Statistical significance for sequence analysis. Illustration of new results on length and position of the local score.

Agnes Lagnoux, Sabine Mercier, Pierre Vallois

▶ **To cite this version:**

Agnes Lagnoux, Sabine Mercier, Pierre Vallois. Statistical significance for sequence analysis. Illustration of new results on length and position of the local score.. 2016. hal-01301246v1

HAL Id: hal-01301246

https://hal.science/hal-01301246v1

Preprint submitted on 13 Apr 2016 (v1), last revised 7 Jan 2019 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Statistical significance for sequence analysis. Illustration of new results on length and position of the local score.

Agnès Lagnoux [1,*], Sabine Mercier [1] and Pierre Vallois [2]

April 13, 2016

1 Introduction

Biological sequence analysis has been largely studied from the 1990's Altschul *et al.* , 1990; Watson , 1995. In practice a biological sequence is considered as a succession of letters which belong to a finite set $\{A_1, \ldots, A_k\}$[1] A score is a function s that gives a real number to any letter A_i. For sequences of a given type, the score permits to determine their physico-chemical properties, such as hydrophobicity. The local score H_n of a sequence $(A_k)_{1\leqslant k\leqslant n}$ of length n, also called Smith and Waterman score, is defined by

$$H_n := \max_{0\leqslant i\leqslant j\leqslant n} \sum_{k=i}^{j} X_k \tag{1}$$

where $X_0 = 0$ and $X_k = s(A_k)$ for $k \geqslant 1$. It is usually supposed that the random variables $(X_i)_{1\leqslant i\leqslant n}$ are independent and identically distributed (i.i.d.). The random variable H_n plays a central role in the analysis of biological sequences and therefore the calculation of its statistical significance is crucial. Using classical tools of Markov chains, the authors in Mercier and Daudin , 2001 have proven that

$$\mathbb{P}(H_n \geqslant a) = (1, 0, \ldots, 0) \cdot \Pi^n \cdot (0, \ldots, 0, 1)', \tag{2}$$

where Π is a $(a+1)$-square matrix linked to the distribution of $(X_i)_{i\geqslant 1}$ and the sign $'$ stands for the transpose of a matrix.
Relation (2) is usable when n is "small". While n goes to infinity, the asymptotic behaviour of H_n depends on the mean score value $\mathbb{E}[X]$. It is a transition phase parameter Arratia and Waterman , 1994 because the score grows as $\log(n)$ for $\mathbb{E}[X] < 0$ Watson , 1995, as n for $\mathbb{E}[X] > 0$ Watson , 1995 and as $\sqrt{n}$ for $\mathbb{E}[X] = 0$ Daudin *et al.* , 2003.
In the case where the mean score is negative and n is large, a Gumbel distribution fits to the asymptotic distribution of H_n minus a logarithmic term, namely

$$\mathbb{P}\left(H_n \leqslant \frac{\log n}{\lambda} + a\right) \underset{n\to\infty}{\approx} \exp(-K^* \cdot e^{-\lambda a}), \tag{3}$$

where λ and K^* depend on the distribution of $(X_i)_{i\geqslant 1}$. The proof of (3) is based on arguments coming from renewal theory Karlin and Altschul , 1990; Karlin and Dembo , 1992.
When $\mathbb{E}[X] = 0$, the asymptotic behaviour of the local score is derived using Brownian motion theory in Daudin *et al.* , 2003 and

$$\mathbb{P}(H_n \leqslant \sqrt{n}a) \underset{n\to\infty}{\approx} \frac{2}{\pi} \sum_{k\in\mathbb{Z}} \frac{(-1)^k}{2k+1} \exp\left\{-\frac{(2k+1)^2\pi^2}{8a^2}\right\}, \tag{4}$$

*to whom correspondence should be addressed
[1]In the case of the DNA, the letters of interest are A, C, G and T.

with the rate of convergence established in Etienne and Vallois , 2003. The authors also derive an asymptotic result whatever the sign of $\mathbb{E}[X]$ for large values of a:

$$\mathbb{P}(H_n \geqslant a) \underset{n\to\infty}{\approx} 2 \cdot \sqrt{\frac{2n}{\pi}}\frac{\sigma}{a} \cdot \exp\left(-\frac{\delta_n - a/\sqrt{n}}{2\sigma^2}\right)^2, \tag{5}$$

where $\delta_n := \sqrt{n} \cdot \mathbb{E}[X]$ and $\sigma := \sqrt{\mathrm{Var}(X)}$.
The statistical analysis of the local score is still an active and challenging field, see reviews Lesk , 2005; Karlin , 2005; Borodovsky and Ekisheva , 2006; Mitrophanov and Borodovsky , 2006 and recent articles Wolfsheimer *et al.* , 2011; Xia *et al.* , 2015. Naturally, the length of the segment which realises the local score is also of interest. More generally, motivated by sequences comparison, Arratia and Waterman Arratia and Waterman , 1989 considered the longest head run larger than a given threshold. An asymptotic behaviour of the length of segments of $(X_k)_k$ with cumulative score exceeding a given threshold is established when $\mathbb{E}[X] < 0$ in Dembo and Karlin , 1991,?. In Karlin and Ost , 1988, the author established a classical extremal type limit law for the length of common words among a set of random sequences. More recently, Reinert and Waterman Reinert and Waterman , 2007 proposed a result on the distribution for the length of the longest exact match for a random sequence across another sequence.
In Chabriac *et al.* , 2014, the authors proposed a slightly different local score H_n^*, defined on adequately truncated sequence, and introduced the associated length L_n^* (see Section 4 for details). When $\mathbb{E}[X] = 0$, using Brownian motion theory, they derived the asymptotic behaviour of

$$\mathbb{P}(H_n^* \geqslant \sqrt{n}a; L_n^* \leqslant n\ell), \quad a \geqslant 0,\ 0 \leqslant \ell \leqslant 1. \tag{6}$$

