Back to the local score in the logarithmic case: a direct and simple proof

J.-N. BACRO*, J.-J. DAUDIN*, S. MERCIER**, S. ROBIN*

* UMR INAPG/INRA 518, 16, rue Cl. Bernard, 75231, Paris Cedex05

** Université de Toulouse II

7th January 2002

Short title for running head:"local score in the logarithmic case"

Abstract

Let $X_1, \ldots, X_n$ be a sequence of i.i.d. integer valued random variables and H_n the local score of the sequence. A recent result shows that H_n is actually the maximum of an integer valued Lindley process. Therefore known results about the asymptotic distribution of the maximum of a weakly dependent process, give readily the expected result about the asymptotic behavior of the local score in the logarithmic case, with a simple way for computing the needed constants. Genomic sequence scoring is one of the most important applications of the local score.

An example of an application of the local score on protein sequences is therefore given in the paper.

Key-words and phrases. Extremal index, Genomic sequence, Lindley process, Local score, Markov chain,

AMS 1991 subject classifications. Primary 60G70 60F10; secondary 62P10.

Acknowledgement : *We acknowledge an anonymous referee for helpful comments and remarks.*

1 Introduction

Let $\mathbb{X} = X_1, \ldots, X_n$ be a sequence of independent and identically distributed random variables of integers with $E[X_1] < 0$ and $P(X_1 > 0) > 0$. 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

$$H_n = \max_{0 \le i \le j \le n} (S_j - S_i) = \max_{0 \le i \le j \le n} (0, X_i + ... + X_j). \tag{1.1}$$

It is well known that $H_n = O(\log n)$ when $n \to \infty$, so that the case $E[X_1] < 0$ is called "the logarithmic case" (see Dembo *et al.* (1994), Arratia and Waterman (1992)). Following Iglehart (1972), Karlin *et al.* (1990) proved the following limit for X_1 non lattice:

$$\lim_{n\to\infty} P[H_n \leq \frac{\log n}{\lambda} + x] = \exp\left(-Ke^{-\lambda x}\right), \tag{1.2}$$

where λ and K depend only on the probability distribution of X_1.

In the lattice case Karlin *et al.* (1990) give the following results:

$$\text{(1.3)} \exp\left(-K^+e^{-\lambda x}\right) \leq \liminf_{n\to\infty} P[H_n \leq \frac{\log n}{\lambda} + x],$$

$$\text{(1.4)} \qquad \limsup_{n\to\infty} P[H_n \leq \frac{\log n}{\lambda} + x] \leq \exp\left(-K^-e^{-\lambda x}\right),$$

where

$$K^- = \frac{\lambda\alpha K}{e^{\lambda\alpha} - 1}, \quad K^+ = \frac{\lambda\alpha K}{1 - e^{-\lambda\alpha}},$$

and α is the span of X_1.

The goal of this paper is to propose a new, simple and direct proof for results as (1.2), (1.3) and (1.4). Moreover, we obtain new bounds for the lattice case which are easier to compute and sharper than those in (1.3) and (1.4).

The results of Karlin *et al.*(1990) were obtained using the fact that H_n is a regenerative process. Therefore a natural way consists in decomposing the process in independent subprocesses. The next step is to obtain the law of the maximum of each subprocess and the fact that the subprocesses are i.i.d. leads to the conclusion. This natural approach has two drawbacks: first the number of regenerative points is not known and must be approximated and secondly the law of the maximum of a subprocess is more complicated than the law of the maximum of H_n itself.

Our approach is completely different, more direct and stands on three different results. The first one is a recent result on the local score: Daudin and Mercier (1999, 2001) have proved that the local score H_n is the maximum of a Lindley process

$$H_n = \max_{0 \leq j \leq n} U_j, \tag{1.5}$$

where $(U_j)_{j \geq 0}$ is the following integer valued Lindley process:

$$U_j = (U_{j-1} + X_j)^+ \quad , \quad U_0 = 0.$$

The second step deals with a work about the asymptotic distribution of the maximum of a Lindley process and more generally about the maximum of weakly dependent processes (see for example Rootzen 1988). The case of a non lattice Lindley process is completely known (Rootzen 1988). The only work we have to do is to adapt these results to the case of a lattice valued process which is not explicitly resolved in the literature. The asymptotic distribution given by Rootzen is linked with the stationary distribution of the Lindley process $(U_j)_{j \geq 0}$ (see Theorem 2 and Theorem 4) whose c.d.f noted F is already given by Mercier (1999) for the lattice case. This third result given in Theorem 3, allows us to compute a constant θ, and two bounds

