View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Scientific Publications of the University of Toulouse II Le Mirail

Exact distribution for the local score of one i.i.d. random sequence

Sabine Mercier, Jean-Jacques Daudin

▶ **To cite this version:**

Sabine Mercier, Jean-Jacques Daudin. Exact distribution for the local score of one i.i.d. random sequence. Journal of Computational Biology, Mary Ann Liebert, 2001, 8 (4), pp.373-380. <hal-00714174>

HAL Id: hal-00714174

https://hal.archives-ouvertes.fr/hal-00714174

Submitted on 12 Jul 2012

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.

Exact distribution for the local score of one i.i.d. random sequence

S. MERCIER* and J.J. DAUDIN**

*Université de Rouen, Analyse et Modèles Stochastiques
UPRES-A CNRS 6085,
76821 Mont Saint Aignan cedex, France.
Tel.: +33 2 35 14 71 00
Fax: +33 2 32 10 37 94
E-mail: Sabine.Mercier@univ-rouen.fr

**Institut National Agronomique Paris-Grigon,
Département OMIP, UMR INAPG-INRA 96021111,
16, rue Cl. Bernard,
75231 Paris cedex 05, France.
Tel.: +33 1 44 08 16 64
Fax: +33 1 44 08 16 66
E-mail: daudin@inapg.inra.fr

ABSTRACT

Let $X_1 ... X_n$ be a sequence of i.i.d. positive or negative integer valued random variables and $H_n = \max_{0 \leq i \leq j \leq n}(X_i + ... + X_j)$ the local score of the sequence. The exact distribution of H_n is obtained using a simple Markov chain. This result is applied to the scoring of DNA and protein sequences in molecular biology.

Key-words: P-value, sequence analysis, local score, Markov chain.

1. INTRODUCTION

The assessment of the statistical significance of scores of DNA and protein sequence is an important stage in the work of molecular biologists. Let $A_1...A_n$ be a nucleic or protein sequence. In order to identify interesting patterns, appropriate scoring values can be assigned to each residue. Scoring assignments for nucleotides or amino acids may arise from a variety of considerations like biochemical categorization, physical properties, or association with secondary structures, (see Kyte *et al.* 1982 and Karlin *et al.* 1990 for examples). The local score of the sequence $A_1...A_n$ according to a scoring scheme σ is defined as follow.

$$H_n = \max_{1\leq i\leq j\leq n} \left(\sum_{k=i}^{j} \sigma(A_k) \right) \ .$$

The local score is a very useful tool for biological sequence analysis in order to identify unusual sequence pattern or similarity that may reflect biological significance. It is desirable to know whether interesting patterns can arise by chance. We are therefore interested in the distribution of H_n under the null hypothesis of only random variation, so that we may judge the statistical significance of a local score of a real biological sequence. For surveys of this subject, see Waterman 1995, or Durbin *et al.* 1998.

Let $\mathbb{X} = (X_i)_{(i=1,...,n)}$ be a sequence of independent and identically distributed random variables of positive or negative integers. Let $S_k = X_1 + ... + X_k$ be the partial sums and $S_0 = 0$. The local score of $\mathbb{X}$ is defined as follows

$$H_n = \max_{0\leq i\leq j\leq n} (S_j - S_i) = \max_{0\leq i\leq j\leq n} (X_i + ... + X_j) \ . \tag{1}$$

The local score has been already studied a lot, and asymptotic approximations have been obtained but its exact distribution is unknown untill now.
Following Iglehart 1972, Karlin and Altschul 1990, Karlin and Dembo 1992, proved the next limit when $E[X] < 0$ and for X non lattice

$$\lim_{n\to+\infty} P[H_n \leq \frac{\log n}{\lambda} + x] = \exp\left(-K^* \cdot e^{-\lambda x}\right) \ , \tag{2}$$

where λ and K^* depend only on the probability distribution of the X_i. For X lattice, there is no limit for $P[H_n \leq \frac{\log n}{\lambda} + x]$ and (2) is replaced by

$$\begin{aligned} \exp\left(-K^- e^{-\lambda x}\right) &\leq \liminf_{n\to+\infty} P[H_n \leq \frac{\log n}{\lambda} + x] \\ &\leq \limsup_{n\to+\infty} P[H_n \leq \frac{\log n}{\lambda} + x] \leq \exp\left(-K^+ e^{\lambda x}\right) \ , \end{aligned} \tag{3}$$

where K^- and K^+ depend on the distribution of the X_i too.
This result applied to sequence alignment, (see Dembo *et al.* 1994(b)), is implemented in BLAST, (see Altschul *et al.* 1990), for gapless alignment between two sequences.