Moreover, it has been proven in Lagnoux *et al.* , 2015 that $\mathbb{P}(H_n = H_n^*)$ converges to an explicit value as $n \to \infty$ that traduces the fact that the probability that H_n is achieved on a final part of the sequence is quite constant when n is large.
The goal of this paper is to illustrate the results based on the pair local score-length and the one on the local score position. In Section 2, we measure with statistical tests how different approximations of the local score distribution fit simulated sequences. In Section 3) we add the local score the length of the segment that realises it and we study the induced changes with numerical simulations. In Section 4, we introduce a new one dimensional statistic which is a function of the two above variables and we test its distribution. Finally in different settings, we compare the classical local score with the one calculated over adequately truncated sequences. This leads us to illustrate the result on the local score position.

2 Accuracy of the results for H_n

2.1 First illustrations

In sequence comparison, the asymptotic behaviour of the empirical distribution of the local score H_n, when n is large, is usually represented by the regression line of $(x, \log(-\log \mathbb{P}(H_n \geqslant x)))$. The approximation (3) of Karlin *et al.* Karlin and Altschul , 1990 implies that

$$\log\left(-\log \mathbb{P}\left(H_n \leqslant \frac{\log n}{\lambda} + a\right)\right) \approx \log(nK^*) - \lambda a.$$

We determine regression lines for different values of n, where $\mathbb{P}(H_n \leqslant x)$ is calculated using the exact method. We recover two facts (see Fig. 1): a common slope for the different lines and the value at the origin that depends on n. However, to our knowledge, goodness-of-fit tests have never been done.

2.2 Goodness-of-fit tests for the local score

Now, we consider a N-sample of sequences of length n. For any sequence $1 \leqslant i \leqslant N$, we determine its local score and then we perform Kolmogorov-Smirnov tests (see reminders in the Appendix) with different

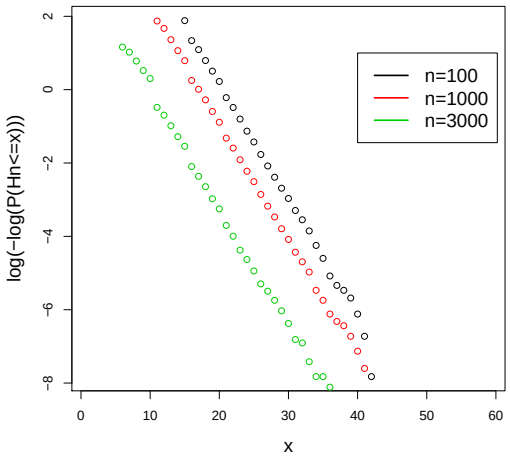

Figure 1: Three regression lines $(x, \log(-\log \mathbb{P}(H_n \geqslant x)))$ obtained for $n = 100$, 1000 and 3000, where $\mathbb{P}(H_n \geqslant x)$ has been calculated with the exact method. We observe a common slope while the value at the origin $\log(nK^*)$ differs according to of the sequence length n.

theoretical cumulative distribution functions F ($F(x) = \mathbb{P}(H_n \leqslant x)$) and several distributions for X. We consider three cases:

i the Karlin *et al.* limit, i.e. $F(x)$ is equal to the right hand side of (3) with $x = \log(n)/\lambda + a$ and $\mathbb{E}[X] < 0$; We also consider its improvement proposed by Cellier *et al.* in Cellier *et al.* , 2003.

ii $F(x)$ is defined as the right hand side of (4) and $\mathbb{E}[X] = 0$;

iii the exact method: F is the distribution function of (2) valid whatever the sign of the mean local score.

In case ii, i.e. when $\mathbb{E}[X] = 0$, we consider the theoretical distribution of (2) defined by the right hand side of (4) and (5). For different values of n, we compare these results with the ones obtained by the exact method and we display them via a graphical representation in Fig. 2.
We now deal with the case $\mathbb{E}[X] \neq 0$, i.e. items i and ii The results are given in Fig. 3, where $N = 10^4$, $\alpha = 99\%$ and $\mathbb{E}[X] = -2$. The horizontal line corresponds to the Kolmogorov-Smirnov 1%-quantile. The graphs show that the exact method is always accepted as expected and the improved method introduced by Cellier *et al.* is accepted as soon as the sequence length n is larger than 10^3. However, when the limit distribution is the one of Karlin *et al.*, the test is surprisingly often rejected even if n is large. For $\mathbb{E}[X] = 0$ and (4) and (5), the Kolmogorov-Smirnov distances for $n \leqslant 4000$ are much larger than in Fig. 3. The adjustment hypothesis to the theoretical distribution is always rejected.
In Fig. 4, we observe how the Kolmogorov-Smirnov distance D_N is realised. For small local scores which are not interesting in our context: the maximum distance (0.12) is attained at x such that the empirical distribution of the sample $\simeq 70\%$ or 90%. Recall that the Kolmogorov-Smirnov distance takes into account the whole distribution: a local tool, centered to the extreme probabilities, will be more adapted.

3 Taking into account the length of the local score

The aim of this section is first to determine the sequences that are statistically interesting and second to compare the results using the single local score H_n and the pair (H_n, L_n) where L_n is the length of the (last) segment that realises the local score H_n.

