+1})(e^t-1)^j \end{aligned}$$

Under the two conditions that $p_i(e^t-1)$ is small (p_i is small) and that $|(\sigma_j - \sigma_{j+1})(e^t-1)^j|$ are upper bounded by some number smaller than 1, - a condition to be checked in the computations -, we get to Equation below where X stands for $e^t - 1$.

$$0 = (\tau_1 - a) + (\tau_1 - \tau_2)X - (\tau_2 - \tau_3)X^2 \ .$$

One choses among the smallest among the two solutions. This yields (30). It satisfies $t_a = 0$ when $a = \tau_1$.

Expanding $\phi(t)$ as an analytic series steadily yields (28) for $h_a(t)$. Substituting the approximate value $\tilde{t}_a$ for t_a in (30) yields (29).

4.2. Experiments and computational precision

We provide here some evidences allowing to compare the results obtained by our large deviation approximation against an exact computation [1] of the same quantity. and by the large deviation approximation. Our goal is to assess the tightness of our approximation, in the domain where an exact computation remains feasible in order to justify large deviation results in the domains that are beyond this scope. Indeed, the required numerical precision in order to obtain some rigorous and exact results is proportional to the number of sequences, making the computation tricky for huge sets of sequences.

We used a set of 3000 sequences, The probability (p-value p_i) to find the motifs in each sequence is reported. It ranges between between 0.0009 and 0.0000001027. The mean σ_1 was 0.0002262 and the variance 0.0002263. The corresponding values for σ_1, σ_2 and σ_3 are

$$\sigma_1 = 0.679, \qquad \sigma_2 = 3.5.10^{-4}, \qquad \sigma_3 = 2.2.10^{-7}.$$

Table 1 synthesized the obtained results. It thus illustrates the tightness of the large deviation approximation.

[1] the exact computations are performed by a companion software available on request.

k	10	15	20	30	50	100	200	500	1000
LD	17.6	32.3	48.6	85	167.5	407.1	965.9	2962.4	6860.7
Exact	19.7	34.5	50.9	87.3	169.4	406.9	956.9	2907.8	6746.2
Relative error (%)	10.6	3.1	6.4	2.6	1.1	< 0.1	< 0.1	< 0.1	< 0.1

Table 1: Large deviations (LD) versus exact computations: log. p-value for different number k of occurrences. The value $MI(a)$ that is the logarithm of the p-value (multiplied by -1) is displayed. Relative error is also given showing an asymptotic convergence of the LD formula to the exact value.

For small values, a little difference arises. It is due to the fact that, using (only) $I(a)$ we neglected a second order term in the development of the p-value. A drift occurs for k occurrences when k is larger than 100. This is due to the approximation done on $\tilde{t}_a$ and $I(a)$. Notice that pushing a little bit further the computations, tighter approximations for $I(a)$ and $\tilde{t}_a$ can be obtained ("pumping method"). For instance, a tighter approximation of $I(a)$ is given as

$$I(a) \sim a\tilde{t_a} - \tau_1 \frac{a-\tau_1}{\tau_1-\tau_2} + \frac{(a-\tau_1)^2}{(\tau_1-\tau_2)^2}(\tau_1\tau_3 - \tau_2^2) \ . \tag{33}$$

5. Further work and conclusion

We have captured important properties and combinatorial constraints for multiple pattern statistics in this paper. Two different cases have been considered. First, asymptotic behaviors for the number of occurrences of patterns in a set containing a small number of very large sequences. We notably proved that several results holding in the case of single word counting extends to multiple patterns providing explicit relations that allow to get large deviation results and precise formulas for the tail distribution of the number of occurrences of multiple patterns. We applied the obtained properties to the double strands counting problem for huge sequences. This problem combines two major difficulties of the actual datasets at disposal. First, dealing with huge sets of sequences implies that exact and precise computations are unrealistic. Second, dependencies implied by a double strands counting cannot be neglected. Here, we solved the first issue by furnishing tractable formulas for computing the p-values in the large deviations domain. The second issue is solved by considering combinatorial properties of sets of patterns, including a correlation between sets of words property that is crucial in the study.

Second, we derive results for the asymptotic behavior of the number of pattern occurrences in a huge number of small sequences. This latter case is important in the context of biological sequence analysis since it corresponds to several realistic cases such as ChIP-chip or ChIP-seq experiments.

Notice that our framework should quite easily extend to multiple sets counting, the case of two sets having a particular interest in our opinion since it provides a solution to the double strands counting problem.