$b_1(F,\theta,x)$ and $b_2(F,\theta,x)$ such that

$$b_1(F,\theta,x) \leq \liminf_{n\to\infty} P[H_n \leq \frac{\log n}{\lambda}+x] \leq \limsup_{n\to\infty} P[H_n \leq \frac{\log n}{\lambda}+x] \leq b_2(F,\theta,x).$$

These bounds can be controlled using the exact distribution of the local score H_n given by Daudin and Mercier (1999).

This article is organized as follows. The general theory about the maximum of weakly dependent processes and the results about the maximum of a non lattice Lindley process are briefly recalled in Section 2. Section 3 is devoted to the case of the lattice valued Lindley process. Simple ways for computing the bounds are proposed and a numerical comparison with (1.3) and (1.4) and the exact values is made in Section 4 on a protein scoring example.

2 General theory about the maximum of a process Application to the Lindley's Process

There is a lot of work about the maxima and exceedances of a stationary process. The work of Leadbetter (1983) summarizes the results obtained in the i.i.d. case and gives more insight about the case of a dependent process. Rootzen (1988) and Perfect (1994) have explicitly given results about the maxima and exceedances of stationary Markov chains and applied them to the continuous Lindley process.

In this section, we recall the results of Rootzen (1988) for the general continuous case and for the Lindley process in particular.

Theorem 1 (Theorem 4.1, Rootzen (1988)) *Let us consider a stationary process Z_t with marginal c.d.f. F verifying the three following conditions:*

(C1) $\forall \tau > 0, \exists u_n(\tau) : \lim_{n\to\infty} n\{1 - F[u_n(\tau)]\} = \tau$

(C2) $D(u_n(\tau))$ *and* $\Delta(u_n(\tau))$ *are verified for all* τ *(see below)*

(C3) $\exists \theta \in (0,1]\,, \tau > 0$ *such that*

$$\lim_{\varepsilon\to 0}\{\limsup_{n\to\infty} \left| P[\max(Z_1, ...Z_{[n\varepsilon]}) \leq u_n(\tau) \mid Z_0 > u_n(\tau)] - \theta \right|\} = 0$$

Then:

(R1) $\forall \tau > 0$, $\lim_{n\to\infty} P\left[\max_{0\leq t\leq n} \mathbb{Z}_t \leq u_n(\tau)\right] = e^{-\theta\tau}$

(R2) $N(n) = \#\{t \leq n : Z_t > u_n(\tau)\}$, *the exceedances process, is a compound Poisson process.*

The condition *(C2)* is a mixing condition which is quite technical (see Rootzen 1988). *(C2)* is not explicitly given here, but it is well known that such a condition is satisfied by any regenerative process with finite return time and in particular by the Lindley's process (Rootzen 1988).

Applying Theorem 1 to the case of the Lindley's process, we have:

Theorem 2 (Rootzen (1988)) *Let* $(X_n)_{n\geq 0}$ *be an i.i.d. sequence of random variables and* $U_{n+1} = (U_n + X_{n+1})^+$ *be the associated Lindley process. Assume the following conditions*

(C1) $E[X_1] < 0$ *and* $P(X_1 > 0) > 0$

(C2) X_1 *is non lattice*

(C3) $\exists \lambda > 0$ *such that* $E[e^{\lambda X_1}] = 1$ *and* $E[|X_1|\, e^{\lambda X_1}] < \infty$

then

(R1) $(U_n)_{n\geq 0}$ *has a stationary distribution with c.d.f.* F_U *whose tail is such that*

$$1 - F_U(u) \overset{u\to\infty}{\sim} Ce^{-\lambda u}$$

(R2) $(U_n)_{n\geq 0}$ *possesses an extremal index* θ*, given by* $\theta = P[V + U' + X_1 \leq 0]$ *where* V *has an exponential p.d.f. with parameter* λ *and* U' *has the c.d.f.* F_U

(R3) $\lim_{n\to\infty} P[\max_{0\leq j\leq n} U_j \leq \dfrac{\log n}{\lambda} + x] = \exp\left(-C\theta e^{-\lambda x}\right)$

