Fourth Colloquium on Mathematics and Computer Science DMTCS proc. **AG**, 2006, 129–140

Statistical Properties of Similarity Score Functions

Jérémie Bourdon and Alban Mancheron

LINA, CNRS FRE 2729 and University of Nantes

In computational biology, a large amount of problems, such as pattern discovery, deals with the comparison of several sequences (of nucleotides, proteines or genes for instance). Very often, algorithms that address this problem use score functions that reflect a notion of similarity between the sequences. The most efficient methods take benefit from theoretical knowledge of the classical behavior of these score functions such as their mean, their variance, and sometime their asymptotic distribution in a given probabilistic model.

In this paper, we study a recent family of score functions introduced in [MR03], which allows to compare two words having the same length. Here, the similarity takes into account all matches and mismatches between two sequences and not only the longest common subsequence as in the case of classical algorithms such as BLAST or FASTA. Based on generating functions, we provide closed formulas for the mean and the variance of these functions in an independent probabilistic model. Finally, we prove that every function in this family asymptotically behaves as a Gaussian random variable.

Keywords: average-case analysis; score functions; sequence comparison.

1 Introduction

In the last few decades, bio-informatics became a full discipline, at the crossroad of computer sciences and biology. Thus, several algorithms were developed in order to solve some biological problems (see [VA03] for a – non-exhaustive – review). In this paper, we focus on methods that try to solve the problem of discovering patterns in a set of nucleic or amino acid sequences, which may have biological properties. The main idea of almost all these algorithms is that a meaningful pattern (from biological point of view) is over-represented in input sequences, according to a similarity notion. Actually, a pattern is almost never perfectly conserved in biological sequences, due to mutations over the evolution process. So, algorithms may decide whenever two patterns are similar or not. But extracting all similar patterns is not sufficient, because the number of solutions may be too much important. For example, TEIRESIAS algorithm [RF98], used with default parameters, provides much more than 30 patterns that appears in all the sequences, for a set of 10 random *i.i.d* sequences having length 100 over DNA alphabet. It is sometime useful to assign a measure to each solution that permits to order the patterns and to output only "promising" solutions. So, methods should define the similarity notion and then, suggest at least one measure in order to refine the solution space. But, even if some measurement may indicate how similar patterns are, it is rarely sufficient. Some methods provide an estimation or an exact calculation of the expected value and the standard deviation of the behavior of the measure. Thus, these methods use the well known so called Z-value, or Z-score (also named "standardized score"). This score expresses how many standard-deviation units above or below the mean of the initial score falls. Thereafter, we will rather use the term "score" than "measure". Just recall that a score is not necessary a resemblance measure (either a distance or a similarity – for more details see [BB95] –).

The Z-score is quite useful when comparing two scores following different distributions (whatever they are) or when the distribution of the score function is known. In this last case, one can efficiently computes the probability of having at least one such score according to its distribution. This probability is also called the p-value, or p-score. Actually, knowing the Z-score is better than just having a score. But the p-value expresses a probability, which accredits the relevance of the result (see [DRV01] for a discussion).

In practice, most of the algorithms do not use Z-score, because the mean and the standard deviation of the score function is not known. Some methods compute the average score and the standard deviation empirically, but since it often requires too much time, they only compute an approximation. More often, the approximation is done under strong assumptions of independent model (for example, by supposing that

in pattern matching, each symbol occurs with a probability which is independent from the context). This appears to provide good approximations on the expected results.

For example, very popular methods in Combinatorial Biology like BLAST [AGM+90, KA90], or like FASTA [WL83, LP85, PL88] (more generally tools based on the Needleman-Wunsch algorithm [NW70]) use score functions, which were shown to follow the *extreme value distribution* (or "Gumbel") laws. This explains why their are so efficient and thus so popular. Another score function, which provides good (biologically speaking) results is the "Information Content" [Sha48, SSGE86]. This kind of score functions follows a *gamma distribution* [HS99]. This function is used in PRATT [EJT99, Jon97].

Mancheron *& al.* [MR03] elaborate an algorithm for pattern extraction problem that could use three families of score function. The first one allows to use substitution matrices, the second one corresponds to the function based on "Information Content" [Sha48], and the third family consists in the cutting in consecutive blocks of "matches" and "mismatches" in the pair-wise compared patterns. This cutting is often refereed as the "similarity" between the two sequences. In this paper, we will show that all the functions in this family asymptotically follow a *normal distribution* under the *Bernoulli* assumption. This family will be fully introduced in the latter section, as well as the method employed. Then, Section 3 concerns the computation of expected value and variance of these score functions, whereas Section 4 is dedicated to their distribution. Finally, we provide some closed formulas of some score functions and experimental results.

2 Generating functions for similarities

In this paper, we study the statistical behavior of a score between two sequences. This score is computed by using a decomposition of the similarity of these sequences.

More formally, we address the following problem: let $s = s_1 \dots s_n$ and $s' = s'_1 \dots s'_n$ be two sequences having length n. We denote by $w(s, s') = w_1 \dots w_n$ the binary word of length n that is defined by $w_i = 1$ if $s_i = s'_i$ and $w_i = 0$ otherwise. The word $w(s, s')$ is called the *similarity* of the two sequences s and s'.