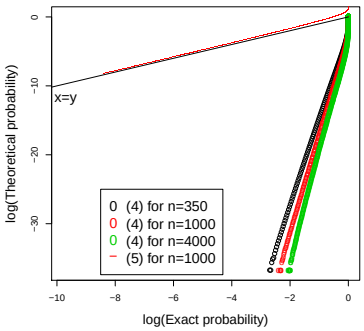

Figure 2: Representation of $\mathbb{P}(H_n \leqslant x)$ for each observed local score x in the sample using the exact distribution given by (2) (x-axis) and the theoretical results based on (4) and (5) (y-axis). The logarithmic scale allows to focus on the distribution tail.

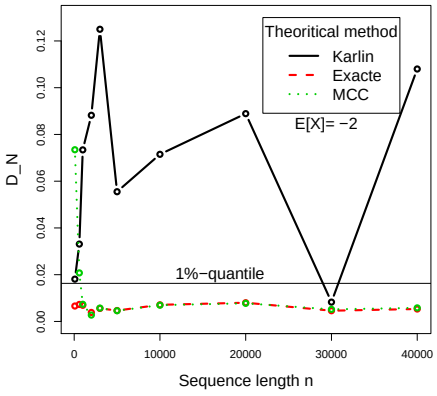

Figure 3: Goodness-of-fit tests for the local score for different theoretical distributions and $\mathbb{E}[X] = -2$. The approximation of Karlin *et al.* is very often rejected. Its improvement (denoted MCC) is rapidly accepted while the exact method is always accepted as expected.

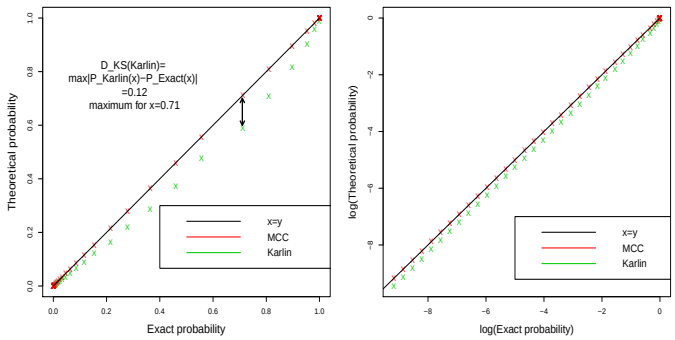

Figure 4: Comparison of the approximated distribution for the local score H_n proposed by Karlin *et al.* and an empirical one based on a Monte Carlo approach for 10^5 i.i.d. simulated sequences of length $n = 3000$. The Kolmogorov distance (0.12) is realised for a local score corresponding to an exact probability about 70%. Notice that considering only local scores with probabilities less than 5% (to focus on the region of interest), the maximum distance is reduced to 0.013. The logarithmic scale allows us to focus on the distribution tail but the Kolmogorov-Smirnov distance is not highlighted.

3.1 Wrongly classified sequences

We compare $\mathbb{P}(H_n \geqslant h; L_n \leqslant \ell)$ and $\mathbb{P}(H_n \geqslant h)$ for some values of h and ℓ using a Monte Carlo scheme. For a sequence $(X_k)_{0\leqslant k\leqslant n}$, it is easy to determine H_n and L_n. The simulation of a sample of N sequences of length n gives rise to $(h_{n,i}, \ell_{n,i})_{1\leqslant i\leqslant N}$. Naturally, $\mathbb{P}(H_n \geqslant h_{n,i})$ (resp. $\mathbb{P}(H_n \geqslant h_{n,i}; L_n \leqslant \ell_{n,i})$) is estimated by its empirical probability i.e.

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{H_n \geqslant h_{n,i}\}} \left(resp.\ \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{H_n \geqslant h_{n,i}\}} \mathbb{1}_{\{L_n \leqslant \ell_{n,i}\}} \right).$$

The results can be seen in Fig. 5. Note that the inequality

$$\mathbb{P}(H_n \geqslant h) \geqslant \mathbb{P}(H_n \geqslant h; L_n \leqslant \ell) \tag{7}$$

implies that all the points of the scatter plot are above first bisector.
According to statistical significance based on the tool of p-values (see the Appendix for more details), a sequence with $H_n = h$ and $L_n = \ell$ is said *α-wrongly classified* if it is α-(H_n, L_n) significant but not α-H_n, i.e.

$$\mathbb{P}(H_n \geqslant h) > \alpha > \mathbb{P}(H_n \geqslant h; L_n \leqslant \ell),$$

for a given level $\alpha \in [0, 1]$. Table 1 gives the percentage of wrongly classified simulated sequences for different levels α and values of n.

3.2 Lists of the best significant sequences

We consider the 606 sequences of the SCOP file[2] and use the hydrophobic scale of Kyte and Doolitle Kyte and Doolittle , 1982. For any sequence i, $1 \leqslant i \leqslant 606$, n_i stands for its length and h_{n_i} (resp. l_{n_i}) denotes its score (resp. the length that realises the local score). In this data set, the minimal value of n_i is 18, the maximal one is 404 and the mean of all the n_i's is 115.3. Fig. 6 gives the 606 observed local score and length $(h_{n_i}; \ell_{n_i})_{1\leqslant i\leqslant 606}$. Then, we estimate the probability $\mathbb{P}(H_{n_i} \geqslant h_{n_i})$ for all i, via the exact method based on