In this paper we give the exact distribution of the local score H_n for X lattice independently of the sign of $E[X]$. This article is organized as follows. The result and its proof is given in Section 2. Section 3 is devoted to a comparison between our exact result and the approximation used by Karlin and Altschul 1990 on the numerical examples given by these authors. Section 4 deals with the ungapped alignment problem. A conclusion is given in section 5.

2. EXACT DISTRIBUTION OF $\mathbf{H_n}$

Let T_k be the time of the k^{th} successive minimum of the process $\{S_k\}$

$$T_0 = 0 \text{ and } T_{k+1} = \inf\{i > T_k : S_i - S_{T_k} < 0\}$$

and $m(j) = \sup\{k : T_k \leq j\}$, the number of successive minimum over the time from 1 to j. The T_k may be infinite for $k \geq 1$ if $E[X] > 0$. These stopping time have already been studied by Karlin and Dembo 1992 to obtain their approximation of the distribution of the local score. See also Karlin and Taylor 1981 (Chapter 17). Let $\mathbb{U}$ be the process defined by :

$$U_j = S_j - S_{T_{m(j)}} \text{ for } j \geq 1 \text{ and } U_0 = 0$$

Lemma 1 $\mathbb{U}$ *possesses the following properties:*

1. $\forall k = 1, ..., m(n)\ U_{T_k} = 0$,
2. $\forall j > 0 \quad U_j \geq 0$,
3. $\forall j > 0\ U_j = (U_{j-1} + X_j)^+$.

Proof

1. By definition of U_j and $m(j)$.
2. If $U_j < 0$ we should have $S_j < S_{T_{m(j)}}$ which contradicts the definition of $T_{m(j)}$.
3. $U_j = S_j - S_{T_{m(j)}} = S_{j-1} + X_j - S_{T_{m(j)}}$. There are two possible cases: $T_{m(j)} = j$ or $T_{m(j)} < j$.
   - $T_{m(j)} = j$. On the one hand $U_j = 0$ (see the first property), and on the other hand

     $$U_{j-1} + X_j = S_{j-1} - S_{T_{m(j-1)}} + X_j = S_j - S_{T_{m(j-1)}} = S_{T_{m(j)}} - S_{T_{m(j-1)}} .$$

     The last term is negative by the definition of T_i. Thus $(U_{j-1}+X_j)^+ = 0$ and we get $U_j = (U_{j-1} + X_j)^+$.
   - $T_{m(j)} < j$. Then $T_{m(j-1)} = T_{m(j)}$ and $S_{T_{m(j-1)}} = S_{T_{m(j)}}$ and we have

     $$U_j = S_j - S_{T_{m(j)}} = S_{j-1} + X_j - S_{T_{m(j-1)}} = U_{j-1} + X_j .$$

     As $U_j \geq 0$ (property 2), $U_{j-1} + X_j \geq 0$. Therefore

     $$(U_{j-1} + X_j)^+ = U_{j-1} + X_j$$

     and $U_j = (U_{j-1} + X_j)^+$.

The process $\mathbb{U}$ and the local score are linked by the following lemma

Lemma 2

$$H_n = \max_{0\leq k\leq n} U_k \ . \tag{4}$$

Proof
With j held fixed and by definition of $T_{m(j)}$ we have $S_j - S_i \leq S_j - S_{T_{m(j)}}$ for all $i \leq j$, with equality for $i = T_{m(j)}$. Then

(5)
$$H_n = \max_{0\leq i\leq j\leq n}(S_j - S_i) = \max_{0\leq j\leq n}[\max_{i\leq j}(S_j - S_i)] = \max_{0\leq j\leq n}(S_j - S_{T_{m(j)}}) = \max_{0\leq j\leq n} U_j.$$

Let a be a positive integer and τ_a the dual variable of U_j:

$$\tau_a = \inf\{j \geq 1, \mathbb{U}_j \geq a\} \ .$$

Using (4), we get

$$\{\tau_a > n\} = \{\max_{k\leq n} U_k < a\} = \{H_n < a\} \ .$$

So the distribution of the local sore H_n is given by

$$P[H_n < a] = P[\tau_a > n] \ . \tag{6}$$

Let $U_j^* = U_j$ for $j < \tau_a$ and $U_j^* = a$ for $j \geq \tau_a$. The Lemma 1 implies that the process $\mathbb{U}^*$ is a Markov chain whose states are $\{0, 1, ..., a\}$. Let $\Pi = (\Pi_{h\ell})_{0\leq h,\ell\leq a}$ be the probability matrix transition of $\mathbb{U}^*$:

$$\Pi_{h,\ell} = P[U_n^* = \ell / U_{n-1}^* = h] \text{ for } 0 \leq h, \ell \leq a \ .$$