The similarity is a sequence of runs of 0's and 1's that corresponds to the different matches and mismatches between the two original sequences. It can be written as

$$w(s, s') = 0^{\ell_0^{(1)}} 1^{\ell_1^{(0)}} 0^{\ell_1^{(1)}} \cdots 1^{\ell_t^{(0)}} 0^{\ell_t^{(1)}} 1^{\ell_{t+1}^{(0)}},$$

where $t \geq 0$, $\forall i \in \{1, \dots, t\}, \ell_i^{(1)}, \ell_i^{(0)} > 0$ and $\ell_0^{(1)}, \ell_{t+1}^{(0)} \geq 0$.

We will now use two functions $f^{(1)}$ and $f^{(0)}$ for scoring the runs of matches (runs of 1's) and the runs of mismatches (runs of 0's). These two functions are called the *component scoring functions*.

The score between two sequences equals the sum of all the values of the function $f^{(1)}$ (resp. $f^{(0)}$) for all the (length of the) runs of matches (resp. all the runs of mismatches) between the two sequences. It is defined by using the similarity $w(s, s')$ between s and s',

$$S(w(s, s')) = \sum_{i=1}^{t+1} f^{(1)}(\ell_i^{(0)}) + \sum_{i=0}^{t} f^{(0)}(\ell_i^{(1)}), \tag{2.1}$$

where $w(s, s') = 0^{\ell_0^{(1)}} 1^{\ell_1^{(0)}} 0^{\ell_1^{(1)}} \cdots 1^{\ell_t^{(0)}} 0^{\ell_t^{(1)}} 1^{\ell_{t+1}^{(0)}}$ is the similarity between s and s'. In fact, the score $S(w)$ is defined for any binary word w since any binary word can be decomposed as a sequence of runs of 0's and 1's.

This scoring function scheme has been used by the motifs extraction algorithm STARS [MR03].

When the sequences s and s' are some random sequences drawn by a memoryless source of respective probabilities $\{p_\alpha\}_{\alpha \in \Sigma}$, the similarity between s and s' is itself is a random binary sequence produced by a memoryless source of probabilities $\{p_0 := 1 - p, p_1 := p\}$, where

$$p = \sum_{\alpha \in \Sigma} p_\alpha p'_\alpha.$$

It is convenient here to introduce some formalism. We denote by $\Omega = \{0, 1\}^{\mathbb{N}}$ (resp.$\Omega_n = \{0, 1\}^n$) the set of infinite binary words (resp. words with n letters). We endow these sets with the memoryless source of probabilities $\{p_0, p_1\}$, that we denote by $\mathbb{P}_p$ and $\mathbb{P}_p^{(n)}$ (or more simply $\mathbb{P}$, and $\mathbb{P}^{(n)}$. In other words, under this distribution the letters are i.i.d. Bernoulli random variables with parameter $p = p_1$. Since the score of any words $w \in \Omega_n$ depends only on the lengths of the runs of 0 and 1 in w, we recall some

classical properties of the runs under $\mathbb{P}_p$. Under $\mathbb{P}_p$, the successive lengths of runs of 0's $L_i^{(0)}, i \geq 1$ and of 1's, $L_i^{(1)}, i \geq 1$ are independent. Furthermore, $L_i^{(0)}$ and $L_i^{(1)}$ are geometric random variables of respective parameters p and $1-p$ and thus, they satisfies $\forall i \geq 1$,

$$E_0 = \mathbf{E}\Big[L_i^{(0)}\Big] = \frac{1}{p}, E_1 = \mathbf{E}\Big[L_i^{(1)}\Big] = \frac{1}{1-p}, V_0 = \mathbf{Var}\Big[L_1^{(0)}\Big] = \frac{p}{(1-p)^2} \text{ and } V_1 = \mathbf{Var}\Big[L_1^{(1)}\Big] = \frac{1-p}{p^2}.$$

Hence, under $\mathbb{P}_p$, the score can be expressed as a sum of terms of the type $f^{(0)}(L_i^{(0)})$ and $f^{(1)}(L_i^{(1)})$. When the function $f^{(0)}$ and $f^{(1)}$ are of polynomial order, all the moments of these random variables are finite; we set

$$E_0' = \mathbf{E}\Big[f^{(0)}(L_1^{(0)})\Big], E_1' = \mathbf{E}\Big[f^{(1)}(L_1^{(1)})\Big], V_0' = \mathbf{Var}\Big[f^{(0)}(L_1^{(0)})\Big] \text{ and } V_1' = \mathbf{Var}\Big[f^{(1)}(L_1^{(1)})\Big]. \quad (2.2)$$

These quantities will be used to express most of our results. To avoid non trivial complications, we restrict our study to component scoring functions that are functions on $\mathbb{N}$ of polynomial order (i.e., there exists $\kappa > 0$ such that $f^{(1)}(m) = O(m^\kappa)$ and $f^{(0)}(m) = O(m^\kappa)$). This condition is quite natural as it ensures the existence of all moments of $f^{(0)}(L_1^{(0)})$ and $f^{(1)}(L_1^{(1)})$.

The constants defined in (2.2) are geometric-like series. Thus, for reasonable score component functions, they are efficiently approximated (in the sense that the computation of the n-th digit requires $O(n)$ terms of the series). Furthermore, for a large class of component functions, these constants admits a closed formula (in section 5 we provide formulas and results for polynomial score component functions).

