# AN IMPROVED APPROXIMATION FOR ASSESSING THE STATISTICAL SIGNIFICANCE OF MOLECULAR SEQUENCE FEATURES

Sabine Mercier, Dominique Cellier, François Charlot

*Applied Probability Trust* (18 April 2003)

# AN IMPROVED APPROXIMATION FOR ASSESSING THE STATISTICAL SIGNIFICANCE OF MOLECULAR SEQUENCE FEATURES

MERCIER, S.,* *University of Toulouse II*

CELLIER, D., CHARLOT, D.,** *University of Rouen*

**Abstract**

Using random walk theory, we first establish explicitly the exact distribution of the maximal partial sum of a sequence of independent and identically distributed random variables. This result allows us to obtain a new approximation of the distribution of the local score of one sequence. This approximation improves the one given par Karlin *et al.*, which can be deduced from this new formula. We obtain a more accurate asymptotic expression with additional terms. Examples of application are given.

*Keywords:* statistical significance, sequence analysis, local score, random walks

AMS 2000 Subject Classification: Primary 60G50

Secondary 60J15, 62P10

## 1. Introduction

Molecular sequence analysis has become an important tool in molecular biology. Determining what is likely or unlikely to occur by chance may help in identifying interesting patterns in sequences. Let $A_1...A_n$ be an observed sequence (DNA, protein, ...) from a finite alphabet (nucleotides or amino acids). Let $\sigma$ be a scoring function taking values in $\mathbb{Z}$. Scoring assignments for nucleotides or amino acids may arise from a variety of considerations like biochemical categorization, physical properties, or association with

* Postal address: GRIMM, Dpt Math Info, University of Toulouse II, 31058 Toulouse cedex 9, France.

** Postal address: LMRS, UMR CNRS 6085, University of Rouen, 76821 Mont-Saint-Aignan, France

secondary structures. The local score of the sequence $A_1...A_n$ according to the scoring scheme $\sigma$ is defined as follows:

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

$H_n$ corresponds to a segment of the sequence with maximal aggregate score. The local score has been already studied many times using a random model when successive letters of a sequence are generated independently and with an identical distribution, or with a Markov chain model. For surveys of this subject, see [21] and [10]. Arratia and Waterman [3], proved that for sequences of length $n$, as $n \to +\infty$, there exists a transition phase : when the average score is positive, there is a "linear growth" of the local score in $n$, $H_n = \mathcal{O}(n)$; and when the average score is negative, it is called the logarithmic case, $H_n = \mathcal{O}(\ln n)$. See also [22], [15] and [8] for studies of the logarithmic domain. Arratia, Gordon and Waterman [2], studied approximations of the distribution of counts of matches in the best matching segment of specified length when comparing two long sequences of independent and identically distributed (i.i.d.) letters. The key tools are large deviations inequality and the Chen-Stein method of Poisson approximation, [1].

Karlin and Altschul [13] and Karlin and Dembo [14] presented for the first time a result on the distribution of the local score for one sequence when the average score per letter is non positive. That is for a sequence $\mathbb{X} = (X_i)_{i\geq 1}$ of independent and identically distributed random variables with $E[X_1] < 0$ and $H_n = \max_{1\leq i\leq j\leq n} \left( \sum_{k=i}^{j} X_k \right)$, they proved that

$$P[H_n \leq \frac{\ln n}{\lambda} + x] \overset{n\to\infty}{\sim} \exp\left(-K^* \cdot e^{-\lambda x}\right) , \tag{1}$$

where $\lambda$ and $K^*$ depend only on the parameters of the model of the sequence. Karlin *et al.* [9], generalized formula (1) to gap-less alignment of two sequences.

The work of Karlin *et al.* [14] stands on two steps : they first established an approximation for the distribution of the maximal partial sum $M = \max_{k\geq 0}(X_1 + ... + X_k)$, using renewal theory and the work of Iglehart [12]. This last approximation is then used to obtain the result (1) on the distribution of the local score $H_n$. But it is an usual result that the distribution of $M$ is the unique stationary distribution of a Windley

process $W$ defined as follows:

$$W_n = (W_{n-1} + X_n)^+ \ ,$$

see for example [5] (Chapter 3), or [11]. This distribution has never been explicitly given and so can't have been directly used to deduce a result for the distribution of the local score $H_n$.