[2]CF scop2dom 20140205aa. *http://scop2.mrc-lmb.cam.ac.uk/downloads/*

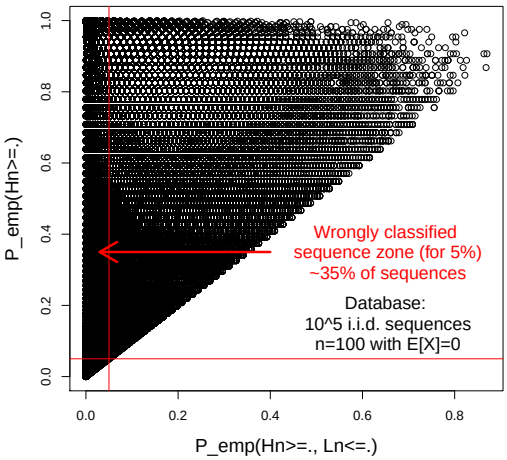

Figure 5: Each point represents one od the $N = 10^5$ simulated sequences of common length $n = 100$ with coordinates the estimations of $\mathbb{P}(H_n \geqslant h_{n,i}; L_n \leqslant l_{n,i})$ and $\mathbb{P}(H_n \geqslant h_{n,i})$. Naturally, all the points are above the $x = y$ line by (7); but many are far from it.

Table 1: Percentage of wrongly classified sequences among simulated sequences for different levels α with a distribution on $\{-5, \ldots, +5\}$ such that $\mathbb{E}[X] = 0$. For $\alpha = 5\%$, more than one third of the simulated sequences are wrongly classified. Simulations have been done for i.i.d. simulated sequences with different length n.

N	n	$\alpha = 0.1\%$	1%	5%
10^5	100	1.62%	11.70%	33.84%
$5 \cdot 10^4$	350	2.05%	12.44%	34.89%
10^4	1000	2.09%	12.61%	34.57%

(2)[3]. In order to compare the test based on the local score and the one making use of the local score-length, we only consider the ten sequences $i_1, \ldots, i_{10}$ with the smallest probabilities for the local score:

$$\mathbb{P}(H_{n_{i_1}} \geqslant h_{n_{i_1}}) \leqslant \ldots \leqslant \mathbb{P}(H_{n_{i_{10}}} \geqslant h_{n_{i_{10}}}).$$

Then, for any k, $1 \leqslant k \leqslant 10$, we simulate $N = 10^5$ i.i.d. sequences of length n_{i_k} and we estimate $\mathbb{P}(H_{n_{i_k}} \geqslant h_{n_{i_k}}; L_{n_{i_k}} \leqslant h_{n_{i_k}})$. The characteristics of the top ten sequences are gathered in Table 2. Three sequences, called i_{11}, i_{12} and i_{13}, are added at the bottom of Table 2. Then we have ordered $\mathbb{P}(H_{n_{i_k}} \geqslant h_{n_{i_k}}; L_{n_{i_k}} \leqslant h_{n_{i_k}})$ for all $1 \leqslant k \leqslant 13$ and mention the rank of each sequence in the last column. The new ordered list based on the local score-length is different from the one considering the local score only. For instance, the sequences of the bottom of Table 2 have a high local score probability but they have a low one considering the local score-length. It would be possible to determine all the α-wrongly classified sequences but it would be an heavy computationally task. Our purpose is just to show that the lists of ranked sequences using either the local score or the local score-length are very different and thus the best significant sequences are not the same.

[3]In that view, we need to estimate the distribution which is done on the 606 sequences.

Table 2: Top10 local score list for the SCOP file[2] (top) and three wrongly classified significant sequences (bottom). Exact proba. refers to the one of $\mathbb{P}(H_{n_i} \geqslant h_{n_i})$, pair estimation means the estimation of $\mathbb{P}(H_{n_i} \geqslant h_{n_i}, L_{n_i} \leqslant \ell_{n_i})$ and (H_n, L_n) order is the order given by the probability of the pair. As those observations are extremes, the estimations are not precised even for 10^5 simulated sequences.

n_i	h_{n_i}	ℓ_{n_i}	p-value	H_n order	Pair Estimation	(H_n, L_n) order
173	185	169	10^{-6}	1	$.< 10^{-6}$	1
103	106	88	$3.13 \cdot 10^{-4}$	2	$5 \cdot 10^{-5}$	2
80	93	76	$4.17 \cdot 10^{-4}$	3	$3.10 \cdot 10^{-4}$	4
94	100	85	$4.03 \cdot 10^{-4}$	4	$2.50 \cdot 10^{-4}$	3
93	88	86	$1.68 \cdot 10^{-4}$	5	$1.24 \cdot 10^{-3}$	5
111	82	107	$6.41 \cdot 10^{-3}$	6	$5.81 \cdot 10^{-3}$	9
129	76	127	$1.75 \cdot 10^{-2}$	7	$1.69 \cdot 10^{-2}$	13
227	93	102	$1.84 \cdot 10^{-2}$	8	$2.94 \cdot 10^{-3}$	8
145	73	130	$2.98 \cdot 10^{-2}$	9	$2.64 \cdot 10^{-2}$	12
109	67	79	$2.56 \cdot 10^{-2}$	10	$1.37 \cdot 10^{-2}$	11
113	49	22	$1.26 \cdot 10^{-1}$	33	$1.37 \cdot 10^{-3}$	6
133	44	18	$2.28 \cdot 10^{-1}$	67	$1.53 \cdot 10^{-3}$	7
227	40	19	$4.96 \cdot 10^{-1}$	192	$8.21 \cdot 10^{-3}$	10

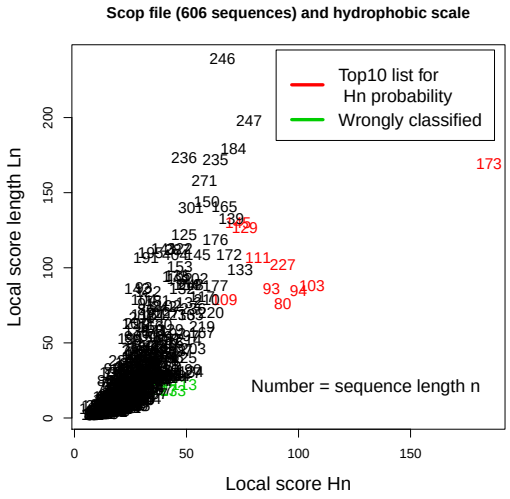

Figure 6: The values of (H_n, L_n) for the 606 sequences of the SCOP file[1] and the 10 sequences with the lowest local score probability (red) and three sequences among the wrongly classified ones (green).