Under $\mathbb{P}_p^n$, the score S_n is the sum of contributions of each blocks, it rewrites as

$$S_n = \sum_{i=1}^{\tau_n} (f^{(0)}(L_i^{(0)}) + f^{(1)}(L_i^{(1)})) + R_n \quad (2.3)$$

where $\tau_n = \max\{k, \sum_{i=1}^k L_i^{(0)} + L_i^{(1)} \leq n\}$ is the number of complete pairs of complete blocks that is needed to attain a word of length n. The quantity R_n is the contribution of the remainder part of the word, i.e., $R_n = f_0(K_n^{(0)}) + f_1(K_n^{(1)})$, where $K_n^{(0)}$ and $K_n^{(1)}$ are the lengths of the last blocks of 0's and 1's and are eventually null.

The remaining part of the paper is devoted to the asymptotic study of S_n when $n \to +\infty$. We derive its mean, its variance and prove that its distribution behaves asymptotically as a Gaussian variable.

2.1 Score generating functions

If the component scoring functions are linear, it is easy to show that the mean and the variance are moments of Bernoulli sums, and that the distribution is then the one of a sum of Bernoulli trials, *i.e.* a binomial distribution, which is asymptotically normal. But in the general case we need another approach.

We now introduce a very useful tool for studying problems on words. Let $\mathcal{L}$ be a set of words. The (probabilistic) *generating function* in one variable associated to the set $\mathcal{L}$ is the formal sum $L(z)$ defined by:

$$L(z) := \sum_{w \in \mathcal{L}} p_w z^{|w|} = \sum_{n \geq 0} z^n \sum_{w \in \mathcal{L}, |w|=n} p_w,$$

where $|w|$ denotes the length of w and p_w is the probability that a random word begins by w (which is commonly called the probability of w). We will denote by $[z^n]L(z)$ the coefficient of z^n in the formal sum $L(z)$.

The *score generating function* associated to the score function $S(w)$ is the bivariate (probabilistic) generating function $L(z,u)$ associated to the set $\mathcal{L}$ defined by:

$$L(z,u) := \sum_{w \in \mathcal{L}} p_w u^{S(w)} z^{|w|} = \sum_{n \geq 0} z^n \sum_{w \in \mathcal{L}, |w|=n} p_w u^{S(w)}.$$

The coefficient of z^n in $L(z,u)$ and its derivatives at $u = 1$ are fundamental in an average-case study.

2.2 Mean and variance of the score.

In the sequel, we will focus on the mean and the variance of a random variable S_n corresponding to a certain score function $S(w)$, when w is a random word of the set $\Sigma^\star = \{0,1\}^\star$ that has length n. These two quantities are easily expressed by means of the derivatives of the score generating function $L(z,u)$. Indeed, we have:

$$\mathbf{E}[S_n] := \sum_{|w|=n} p_w S(w) = [z^n]\frac{\partial}{\partial u}L(z,u)|_{u=1}, \text{ and } \mathbf{Var}[S_n] = \mathbf{E}\big[S_n^2\big] - (\mathbf{E}[S_n])^2,$$

$$\text{with } \mathbf{E}\big[S_n^2\big] := \sum_{|w|=n} p_w S(w)^2 = [z^n]\left(\frac{\partial^2}{\partial u^2}L(z,u)|_{u=1} + \frac{\partial}{\partial u}L(z,u)|_{u=1}\right).$$

Thus, obtaining tractable expressions for the score generating function and its derivatives allows to easily extract the coefficient of z^n.

With some natural assumptions on the score function (the score function is additive), we can use a "dictionary" that translates relations on sets to relations on their generating functions (*cf.* [FSar] for a detailed presentation of generating functions).

Sets	Generating functions
Σ	$z \cdot (\sum_{m\in\Sigma} p_m u^{S(m)})$
$\mathcal{A} \cup \mathcal{B}$	$A(z,u) + B(z,u)$
$\mathcal{A} \times \mathcal{B}$	$A(z,u) \times B(z,u)$
$\mathcal{A}^\star := \bigcup_{i\geq 0} \mathcal{A}^i$	$\frac{1}{1}$

3 The average score and its variance

First, notice that any word on alphabet $\{0,1\}$ decomposes as sequences of runs of 0's and runs of 1's. Formally, in a regular expression language, one has

$$\{0,1\}^\star = 0^\star(1^+0^+)^\star 1^\star,$$

where 0^+ and 1^+ denotes respectively the sets of runs of 0's (resp. 1's) of any strictly positive length.

This decomposition respects blocs. Furthermore, the score function is additive "by blocs" (*i.e.,* $S(u\cdot v) = S(u) + S(v)$ if the last symbol of u differs from the first symbol of v). We are thus able to apply the dictionary that translates relations on languages on relations on generating functions and

$$L(z,u) = (1 + S_0(z,u)) \cdot \frac{1}{1 - S_1(z,u)S_0(z,u)} \cdot (1 + S_1(z,u)),$$

where $S_0(z,u) := \sum_{k>0} p_0^k u^{f^{(0)}(k)} z^k$ and $S_1(z,u) := \sum_{k>0} p_1^k u^{f^{(1)}(k)} z^k$ are the score generating functions associated to the sets 0^+ and 1^+.

