Assessing statistical significance of overrepresented oligonucleotides

Alain Denise[1], Mireille Régnier[2], and Mathias Vandenbogaert[2,3]

[1] LRI, UMR CNRS 8623 and IGM,UMR CNRS 8621, Université Paris-Sud, 91405 Orsay cedex, France
[2] INRIA-BP 105. 78 153 Le Chesnay, France
Mireille.Regnier@inria.fr
[3] LaBRI, Université Bordeaux I, 351, Cours de la Liberation, 33405 Talence, France †

Abstract. Assessing statistical significance of overrepresentation of exceptional words is becoming an important task in computational biology. We show on two problems how large deviation methodology applies. First, when some oligomer H occurs more often than expected, e.g. may be overrepresented, large deviations allow for a very efficient computation of the so-called p-value. The second problem we address is the possible changes in the oligomers distribution induced by the overrepresentation of some pattern. Discarding this noise allows for the detection of weaker signals. Related algorithmic and complexity issues are discussed and compared to previous results. The approach is illustrated with three typical examples of applications on biological data.

Keywords: Promoter sequences, regulatory sites, coregulation, motif statistics, large deviations, p-values.

1 Introduction

Putative DNA recognition sites can be defined in terms of an idealized sequence that represents the bases most often present at each position. Conservation of only very short consensus sequences is a typical feature of regulatory sites (such as promoters) in both prokaryotic and eukaryotic genomes. Structural genes are often organized into clusters that include genes coding for proteins whose functions are related. Data from the *Arabidopsis* genome project suggest that more than 5% of the genes of this plant encode transcription factors. The necessity for the use of genomic analytical approaches becomes clear when it is considered that less than 10% of these factors have been genetically characterized. Transcription-factor genes comprise a substantial fraction of all eukaryotic genomes, and the majority can be grouped into a handful of different, often large, gene families

† This research was partially supported by IST Program of the EU under contract number 99-14186 (ALCOM-FT), REMAG Action from INRIA and IMPG French program.

according to the type of DNA-binding domain that they encode. Functional redundancy is not unusual within these families; therefore the proper characterization of particular transcription-factor genes often requires their study in the context of a whole family. The scope of genomic studies in this area is to find cis-acting regulatory elements from a set of co-regulated DNA sequences (e.g. promoters). The basic assumption is that a cluster of co-regulated genes is regulated by the same transcription factors and the genes of a given cluster share common regulatory motifs.

On the other hand, the importance of whole-genome studies is highlighted by the fact that approximately 50% of the *Arabidopsis* genes have no proposed function, and that the *Arabidopsis* genome, like those of *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Drosophila*, contains extensive duplications. These include many tandem gene duplications as well as large-scale duplications on different chromosomes. The prevalence of gene duplication in Arabidopsis implies that redundancy is a problem that will have to be dealt with in the functional analysis of genes. In order to find long approximate repeats, some approaches consist in searching for short exact motifs that appear in the sequences.

In both problems, there is a need for identification of over-represented motifs in the considered sequences. Research is very active in this area [4, 23, 15, 21, 24, 25, 22, 14, 7, 1, 9]. In these works, one searches for exceptional patterns in nucleotidic sequences, using various tools to assess the significance of such rare events.

Large deviation is a mathematical area that deals with rare events; to our knowledge, it has not been used in computational biology, although the extremal statistics on alignments [6] can be viewed as large deviation results. Nevertheless, our recent results in [3], that extend preliminary results in [17] show it may be a very powerful method to assess statistical significance of very rare events.

The first problem we address is the following. One considers a *candidate*, e.g. a word that occurs more often than expected. One needs to quantify this difference between the observation and the expectation. Among the classical statistical tools, the so-called p-values are much more precise than the Z-scores (or the χ-scores). The drawback is that their computation is considered as much harder. Large deviations provide a very efficient way to compute them in some cases.

As a second problem, we consider some consequences of the overrepresentation of a word on a sequence distribution. In particular, it has been observed that, whenever a word is overrepresented, its subwords or the words that contain it, look overrepresented. Such words are called below *artefacts* [1]. It is a desirable goal to choose the best element in the set composed of a word and its artefacts. It is also important to discard automatically the "noise" created by the artefacts, in order to detect other words that are potentially overrepresented. An important example is the noise introduced by the *Alu* sequences. Another one is the χ-sequence `GNTGGTGG` in *H. influenzae* [11]. We provide some mathematical results and the algorithmic consequences.

Efficiency of this approach comes from the existence of explicit formulae for the (conditioned) distribution. Large deviations allow for a very fast computation. Moreover, due to the "simplicity" of the result -if not of the proof-, their implementation is easy and provides numerically stable and guaranteed computations. The reason is that the problem reduces to the numerical computation of the root of a polynomial. Other approaches need the numerical computation(s) of $\exp, \log$ functions. The implementation is delicate and machine dependent. Hence, large deviation approach occasionnally correct commonly used approximations. Interestingly, they apply with the same cost for self-overlapping patterns. This is a great improvement on the approaches based on binomial or related formulae [26, 23] where autocorrelation effects are neglected. Still, our computation that takes into account the correlations is much faster and precise than computing the approximation. Approach is valid for various counting models. For a sake of clarity, we present it for the most commonly used, the overlapping model [27].