4 Accuracy of the results for the pair

4.1 Background and notation

The local score of Smith and Waterman defined by (1) can be rewritten as

$$H_n = \max_{0 \leqslant j \leqslant n} U_j \tag{8}$$

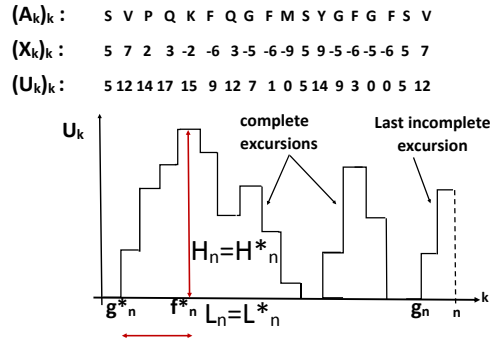

Figure 7: Here, the local score is realised in a complete excursion: $H_n = H_n^*$.

where $U_j = \max_{0 \leqslant i \leqslant j} \sum_{k=i+1}^{j} X_k = \sum_{k=1}^{j} X_k - \min_{1 \leqslant i \leqslant j} \sum_{k=1}^{i} X_k$. $(U_k)_{k \geqslant 0}$ is the Lindley process defined recursively by

$$U_0 := 0 \quad \text{and} \quad U_k := (U_{k-1} + X_k)^+, \ k \geqslant 1.$$

Let us define the maximum on complete excursions up to time n:

$$H_n^* := H_{g_n} = \max_{0 \leqslant k \leqslant g_n} U_k, \tag{9}$$

where $g_n := \max\{k \leqslant n;\ U_k = 0\}$. The last time that achieves the maximum of U before g_n is:

$$f_n^* := \max\{k \leqslant g_n;\ U_k = H_n^*\},$$

the left end-point of the excursion straddling f_n^* is:

$$g_n^* := g_{f_n^*} = \max\{k \leqslant f_n^*\ ;\ U_k = 0\}$$

and $L_n^* := f_n^* - g_n^*$ stands for the length of the (last) segment that realises the local score on complete excursions. Obviously, the local score is realised on a complete excursion if and only if $H_n^* = H_n$. All the r.v.'s introduced above can be viewed in Fig. 7.
The approximation of the distribution of (H_n^*, L_n^*) given in Chabriac *et al.* , 2014, Theorem 2.4 is based on Donsker's theorem which tells us that the random walk $(\sum_{i=1}^{k} X_i)_{0 \leqslant k \leqslant n}$ normalised by the factor $1/\sqrt{n}$ converges in distribution as $n \to \infty$ to the Brownian motion $(B(s))_{0 \leqslant s \leqslant 1}$. Then, let us define the process $(U^n(s))_{0 \leqslant s \leqslant 1}$:

$$U^n\left(\frac{k}{n}\right) := \frac{1}{\sqrt{n}} U_k, \quad 0 \leqslant k \leqslant n; \tag{10}$$

and $U^n(t)$ is extended to $[0,1]$ through a linear interpolation. Then $(U^n(t))_{0 \leqslant t \leqslant 1}$ converges[4] weakly to the reflected Brownian motion $(U(t))_{0 \leqslant t \leqslant 1}$ started at 0 Chabriac *et al.* , 2014, Proposition 3.2:

$$U(t) := |B(t)|, \qquad t \geqslant 0.$$

Similarly to the random walk setting, let $H^*(1)$ be the "local score" evaluated over all the complete excursions from 0 up to time 1. Then $H^*(1) = H(g(t))$ where $H(s)$ is the classical local score over $[0, s]$ and $g(t)$ is the last zero of U up to time t. Proposition 2.1 and Theorem 2.4 in Chabriac *et al.* , 2014 give the density of the pair $(H^*(1), L^*(1))$ providing an approximation of the distribution of $(H_n^*/\sqrt{n}, L_n^*/n)$ as n goes to infinity.

[4]The accuracy of this convergence had been illustrated in Appendix.

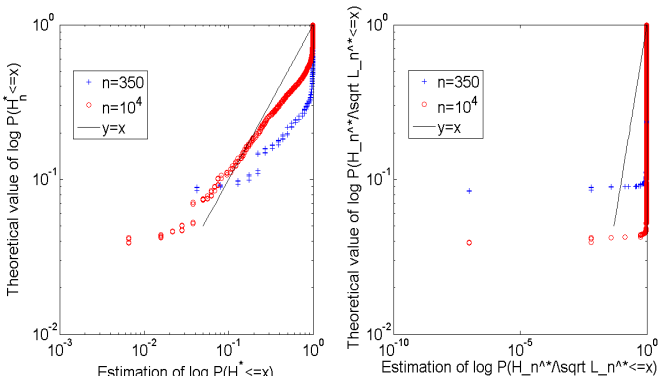

Figure 8: Log-log plot of the empirical value of $\mathbb{P}(H_n^* \leqslant h_n^*)$ (left) and $\mathbb{P}(H_n^*/\sqrt{L_n^*} \leqslant h_{n,i}^*/\sqrt{l_{n,i}^*})$ (right) versus the theoretical value of its limit derived from Chabriac *et al.* , 2014, Theorem 2.4 with $n = 350$ and $n = 10^4$.