In the sequel, it proves useful to introduce the series

$$G_0(z,u) := \sum_{k>0} p_1 p_0^{k-1} u^{f^{(0)}(k)} z^k \text{ and } G_1(z,u) := \sum_{k>0} p_0 p_1^{k-1} u^{f^{(1)}(k)} z^k.$$

They are the generating functions of random geometric variables of respective probabilities p_1 and p_0. These series are closely related to S_0 and S_1 and one has $S_0 = p_0 G_0/p_1$ and $S_1 = p_1 G_1/p_0$. Thus $L(z,u)$ rewrites as,

$$L(z,u) = (1 + p_0 G_0(z,u)/p_1) \cdot \frac{1}{1 - G_1(z,u)G_0(z,u)} \cdot (1 + p_1 G_1(z,u)/p_0).$$

In order to obtain the average score, we compute the first derivative (according to variable u) of $L(z,u)$. One has

$$\frac{\partial}{\partial u} L(z,u) = \frac{\frac{p_0}{p_1}\frac{\partial G_0(z,u)}{\partial u}(1 + \frac{p_1}{p_0} G_1(z,u))^2 + \frac{p_1}{p_0}\frac{\partial G_1(z,u)}{\partial u}(1 + \frac{p_0}{p_1} G_0(z,u))^2}{(1 - G_0(z,u)G_1(z,u))^2}. \tag{3.1}$$

When evaluated at point $u = 1$, most of the quantities simplify. Indeed, one obtain

$$1 + \frac{p_0}{p_1} G_0(z,1) = \frac{1}{1 - zp_0}, \quad 1 + \frac{p_1}{p_0} G_1(z,1) = \frac{1}{1 - zp_1}, \text{ and}$$

$$\frac{1}{1 - G_0(z,1)G_1(z,1)} = \frac{(1 - zp_0)(1 - zp_1)}{1 - z}.$$

Furthermore, notice that the functions $(\frac{\partial G_0(z,u)}{\partial u}/z)$ and $(\frac{\partial G_1(z,u)}{\partial u}/z)$ are analytic functions on $\mathbb{R}$, non null for $z = 0$ and $z = 1$. The first derivative thus equals

$$\frac{\partial}{\partial u} L(z,u)\big|_{u=1} = \frac{z}{(1-z)^2}\frac{1}{p_0 p_1}[p_0^2 \frac{\partial G_0(z,u)/z}{\partial u}\big|_{u=1}(1 - zp_0)^2 + p_1^2 \frac{\partial G_1(z,u)/z}{\partial u}\big|_{u=1}(1 - zp_1)^2].$$

Finally, Lemma 1 clearly applies. This provides the following expression for the average score:

$$\mathbf{E}[S_n] = (n-1)p_0 p_1 (E_0' + E_1') + 2(p_0^2 E_0' + p_1^2 E_1') - p_0 p_1 (\overline{C_0} + \overline{C_1}) + o(1), \tag{3.2}$$

where $\overline{C_0}$ and $\overline{C_1}$ are given by

$$\overline{C_0} = \mathbf{E}\big[L^{(0)} f^{(0)}(L^{(0)})\big], \text{ and } \overline{C_1} = \mathbf{E}\big[L^{(1)} f^{(1)}(L^{(1)})\big].$$

The determination of the variance of the score involves the second derivative of $L(z,u)$ at point $u = 1$. Although this computation is more intricate, it does not imply additional technical improvements.

Finally we prove the following theorem.

Theorem 1 *The expectation and the variance of the score function $S(w)$ when w is a word of length n produced by a binary Bernoulli source of probabilities $p_0 = 1 - p$ and $p_1 = p$ satisfy*

$$\boldsymbol{E}[S_n] = (n-1)p_0 p_1 (E_0' + E_1') + 2(p_0^2 E_0' + p_1^2 E_1') - p_0 p_1 (\overline{C_0} + \overline{C_1}) + o(1), \tag{3.3}$$

$$\boldsymbol{Var}[S_n] = np_0 p_1 (V_0' + V_1' + (E_0' + E_1')^2 (p_0^3 + p_1^3 - 2) + 2p_0 p_1 (E_0' + E_1')(\overline{C_0} + \overline{C_1}) + o(n) \tag{3.4}$$

where the constants are closely related to moments of two independent geometric random variables $L^{(1)}$ and $L^{(0)}$ of respective success probabilities p_0 and p_1,

At this point, we have proven a linear behavior for the mean and the variance of the score for random strings. Thus, Bienaymé-Tchebyshev inequality allows to express a concentration property for the distribution of the score.The next step is to study the distribution of the score. We prove in the following section that this distribution follows asymptotically a normal law.

4 Distribution

The asymptotic distribution is one of the most informative results in the study of a sequence of random variables. For instance, Z-scores (i.e., the centered and normalized version of scores) that are used in a large amount of probabilistic heuristics, are especially meaningful when the refereed parameter possesses a normal distribution with mean 0 and variance 1.

When the component functions are linear functions, the score is a linear function of a binomial random variable and converges in distribution to a Gaussian random variable. In this section, we extend this result to any pair of component score functions of any type.

Theorem 2 *Under $\mathbb{P}_p^n$, the score S_n admits the following convergence in law*

$$\frac{S_n - \frac{nc'}{c}}{\sqrt{\textbf{\textit{Var}}[S_n]}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1),$$

where $c = E_0 + E_1 = \frac{1}{p(1-p)}$, $c' = E'_0 + E'_1$ and $\textbf{\textit{Var}}[S_n]$ admits the expressions (3.2) *and* (3.4).