In Section 2, we present and discuss the statistical criteria commonly used for evaluating over-representation of patterns in sequences. Section 3 is devoted to our mathematical results. In Sections 4, 6 and 5, we validate our approach by a comparison with published results derived by other methods that are computationally more expensive. Finally, in Section 7, we discuss possible improvements and present further work.

2 Statistical tools in computational biology

In the present section, we present basic useful definitions for statistical criteria and we briefly discuss their limits, e.g the validity domains and the computational efficiency. Below, we denote by $O(\mathrm{H})$ the number of observations of a given pattern H in a given sequence. Depending of the application, it may be either the number of occurrences [14, 7] or the number of sequences where it appears [1, 23].

Z-scores Many definitions of this parameter can be found in the literature. Other names can be used: see for instance the so-called *contrast* used in [14]. A common feature is the comparison of the observation with the expectation, using the variance as a normalization factor. A rather general definition is

$$Z(\mathrm{H}) = \frac{E(H) - O(\mathrm{H})}{\sqrt{V(H)}} \tag{1}$$

where H is a given pattern or word, $O(\mathrm{H})$ is the *observed* number of occurrences, $E(H)$ the expectation and $V(H)$ the variance. Many recent works allow for a fast computation of E and V, hence Z. Relevant approximations are discussed in [16], notably the Poisson approximation $V = E$. Nevertheless, if Z-scores are a very efficient filter to detect potential candidates, they are not precise enough. Notably, this parameter is not stable enough for very exceptional words, e.g. when the expectation is much smaller than 1. This will be detailed in Section 4.

Moreover, it is relevant only for large sequences, and does not adapt easily to the search in several small sequences.

p-values For each word that occurs r times in a sequence or in a set of N (related) sequences, one computes the probability that this event occurs just "by chance":

$$pval(\mathrm{H}) = P(O(\mathrm{H}) \geq r) \ . \tag{2}$$

When the expectation of a given word is much smaller than 1, a single occurrence is a rare event. In this case, the p-value is defined as: $P(O(\mathrm{H}) \geq r$ knowing that $O(\mathrm{H}) \geq 1)$, e.g.:

$$pval(\mathrm{H}) = \frac{P(O(\mathrm{H}) \geq r)}{P(O(\mathrm{H}) \geq 1)} \ . \tag{3}$$

The computation is performed in two steps. First, the probability that H occurs in a given sequence, e.g. $P(O(\mathrm{H}) \geq 1)$, is known. An exact formula is provided in [17] and used in [7]. An approximated formula is often used, for instance in software RSA-tools (http://www.ucmb.ulb.ac.be/bioinformatics/rsa-tools/) or in [1]. Then two different cases occur.

Set of small sequences Then,the p-value in (2) is the probability that r sequences out of N contain H; when they are independent, it is given by a binomial formula:

$$pval(\mathrm{H}) = \binom{N}{r}(P(O(\mathrm{H}) \geq 1))^r(1 - P(O(\mathrm{H}) \geq 1))^{N-r} \tag{4}$$

Large sequences One needs to compute $P(O(\mathrm{H}) \geq r$, through the exact formulae in [17] or an approximation.

This p-value is evaluated in [23] through a *significance coefficient*. Given a motif H, the significance coefficient is defined as

$$\mathrm{Sig} = -log_{10}[P(O(\mathrm{H}) \geq r) * D] \ ,$$

which takes into account the number of different oligonucleotides D. The number of distinct oligonucleotides depends whether one counts on a single strand or on both strands.

3 Main results

3.1 Basic notations

The model of random text that we handle with is the *Bernoulli model*: one assumes the text to be randomly generated by a memoryless source. Each letter s of the alphabet has a given probability p_s to be generated at any step. Generally, the p_s are not equal.

Definition 1. *Given a pattern* H *of length* m *on the alphabet* $\mathcal{S}$ *and a Bernoulli distribution on the letters of* $\mathcal{S}$*, the probability of* H *is defined as*

$$P(\mathrm{H}) = \prod_{i=1}^{m} p_{h_i}$$

where h_i *denotes the* i*-th character of* H*. By convention, empty string* ϵ *has probability* 1*.*

Finding a pattern in a random text is, in some sense, correlated to the previous occurrences of the same or other patterns [13]. Hence for example, the probability of finding $\mathrm{H}_1 = \mathtt{ATT}$ knowing that one has just found $\mathrm{H}_2 = \mathtt{TAT}$ is - intuitively - rather good since a `T` right after H_2 is enough to give H_1. *Correlation polynomials* and *correlation functions* give a way to formalize this intuition.