Note that $V + U' + X_1$ is a three fold convolution between independent variables and that we have

$$U' = \sup(0, X_1, X_1 + X_2, ...).$$

The notation U' refers to the fact that its distribution is exactly the stationary distribution F_U of the Lindley process $(U_j)_{(j>0)}$ (see Remark 1). However U' is **not** the random variable corresponding to the studied Lindley process. V is the excess of the stationary Lindley process U whose c.d.f is defined by

$$P(V \leq v) = \lim_{u \to \infty} P(U \leq v + u \mid U > u).$$

As the tail of U is exponential, the distribution of V is also exponential with the same parameter. However U and V do not have the same distribution except in the very peculiar case when U has itself an exponential distribution. Daudin and Mercier (1999) have proved that $H_n = \max_{0 \leq j \leq n} U_j$, where the Lindley process U_j only depends on the p.d.f. of X_1. Therefore (1.2) is directly proved in the non-lattice case by Theorem 2. The constant λ is the same as in (1.2) and the constant K of Karlin *et al.* (1990) is related to the extremal index of U_j: we have $K = C\theta$. For computational purpose, one only needs a good evaluation of C and of F_U in order to compute θ.

Note : *in section 4, we shall be interested in bounded random variables X_1. In this case (C1) implies (C2) (see Dembo and Karlin (1991)).*

3 The maximum of a lattice Lindley process

The case of the discrete valued process is more difficult because of the discontinuities of the distribution function: it is well known for example, that the geometric and Poisson distributions do not possess a limit for the distribution of extremes values, in contrast to most of the continuous distributions. In the context of queuing theory, Iglehart (1972) noted that it was possible to obtain a limit distribution for the maximum of waiting times, but that no such limit exists for the size of the queue. Anderson (1970) has given some results for some discrete distributions and we shall go further on his line. More recently, McCormick and Park (1992) and McCormick and Sun (1993) have obtained asymptotic bounds for some class of discrete distribution such as autoregressive negative binomial process. However, their results can not be applied in the case of the Lindley process because their mixing conditions are too stringent.

Let X_1 take its values in expression $\mathbb{Z}$ and assume that the g.c.d. of the absolute values of X_1 is equal to one (if it is not the case, we can divide each value by the g.c.d. without loss of generality). Therefore, one can see by using the Bezout theorem that the span of $(U_j)_{j\geq 0}$ (introduced in expression 1.3 and 1.4) is equal to one, so there is no need for using a general span possibly greater than one as in Karlin *et al.* (1990). For if $x_1 < 0$ and $x_2 > 0$

are two values of X_1, prime with each other, it is possible to find two integers a_1 and a_2 such that $a_1x_1 + a_2x_2 = 1$. Therefore if $X_1 = X_2 = \ldots = X_{a_1} = x_1$ and $X_{a_1+1} = X_{a_1+2} = \ldots = X_{a_1+a_2} = x_2$ and if $U_j > -a_1x_1$,we obtain $U_{j+a_1+a_2} = U_j + 1$. Therefore, one can see that the minimum increase of $(U_j)_{j\geq 0}$ is 1 for U_j sufficiently high.

It is obvious that $(U_j)_{j\geq 0}$ is a regenerative process with the state $\{0\}$ as regenerative point, and $E[T] < \infty$ if $E[X_1] < 0$ where T is the time between two successive returns to $\{0\}$. The Lindley process is an infinite Markov chain whose transition probability matrix depends only of the p.d.f. of X_1. Mercier (1999) has obtained the following results about the stationary distribution of $(U_j)_{j\geq 0}$.

Theorem 3 *$(U_j)_{j\geq 0}$ possesses a stationary distribution P_U (with c.d.f. F_U) whose probabilities satisfy a linear recurrence relation of order $d \equiv \max(X_1) - \min(X_1)$. Therefore*

(R1) $\exists(\delta_i, R_i)_{(i=1,\ldots,d)}, P_U(k) = \sum_{i=1}^{d} \delta_i R_i^k$

(R2) Let $R = \max_{i=1,\ldots,d}(R_i \in \mathbb{R}, R_i < 1)$ *and δ its associated coefficient. We have*

$$1 - F_U(x) \overset{x\to\infty}{\sim} [\delta/(1-R)]\, R^{\lfloor x \rfloor + 1}$$

where $\lfloor x \rfloor$ is the integer part of x.