We recall the following classical lemma:

Lemma 2 *Let (X_n) and (Y_n) be two sequences of random variables in $\mathbb{R}^d$, if $X_n \xrightarrow[n]{\mathcal{L}} X$ and $||X_n - Y_n|| \xrightarrow[n]{proba} 0$ (for a given norm on $\mathbb{R}^d$), then $Y_n \xrightarrow[n]{\mathcal{L}} X$.*

Proof of Theorem 2:
We introduce two centered random walks that will be useful to decompose the quantities of interests.

$$Z_k = \sum_{i=1}^{k} L_i^{(0)} + L_i^{(1)} - c, \qquad \text{and } Z'_k = \sum_{i=1}^{k} f^{(0)}(L_i^{(0)}) + f^{(1)}(L_i^{(1)}) - c',$$

where $c = E_0 + E_1 = \frac{1}{p(1-p)}$ and $c' = E'_0 + E'_1$.
The random variable S_n admits the following representation:

$$S_n - \frac{nc'}{c} = \left(Z'_{\tau_n} - Z'_{n/c}\right) + Z'_{n/c} + \left(\tau_n - \frac{n}{c}\right) c' + R_n.$$

Theorem 2 is a consequence of the following proposition:

Proposition 1 *The following convergences holds*

(1) $n^{-1/2} R_n \xrightarrow[n]{proba} 0$*;*

(2) $n^{-1/2}\sqrt{c}\left(Z'_{n/c}\,,\ (\tau_n - \frac{n}{c})c'\right) \xrightarrow[n]{\mathcal{L}} \mathcal{N}(0, M)$*, the centered Gaussian distribution in $\mathbb{R}^2$ with covariance matrix*

$$M = \begin{pmatrix} V'_0 + V'_1 & (c'/c)\rho \\ (c'/c)\rho & (c'/c)^2(V_0 + V_1) \end{pmatrix} \quad \textit{where } \rho = \frac{c'}{c} cov\left(f^{(0)}(L_1^{(0)}) + f^{(1)}(L_1^{(1)}), L_1^{(0)} + L_1^{(1)}\right),$$

(3) $n^{-1/2}(Z'_{\tau_n} - Z'_{n/c}) \xrightarrow[n]{proba} 0.$

Indeed, thanks to (2),

$$n^{-1/2}\left(Z'_{n/c} + (\tau_n - \frac{n}{c})c'\right) \xrightarrow[n]{\mathcal{L}} \frac{1}{\sqrt{c}}\mathcal{N}(0, \Sigma^2),$$

where the variance $\Sigma^2 = \frac{1}{c}((c'/c)^2(V_0+V_1)+(V_0'+V_1')+2(c'/c)\rho$. Now it is easy to check that this quantity is the same as the first order term of $\mathbf{Var}[S_n]$ given in (3.4).. Lemma 2 concludes the proof of theorem 2. □

We now prove all the points in the proposition.

Proof: (1) Since $f^{(0)}$ and $f^{(1)}$ are of polynomial order, there exists κ such that $f^{(0)}(m) \leq m^\kappa$ and $f^{(1)}(m) \leq m^\kappa$. Thus, one has

$$|R_n| = |f^{(0)}(K_n^{(0)}) + f^{(1)}(K_n^{(1)})| \leq |(K_n^{(0)})^\kappa| + |(K_n^{(1)})^\kappa| \leq |\sup_{1\leq n/2}(L_i^{(0)})^\kappa| + \sup_{1\leq n/2}|(L_i^{(1)})^\kappa|,$$

since $K_n^{(0)}$ and $K_n^{(1)}$ are included in one of the n first blocks.

Now, it is easy to prove that the probability that the maximum of $n/2$ i.i.d. geometrical random variable is larger than $\varepsilon\sqrt{n}$ goes to 0 when $n \to +\infty$ (the right order of this maximum is $\log n$).

(2) First, the definition of τ_n implies that

$$\left\{\frac{(\tau_n - n/c)c'}{\sqrt{n}} \leq y\right\} = \left\{Z_{\frac{n}{c}+y\frac{\sqrt{n}}{c'}} \geq -y\sqrt{n}\frac{c}{c'}\right\}.$$

By a simple application of the Bienaymé-Tchebyshev inequality $n^{-1/2}(Z_{\frac{n}{c}} - Z_{\frac{n}{c}+y\frac{\sqrt{n}}{c'}}) \xrightarrow[n]{proba.} 0$, and then applying lemma 2, $X_n = (Z'_{\frac{n}{c}}, Z_{\frac{n}{c}})$ and $Y_n = (Z'_{\frac{n}{c}}, Z_{\frac{n}{c}+y\frac{\sqrt{n}}{c'}})$ have the same limit in distribution in $\mathbb{R}^2$, if any. Now the vector $X_n = (Z'_n, Z_n)$ is clearly a sum of n i.i.d. (centered) random variables Γ_i with

$$\Gamma_i = (f^{(0)}(L_i^{(0)}) + f^{(1)}(L_i^{(1)}) - c', L_i^{(0)} + L_i^{(1)} - c).$$

The result is now a consequence of the central limit theorem applied to $X_{n/c}$.

