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 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-Grignon,
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 $X_1 \ldots X_n$ be a sequence of i.i.d. positive or negative integer valued random variables and $H_n = \max_{0 \leq k \leq n} (X_1 + \ldots + X_k)$ 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.
The assessment of the statistical significance of scores of DNA and protein sequence is an important stage in the work of molecular biologists. Let $A_1 \ldots A_n$ 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 $A_1 \ldots A_n$ according to a scoring scheme $\sigma$ 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 $X = (X_i)_{i=1, \ldots, n}$ be a sequence of independent and identically distributed random variables of positive or negative integers. Let $S_k = X_1 + \ldots + X_k$ be the partial sums and $S_0 = 0$. The local score of $X$ is defined as follows

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

The local score has been already studied a lot, and asymptotic approximations have been obtained but its exact distribution is unknown until 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

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

where $\lambda$ 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

$$\exp (-K^- e^{-\lambda x}) \leq \lim_{n \to +\infty} \inf P[H_n \leq \frac{\log n}{\lambda} + x] \leq \lim_{n \to +\infty} \sup P[H_n \leq \frac{\log n}{\lambda} + x] \leq \exp (-K^+ e^{\lambda x}),$$

$$3$$
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 $H_n$

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 $U$ 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 $U$ possesses the following properties:

1. $\forall k = 1, \ldots, m(n)$ $U_{T_k} = 0$,
2. $\forall j > 0$ $U_j \geq 0$,
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 $U_j = (U_{j-1} + X_j)^+$.\n
5
The process \( U \) and the local score are linked by the following lemma

**Lemma 2**

\[
H_n = \max_{0 \leq k \leq n} U_k .
\]

**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

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

Let \( a \) be a positive integer and \( \tau_a \) the dual variable of \( U_j \): 

\[
\tau_a = \inf \{ j \geq 1, U_j \geq a \} .
\]

Using (4), we get

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

So the distribution of the local score \( H_n \) is given by

\[
P[H_n < a] = P[\tau_a > n] .
\]

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 \( U^* \) 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 \( U^* \):

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

By construction, \( a \) is an absorbing state, thus we have

\[
\Pi_{a,a} = 1 \text{ and } \Pi_{a,\ell} = 0 \text{ for } \ell = 0, ..., a-1 .
\]

Moreover,

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


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

\begin{equation}
\Pi_{h,a} = P[X_n \geq a-h] = P[X_1 \geq a-h] \text{ for } h = 0, ..., a-1.
\end{equation}

For the state 0, we have

\begin{equation}
\Pi_{h,0} = P[X_1 \leq -h] \text{ for } h = 0, ..., a-1,
\end{equation}

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

\begin{equation}
\Pi_{h,\ell} = P[X_n = \ell - h] = P[X_1 = \ell - h].
\end{equation}

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), $\Pi$ is given by

\[ \Pi = \begin{pmatrix}
  f(0) & p(1) & p(a-1) & 1 - f(a-1) \\
  \vdots & \vdots & \vdots & \vdots \\
  f(-h) & \ldots & p(\ell - h) & 1 - f(a - h - 1) \\
  \vdots & \vdots & \vdots & \vdots \\
  f(1-a) & p(2-a) & p(0) & 1 - f(0) \\
  0 & 0 & \ldots & 0 & 1
\end{pmatrix}. \]

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).

\begin{equation}
P_n = P_0 \Pi^n \text{ with } P_0 = (1, 0, 0, ..., 0).
\end{equation}

Then

\begin{equation}
P[\tau_a > n] = 1 - P[\tau_a \leq n] = 1 - P[U_n^* = a].
\end{equation}

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

\begin{equation}
P[H_n < a] = 1 - P_n(a) = 1 - P_0 \Pi^n P'_a \text{ with } P_a = (0, 0, 0, ..., 1).
\end{equation}
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 $\lambda$ 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 : $3.56 \times 10^{-2}$ 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 \( A \) and \( B \), 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

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

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 \( m \geq n \). 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

\[
H_{n,m} = \max_{\substack{0 \leq \ell \leq \min(n,m) - 1, \ 1 \leq i \leq n - \ell, \ 1 \leq j \leq m - \ell}} \left( \sum_{k=0}^{\ell} \sigma(A_{i+k}, B_{j+k}) \right)
\]