References

[BCRV05] V. Boeva, J. Clément, M. Régnier, and M. Vandenbogaert. Assessing the significance of sets of words. In *Combinatorial Pattern Matching 05*, volume 3537 of *Lecture Notes in Computer Science*, page 358–370. Springer Verlag, 2005. In Proceedings CPM'05, Jeju Island, Korea.

[DZ98] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*. Springer, New York, 2nd edition, 1998.

[FS09] Philippe Flajolet and Robert Sedgewick. *Analysis of Algorithms*. Cambridge University Press, 2009.

[GO81] Leonidas J. Guibas and Andrew M. Odlyzko. String overlaps, pattern matching, and nontransitive games. *J. Comb. Theory, Ser. A*, 30(2):183–208, 1981.

[MRSKL04] L. Marino-Ramirez, J.L. Spouge, G.C. Kanga, and D. Landsman. Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acid Research*, 32(3):949–958, 2004.

[NBM+11] N. Negre, C. D. Brown, L. Ma, C. A. Bristow, S. W. Miller, U. Wagner, P. Kheradpour, M. L. Eaton, P. Loriaux, R. Sealfon, Z. Li, H. Ishii, R. F. Spokony, J. Chen, L. Hwang, C. Cheng, R. P. Auburn, M. B. Davis, M. Domanus, P. K. Shah, C. A. Morrison, J. Zieba, S. Suchy, L. Senderowicz, A. Victorsen, N. A. Bild, A. J. Grundstad, D. Hanley, D. M. MacAlpine, M. Mannervik, K. Venken, H. Bellen, R. White, M. Gerstein, S. Russell, R. L. Grossman, B. Ren, J. W. Posakony, M. Kellis, and K. P. White. A cis-regulatory map of the Drosophila genome. *Nature*, 471(7339):527–531, Mar 2011.

[Nue05] G. Nuel. Ld-spatt: Large deviations statistics for patterns on markov chains. *Journal of Computational Biology*, 11(6):1023–1033, 2005.

[RD04] Mireille Régnier and Alain Denise. Rare events and conditional events on random strings. *Discrete Mathematics and Theoretical Computer Science*, 6(2):191–214, 2004.

[Rég00] M. Régnier. A unified approach to word occurrences probabilities. *Discrete Applied Mathematics*, 104(1):259–280, 2000. Special issue on Computational Biology.

[RKFR09] Mireille Regnier, Zara Kirakossian, Eugenia Furletova, and Mikhail Roytberg. A Word Counting Graph. In Jacqueline W. Daykin Joseph Chan and M. Sohel Rahman, editors, *London Algorithmics 2008: Theory and Practice (Texts in Algorithmics)*, page 31 p. London College Publications, 2009.

[RV06] M. Régnier and M. Vandenbogaert. Comparison of statistical significance criteria. *Journal of Bioinformatics and Computational Biology*, 4(2):537–551, 2006.

[Szp01] W. Szpankowski. *Average Case Analysis of Algorithms on Sequences*. John Wiley and Sons, New York, 2001.

[TLB+05] M. Tompa, N. Li, T.L. Bailey, G.M. Church, B. De Moor, E. Eskin, A.V. Favorov, M.C. Frith, Y. Fu, J.W. Kent, V.J. Makeev, A.A. Mironov, W.S. Noble, G. Pavesi, G. Pesole, M. Régnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. An assessment of computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23(1):137 – 144, January 2005.

[TV07] H. Touzet and J.-S. Varré. Efficient and accurate p-value computation for position weight matrices. *Algorithms for Molecular Biology*, 15(2), 2007. 12 pages.

[Van04] M. Vandenbogaert. *Algorithmes et mesures statistiques pour la recherche de signaux fonctionnels dans les zones de régulation*. Thèse de doctorat, Université de Bordeaux, 2004.

[Wat95] M. Waterman. *Introduction to Computational Biology*. Chapman and Hall, London, 1995.

6. Appendix 1

In this appendix, we prove that Proposition 1 can be established from equations (10) and (9) by using simple algebraic manipulations.

Matrix $\mathbb{A}(z)$ is regular for z around 0 [RD04]. Equation (9) leads to

$$\mathbb{D}^{-1}(z) = \frac{\mathbb{A}^{-1}(z)}{1-z}[\mathbb{I} + \frac{\mathbb{H}\mathbb{A}^{-1}(z)}{1-z}]^{-1}$$

The rank of matrix $\mathbb{H}$ being 1, the rank of matrix $\mathbb{B} = \mathbb{H}\mathbb{A}^{-1}(z)$ is 1, too. For any integer k, equation $\mathbb{B}^k = Trace(\mathbb{B})^{k-1}\mathbb{B}$ holds and

$$(\mathbb{I} + \frac{\mathbb{B}}{1-z})^{-1} = \mathbb{I} + \sum_{k\geq 1} \frac{(-1)^k Trace(\mathbb{B})^{k-1}}{(1-z)^k}\mathbb{B} = \mathbb{I} - \frac{1}{1-z+Trace(\mathbb{B})}\mathbb{B} \ .$$

Equation (10) steadily rewrites into (19).