Proof Let Π be the infinite transition matrix of the Lindley Markov chain. It is well known (Asmussen (1987), Borovkov (1976)) that this Markov chain possesses a stationary distribution μ which satisfies $\mu\Pi = \mu$. The non null terms of the columns of Π are composed with the same finite vector of size d (except the first $|\min(X_1)|$ ones). This implies that μ satisfies a linear recurrence of order $d = \max(X_1) - \min(X_1)$. Standard theory about linear recurrence equations gives the result.

Remark 1 *The invariant distribution of the Lindley process is also the distribution of the maximum of the partial sums $S_k = X_1 + ... + X_k$ and $S_0 = 0$ (see Mercier (1999) for example). It is easy to prove that R is the only positive root different to 1 of the polynomial linked to the linear recurrence relation, and that $\lambda = -\log(R)$. So (R2) is in agreement with the result of Karlin and Dembo (1992) on the distribution of the maximum of the partial sums.*

Remark 2 *Note that $\mu = \lim_{n \longrightarrow \infty} (1, 0...0)\Pi^n$. Therefore, the stationary distribution may easily be numerically approximated by*

$$\mu \approx (1, 0...0)\Pi_t^n \tag{3.1}$$

using a North-West truncation Π_t of the infinite transition matrix Π and n sufficiently high. It is also possible to use the first eigenvector of Π_t, but there *may be some numerical stability problems when t is high and the numerical method for the diagonalization is not well suited. Note that we are interested in the tail probabilities of μ.*

We now express θ:

Theorem 4 *Let $\theta = P(V + U' + X_1 \leq 0)$ where $V - 1$ has a geometrical p.d.f. with parameter R and U' has the stationary distribution of $(U_j)_{j\geq 0}$. Then*

$$\lim_{\varepsilon\to 0}\{\limsup_{n\to\infty}\left|P[\max(U_1, ...U_{[n\varepsilon]}) \leq u_n \mid U_0 > u_n] - \theta\right|\} = 0$$

for a normalizing sequence u_n.

Proof It is a simple adaptation of the Rootzen proof given in the continuous case with the geometric distribution in place of the exponential one (Rootzen, 1988). The geometric distribution of $V - 1$ (and not V) comes from the following argument: actually θ is defined (Rootzen, 1988) by

$$\theta = P(V' + U' + X_1 \leq v' \mid V' > v') = P(V' - v' + U' + X_1 \leq 0 \mid V' - v' > 0)$$

and the distribution of $V = V' - v'$ conditionally to $V' - v' > 0$ is a shifted geometric distribution, for V' has a geometric distribution. This comes from the lattice structure of X_1 and is a difference with the exponential continuous case. Note that the convolution is easy to compute numerically for its range is restricted to the integers between $\min(X_1)$ and 0. The p.d.f. of V and U are easily obtained numerically using (3.1).

Remark 3 *Note that one can use an alternative way for computing θ using the analogy with the extremal index in the non-lattice case: the following expression where x_n is a normalizing sequence*

$$P(H_n \leq x_n) \overset{n\to\infty}{\sim} [F_U(x_n)]^{\theta n} \tag{3.2}$$

may be approximated for n_0 and x sufficiently large by

$$P(H_{n_0} \leq x) \approx [F_U(x)]^{\theta n_0} \tag{3.3}$$

We can compute the left hand side by exact computation (Daudin and Mercier (1999)) and $F_U(x)$ is also known.

Therefore using (3.2) and (3.3):

$$\theta = \lim_{n\to\infty} \frac{1}{n} \frac{\ln(P[H_n \leq x_n])}{\ln(F_U(x_n))} \approx \frac{1}{n_0} \frac{\ln(P[H_{n_0} \leq x])}{\ln(F_U(x))}.$$

Practically, for numerical reasons, the right hand side term does not depend on x if $10^{-3} > F_U(x) > 10^{-8}$. The two preceding methods for computing θ have been evaluated and give very similar results, the second one being generally lower than the first one.