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

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

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


Figure 1: There is a total of $m + n - 1$ different positions between the two sequences that can be regrouped in three main cases.
<table>
<thead>
<tr>
<th>Index</th>
<th>Sequence and frequencies</th>
<th>n</th>
<th>a</th>
<th>KA P-value</th>
<th>Exact P-value</th>
<th>Relative error</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) Mixed charge</td>
<td>Human 67-kDa keratin cytoskeletal type II. f(s=2)=20, 1%</td>
<td>643</td>
<td>21</td>
<td>&lt; $8 \times 10^{-3}$</td>
<td>6, 99 $\times 10^{-3}$</td>
<td>14%</td>
</tr>
<tr>
<td></td>
<td>f(s=2)=20, 1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X: 2 for DEKRH</td>
<td>Human c-junc nuclear transcription factor f(s=2)=20, 2%</td>
<td>331</td>
<td>29</td>
<td>&lt; $2 \times 10^{-4}$</td>
<td>1, 04 $\times 10^{-4}$</td>
<td>92%</td>
</tr>
<tr>
<td></td>
<td>Human c-junc nuclear transcription factor f(s=2)=9, 4%; f(s=2)=8%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1 for others</td>
<td>Drosophila zest protein nuclear transcription factor f(s=2)=9, 4%</td>
<td>575</td>
<td>11</td>
<td>3, 7 $\times 10^{-3}$</td>
<td>2, 75 $\times 10^{-3}$</td>
<td>34%</td>
</tr>
<tr>
<td></td>
<td>Drosophila zest protein nuclear transcription factor f(s=2)=9, 4%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b) Acidic charge</td>
<td>Human c-mas, angiotensin receptor protein f(s=1)=31, 7%; f(s=1)=31, 9%</td>
<td>552</td>
<td>17</td>
<td>1, 8 $\times 10^{-5}$</td>
<td>1, 73 $\times 10^{-5}$</td>
<td>4%</td>
</tr>
<tr>
<td></td>
<td>Human c-mas, angiotensin receptor protein f(s=1)=46, 8%; f(s=1)=29, 8%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X: 2 for DE, -2 for KR</td>
<td>Human c-mas, angiotensin receptor protein f(s=1)=46, 8%; f(s=1)=29, 8%</td>
<td>325</td>
<td>15</td>
<td>8, 1 $\times 10^{-2}$</td>
<td>7, 47 $\times 10^{-2}$</td>
<td>7%</td>
</tr>
<tr>
<td></td>
<td>Cystic fibrosis (CF) gene product f(s=1)=41, 6%; f(s=1)=26, 8%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1 for others</td>
<td>Cystic fibrosis (CF) gene product f(s=1)=41, 6%; f(s=1)=26, 8%</td>
<td>1480</td>
<td>21</td>
<td>10 $\times 10^{-3}$</td>
<td>9, 31 $\times 10^{-4}$</td>
<td>7%</td>
</tr>
<tr>
<td>c) Basic charge</td>
<td>Drosophila sodium ion channel protein f(s=2)=10, 2%; f(s=2)=9, 9%</td>
<td>1320</td>
<td>10</td>
<td>3, 4 $\times 10^{-2}$</td>
<td>3, 56 $\times 10^{-2}$</td>
<td>-5%</td>
</tr>
<tr>
<td></td>
<td>Drosophila sodium ion channel protein f(s=2)=10, 2%; f(s=2)=9, 9%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X: 2 for KRH, -2 for DE, -1 for others</td>
<td>U1 70-kDa small nuclear ribonucleoprotein f(s=2)=25, 1%; f(s=2)=18, 5%</td>
<td>575</td>
<td>12</td>
<td>4, 10 $\times 10^{-3}$</td>
<td>3, 97 $\times 10^{-3}$</td>
<td>0, 7%</td>
</tr>
<tr>
<td></td>
<td>U1 70-kDa small nuclear ribonucleoprotein f(s=2)=25, 1%; f(s=2)=18, 5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d) Hydrophobic</td>
<td>Drosophila engrailed f(s=1)=31, 7%; f(s=1)=31, 9%</td>
<td>552</td>
<td>17</td>
<td>1, 8 $\times 10^{-5}$</td>
<td>1, 73 $\times 10^{-5}$</td>
<td>4%</td>
</tr>
<tr>
<td></td>
<td>Drosophila engrailed f(s=1)=31, 7%; f(s=1)=31, 9%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X: 1 for ILVFMCA, -1 for GSTWYP, -2 for others</td>
<td>Human thrombomodulin f(s=5)=8, 5%</td>
<td>575</td>
<td>12</td>
<td>9, 1 $\times 10^{-1}$</td>
<td>9, 50 $\times 10^{-1}$</td>
<td>-4%</td>
</tr>
<tr>
<td></td>
<td>Human thrombomodulin f(s=5)=8, 5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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