This rewriting allows for a computation of the trace and the determinant of matrix $\mathbb{M}(z)$. Trace computation makes use of technical property (34). Given two 2×2 matrices, $\mathbb{M}$ and $\mathbb{N}$, the rank of $\mathbb{N}$ being 1, it is easy to establish that

$$Trace(\mathbb{M}\mathbb{N}\mathbb{M}) = Trace(\mathbb{M})Trace(\mathbb{N}\mathbb{M}) - Trace(\mathbb{N})determinant(\mathbb{M}) \ . \qquad (34)$$

Using (34) for $\mathbb{A}^{-1}(z)\mathbb{H}\mathbb{A}^{-1}(z)$ yields

$$Trace(\mathbb{A}^{-1}(z))Trace(\mathbb{B}) - Trace(\mathbb{H})det(\mathbb{A}^{-1}(z)) \ .$$

As $Trace(\mathbb{A}^{-1}(z)) = Trace(\mathbb{A}(z)) \cdot det(\mathbb{A}^{-1}(z))$, this yields

$$\begin{aligned} Trace(\mathbb{M} - \mathbb{I}) = &-\frac{Trace(\mathbb{A}(z))}{det(\mathbb{A}(z))} \\ &+ \frac{1}{det(\mathbb{A}(z))}[Trace(\mathbb{A}(z))Trace(\mathbb{B}) - Trace(\mathbb{H})]\frac{1}{1-z+Trace(\mathbb{B})} \end{aligned}$$

that simplifies into $-\frac{\theta(z)}{K(z)}$. Therefore, $Trace(\mathbb{M}(z)) = 2 - \frac{\theta(z)}{K(z)}$. We now observe that $det(\mathbb{M}) = det(\mathbb{M} - \mathbb{I}) - 1 + Trace(\mathbb{M})$. Using it for the singular matrix $\frac{\mathbb{B}}{1-z+Trace(\mathbb{B})}$ leads to

$$det(\mathbb{M}(z) - \mathbb{I}) = -\frac{1}{det(\mathbb{A})}det(\mathbb{I} - \frac{\mathbb{B}}{1-z+Trace(\mathbb{B})}) = \frac{1-z}{K(z)} \ ,$$

and (21) follows from (20).

7. Appendix 2

Here is a sketch of the proof of saddle point approach. The aim is to show that the contribution is given by the root λ that traverses saddle point $(z_a, \lambda(z_a))$.

We proceed to the proof of Lemma 1 and 2

PROOF. Functions $log(\lambda(z))$ and $log(z)$ are defined and analytical in $]0.\rho]$. Therefore. the same property holds for function $h_a(z)$. As

$$\lim_{z\to 0} h_a(z) = \lim_{z\to \rho} h_a(z) = +\infty \ ,$$

this function admits in this interval a root z_a. It follows from P and Q definitions that $\lambda(z_a)$ isa root of fundamental equation. This establishes Lemma 1.

PROOF. $\mathbb{M}(z)$ and $\mathbb{I} - \mathbb{M}(z) = \mathbb{A}^{-1}(z)(\mathbb{I} - \mathbb{B})$ represent the generating function of languages. Therefore, all coefficients are positive. Consequently, for z real and positive in the disk of convergence, eigenvalues $\lambda(z)$ and $\mu(z)$ are real positive and their modulus is smaller than 1.

One now studies the difference $\lambda(z) - \mu(z)$.

$$\lambda(z) - \mu(z) = 2\lambda(z) - \theta(z) = \frac{2\psi(z)K(z) + \theta(z)\phi(z)}{\phi(z)K(z)} \ . \tag{35}$$

Equation 14 at point z_a yields

$$\lambda(z_a) - \mu(z_a) = \frac{\psi^2(z_a)K(z_a) - (1 - z_a)\phi^2(z_a)}{\phi(z_a)\psi(z_a)K(z_a)} \ . \tag{36}$$

Clearly $\psi^2(z_a), \phi^2(z_a)$ and $(z_a - 1)$ are greater than 0. As $\lambda(z_a) < 1$, the same property holds for $\phi(z_a)\psi(z_a)$. As $z_a < \rho$, the same property holds for $K(z_a)$ and inequality $\mu(z_a) < \lambda(z_a)$ is established.