Definition 2. *The* correlation set *of two patterns* $\mathrm{H_i}$ *and* $\mathrm{H_j}$ *is the set of words* w *which satisfy: there exists a non-empty suffix* v *of* $\mathrm{H_i}$ *such that* $vw = \mathrm{H_j}$*. It is denoted* $\mathcal{A}_{i,j}$*. If* $\mathrm{H_i} = \mathrm{H_j}$*, then the correlation set is called the* autocorrelation set *of* $\mathrm{H_i}$*.*

Thus for example, the correlation set of $\mathrm{H}_1 = \mathtt{ATT}$ and $\mathrm{H}_2 = \mathtt{TAT}$ is $\mathcal{A}_{1,2} = \{AT\}$; the autocorrelation set of H_1 is $\{\epsilon\}$, while the autocorrelation set of H_2 is $\{\epsilon, \mathtt{AT}\}$. Empty string always belong to the autocorrelation set of any pattern.

Definition 3. *The* correlation polynomial *of two patterns* $\mathrm{H_i}$ *and* $\mathrm{H_j}$ *of length* m_i *and* m_j *is defined as:*

$$A_{i,j}(z) = \sum_{w \in \mathcal{A}_{i,j}} P(w) z^{|w|} \quad ,$$

where $|w|$ *denotes the length of word* w*. If* $\mathrm{H_i} = \mathrm{H_j}$*, then this polynomial is called the* autocorrelation polynomial *of* $\mathrm{H_i}$*. The* correlation function *is:*

$$D_{i,j}(z) = (1 - z) A_{i,j}(z) + P(\mathrm{H_j}) z^{m_j} \quad .$$

. When $H_i = H_j$*, the correlation function can be written* D_i*.*

The most common counting model is the *overlapping model*: overlapping occurrences of patterns are taken into account. It is as follows. For example, consider two oligonucleotides $\mathrm{H}_1 = \mathtt{ATT}, \mathrm{H}_2 = \mathtt{TAT}$ and a sequence `TTATTATATATT`. This sequence contains 2 occurrences of H_1 and 4 occurrences of H_2, as shown below:

$$\mathtt{T}\ \underbrace{\mathtt{T}\ \rlap{$\overbrace{\phantom{\mathtt{A\ T\ T}}}^{\mathrm{H}_1}$}\mathtt{A\ T}}_{\mathrm{H}_2}\ \underbrace{\mathtt{T\ A}\ \rlap{$\overbrace{\phantom{\mathtt{T\ A\ T}}}^{\mathrm{H}_2}$}\mathtt{T}}_{\mathrm{H}_2}\ \mathtt{A}\ \underbrace{\mathtt{T}\ \rlap{$\overbrace{\phantom{\mathtt{A\ T\ T}}}^{\mathrm{H}_1}$}\mathtt{A\ T}}_{\mathrm{H}_2}\ \mathtt{T\ T\ T}$$

It turns out [3] that our main results rely on the computation of the (real) roots of a polynomial equation:

Definition 4. *Let a be a real number such that $a > P(\mathrm{H}_1)$. Let (E_a) be the* fundamental equation*:*

$$D_1(z)^2 - (1 + (a-1)z)D_1(z) - az(1-z)D_1'(z) = 0 \ . \qquad (5)$$

Let z_a be the largest real positive solution of Equation (E_a) that satisfies $0 < z_a < 1$. The number z_a is called the fundamental root *of (E_a).*

3.2 p-value for a single pattern

The main result of this section is the theorem below, proven in [3], that provides the probability for the observed number of occurrences to be much greater than the expectation.

Theorem 1. *Let H_1 be a given pattern, and k be its observed number of occurrences in a random sequence of length n. Denote $a = \frac{k}{n}$ and assume that $a > P(\mathrm{H}_1)$. Then:*

$$pval(\mathrm{H}_1) = Prob(O(\mathrm{H}_1) \geq k) \approx \frac{1}{2\sigma_a\sqrt{n}} e^{-nI(a)+\delta_a} \qquad (6)$$

where

$$I(a) = a \ln\left(\frac{D_1(z_a)}{D_1(z_a) + z_a - 1}\right) + \ln z_a \ , \qquad (7)$$

$$\sigma_a^2 = a(a-1) - a^2 z_a \left(\frac{2D_1'(z_a)}{D_1(z_a)} - \frac{(1-z_a)D_1''(z_a)}{D_1(z_a) + (1-z_a)D_1'(z_a)}\right) \ , \qquad (8)$$

$$\delta_a = \log\left[\frac{P(\mathrm{H})z_a^m}{D_1(z_a) + (1-z_a)D_1'(z_a)}\right] \ , \qquad (9)$$

and z_a is the fundamental root of E_a. $I(a)$ is called the rate function*. Additionnally:*

$$Prob(O(\mathrm{H}_1) = k) \approx \frac{1}{\sigma_a\sqrt{2\pi n}} e^{-nI(a)+\delta_a} \ . \qquad (10)$$

Remark 1. When $a = P(\mathrm{H}_1)$, the number of H_1-occurrences is equal to its expected value. Conditional variance σ_a in (8) becomes: $\sigma = P(\mathrm{H}_1)(2A_1(1) - 1 + (1-2m)P(\mathrm{H}_1))$ (where m denotes the length of H_1),, e.g. the unconditional variance computed by various authors [27, 19].