4.2 First illustrations

First, we consider the first marginal H_n^* of the pair (H_n^*, L_n^*) and a sample of $N = 10^4$ sequences of length n with $\mathbb{E}[X] = 0$. The observed local scores on complete excursions are $\left(h_{n,i}^*\right)_{1\leqslant i\leqslant N}$. We represent in Fig. 8 (left) the log-log plot associated to the estimation of $\mathbb{P}(H_n^* \leqslant h_{n,.}^*)$ and the theoretical value of its limit derived from Chabriac *et al.* , 2014, Theorem 2.4. This picture may be linked to Fig. 2. We observe that the "distance" between the "curve" and the diagonal decreases as n grows.

4.3 Goodness-of-fit tests for the pair

A statistical test based on the two-dimensional random variables (H_n^*, L_n^*) is not easy to perform because there are no satisfactory 2D-extensions of the Kolmogorov-Smirnov tests Peacock , 1983; Fasano and Franceschini , 1987; Lopes *et al.* , 2007; Justel *et al.* , 1997. All proposals fail on at least one of the points: being independent of a reference basis on the space and/or being distribution-free.
One way to tackle the problem is to go back to dimension one using random projections Cuesta-Albertos *et al.* , 2006; considering for instance the random variable $H_n^*/\sqrt{L_n^*}$. The first reason comes from the scaling property of the Brownian motion: for any $\lambda > 0$, $\left(B(\lambda t)/\sqrt{\lambda};\ t \geqslant 0\right)$ remains a Brownian motion. Roughly speaking the normalisation is in space over the square root of time. In our context, we are interested by sequences which have a high local score and a small length. In that case, the ratio $H_n^*/\sqrt{L_n^*}$ is large. Consequently, a statistical test based on $H_n^*/\sqrt{L_n^*}$ takes into account exceptional sequences with respect to the criterion local score/length.
We consider a N-sample of (H_n^*, L_n^*) with $N = 10^4$. As done previously, we represent in Fig. 8 (right) the log-log plot associated to the empirical value of $\mathbb{P}(H_n^*/\sqrt{L_n^*} \leqslant h_{n,.}^*/\sqrt{l_{n,.}^*})$ and the theoretical value of its limit Chabriac *et al.* , 2014, Theorem 2.4. The convergence appears to be even slower than considering H_n^*. Then, we perform a classical Kolmogorov-Smirnov test to check the adjustment of $H_n^*/\sqrt{L_n^*}$ to $H^*(1)/\sqrt{L^*(1)}$ using the theoretical result in Chabriac *et al.* , 2014, Theorem 2.4. The results are summarized in Table 3. Working with $\alpha = 1\%$, the threshold is 1.62810^{-2} and the null hypothesis is always rejected as in Subsection 2.2 and cases i with Karlin approximation and ii.

5 Local score position

Here we study how H_n^* and the usual local score H_n may differ i.e. when the local score is realised on the last incomplete excursion. Let us note $p_c = \mathbb{P}(H_n = H_n^*)$. When $\mathbb{E}[X] = 0$, Theorem 1.1 in Lagnoux *et al.* , 2015 proves that $p_c \approx 30\%$: the probability that the local score is achieved at the final part of the sequence

Table 3: Kolmogorov-Smirnov distances for $H_n^*/\sqrt{L_n^*}$ and the theoretical distribution derived in Chabriac *et al.* , 2014, Theorem 2.4.

KS-stat. / n	100	350	500	1000	2000	5000	10000
$H_n^*/\sqrt{L_n^*}$	0.80	0.87	0.89	0.90	0.91	0.92	0.93

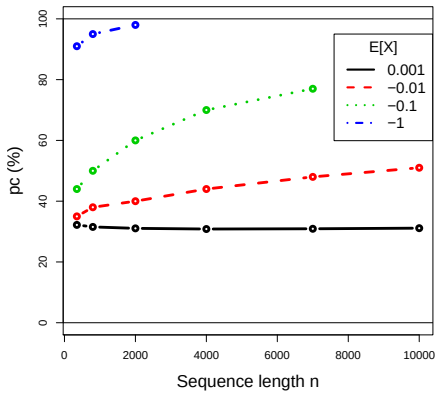

Figure 9: Evolution of p_c for different sequence lengths and $\mathbb{E}[X]$ values.

Table 4: **Percentage of sequences achieving their local score on a complete excursion (SCOP files[2,5]).**

Data base	Nb of sequences	Average length (max)	$\widehat{\mathbb{E}[X]}$	Percentage
SCOP1[5]	780	292 (1506)	-0.02	40%
SCOP2[2]	606	115 (404)	-0.23	64%

is about 70%. Note that it is also non intuitive that p_c is quite constant for the centered case even if the sequence length n increases and the number of complete excursions increases.
Here we illustrate this non-intuitive result when $\mathbb{E}[X] = 0$. We also investigate numerically the case $\mathbb{E}[X] < 0$. We consider both simulated and real sequences: SCOP1[5] and SCOP2[2] datasets using the hydrophobic scale Kyte and Doolittle , 1982. The results are represented in Table 4 and Fig. 9.