By construction, a is an absorbing state, thus we have

$$\Pi_{a,a} = 1 \text{ and } \Pi_{a,\ell} = 0 \text{ for } \ell = 0, ..., a-1 \ . \tag{7}$$

Moreover,

$$\begin{aligned}
\Pi_{h,a} &= P[U_n^* = a / U_{n-1}^* = h] \\
&= P[U_n \geq a / U_{n-1} = h] \\
&= P[(U_{n-1} + X_n)^+ \geq a / U_{n-1} = h] \\
&= P[U_{n-1} + X_n \geq a / U_{n-1} = h]
\end{aligned}$$

$$= \quad P[X_n \geq a - h / U_{n-1} = h] \ .$$

As the X_k are iid and U_{n-1} only depends on the $(X_k, k \leq n-1)$, we obtain

$$\Pi_{h,a} = P[X_n \geq a - h] = P[X_1 \geq a - h] \text{ for } h = 0, ..., a-1 \ . \tag{8}$$

For the state 0, we have

$$\Pi_{h,0} = P[X_1 \leq -h] \text{ pour } h = 0, ..., a-1 \ , \tag{9}$$

and for the general transition from h, $h = 1, ..., a-1$ to ℓ, $\ell = 1, ..., a-1$, we obtain

$$\Pi_{h,\ell} = P[X_n = \ell - h] = P[X_1 = \ell - h] \ . \tag{10}$$

Let $f(k) = P[X_i \leq k]$ and $p(k) = P[X_i = k]$ for $0 \leq h, \ell \leq a$. According to (7), (8), (9) and (10), Π is given by

$$\Pi = \left(\begin{array}{c|ccc|c} f(0) & p(1) & & p(a-1) & 1 - f(a-1) \\ \vdots & & \vdots & & \vdots \\ f(-h) & \dots & p(\ell - h) & & 1 - f(a-h-1) \\ \vdots & & & & \vdots \\ f(1-a) & p(2-a) & & p(0) & 1 - f(0) \\ \hline 0 & 0 & \dots & 0 & 1 \end{array} \right) \ .$$

The distribution of U_n^*, denoted by $P_n = (P[U_n^* = 0], ..., P[U_n^* = a])$ is given by an elementary result on Markov chains, (see Freedman 1971).

$$P_n = P_0 \Pi^n \text{ with } P_0 = (1, 0, 0, ..., 0) \ . \tag{11}$$

Then

$$P[\tau_a > n] = 1 - P[\tau_a \leq n] = 1 - P[U_n^* = a] \ . \tag{12}$$

Taking into account (11), (12) and (6), we obtain the following theorem

Theorem 1 *The c.d.f. of the local score is given by*

$$P[H_n < a] = 1 - P_n(a) = 1 - P_0 \Pi^n P_a' \text{ with } P_a = (0, 0, 0, ..., 1) \ . \tag{13}$$

3. COMPARISON BETWEEN THE EXACT AND THE ASYMPTOTIC RESULTS ON EXAMPLES

Karlin and Altschul 1990 ([KA]) have presented representative examples of maximal scoring segments of protein sequences with scores based on charge, hydropathy, cysteine clusters and amino acid similarity. We use the same examples and give the exact result in comparison with the approximation given by [KA]. The following table (*Table 1*) summarize the results. The indexing of the examples is the one used in [KA]. Even if the relative error may be large in some case, one can see that the [KA] approximation is correct and does not lead to dramatic errors of decision. However the relative error is high in two cases (Examples a) ii) and c) iii)) where a is the highest among all examples. Generally the relative error is positive, a fact which is in agreement with [KA] remark about the use of the constant K^+ which leads to conservative P-values. However the error is negative in two cases which proves that the [KA] approximation is not always conservative.
This formula is good for the tail of the distribution. One can use two different ways in order to judge of its accuracy.
First, when the P-value obtained is small enough ($< 10^{-3}$), the asymptotic bounds (see (3)) can be use with confidence.
The second way stands on x, where $x = a - \log(n)/\lambda$. The ratio $H_n/\log(n)$ converges almost surely to $1/\lambda$ where λ depends on the distribution of X (see Dembo *et al.* 1994(a), Arratia and Waterman *et al.* 1985), and x corresponds to the deviation between the calculated local score a and the almost surely "limit". Larger is x, better are the conditions. For a safe use of (2), empirical tests lead to consider that $x > 3$ is a sufficient condition if n is enough high and $E[X] << 0$.
The two negative relative errors in the Table 1, c)(i) and e), are linked to the too large P-value : $3,56.10^{-2}$ for c)(i) and 9.5^{-1} for e). We are not in the tail of the distribution and the upper bound given by Karlin *et al.* can be less than the exact P-value. The deviation x equals to 2,39 for c)(i) and -14,88 for e). However note that the case d)(ii) is an example of high P-value but (3) is correct.