Remark 2. The two probabilities $Prob(O(\mathrm{H}_1) \geq k)$ and $Prob(O(\mathrm{H}_1) = k)$ appear to be very similar in magnitude.

These results turn out to be very precise. It is shown in [3] that the *relative error* is $O(1/n)$; that is to say, the *neglected term* is upper bounded by $e^{-nI(a)}\frac{1}{\sigma_a n^{3/2}}$, which is exponentially small. A numerical comparison with the exact computation implemented, for the Bernoulli model, in [7] is given in Section 4. It appears that, when the random sequence is large, the formulae above provide an attractive alternative to the exact computation. Morevover, they also hold for the Markov model [3].

Another approach [5, 18] is the approximation of the word counting distribution by a compound Poisson distribution with the same mean, e.g. $nP(\mathrm{H})$ in the overlapping counting model. For the compound Poisson distribution defined in [18], the variance is not asymptotically equal to the variance of the process. As a consequence, the validity domain is restricted to the domain where the difference is small. More important, the normalization factor is improper. This implies an error on the *rate function* $I(a)$, and the neglected term in the validity domain, is not exponentially small [18]. Numerical evaluation of the compound Poisson distribution can be found in [12].

3.3 Conditioning by an overrepresented word

In this subsection, we assume that a pattern H_1 has been detected as an overrepresented word and we provide mathematical results to investigate the changes induced on the sequence distribution. Intuitively, the artefacts of an overrepresented word should look overrepresented. For example, if $\mathrm{H}_1 = AATAAA$, any word H2 $= ATAAAN$ is an artefact. A rough approximation of its expected value is $O(\mathrm{H}_1) \times \frac{P(N)}{P(A)}$. As $O(\mathrm{H}_1) >> E(\mathrm{H}_1)$, this is much greater than unconditioned expectation $E(\mathrm{H}_1) \times \frac{P(N)}{P(A)}$. The theorem below, proven in [3], establishes the precise formulae:

Theorem 2. *Given two patterns* H_1 *and* H_2*, assume the number of* H_1*-occurrences,* $O(\mathrm{H}_1)$*, is known and equal to* k*, with* $a = \frac{k}{n} \geq P(\mathrm{H}_1)$*. Then, the conditional expectation of* $O(\mathrm{H}_2)$ *is:*

$$E(O(\mathrm{H}_2)/O(\mathrm{H}_1) = k) \sim n\alpha \tag{11}$$

where α *is a function of the autocorrelation functions, the probabilitiesand* a*:*

$$\alpha = a\frac{D_{1,2}(z_a) \times D_{2,1}(z_a)}{D_1(z_a)(D_1(z_a) + z_a - 1)} \tag{12}$$

and z_a *is the fundamental root of Equation (5). Moreover, the variance is a linear function of* n*.*

Remark 3. In the central region, e.g. $k = nP(\mathrm{H}_1)$, substitutions $a = P(\mathrm{H}_1)$ and $z_a = 1$ in (11) yield $\alpha = P(\mathrm{H}_2)$, if H_1 and H_2 do not overlap.

Once a dominating signal has been detected, one looks for a weaker signal by a comparison of the number of observed occurrences of patterns with their conditional expectations. This procedure automatically eliminates artefacts. An example is provided in Section 6. It also allows for a choice of the best candidate between a word and its artefacts.

Computational complexity. Another approach is used in *Regexpcount* [11]. Although our formal proof of Theorem 2 relies on similar mathematical tools, our *explicit formulae* allow for skipping the expensive intermediate computations (bivariate generating functions, recurrences,...), hence provide a much faster algorithm.

4 Tandem repeats in B. subtilis and A. thaliana

In [7], authors search for localized repeats with a statistical filter. Software *Except* relies on a simple basic idea: long approximate repeats are likely to contain multiple exact occurrences of shorter words. DNA sequences are divided into overlapping fragments of size n. This size n is a parameter of the algorithm chosen for each run. Typically, n ranges from 250 to 5000. In each window, the p-value is computed for any pattern that occurs more than once. As the total number of occurrences, r, remains relatively small (typically 3 to 5), exact computation through generating functions is (theoretically) possible. Nevertheless, this approach, chosen by the authors, is computationally expensive. Typically, r repeated multiplications of polynomials of degree n. This gives a time complexity $O(n \log n \log r)$, if a Fast Fourier Transform is used, and numerical stability is rather delicate.