(3) Let $\varepsilon > 0$. We shall establish that $\lim_{n\to\infty} \text{Prob}\Big[|Z'_{\tau_n} - Z'_{n/c}| \geq \varepsilon\sqrt{n}\Big] = 0$. We distinguish two cases whether the condition $|\tau_n - n/c| \leq n^{2/3}$ is satisfied or not. One has

$$\text{Prob}\Big[\tfrac{|Z'_{\tau_n} - Z'_{n/c}|}{\sqrt{n}} \geq \varepsilon\Big] \quad \leq \quad \text{Prob}\Big[|Z'_{\tau_n} - Z'_{n/c}| \geq \varepsilon\sqrt{n}, |\tau_n - n/c| \leq n^{2/3}\Big] + \text{Prob}\big[|\tau_n - n/c| > n^{2/3}\big].$$

The second probability tends to zero when $n \to \infty$ by (2). The first probability, denoted from now on by a_n satisfies

$$\begin{aligned} a_n &\leq \text{Prob}\big[\Delta(Z', [n/c - n^{2/3}, n/c + n^{2/3}]) \geq \varepsilon\sqrt{n}\big] && (4.1)\\ &= \text{Prob}\big[\Delta(Z', [0, 2n^{2/3}] \geq \varepsilon\sqrt{n}\big] && (4.2)\\ &\leq \text{Prob}\big[\max\{2|Z'_k|, k \in [0, 2n^{2/3}]\} \geq \varepsilon\sqrt{n}\big] && (4.3)\end{aligned}$$

where for any interval I, $\Delta(Z', I) = \max\{Z'_k, k \in I\} - \min\{Z'_k, k \in I\}$. Formula (4.1) is a consequence of the following considerations. First, if $I \subset J$ then $\Delta(Z', I) \leq \Delta(Z', J)$. Secondly $|Z'_{\tau_n} - Z'_{n/c}| \leq \Delta(Z', [\tau_n \wedge n/c, \tau_n \vee n/c])$. Hence $|Z'_{\tau_n} - Z'_{n/c}| \leq \Delta(Z', [n/c - n^{2/3}, n/c + n^{2/3}])$ when $|\tau_n - n/c| \leq n^{2/3}$. Equation (4.2) follows the Markov property of the random walk Z', and (4.3) is clear.

The classical Doob's inequality applied to the martingale (Z'_k), yields that for any $q > 1$,

$$\mathbf{E}[\max_{0\leq k\leq m} |Z'_k|^q] \leq (\frac{q}{q-1})^q \mathbf{E}[|Z'_m|^q].$$

Therefore, by (4.3) and the Markov inequality, taking $m = 2n^{2/3}$ and $q = 2$,

$$a_n \leq \text{Prob}\big[\max_{k\in[0,m]} |Z'_k|^2 \geq \varepsilon^2 n/4\big] \leq C\frac{\mathbf{E}\big[|Z'_m|^2\big]}{\varepsilon n} = \frac{Cm\sigma'^2}{\varepsilon n},$$

for some constant C. This latter quantity converges to zero when $n \to \infty$. □

5 Application to traditional functions and numerical examples

5.1 STARS*'s classical functions*

Now, we present several example of score functions. Here, our aim is to give closed expressions for the constants E'_0, $\overline{C}_0$, V'_0, E'_1, $\overline{C}_1$ and V'_1 for each score function used by the pattern extraction algorithm STARS.

Let X be a geometric random variable of parameter p. We compute the moments

$$\mathbf{E}[f(X)] = \sum_{k>0} f(k)p(1-p)^{k-1}, \qquad \mathbf{E}[Xf(X)] = \sum_{k>0} kf(k)p(1-p)^{k-1}, \qquad \text{and}$$

$$\mathbf{E}\big[(f(X))^2\big] = \sum_{k>0} f(k)^2 p(1-p)^{k-1},$$

for almost all component scoring functions (that could be) implemented in STARS.

For a monomial score, the computation can be easily done by using function $D(z) = 1/(1-z)$ (notice that $\sum_{k\geq 0} p^k = 1/(1-p)$) and the derivative operator $z\frac{\partial}{\partial z}$ (abbreviated by Δ). Indeed, when f is the monomial function $k \mapsto k^\ell$, the three moments express as

$$\mathbf{E}[f(X)] = \frac{p}{1-p}\Delta^\ell D(z)|_{z=1-p}, \qquad \mathbf{E}[Xf(X)] = \frac{p}{1-p}\Delta^{\ell+1} D(z)|_{z=1-p}, \qquad \text{and}$$

$$\mathbf{Var}[f(X)] = \frac{p}{1-p}\Delta^{2\ell} D(z)|_{z=1-p}.$$

Table 1 provides closed formulas for small values of ℓ. With the help of any kind of computer algebra software, it is obvious to obtain a closed formula for a monomial function of any degree. This formula corresponds to sums of Stirling numbers of second kind. Furthermore, linear property of moments allows to derive closed formulas for any polynomial function.

Table 1: Closed formulas some component score functions.