In this article, we are interested in theoretical and explicit formula to assess the statistical significance of sequence analysis. Using the random walk theory, we establish (see Proposition 1) the exact distribution of the maximum of the partial sums of a sequence of independent and identically distributed random variables when the average score per letter is non positive. We don't have found this result in the litterature ; maybe it is due to the fact that our particular case (biological sequence analysis, sequences comparison) is not a very interesting problem for the specialists of waiting file. Some works seem to be very similar, see [5] (Chapter 3), or Wald [20], but they focus on stopping time identities or first passage time probability and don't go further for the distribution of the maximal partial sums. The explicit distribution of $M$ in Proposition 1, allows us to obtain a more precised approximation for the distribution of the local score of one sequence. This result improves the one given par Karlin *et al.*, [13], [14]: we obtain additional terms which give us a more accurate asymptotic expression.

Daudin and Mercier [19], give a theoretical method to calculate directly the exact distribution of the local score $H_n$ of an i.i.d. sequence in linear or logarithmic domain: using Markov chain theory, the distribution of $H_n$ is given by a power of a matrix. This work does not stand on the study of the partial sum and is independent of the sign of $E[X_1]$. The method is simple and largely usable but does not give the explicit formula of the distribution. It is a very practical method to obtain the statistical significance of local score of short sequences, but the computation for very long sequences can become more fastidious. This work must be put together in a complementary way with the asymptotic result of this article to study the statistical problem of sequence analysis for both long or short sequences.

This article is organized as follows. Section 2 deals with the distribution of the maximum partial sums and an approximation of the distribution of the local score. These results are proved in Section 3. Several examples are developed in Section 4 for scoring functions with non positive drift (see (2)). We give a conclusion in Section 5.

## 2. Notations and results

The following notations will be used. Let $\mathbb{X} = (X_i)_{i\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables from the alphabet $\mathcal{A} = \{+u, ..., -v\}$. Let $p_i = P[X_0 = i]$ for $i = 0, 1, ..., u$, and $q_j = P[X_0 = -j]$ for $j = 1, ..., v$, with $\sum_{i=0,...,u} p_i + \sum_{j=1,...,v} q_j = 1$. The sequence $\mathbb{X}$ corresponds to the score sequence: $X_i$ modelizes $\sigma(A_i)$. We suppose

$$E[X_0] = \sum_{i=1}^{u} i \cdot p_i - \sum_{j=1}^{v} j \cdot q_j < 0 \ , \tag{2}$$

so we are in the logarithmic domain. For all $k \in \mathbb{N}^*$, we denote $S_k = \sum_{i=1}^{k} X_i$ the partial sums associated to the sequence $\mathbb{X}$ and we set $S_0 = 0$. Let $M = \sup_{k\geq 0} S_k$ be the maximal partial sum.

For a real $x$, $\lfloor x \rfloor = \max\{k \in \mathbb{Z} : k \leq x\}$ will denote the integer part of $x$.

For part of our work, we are going to adopt the same blueprint than Karlin *et al.*; that is, in order to get results on the local score distribution, we are first going to work on the distribution of $M$.

Let $P$ be the polynomial

$$P(x) = \sum_{i=1}^{u} p_i \cdot x^{u-i} + (p_0 - 1) \cdot x^u + \sum_{j=1}^{v} q_j \cdot x^{u+j} \ , \tag{3}$$

of degree $u + v$. We have $P(x) = x^u \cdot (E[x^{-X_0} - 1])$.

**Proposition 1.** (Distribution for the maximal partial sums.)

1. *The sum of multiplicities of the roots with modulus stricly smaller than 1 is equal to u, and equal to v for the ones with modulus equal or more than 1.*

2. *The polynomial $P$ defined in (3) has only two real positive roots that are of simple multiplicity: 1 and $R$ with $0 < R < 1$. For any other root $R_i$ of $P$ such that $|R_i| < 1$, we have $|R_i| < R$.*

3. *The probability of $M$ is as follows*

$$\gamma_k = P[M = k] = \sum_i \delta_i \cdot R_i^k \ , \tag{4}$$

*with $R_i$ the real roots and $C_j$ the complex roots of $P$ with modulus smaller than 1. Letting $p = P[X_0 \geq 0]$, the $(\gamma_i)$ (and the $(\delta_i)$, $\delta_{Rj}$, $\delta_{Ij}$ implicitly) are computed with the following linear equations systems:*

$$\gamma_k = \sum_{j=1}^{v} (q_j \cdot \gamma_{k+j}) + \sum_{i=0}^{k} (p_i \cdot \gamma_{k-i}) \qquad (1 \leq k \leq u-1)$$

*and*

$$\sum_{k \geq 0} \gamma_k = 1 \ .$$

The method we use to prove this result is also totally different from the way Karlin *et al.* established their approximation on the distribution of $M$. Our method stands on Markov chains and random walk theory and brings additional terms that allow us to confirm their approximation, see [14] and also [16]. We then apply our result to the distribution of the local score. Let us define $T_1$ as the stopping time of the first non positive partial sum. We denote by $S^- = S_{T_1}$ the first non positive partial sum and $\mu = E[T_1]$. Wald's identity gives us $\mu = E[S^-]/E[X_0]$. The negative drift of $X_i$, (cf (2)) assures that $\mu$ is finite.