Oligomer	Obs.	p-val. (large dev.)	p-val. [7]	Z-sc.
AATTGGCGG	2	8.059×10^{-4}	8.343×10^{-4}	48.71
TTTGTACCA	3	4.350×10^{-5}	4.611×10^{-5}	22.96
ACGGTTCAC	3	2.265×10^{-6}	1.458×10^{-6}	55.49
AAGACGGTT	3	2.186×10^{-6}	2.780×10^{-6}	48.95
ACGACGCTT	4	1.604×10^{-9}	0.982×10^{-9}	74.01
ACGCTTGG	4	5.374×10^{-10}	4.391×10^{-10}	84.93
GAGAAGACG	5	0.687×10^{-14}	1.180×10^{-14}	151.10

Table 1. Measures on the 7 oligonucleotides considered in [7]

Large deviation computation for rare events reduces to the numerical computation of real roots of a polynomial equation, namely Equation (5). Hence, it is easier to program, faster and much more stable numerically. This was efficiently implemented in *Maple*, and compared with the results published in [7]. This is presented in Table 1. The results in [7] are given for one 2008 nucleotides long fragment in *A. thaliana* where 5 approximate tandem repeats of a 40-uple were found. For all patterns, the occurrence probability $nP(\mathrm{H})$ ranges between 10^{-6} and 10^{-7}. For each oligonucleotide, the first value is the number of occurrences in the window and the second one is the p-value computed by our large deviation formulae, where the correcting term δ_a has been neglected. The third one is the p-value computed in [7] with a generating function method and the last one is the Z-score. We notice that, for any pattern, the p-values computed with two different methods are of the same magnitude order (this is illustrated in Figure 4, where the logs of p-values are plotted.) However, they can differ up to a factor 1.72. This is due to the approximation done in our calculations. When a increases, the difference $z_a - 1$ also increases as well as the contribution of $e^{-\delta_a}$. Nevertheless, it is worth noticing that the p-value order is almost the same. One inversion occurs between patterns ACGGTTCAC and AAGACGGTT, that have similar p-values.

On the other hand, the last column of the table confirms that Z-score is not adequate for very rare events. Patterns AAGACGGTT and AATTGGCGG have the same Z-score 48, while p-values have a ratio 100. For patterns ACGACGCTT and ACGCTTGG, the two parameters define a different order. The same inversion appears between AATTGGCGG and TTTGTACCA.

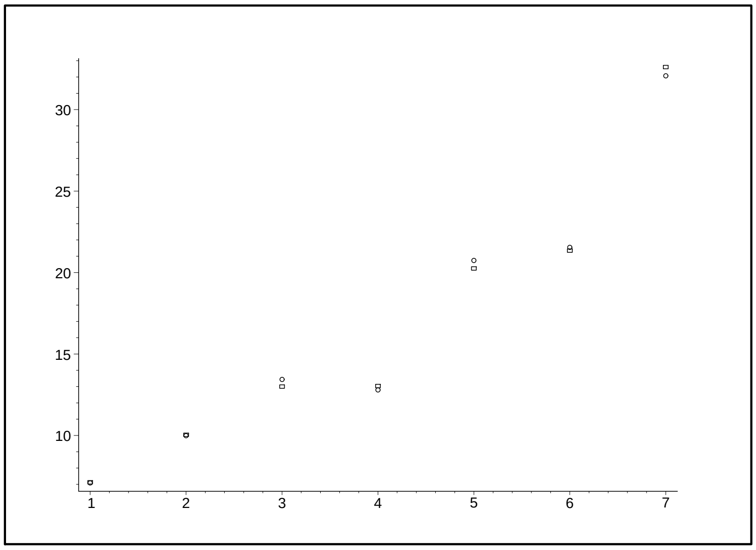

Fig. 1. Graphical comparison of -logs of p-values of Table 1. Values in abscissa refer to the ordering of motifs in the table. Circles: Large deviations formula.. Squares: Results of [7]

5 Oligonucleotide frequencies from yeast upstream regulatory regions

In [23], van Helden et al. study the frequencies of oligonucleotides extracted from regulatory sites from the upstream regions of yeast genes. Statistical significance of the oligonucleotides occuring in the 800 bp upstream sequences of regulatory regions is assessed by evaluating the probability of observing r or more occurrences of the oligonucleotide in the regulatory sequence, using the binomial formula. In [23], the probabilities are not computed in the Bernoulli model. For a given oligonucleotide, the authors count the number f of its occurrences in the non-coding sequences of yeast. Then, an approximate formula for $P(O(\mathrm{H}) \geq 1)$ is given and p-value follows through a computation of binomial formula (4). It is observed in [23] that these binomial statistics prove to be appropriate, except for self-overlapping patterns such as `AAAAAA`, `ATATAT`, `ATGATG`. As a matter of fact, auto-correlation does not affect the expected occurrence number, but increases the variance [8].In other words, the probability to observe either very high or very low occurrence values is increased for auto-correlated patterns.