These numerical examples are only illustrations about the quality of the [KA] approximation. A more complete numerical study must be made in order to give more general results and to produce advices about its safe use specially with short sequences.

Table 1

4. UNGAPPED ALIGNMENT PROBLEM

In order to compare two biological sequences $\mathbb{A}$ and $\mathbb{B}$, scores can be assigned to couple of amino acids or nucleotides reflecting the similarity between the two components : $\sigma(A_i, B_j)$. We can define the local score of two sequences as

$$H_n = \max_{1 \le i \le j \le n} \left(\sum_{k=i}^{j} \sigma(A_i, B_j) \right) .$$

The statistical significance of this problem is already solve by the case of one sequence (see previous sections). But this approach of evolution between two sequences implies that they have the same length, and does not consider shifted couple of segments. It does not consider as well the fact that components can be inserted or deleted, what we call indel, or gap.

For the ungapped alignment problem, but considering shifts, see R Mott and R. Tribe 1999, we can use the exact probabilities in the following way. Let us consider two sequences of respective size m and n with $m \ge n$. We can align the first sequence with $m+n-1$ sequences obtained by slipping the second sequence (see Figure 1).

The local score is then defined as

Figure 1

$$H_{n,m} = \max_{\substack{0 \le \ell \le \min(n,m)-1, \\ 1 \le i \le n-\ell \\ 1 \le j \le m-\ell}} \sum_{k=0}^{\ell} \sigma(A_{i+k}, B_{j+k})$$

As most authors we do not take into account the dependence between them. Their weak dependence is ignored. Therefore the probability that the local score does not exceed a is approximately the product of the corresponding probabilities for each of the $m+n-1$ couple of sequences.

The length of each couple is equal to the common length, between 1 and n. There are 2 couples of length i for $i = 1, ..., n-1$ (see the first case and the third cases of Figure 1) and $m-n+1$ couples of length n (see the middle case of the figure). Therefore the probability of exceeding a is approximatively equal to

$$P(H_{n,m} \ge a) = 1 - \prod_{i=1,...,n-1} P(H_i < a)^2 P(H_n < a)^{m-n+1}$$

Note that we are able to take into account the size of each couple of sequences, which is not possible when an asymptotic method is used.

5. CONCLUSIONS

Our exact method is very easy to implement. A MATLAB computer program (available upon request) contains only thirty lines of code. It is not computationally intensive if one possesses a good subroutine for computing n^{th}-powers of a matrix. For example the computation of the P-value a 100,000-long sequence with $a = 100$ is obtained in less than one second using a Pentium-II PC. Note that the [KA] approximation also needs a computer program in order to obtain the constant K^*. The imprecision due to computation errors can be controlled using the fact that the sum of the probability of the states of the Markov chain must be equal to one. The first wrong finger in the above example is the 12th and we think that the computation errors are small for P-values greater than 10^{-10}. In our mind it would be useful to implement the computation of the exact P-values in any scoring sequence software. It would give exact P-values even if n is small and if $E[X]$ is near from zero, which is not the case for the asymptotic expression (3).
Some work remains to deal with the alignment scoring problem with gap and with the case of Markovian dependent sites.

REFERENCES

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

Arratia, R. and Waterman, M.-S. 1985. Critical phenomena in sequence matching. *Ann. Prob.* 13, 1236–1249.

Dembo, A., Karlin, S. and Zeitouni, O. 1994(a). Critical phenomena for sequence matching with scoring. *Ann. Prob.* 22(4), 1993–2021.

Dembo, A., Karlin, S., and Zeitouni, O. 1994(b). Limit distribution of maximal non-aligned two-sequence segmental score. *Ann. Prob.*, 22, 2022–2039.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. 1998. *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, Cambridge, U.K.

Freedman, D. 1971. *Markov chains.* Holden-Day.

Iglehart, D.-L. 1972. Extremes values in the GI/G/1 queues. *Ann. Math. Statist.*, 43(2), 627–635.