$k \mapsto f(k)$	$\mathbf{E}[f(X)]$	$\mathbf{E}[Xf(X)]$	$\mathbf{Var}[f(X)]$
$k \mapsto \alpha$	α	$\alpha\,\frac{1}{p}$	0
$k \mapsto k$	$\frac{1}{p}$	$\frac{2-p}{p^2}$	$\frac{1-p}{p^2}$
$k \mapsto k^2$	$\frac{2-p}{p^2}$	$\frac{6-6p+p^2}{p^3}$	$\frac{20-32p+13p^2-p^3}{p^4}$

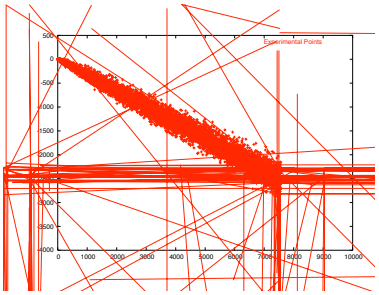
For functions that are polynoms, it is not obvious to obtain a closed formula. Nevertheless, for functions that are of order $o(k)$, such as $k \mapsto \sqrt{k}$ or $k \mapsto \log(1+k)$, the remainder $R_n := \sum_{k>n} f(k)p^k$ is of order $o(np^n)$. Thus, the computations of the first n terms of the sum give access to the first $\lfloor n \log p \rfloor$ digits of the constant.

5.2 Numerical experiments

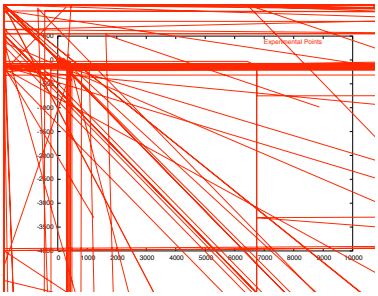
These experiments illustrate and confirm the theoretical results of previous sections. We consider two cases: first, a memoryless source with probability $\{p_0 = 3/4, p_1 = 1/4\}$ and second, words taken from the *Bacillus Subtilis* DNA sequence ($\simeq 4M$ nucleic acids). In this last case, we compute the match probability using the four bases frequencies ($p_1 = 0.254188$). We compare bases (from left to right) of two randomly chosen sub-sequences of size n from the whole genome. Then we build the sequence of size n over $\{0,1\}$ that corresponds to mismatches and matches.

We use the following score functions: $f^{\neq}(k) = -k$ and $f^{=}(k) = k^2$.

The two following figures are respectively the theoretical results and the experimental results. The X-axes and Y-axes corresponds respectively to the length of the words and to their scores. We trace the theoretical mean and the bandwidth given by the standard deviation.

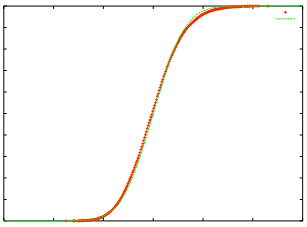


Memoryless sources

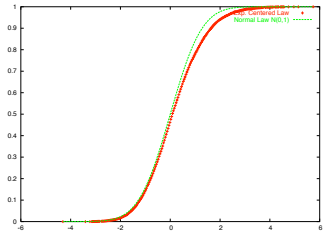


Bacillus Subtilis

Then, we provide the theoretical and experimental laws computed using a 20000 score measure sample for sequences of size 1000. We obtain the following results:



Memoryless sources



Bacillus Subtilis

We can notice that for real sequences, the distribution remains Gaussian. A small bias concerning scores grower than the average appears probably because bases are not randomly uniformly distributed in real sequences. Nevertheless, this bias does not really break the Gaussian behavior of the distribution.

6 Conclusion and Future Works

6.1 More general sources

Our study lies in a context of simple memoryless sources. It provides a precise approximation of the expected score and its variance. Nevertheless, biological (and more generally real) sequences are (fortunately) built in a more complex manner. Thus, a similar study in a context where sources admits correlations between symbols (such as Markov chains or, in general, dynamical sources [Val01]) should certainly be meaningful. Our study can entirely be performed in the context of dynamical sources. Indeed, the key fact of the study consists in the simplicity of the expression of S_0, S_1 and its derivatives at point $u = 1$. This also holds when the generating functions are replaced by generating operators (for details on generating operators, the reader can refer to [BV02]). Similar results, that moreover emphases the influence of correlations, can be deduced.

6.2 Non binary alphabet

To compute the score between the words $w^{[1]}$ and $w^{[2]}$, we considered only two kind of events, the "matches" and the "mismatches" for each position. Among the functions described in [MR03], those we have presented in this paper correspond to a sub-family, there is no difficulty to extend the previous results to the whole family. Indeed, among the functions based on cutting in consecutive blocks of "matches" and "mismatches", some depend on a "quorum". Within this context, a "match" is to be considered only if it occurred in more than $Q\%$ of the words already processed, while a "mismatch" is to be considered only if it occurred in less than $Q\%$ of the cases. In the other situations, the comparison has to be ignored. Thus, if we consider the two words $w^{[1]}$ and $w^{[2]}$, as well as the boolean vector of "presence" $\overrightarrow{q}$ of size n associated with the quorum Q (*ie.* , $\overrightarrow{q}[i]$ is true if, and only if, the constraint quorum is satisfied), we can build the word $w^Q \in \{0,1,x\}^n$ (Q being the quorum) such that: $w_i^Q = 1$ if $w_i^{[1]} = w_i^{[2]} \wedge \overrightarrow{q}[i]$, $w_i^Q = 0$ if $w^{[}1]_i \neq w^{[}2]_i \wedge \neg\overrightarrow{q}[i]$, and then $w_i^Q = x$ otherwise.