6 Conclusion

We realise goodness-of-fit tests for the distribution of $H_n/\sqrt{n}$ using simulated sequences. The numerical results show that the limit distribution proposed by Karlin *et al.* and the one given by (4) are often rejected. In the 2D-setting, we introduce a new r.v. that allows us to apply the 1D-methodology. Using biological sequences and considering both the (modified) local score H_n^* and the length of the segment that realises it permits to exhibit new significant sequences. We observe on simulated sequences, that the local score is achieved on a complete excursion with a probability that is an increasing function of n and $\mathbb{E}[X]$ (with

[5]SP scop2dom 20140205aa. *http://scop2.mrc-lmb.cam.ac.uk/downloads/*

$\mathbb{E}[X] < 0$) as well. In the biological setting of SCOP files, around half of the sequences realises its local score on a complete excursion. Such new results are really promising for biologists and will lead to go further in theoretical understanding.

Acknowledgement

Appendix

Statistical significance and p-values

In our context, a statistically significant or a biologically interesting sequence is a sequence presenting an atypical segment that cannot be attributed to chance. It is then natural to introduce the following testing procedure:

$$\begin{aligned} \mathcal{H}_0 &: \text{“The sequence is common and ordinary.”} \\ \mathcal{H}_1 &: \text{“The sequence contains an atypical segment.”} \end{aligned}$$

Here, an exceptional local score H_n or a non-common value of the pair (H_n, L_n) means that the related segment is atypical and contains relevant biological information. Significance in statistical hypothesis testing is classically measured by the p-value. It is a function of the observed sample which is defined as the probability of obtaining a result more “extreme” than what was actually observed, assuming that the null hypothesis under consideration is true. In practice, we choose a level α and reject the null hypothesis $\mathcal{H}_0$ as soon as $p \leqslant \alpha$.

Approximation validations

The goal is to give numerical values of the sample size N and the sequence length n necessary to achieve “good” approximations in a sense to be precised.
First, consider a sample $(X_i)_{1 \leqslant i \leqslant N}$ of one-dimensional r.v. X and

$$\mu_N := \sum_{i=1}^{N} \delta_{X_i} \tag{11}$$

be the associated empirical measure, where δ_x is the Dirac measure at x. According to Theorems 3.1 and 3.2 in Bobkov and Ledoux , 2014,

$$\mathbb{E}[W_1(\mu_N, \mu)] = O\left(\frac{1}{\sqrt{N}}\right), \tag{12}$$

where μ is the law of X and $W_1(\mu_N, \mu)$ stands for the Wasserstein distance between μ_N and μ. Note that (12) implies that $\mathbb{P}(W_1(\mu_N, \mu) \geqslant \alpha) = O\left(\frac{1}{\sqrt{N}}\right)$. When X is a 2D-r.v., the rate of convergence of $\mathbb{E}[W_1(\mu_N, \mu)]$ is $\log(1+N)/\sqrt{N}$ Fournier and Guillin , 2015, Theorem 1.
Second, we study the linear interpolation $(U^n(t))_{0 \leqslant t \leqslant 1}$ of $(U_k)_{0 \leqslant k \leqslant n}$ defined by (10). We perform a two-sample Kolmogorov-Smirnov test and consider three different distributions driving the r.v. X:

i a standard Gaussian r.v.;

ii a uniform distribution on [0,1], that is then standardized;

iii a r.v. on $[\![-2, 2]\!]$ with probabilities (0.075,0.2,0.45,0.2,0.075).

Then we simulated three different processes $(U^n(t))_{0 \leqslant t \leqslant 1}$ according to (10). In each case, we compute the local score associated to the i-th interpolated sequence $((U_i^n(t))_{0 \leqslant t \leqslant 1}$ and the empirical c.d.f. F_N (cf. (13)) associated with the sample. Then we perform a two sample Kolmogorov-Smirnov test as recalled below. In practice, the simulations suggest to take a sample size $N = 5000$ and a sequence with length $n = 2000$ at level 5%.

Kolmogorov-Smirnov test

Goodness-of-fit tests of Kolmogorov–Smirnov type are the most widely used to decide whether it is reasonable to assume that some one-dimensional data come from a given distribution. The problem is the following: given a N-sample $(X_i)_{1\leqslant i\leqslant N}$ of a r.v. X, can we accept that their underlying common distribution is a given F? The null hypothesis $\mathcal{H}_0$ is naturally "The sample $(X_i)_{1\leqslant i\leqslant N}$ is distributed as F". To carry out this test, Kolmogorov Kolmogorov , 1933 introduced the following distance between cumulative distribution functions:

$$D_N = \sup_x |F_N(x) - F(x)|$$

where F_N is the empirical c.d.f. associated to the sample

$$F_N(x) := \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{X_i\leqslant x\}} \tag{13}$$

and F is the cumulative distribution function we want to test the goodness-of-fit with. $\mathcal{H}_0$ is then rejected at level α if $\sqrt{N}D_N > K_{\alpha,N}$ where $K_{\alpha,N}$ is the α-quantile of $\sqrt{N}D_N$.

The Kolmogorov-Smirnov test is also used to test whether two samples share the same distribution Kolmogorov , 1941; Smirnov , 1934. In this case, the test statistic is

$$D_{N,N'} = \sup_x |F_{1,N}(x) - F_{2,N'}(x)|,$$

where $F_{1,N}$ and $F_{2,N'}$ are the empirical distribution functions of the first and second sample respectively of size N and N'. The null hypothesis is then rejected at level α if $D_{N,N'} > c(\alpha)\sqrt{\frac{N+N'}{NN'}}$. The value of $c(\alpha)$ can be computed numerically.