Finally we conclude by the following theorem which is the main result of this paper:

Theorem 5 *Let $(X_n)_{n\geq 0}$ be an i.i.d. $\mathbb{Z}$-valued sequence and $(U_n)_{n\geq 0}$ be the associated Lindley process. Assume the following conditions:*

(C1) $E[X_1] < 0$

(C2) $P[X_1 > 0] > 0$

then

$$\exp\left(-\frac{\delta}{1-R}\theta R^x\right) \leq \liminf_{n\to\infty} P[H_n \leq \frac{\log n}{-\log(R)} + x],$$
$$\limsup_{n\to\infty} P[H_n \leq \frac{\log n}{-\log(R)} + x] \leq \exp\left(-\frac{\delta}{1-R}\theta R^{x+1}\right).$$

Proof This result is a simple adaptation of the continuous case given in Leadbetter (1983) and Rootzen (1988). The limits of $P[U_n \leq \frac{\log n}{-\log(R)} + x]$ and $P[H_n \leq \frac{\log n}{-\log(R)} + x]$ do not exist and are replaced at each step of the proof by the $\liminf$ and $\limsup$.

There are two possible ways for computing the constants δ and R. The first one consists in resolving the linear recurrence satisfied by μ. It is not difficult to obtain the largest real root of modulus less than 1 of the characteristic polynomial of the recurrence. The computation of δ is more cumbersome if the degree of the polynomial (i.e. $\max(X_1) - \min(X_1)$) is high for it implies to have as many side conditions as the number of roots and to solve the linear system associated. In such a case the following way may be better. It consists in using (3.1) for obtaining μ and then to adjust the values of δ

and R on some values of μ in its tail distribution by any numerical method such as linear regression, for example: $\mu_x = P_U(x) \approx \delta R^x$ implies that $\log \mu_x \approx \log \delta + x \log R$ for $x \in \mathbb{N}$.

4 Examples and applications

Here are two examples for the lattice case. The first deals with the standard random walk and the results of the different methods applied to this case are given. Applications to biological sequences analysis are tackled in the second subsection.

4.1 The standard random walk

When $X_1 = \pm 1$ with $p = P[X_1 = 1]$, $q = P[X_1 = -1]$ and $p < q$, we have the well known result (Feller (1968)) about the stationary distribution of $(U_j)_{j \geq 0}$ which is geometrical

$$P_U(k) = (1 - \frac{p}{q})(\frac{p}{q})^k$$

therefore

$$\delta = 1 - \frac{p}{q} \;\; ; \;\; R = \frac{p}{q} < 1 \; .$$

So

$$C = \frac{\delta}{1 - R} = 1,$$

and using the c.d.f. and independence property of V, U', X_1, ($V-1$ and U have the same geometric distribution) we obtain $\theta = P[V + U' + X_1 \leq 0]$:

$$\theta = P(V=1, U'=0, X_1=-1) = (1-\frac{p}{q})^2 q = \frac{(q-p)^2}{q},$$

which is in accordance with the value of K given by Karlin and Altschul (1990):

$$K = C\theta = \theta = \frac{(q-p)^2}{q}.$$

Using Theorem 5, we have:

$$\begin{aligned} \exp\left(-\frac{(q-p)^2}{q}\left(\frac{p}{q}\right)^x\right) &\leq \liminf_{n\to\infty} P[H_n \leq \frac{\log n}{-\log(\frac{p}{q})} + x], \\ \limsup_{n\to\infty} P[H_n \leq \frac{\log n}{-\log(\frac{p}{q})} + x] &\leq \exp\left(-\frac{(q-p)^2}{q}\left(\frac{p}{q}\right)^{x+1}\right). \end{aligned}$$

4.2 Protein scoring

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. The local score of the sequence $A_1 \ldots A_n$

according to a scoring scheme σ is defined as follows.

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

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. See Karlin and Altschul (1990), Karlin *et al.* (1990) for examples of simple scoring functions. 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. We consider the examples given in Karlin and Altschul (1990). The two following tables give the values obtained for the constants in their 10 examples where

$$P_{\max} = 1 - \exp\left(-\frac{\delta}{1-R}\theta R^{x}\right), \tag{4.1}$$

$$P_{\min} = 1 - \exp\left(-\frac{\delta}{1-R}\theta R^{x+1}\right) \tag{4.2}$$

are respectively the lower and upper bounds given by Theorem 5, P is the exact value of $P(H_n \ge a)$ given by Daudin and Mercier (1999), where a is the maximal local score of the sequence and $x = a + \log n/\log R$. The last row of each table gives the upper bound obtained by Karlin and Altschul (1990). One can see that the bounds are correct unless in two cases (in bold) where the value of x is too small ($x = 2.39$ in the case c(i), and $x = -14.8$

in case e) or equivalently P is too large ($P = 0.0356$ in the case c(i), and $P = 0.95$ in case e).