**Theorem 2.1.** (New approximation for local score distribution.) *Let*

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

*We have the following asymptotic distribution*

$$P\left[ H_n \leq \frac{\ln n}{\lambda} + x \right] \overset{n \to +\infty}{\sim} \left[ 1 - \sum_i \Delta_a^{(i)} \frac{K_i \cdot {\rho_i}^x}{n^{\Gamma_i}} \right]^{\frac{n}{\mu}+1} \ , \tag{5}$$

*with*

$$\lambda = ln(1/R), \rho_i = |R_i|, \Gamma_i = ln(\rho_i)/ln(R) \ ,$$

$$\Delta_a^{(i)} = (-1)^{a_n(x)} \text{ for } \mathrm{R_i} \neq \mathrm{R} \text{ and }, 1 \text{ for R} ,$$

$$a_n(x) = \lfloor \frac{\ln n}{\lambda} + x \rfloor ,$$

$$K_i = \frac{\delta_i \cdot R_i}{(1 - R_i)} \cdot (1 - E[R_i^{-S^-}]) ,$$

*$R_i$ and $\delta_i$ being defined in Proposition 1.*

The term in the sum corresponding to the root $R$ (see Proposition 1 Point 2) is the main term due to the fact that $|R|$ is the greatest modulus smaller than 1 among the roots of $P$. Considering only this root corresponds to the result of Karlin *et al.*, and we have $\lambda = \ln(1/R)$. This will be shown in Section 5.

## 3. Proofs

### 3.1. Proposition

3.1.1. *Point 1 of Proposition 1* Point 1 can be deduced from [5]. As it is written in [5], Wald [20] has worked on the problem of computing the first passage time probability of a random walk. He suggested an approach using the following martingale $\hat{f}(x) = E[x^{X_0}]$. We have : $P(x) = x^u \cdot (\hat{f}(1/x) - 1)$.

3.1.2. *Point 2 of Proposition 1* Let $f$ be the following function defined on $\mathbb{R}^+ \setminus \{0\}$ by:

$$f(x) = E[x^{-X_0}] .$$

We have the following equivalence: $x$ is a root of $P \Leftrightarrow x$ is solution of $f(x) = 1$. The function $f$ is strictly convex. We also have $f(1) = 1$, $f'(1) = -E[X_0] > 0$ and $\lim_{x \to 0} f(x) = +\infty$. So there exists a unique root $R$ in $]0, 1[$.

3.1.3. *Point 3 of Proposition 1* As we have already said in the introduction, it is an usual result that the distribution of $M$ is the unique stationary distribution of the Markov chain $W$, a Windley process, defined as follows

$$W_n = (W_{n-1} + X_n)^+ .$$

See for example [6], [4] and [5].
The equations given in Proposition 1 come from the translation of this result. Let $\gamma = (\gamma_k)_{k \in \mathbb{N}}$ be the distribution of $M$ and $\pi = (\pi_{ij})_{i,j \in \mathbb{N}}$ the probability transition

of $W$. Due to the distribution of the $X_i$, and with $p = P[X_0 \geq 0] = \sum_{i=0}^{u} p_i$, the probability $\pi$ is given by

$$\begin{aligned}
\pi_{0,0} &= P[X_0 \leq 0] = \sum_{i=1}^{v} q_j + p_0, \\
\pi_{0,i} &= p_i \qquad (\forall i = 1, ..., u), \\
\pi_{k,0} &= \sum_{j=k}^{v} q_j \qquad (\forall 1 \leq k \leq v), \\
\pi_{k,k-j} &= q_j \qquad (\forall 1 \leq j \leq v)\ (\forall k > j), \\
\pi_{k,k+i} &= p_i \qquad (\forall k \geq 1)\ (\forall i = 0, ..., u), \\
\pi_{k,l} &= 0 \qquad \text{else.}
\end{aligned}$$

The stationarity of the distribution $\gamma$ brings us $\gamma\pi = \gamma$, with $\gamma\pi = (\sum_{i\in\mathbb{N}} \gamma_i \pi_{ik})_{k\in\mathbb{N}}$.