References

Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. Basic Local Alignment Search Tool. *JMB*, 215, 403–410, 1990.

Arratia, R and Waterman, M.-S. The Erdos-Renyi Strong Law for Pattern Matching with a Given Proportion of Mismatches. *The Annals of Probability*, 17 (3), 1157–1162, 1989.

Arratia, R and Waterman, M.-S. A phase transition for the score in matching random sequences allowing deletions. *Adv. in Applied Probabilities*, 4, 200–225, 1994.

Borodovsky, M. and Ekisheva, S. Problems and solutions in biological sequence analysis, *Cambridge University Press, Cambridge*, 2006.

Bobkov, S. and Ledoux, M. *One dimensional empirical measures, order statistics, and Kantorovich transport distances*. Preprint, 2014.

Cellier D., Charlot, F. and Mercier, S. An improved approximation for assessing the statistical significance of molecular sequence features *Jour. Appl. Prob.*, 40, 427–441, 2003.

Chabriac, C., Lagnoux, A., Mercier, S. and Vallois, P. Elements related to the largest complete excursion of a reflected Brownian motion stopped at a fixed time. Application to local score. *Stochastic Processes and their Applications*, 124(12), 2014.

Cuesta-Albertos, J.A. and Fraiman, R. and Ransford, T. Random projections and goodness-of-fit tests in infinite-dimensional spaces. *Bull. Braz. Math. Soc. (N.S.)*, 37(4):477–501,2006.

Daudin, J.-J. and Etienne, M.-P. and Vallois, P. Asymptotic behavior of the local score of independent and identically distributed random sequences. *Stochastic Process. Appl.*, 107(1):1–28, 2003.

Dembo, A. and Karlin, S. Strong Limit Theorems of Empirical Functionals for Large Exceedances of Partial Sums of i.i.d. Variables. *Ann. Probab.*, 19(4):1737–1755, 1991.

Dembo, A. and Karlin, S. Strong Limit Theorems of Empirical Distributions for Large Segmental Exceedances of Partial Sums of Markov Variables. *Ann. Probab.*, 19(4):1756–1767, 1991.

Etienne, M.-P. and Vallois, P. Approximation of the distribution of the supremum of a centered random walk. Application to the local score. *Methodology and Computing in Applied Probability*, 6:255–275, 2004.

Fasano, G. and Franceschini, A. A multidimensional version of the Kolmogorov-Smirnov test. *Mon. Not. R. ast. Soc.*, 225:155–170, 1987.

Fournier, N. and Guillin, A. On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields*, 162(3-4):707–738, 2015.

Justel, A. and Peña, D. and Zamar, R. A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statist. Probab. Lett.*, 35(3):251–259, 1997.

Karlin, S. and Altschul, S.-F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *PNAS*, 87:2264–2268, 1990.

Karlin, S. and Dembo, A. Limit distributions of maximal segmental score among Markov-dependent partial sums. *AdAP*, 24:113–140, 1992.

Kyte J. and Doolittle R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157(1):105-132, 1982.

Karlin, S. and Ost, F. Maximal length of common words among random letter sequences. *Ann. Probab.*, 16(2):53-563, 1988.

Karlin, S. Statistical signals in bioinformatics. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38):13355-13362, 2005.

Kolmogorov, A.N. Sulla determinazione empirica di una legge di distribuzione. *Giorn. dell'Instituto Ital. degli Attuari.*, 4, 83–91, 1933.

Kolmogorov, A.N. Confidence limits for an unknown distribution function. *Ann. Math. Stat.*, 12, 461–463, 1941.

Smirnov, N.V. Estimate of deviation between empirical distribution functions in two independent samples. *Bulletin Moscow Univ. 2*, 2, 3–16, 1934.

Lagnoux, A. and Mercier, S. and Vallois, P. Probability that the maximum of the reflected Brownian motion over a finite interval $[0, t]$ is achieved by its last zero before t. *Electronic Communications in Probability*, 20(62):1-9, 2015.

Lesk, A.M. An introduction to Bioinformatics. *Oxford University Press*, 2005.

Lopes, R.H.C. and Reid, I.D. and Hobson, P.R. The two-dimensional Kolmogorov-Smirnov test. *Proceedings of Science - XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research, Nikhef, Amsterdam, the Netherlands, April 23-27*, 615–627, 2007.

Mitrophanov, A.Y. and Borodovsky M. Statistical significance in biological sequence analysis. *Briefing in Bioinformatics*, 2006.

Mercier, S. and Daudin, J.J. Exact distribution for the local score of one i.i.d. random sequence. *Jour. Comp. Biol*, 8(4):373–380, 2001.

Peacock, J.A. Two-dimensional goodness-of-fit testing in astronomy. *Mon. Not. R. Ast. Soc.*, 202:615–627, 1983.

Reinert, G. and Waterman, M.S. On the length of the longest exact position match in a random sequence. *EEE/ACM Trans Comput Biol Bioinform*, 4(1):153–156, 2007.

Waterman, M.S. Introduction to Computational Biology: Maps, Sequences and Genomes. *Chapman & Hall*, 1995.

Wolfsheimer S., Herms I., Rahmann S. and Hartmann A.K. Accurate statistics for local sequence alignment with position-dependent scoring by rare-event sampling. *BMC Bioinformatics*, 12–47, 2011.

Xia L.C., Ai D., Cram J.A., Liang X., Fuhrman J.A. and Sun F. Statistical significance approximation in local trend analysis of high-throughput timeseries data using the theory of Markov chains *BMC Bioinformatics*, 16:301, 2015.