These results are not unexpected because the bounds are only valid for extreme values. Our upper bound is generally better (closer to the true value) than the Karlin and Altschul one. In three cases, there is a large difference between these two bounds: a(ii), b, c(iii). It is not possible to give more insight about this difference for the Karlin and Altschul (1990) paper does not give the numerical values of K. Note that P is nearer from $P_{\max}$ than from $P_{\min}$ for $a = x - \log n / \log R$ is an integer. For example, if we take in the first case a(i) $a = 20.2$ in place of $a = 21$, there are only four modified terms in the associate column of the table: $a, x, P_{\min}$ and $P_{\max}$. We have obviously $P[H_n \geq 21] = P[H_n \geq 20.2] = 0.00699$ but the two bounds are changed. We obtain $a = 20.2, x = 5.56, P_{\min} = 0.00672$ and $P_{\max} = 0.0104$, and the true value is thus nearer from the lower bound. In practice we are only interested by $P[H_n \geq a]$ for integer values of a, so the continuous function $1 - \exp\left(-\frac{\delta}{1-R}\theta R^x\right)$ is somewhat artificial for we use it only for discrete values x such that $a = x - \log n / \log R \in \mathbb{N}$.

We can use two alternative methods for computing $P[H_n \geq a]$: the exact and the asymptotic bounds. What method can we advise? We have made some tests for very low values of $P[H_n \geq a]$ which show that the exact

Example	a (i)	a (ii)	b	c (i)	c (ii)
values of X_1	-1,2	-1,2	-2,-1,2	-2,-1,2	-2,-1,2
Probabilities	0.799, 0.201	0.798, 0.202	0.08, 0.8296, 0.0904	0.099, 0.799, 0.102	0.094, 0.796, 0.11
n	643	331	575	1320	575
a	21	29	11	10	12
x	6.364	15.75	4.740	2.393	4.900
θ	0.1418	0.1397	0.4790	0.4632	0.4271
δ	0.3088	0.3072	0.4596	0.4542	0.4413
R	0.6429	0.6453	0.3624	0.3889	0.4086
$K = \frac{\delta}{1-R}\theta$	0.1226	0.1210	0.3476	0.3442	0.3210
P_{min}	4.73e-3	7.88e-5	1.024e-3	1.39e-2	1.63e-3
P_{max}	7.34e-3	1.22e-4	2.82e-3	**3.53e-2**	3.99e-3
P	6.99e-3	1.04e-4	2.75e-3	3.56e-2	3.97e-3
$P_{\max} K\&A$	8.0e-3	2e-4	3.7e-3	**3.4e-2**	4.0 e-3

Table 1: *Numerical comparisons between the bounds of Karlin et al. (90,92) given by (1.3) and (1.4), the bounds we propose (see Theorem 5) and the exact value of the distribution of the local score calculated with the method of Daudin and Mercier (1999, 2001). The examples of the scoring function are the ones proposed by Karlin and Altschul (1990).*

Examples	c (iii)	d (i)	d (ii)	d (iii)	e
values of X	-2,-1,2	-2,-1,1	-2,-1,1	-2,-1,1	-1,5
Probabilities	0.185, 0.564, 0.251	0.364, 0.319, 0.317	0.234, 0.298, 0.468	0.316, 0.268, 0.416	0.915, 0.085
n	614	552	325	1480	575
a	37	17	15	21	12
x	19.25	10.39	0.1345	8.821	-14.88
θ	0.1236	0.3586	0.0882	0.1892	0.1011
δ	0.2670	0.6149	0.3224	0.4508	0.1523
R	0.6965	0.3851	0.6777	0.5491	0.7895
$K = \frac{\delta}{1-R}\theta$	0.1087	0.3642	0.0882	0.1892	0.0732
P_{min}	7.17e-5	6.98e-6	5.52e-2	5.25e-4	0.857
P_{max}	1.03e-4	1.81e-5	8.03e-2	9.56e-4	**0.912**
P	9.22e-5	1.73e-5	7.47e-2	9.31e-4	0.950
$P_{\mathrm{max}} K\&A$	2.0 e-4	1.8 e-5	8.0e-2	1.0e-3	**0.91**