Table 2 compares the results of several methods to compute the significance coefficient defined above. Figure 2 presents a graphical view of this comparison. The sequence considered is the 800 bp upstream region of ORF YGR022C of *Saccharomyces cerevisiæ* chromosome VII. By comparing *BF2*, *LDo* and *GF* (this last one representing the "more exact" result), we see that the large deviations values are very close to the GF ones (while their computation is much faster). The table also confirms that the overlaps must be taken into account when counting the significance coefficients : the three first patterns, for which the difference between *BF2* and *GF* is huge, are the periodic ones.

6 Polyadenylation signals in human genes

In [1], Beaudoing et al. study polyadenylation signals in mRNAs of human genes. One of their aims is to find several variants of the well known `AAUAAA` signal. For this purpose, they select 5646 putative mRNA 3' ends of length 50 nucleotides and seek for overrepresented hexamers. Pattern `AAUAAA` is clearly the most represented: it occurs in 3286 sequences, for a total number of 3456 occurrences. Seeking for other (weaker) signals involves searching for other overrepresented hexanucleotides. Nevertheless, it is necessary to avoid *artefacts*, e.g. patterns that appear overrepresented because they are similar to the first pattern. The algorithm designed by Beaudoing et al. consists in cancelling all sequences where the overrepresented hexamer has been found. Hence, they search for the most represented hexamer in the 2780 sequences which do not contain the strong signal `AAUAAA`.

Here we show how Theorem 2 gives a procedure for dropping the artefacts of a given pattern without cancelling the sequences where it appears. Table 3 presents the 15 most represented hexamers in the sequences considered in [1]. Columns 2 and 3 respectively give the observed number of occurrences and the

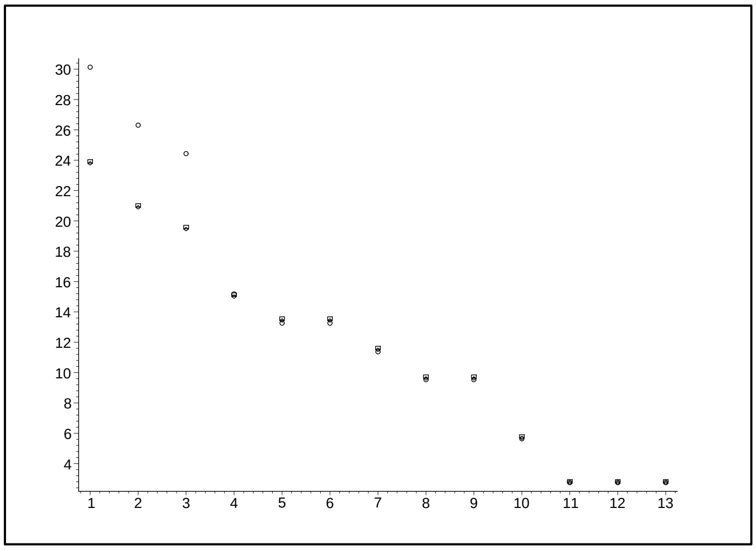

Fig. 2. Graphical comparison of significance coefficients (Sig) of Table 2. Values in abscissa refer to the ordering of motifs in the table. Circles: Binomial formula, Bernoulli probabilities (BF2). Diamonds: Large deviations considering overlaps. Boxes: generating function approach (GF).

Motif	O(H)	BF1	BF2	LD	LDo	GF
`TGATGA`	22	20	30.13	31.31	23.79	23.92
`GATGAT`	20	20	26.31	27.25	20.89	21.02
`ATGATG`	19	10.03	24.43	25.26	19.46	19.59
`GATGAG`	12	10.21	15.18	15.41	15.01	15.14
`GGATGA`	11	20	13.26	13.43	13.43	13.56
`ATGAGG`	11	20	13.26	13.43	13.43	13.56
`TGAGGA`	10	10.27	11.37	11.50	11.50	11.62
`AGGATG`	9	9.18	9.53	9.61	9.61	9.73
`GAGGAT`	9	9.05	9.53	9.61	9.61	9.73
`TGAAGA`	8	4.54	5.64	5.68	5.68	5.79
`AAGATG`	6	2.55	2.75	2.72	2.72	2.83
`GAAGAT`	6	2.39	2.75	2.72	2.72	2.83
`GATGAA`	6	2.35	2.75	2.72	2.72	2.83

Table 2. Computations of significance coefficient (Sig) of some hexanucleotides according to several methods, in the 800 bp upstream region of ORF YGR022c of *Saccharomyces cerevisiae* chromosome VII. *O(H)*: number of observed occurrences. *BF1*: Binomial formula with expected occurrences computed as in [23]. *BF2*: Binomial formula, Bernoulli probabilities. *LD*: Large deviations, without considering overlaps. *LDo*: Large deviations considering overlaps. *GF*: Generating function approach [7]. Column *GF* was computed using EXCEP software [7], based on formulas of [20, 17].