Kyte, J., and Doolittle, R.F. 1982. A simple method for displaying the hydrophatic character of a protein. *J. Mol. Biol.*, 157, 105–132.

Karlin, S., and Altschul, S.F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA*, 87, 2264–2268.

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

Karlin, S., and Taylor, H.M. 1981. *A second course in stochastic processes.* Academic Press.

Mott, R., and Tribe, R. 1999. Approximate Statistics of Gapped Alignments. *J. Comput. Biol.* 6, 91–112.

Waterman, M.S. 1995. *Introduction to Computational Biology.* Chapman and Hall, London.

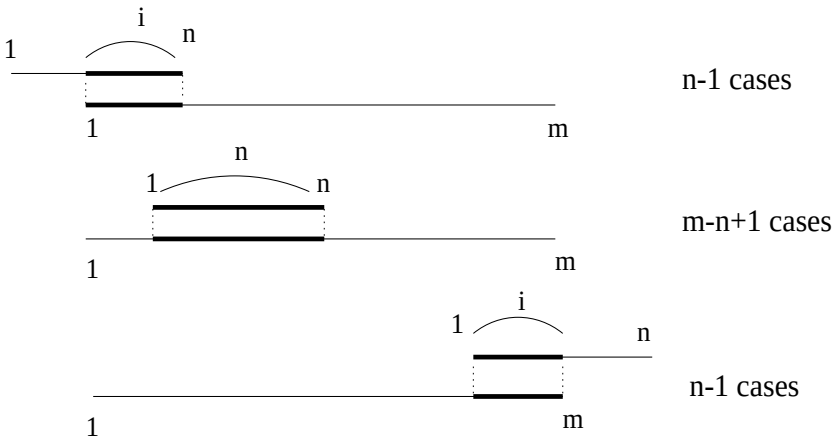

Figure 1: There is a total of $m + n - 1$ different positions between the two sequences that can be regrouped in three main cases.

Index	Sequence and frequencies	n	a	KA P-value	Exact P-value	Relative error
a) Mixed charge X: 2 for DEKRH -1 for others	(i) Human 67-kDa keratin cytoskeletal type II. f(s=2)=20,1%	643	21	$< 8.10^{-3}$	$6,99.10^{-3}$	14%
	(ii) Human c-junc nuclear transcription factor f(s=2)=20,2%	331	29	$< 2.10^{-4}$	$1,04.10^{-4}$	92%
b) Acidic charge X: 2 for DE, -2 for KR -1 for others	Drosophila zeste protein nuclear transcription factor f(s=2)=9,4%; f(s=-2)=8%	575	11	$3,7.10^{-3}$	$2,75.10^{-3}$	34%
c) Basic charge X: 2 for KRH, -2 for DE, -1 for others	(i) Drosophila sodium ion channel protein f(s=2)=10,2%;f(s=-2)=9.9%	1320	10	$3,4.10^{-2}$	$3,56.10^{-2}$	-5%
	(ii) Zeste protein f(s=2)=11,0%; f(s=-2)=9.4%	575	12	4.10^{-3}	$3,97.10^{-3}$	0,7%
	(iii) U1 70-kDa small nuclear ribonucleoprotein f(s=2)=25,1%; f(s=-2)=18,5%	614	37	$< 2.10^{-4}$	$9,22.10^{-5}$	117%
d) Hydrophobic X: 1 for ILVFMCA, -1 for GSTWYP, -2 for others	(i) Drosophila engrailed f(s=1)=31,7%; f(s=-1)=31,9%	552	17	$1,8.10^{-5}$	$1,73.10^{-5}$	4%
	(ii) Human c-mas, angiotensin receptor protein f(s=1)=46,8%; f(s=-1)=29,8%	325	15	8.10^{-2}	$7,47.10^{-2}$	7%
	(iii) Cystic fibrosis (CF) gene product f(s=1)=41,6%; f(s=-1)=26,8%	1480	21	10^{-3}	$9,31.10^{-4}$	7%
e) Cysteine Cluster X: 5 for C -1 else	Human thrombomodulin f(s=5)=8,5%	575	12	$9,1.10^{-1}$	$9,50.10^{-1}$	-4%

Table 1: Comparison of exact and asymptotic values of $P(H_n \geq a)$ on KA (Karlin and Altschul) examples: a) research for high-scoring mixed charge segments (of basic and acidic residues); b) research for high-scoring acidic charge segment; c) high-scoring basic charge segment; d) Strong hydrophobic segments; e) Cysteine cluster