Here, we are thus interested in the score of a word built over the 3-ary alphabet $\{0,1,x\}$ where the consecutive matches of scored by the functions f_0, f_1 and $f_x \equiv 0$. We can state the most general problem

of studying the score functions defined by using a decomposition in blocs over an alphabet Σ with scoring functions $\{f_m\}_{m\in\Sigma}$. The basic decomposition of $\Sigma^\star$ has to be adapted for this context but the core of the study remains valid. For alphabets with more than two letters, we can make use of a recurrence between the set of words built over an alphabet with ℓ symbols and set of words built over an alphabet with $\ell+1$ symbols. Indeed, if $\Sigma_\ell := \{0,1,\ldots,\ell-1\}$ and $\Sigma_{\ell+1} := \{0,1,\ldots,\ell\}$ denote two alphabets of respectively ℓ and $\ell+1$ symbols, one has

$$\varepsilon + \Sigma_{\ell+1}^+ = (\varepsilon + \Sigma_\ell^+)\cdot(\ell^+\cdot\Sigma_\ell^+)^\star\cdot(\varepsilon+\ell^+).$$

This recurrence directly translates into a recurrence on score generating functions as shown in Section 2.

Aknowledgments. We wish to thank Jean-François Marckert for its helpful comments and advices. We also thank the anonymous referees for their precious reading of this paper.

References

[AGM+90] Stephen F. Altschul, Warren R. Gish, Webb Miller, Eugène W. Myers, and David J. Lipman. A Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[BB95] Vladimir Batagelj and Matev¸ Bren. Comparing Resemblance Measures. *Journal of Classification*, 12(1):73–90, 1995.

[BV02] Jérémie Bourdon and Brigitte Vallée. Generalized Pattern Matching Statistics. In Trends in Mathematics Birkhauser, editor, *Mathematics and Computer Science II*, pages 1–16, 2002.

[DRV01] Alain Denise, Mireille Régnier, and Mathias Vandenbogaert. Assessing the Statistical Significance of Overrepresented Oligonucleotides. In Olivier Gascuel and Bernard M. E. Moret, editors, *Algorithms in Bioinformatics. Proceedings of the 1st International Workshop on Algorithms in BioInformatics (*WABI*)*, volume 2149 of *Lecture Notes in Computer Science (*LNCS*)*, pages 85–97. Springer-Verlag, 2001.

[EJT99] Ingvar Eidhammer, Inge Jonassen, and William R. Taylor. Structure Comparison and Structure Patterns. Technical Report 174, Department of Informatics, University of Bergen, Norway, 1999.

[FO90] Philippe Flajolet and Andrew Odlyzko. Singularity analysis of generating functions. *SIAM J. Discrete Math.*, 3(2):216–240, 1990.

[FSar] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics—Symbolic Combinatorics*. Research Report of the INRIA, to appear. http://algo.inria.fr/flajolet/Publications/books.html.

[HS99] Gerard Z. Hertz and Gary D. Stormo. Identifying DNA and Protein Patterns with Statistically Significant Alignments of Multiples Sequences. *Bioinformatics*, 15(7–8):563–577, 1999.

[Jon97] Inge Jonassen. Efficient discovery of conserved patterns using a pattern graph. *Computer Applications in the Biosciences (*CABIOS*)*, 13:509–522, 1997.

[KA90] Samuel Karlin and Stephen F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. In *Proceedings of National Academy of Science (*PNAS*)*, volume 87 of *6*, pages 2264–2268, 1990.

[LP85] David J. Lipman and William R. Pearson. Rapid and Sensitive Protein Similarity Search. *Science*, 227(4693):1435–1441, 1985.

[MR03] Alban Mancheron and Irena Rusu. Pattern discovery allowing gaps, substitution matrices and multiple score functions. In Gary Benson and Roderic Page, editors, *Algorithms in Bioinformatics. Proceedings of the 3rd International Workshop on Algorithms in BioInformatics (*WABI*)*, volume 2812 of *Lecture Notes in Bioinformatics (*LNBI*)*, pages 129–145. Springer-Verlag, 2003.

[NW70] Saul B. Needleman and Christian D. Wunsch. A General Method Applicable to the Search for Similarities in the Amino Acid Sequences of Two Proteins. *Journal of Molecular Biology*, 48:443–453, 1970.

[PL88] William R. Pearson and David J. Lipman. Improved tools for biological sequences comparison. In *Proceedings of National Academy of Science (*PNAS*)*, volume 85, pages 2444–2448, 1988.

[RF98] Isidore Rigoutsos and Aris Floratos. Combinatorial pattern discovery in biological sequences: The TEIRESIA algorithm. *Bioinformatics*, 14(1):55–67, 1998.

[Sha48] Claude E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[SSGE86] Thomas D. Schneider, Gary D. Stormo, Larry Gold, and Andzej Ehrenfeuch. The Information Content of Binding Sites on Nucleotide Sequences. *Journal of Molecular Biology*, 188:415–431, 1986.

[VA03] Susana Vinga and Jonas S. Almeida. Alignment-free sequence comparison – a review. *Bioinformatics*, 19(4):513–523, 2003.

[Val01] Brigitte Vallée. Dynamical sources in information theory: fundamental intervals and word prefixes. *Algorithmica*, 29(1-2):262–306, 2001.

[WL83] W. John Wilbur and David J. Lipman. Rapid similarity searches of nucleic acid and protein data banks. In *Proceedings of National Academy of Science (*PNAS*)*, volume 80, pages 726–730, 1983.