Hexamer	Obs.	Rk	Exp.	Z-sc.	Rk	Cd.Exp.	Cd.Z-sc.	Rk
AAUAAA	3456	1	363.16	167.03	1			**1**
AAAUAA	1721	2	363.16	71.25	2	1678.53	1.04	1300
AUAAAA	1530	3	363.16	61.23	3	1311.03	6.05	404
UUUUUU	1105	4	416.36	33.75	8	373.30	37.87	**2**
AUAAAU	1043	5	373.23	34.67	6	1529.15	-12.43	4078
AAAAUA	1019	6	363.16	34.41	7	848.76	5.84	420
UAAAAU	1017	7	373.23	33.32	9	780.18	8.48	211
AUUAAA	1013	8	373.23	33.12	10	385.85	31.93	**3**
AUAAAG	972	9	184.27	58.03	4	593.90	15.51	34
UAAUAA	922	10	373.23	28.41	13	1233.24	-8.86	4034
UAAAAA	922	11	363.16	29.32	12	922.67	9.79	155
UUAAAA	863	12	373.23	25.35	15	374.81	25.21	**4**
CAAUAA	847	13	185.59	48.55	5	613.24	9.44	167
AAAAAA	841	14	353.37	25.94	14	496.38	15.47	36
UAAAUA	805	15	373.23	22.35	21	1143.73	-10.02	4068

Table 3. Table of the most frequent hexanucleotides. *Obs*: number of observed occurrences. *Rk*: Rank. *Exp.*: (non-conditional) expectation. *Cd.Exp.*: expectation conditioned by number of occurrences of AAUAAA.

rank according to this criteria. Columns 4, 5 and 6 present the (non-conditioned) expected number of occurrences, the corresponding Z-score and the rank of the hexamer according to this Z-score. Here, the variance has been approximated by the expectation; this is possible as stated in [16]. Remark that rankings of columns 3 and 6 are quite similar: only patterns UAAAAA and UAAAUA do not belong to both rankings. A number of motifs look like the canonical one: they may be artefacts. This is confirmed by the three last columns which present, respectively, the expected number of occurrences conditioned by the observed number of occurrences of AAUAAA, the corresponding conditioned Z-score and the rank according to this criteria. It is clear that artefacts are dropped out, generally very far away in the ranking. It is worth noticing that some patterns which seemed overrepresented are actually avoided: this is the case for AUAAAU which goes down from 5th to last place (among the 4096 possible hexamers, only 4078 are present in the sequences). As AUAAAU is an artefact of the strong signal, this means that U is rather avoided right after this signal.

The case of UUUUUU in rank 2 is particular: this pattern is effectively overrepresented, but was not considered by Beaudoing et al. as a putative polyadenylation signal because its position does not match with observed requirements (around -15/-16 nucleotides upstream of the putative polyadenylation site.) It should also be stated that the approximation of the variance by the expectation that we do here for all patterns is not as good for periodic patterns like UUUUUU as for others [16]. By this way, variance of UUUUUU is underevaluated; so its actual Z-score is significantly lower than the one given in the table.

Now overrepresentation of AUUAAA (rank 3) is obvious; this is the known first variant of the canonical pattern. We remark that the following hexamer, UUAAAA, is an artefact of AUUAAA. It suggests to define a conditional expectation, or, even better, a p-value that takes into account the overrepresentation of two or more signals instead of one: in this example, AAUAAA *and* AUUAAA. This extension of Theorem 2 is the subject of a future work.

As it is mentioned above, the Z-score is not precise enough, and this remark also holds for conditioned Z-scores. In a second step, the authors of [1] computed a p-value defined by formula (4). This formula is approximated by the incomplete β-function. Nevertheless, any computation is rather delicate, and machine dependent due to numerous call to *exp* and *log* functions. The numerical stability necessitates a very careful use of real precision. It is worth noticing that large deviation principle applies for a Bernoulli process, with explicit values for the rate function and σ_a [3].

7 Conclusion and Perspectives

In this paper, we illustrated a possible use of large deviation methods in computational biology. These results allow, in some cases, a very fast computation of p-values that is numerically stable. These preliminary results are quite appealing and should be extended in several directions. First, it may be necessary to eliminate several strong independent signals [1]. A second task is the simplification of our formulae for artefacts: this would allow to achieve automatically the choice between a word and its subwords. A third task is the extension to the computation of the p-value for the conditioned case. Finally, regulatory sites may also be associated with structured motifs [10] or spurious motifs [2] and extension to this case should be realized.

8 Acknowledgements

We thank Eivind Coward for making the EXCEP software available for us and Jacques Van Helden for fruitful electronic discussions.

References

1. E. Beaudoing, S. Freier, J. Wyatt, J.M. Claverie, and D. Gautheret. Patterns of Variant Polyadenylation Signal Usage in Human Genes. *Genome Research.*, 10:1001–1010, 2000.
2. J. Buhler and M. Tompa. Finding Motifs Using Random Projections. In *RECOMB'01*, pages 69–76. ACM-, 2001. Proc.RECOMB'01, Montréal.
3. A. Denise and M. Régnier. Word statistics conditioned by overrepresented words, 2001. in preparation; http://algo.inria.fr/regnier/index.html.
4. M.S. Gelfand and E.V. Koonin. Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymess. *Nucleic Acids Research*, 25(12):2430–2439, 1997.