Table 2: *Numerical comparisons between the bounds of Karlin et al. (90,92) given by (1.3) and (1.4), the bounds we propose (see Theorem 5) and the exact value of the distribution of the local score calculated with the method of Daudin and Mercier (1999, 2001). The examples of the scoring function are the ones proposed by Karlin and Altschul (1990).*

method gives very precise result if $P[H_n \geq a] > 10^{-30}$. It takes into account the discrete nature of a and is not computationaly intensive. Its drawback is that it gives only one value for a given a and we have to start again for any new value. The asymptotic bounds are precise when $P[H_n \geq a] < 10^{-10}$; however even if the interval between the bounds is narrow in absolute value, the relative precision $\frac{P\max - P\min}{P\max}$ does not increase with n. However once the constants δ, θ and R have been computed, the bounds can be readily obtained using expressions (4.2) and (4.1) for any value of a under the condition that $x = a + \log n / \log R$ is sufficiently large.

5 Conclusion

Combining new results on the local score (Mercier and Daudin 2001) with known results on the extremes of the Lindley process (Rootzen 1988), allows us to propose a new approach to describe the asymptotic behavior of the local score in the logarithmic case. This approach has the following advantages:

- in the non lattice case, the limit distribution result of Karlin *et al.* (1.2) is straightforward;
- in the lattice case, this new method brings similar results as (1.3) and (1.4) but with slightly different bounds which are more directly obtained. Moreover, they also seem to be more precise.

6 References

Anderson, C. W. (1970). Extreme value theory for a class of discrete distributions with applications to some stochastic processes, *J. Appl. Prob.*, 7, 99-113.

Arratia, R., and Waterman, M.-S. (1992). A phase transition for the score in matching random sequences allowing deletions, *Ann. Appl. Prob.*, 4, 200-225.

Asmussen, S. (1987). *Applied Probability and Queues.* Wiley, New-York.

Borovkov, A.-A. (1976). *Stochastic processes in queuing theory.* Springer Verlag.

Daudin, J.-J., and Mercier, S. (1999). Distribution exacte du score local d'une suite de variables indépendantes et identiquement distribuées, *C.R. Acad. Sci. Paris*, t 329, 815-820.

Dembo, A., and Karlin, S. (1991). Strong limit theorems of empirical functionals for large exceedences of partial sums of i.i.d. variables. *Ann. Probab.* 19, 4, 1737-1755.

Dembo, A., Karlin, S., and Zeitouni, O. (1994). Critical phenomena for sequences matching with scoring. *Ann. Prob.* 19, 4, 1737-1755.

Feller, W. (1968). *An introduction to Probability Theory and its Applications*, Wiley, New-York.

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

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., Dembo, A., and Kawabata, T. (1990). Statistical composition of high-scoring segments from molecular sequences, *Ann. Stat.*, 18, 2, 571-581.

Leadbetter, M. R. (1983). Extremes and local dependance in stationary sequences, *Z. Wahr. verw. Gebiete*, 65, 291-306.

McCormick, W. P., and Park Y. S. (1992). Asymptotic analysis of extremes from autoregressive negative binomial processes, *J. Appl. Prob.*, 29, 904-920.

McCormick, W.P., and Sun J. (1993). Sums and maxima of discrete stationary processes, *J. Appl. Prob.*, 30, 863-876.

Mercier, S. (1999). Statistique des scores pour l'analyse et la comparaison de séquences biologiques, *PHD thesis*, Rouen University.

Mercier, S. and Daudin, J.-J. (2001). Exact distribution for the local score of one i.i.d. random sequence, *Journal of Computational Biology*,8, 4,373-380.

Perfect, R. (1994). Extremal behavior of stationary Markov chains with applications, *Ann. Appl. Prob.*, 4, 2, 529-548.

Rootzen, H. (1988). Maxima and exceedance of stationary markov chains, *Adv. Appl. Prob.*, 20, 371-390.

e-mail address : daudin@inapg.inra.fr