5. M. Geske, A. Godbole, A. Schafner, A. Skolnick, and G. Wallstrom. Compound Poisson Approximations for Word Patterns Under Markovian Hypotheses. *J. Appl. Prob.*, 32:877–892, 1995.
6. R. Karlin and S.F. Altschul. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci. U.S.A.*, 90:5873–5877, 1993.
7. Maude Klaerr-Blanchard, Hélène Chiapello, and Eivind Coward. Detecting localized repeats in genomic sequences: A new strategy and its application to *B. subtilis* and *A. thaliana* sequences. *Comput. Chem.*, 24(1):57–70, 2000.
8. J. Kleffe and M. Borodovsky. First and second moment counts of words in random texts generated by Markov chains. *Comput. Appl. Biosci.*, 8, 433-441, 1992.
9. X. Liu, D.L. Brutlag, and J. Liu. Bioprospector: Discovering conserved dna motifs in upstream regulatory regions of co-expressed gene. In *6-th Pacific Symposium on Biocomputing*, pages 127–138, 2001.
10. L. Marsan and M.F. Sagot. Extracting structured motifs using a suffix tree-algorithms and application to promoter consensus identification. In *RECOMB'00*, pages 210–219. ACM-, 2000. Proceedings RECOMB'00, Tokyo.
11. P. Nicodème. The symbolic package Regexpcount. In *GCB'00*, 2000. presented at GCB'00, Heidelberg, October 2000; available at http://algo.inria.fr/libraries/software.html.
12. G. Nuel. *Grandes déviations et chaines de Markov pour l'étude des mots exceptionnels dans les séquences biologiques.* Phd thesis, Université René Descartes, Paris V, 2001. to be defended in July,2001.
13. P.A. Pevzner, M. Borodovski, and A. Mironov. Linguistic of Nucleotide sequences:The Significance of Deviations from the Mean: Statistical Characteristics and Prediction of the Frequency of Occurrences of Words. *J. Biomol. Struct. Dynam.*, 6:1013–1026, 1991.
14. E.M. Panina, A.A. Mironov, and M.S. Gelfand. Statistical analysis of Complete Bacterial Genomes:Avoidance of Palindromes and Restriction-Modification Systems. *Genomics. Proteomics. Bioinformatics*, 34(2):215–221, 2000.
15. F.R. Roth, J.D. Hughes, P.E. Estep, and G.M. Church. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnol.*, 16:939–945, 1998.
16. M. Régnier, A. Lifanov, and V. Makeev. Three variations on word counting. In *GCB'00*, pages 75–82. Logos-Verlag, 2000. Proc. German Conference on Bioinformatics, Heidelberg; submitted to BioInformatics.
17. M. Régnier and W. Szpankowski. On Pattern Frequency Occurrences in a Markovian Sequence. *Algorithmica*, 22(4):631–649, 1998. preliminary draft at ISIT'97.
18. G. Reinert and S. Schbath. Compound Poisson Approximation for Occurrences of Multiple Words in Markov Chains. *Journal of Computational Biology*, 5(2):223–253,
19. G. Reinert, S. Schbath, and M. Waterman. Probabilistic and Statistical Properties of Words: An Overview. *Journal of Computational Biology*, 7(1):1–46, 2000.
20. S. Robin and J. J. Daudin. Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Prob.*, 36(1):179–193, 1999.
21. E. Rocha, A. Viari, and A. Danchin. Oligonucleotides bias in *bacillus subtilis*: general trands and taxonomic comparisons. *Nucl. Acids Research*, 26:2971–2980, 1998.
22. A.T. Vasconcelos, M.A. Grivet-Mattoso-Maia, and D.F. de Almeida. Short interrupted palindromes on the extragenic DNA of Escherichia coli K-12, Haemophilus influenzae and Neisseria meningitidis. *BioInformatics*, 16(11):968–977, 2000.

23. J. van Helden, B. Andre, and J. Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, 281:827–842, 1998.
24. Martin Tompa. An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. In *ISMB'99*, pages 262–271. AAAI Press, 1999. Seventh International Conference on Intelligent Systems for Molecular Biology, Heidelberg,Germany.
25. A. Vanet, L. Marsan, and M.-F. Sagot. Promoter sequences and algorithmical methods for identifying them. *Res. Microbiol.*, 150:779–799, 1999.
26. M.S. Waterman, R. Arratia, and D.J. Galas. Pattern recognition in several sequences: consensus and alignment. *Bull. Math. Biol.*, 45, 515-527, 1984.
27. M. Waterman. *Introduction to Computational Biology.* Chapman and Hall, London, 1995.