Thus, $\gamma$ satisfies

$$\gamma_0 = \sum_{k=1}^{v} \left[ \left( \sum_{j=k}^{v} q_j \right) \cdot \gamma_k \right] + \left( \sum_{j=1}^{v} q_j + p_0 \right) \cdot \gamma_0, \tag{6}$$

$$\gamma_k = \sum_{j=1}^{v} (q_j \cdot \gamma_{k+j}) + \sum_{i=0}^{k} (p_i \cdot \gamma_{k-i}) \qquad (1 \leq k \leq u-1), \tag{7}$$

$$\gamma_k = \sum_{j=1}^{v} (q_j \cdot \gamma_{k+j}) + \sum_{i=0}^{u} (p_i \cdot \gamma_{k-i}) \qquad (k \geq u). \tag{8}$$

$$\tag{9}$$

From the equation (8), we deduce that $(\gamma_k)_{k\in\mathbb{N}}$ is a linear combination of sequences

$$\left( (R_i^k)_{k\geq 0} \right) \ ,$$

where $R_i$ are the distinct roots of the polynomial defined in (3), with module less than 1. As $\gamma$ is a probability, we only consider roots with a modulus smaller than 1.
Point 3. of Proposition 1 is then proved.
The conditions for the $\delta_i$ and $(\delta_{\mathcal{R}_j}, \delta_{\mathcal{I}_j})$ are deduced from the equations (6) and (7) and the fact that $\sum_k \gamma_k = 1$. We can show that (6) is a combination of the $u-1$ equalities of (7), so we correctly have $u$ equations to solve the $u$ coefficients of the interesting roots of $P$.

### 3.2. Proof of Theorem 1

3.2.1. *Notations* Using Proposition 1, we are going to establish an approximation for the distribution of the local score $H_n$. Let $T_k$ be the successive decreasing excursions of the process $\{S_k\}$: $T_0 = 0$ and $T_{k+1} = \inf\{i > T_k : S_i - S_{T_k} < 0\}$. According to the negative drift, the independence and the identical distribution of the $X_i$, we have

$$(\Sigma_k)_{k\geq 0} = (S_i - S_{T_k},\ T_k \leq i < T_{k+1})_{k\geq 0}$$

that are independent and identically distributed. Let's define as well

$$Q = Q_1 = \max_{0\leq k<T_1} S_k \qquad \text{and} \qquad Q_i = \max_{T_i\leq k<T_{i+1}} (S_k - S_{T_i}).$$

The $Q_i$ are i.i.d. random variables. We have

$$\forall m \geq 0 \qquad H_{T_m} = \max_{1\leq i\leq m} Q_i\ . \tag{10}$$

As in [14], the determination of the distribution of $H_{T_m}$ is obtained in two steps: we first establish the distribution of $Q_1$ according to the one of $M$, and then we use the result on $Q_1$ to get the distribution of $H_{T_m}$. The second step is easy due to (10) and to the fact that the $(Q_i)_i$ are i.i.d. We have

$$\begin{aligned} P[H_{T_m} \leq a] &= P[\max_{i=1,\ldots,m} Q_i \leq a] \\ &= P[\forall i = 1,...,m \quad Q_i \leq a] \\ &= (P[Q_1 \leq a])^m\ . \end{aligned} \tag{11}$$

With the notations of Theorem 1 and $S^- = S_{T_1}$ the first non positive partial sum, we have

**Lemma 1.**

$$P[Q > a] \overset{a\to\infty}{\sim} \sum_i K_i R_i^a \text{ where } K_i = \frac{\delta_i \cdot R_i}{(1-R_i)} \cdot (1 - E[R_i^{-S^-}])\ . \tag{12}$$

*Proof*

Let $F(y)$ denote the distribution of the maximum $Q = \max_{0\leq k<T_1} S_k$ and $G(y)$ the one of $M$: $F(y) = P[Q \leq y]$, $G(y) = P[M \leq y]$. To prove Lemma 1, we employ the

method of Karlin *et al.* [14]. For $y > 0$, let $\sigma_y = \inf\{k, S_k < 0 \text{ or } S_k > y\}$ be another stopping time. We have

$$\begin{aligned} 1 - G(y) &= P[M > y, S_{\sigma_y} > y] + P[M > y, S_{\sigma_y} < 0] \\ &= P[S_{\sigma_y} > y] + P[M > y, S_{\sigma_y} < 0] \ . \end{aligned}$$

By definition, we have $\{S_{\sigma_y} > y\} = \{Q_1 > y\}$. Thus

$$1 - G(y) = 1 - F(y) + P[M > y, S_{\sigma_y} < 0] \ ,$$

and

$$\begin{aligned} P[M > y, S_{\sigma_y} < 0] &= \sum_{z \geq 1} P[M > y, S_{\sigma_y} = -z] = \sum_{z \geq 1} P[\max_{k \geq 0} S_k > y, S_{\sigma_y} = -z] \\ &= \sum_{z \geq 1} P[\max_{k \geq \sigma_y} S_k - S_{\sigma_y} > y + z, S_{\sigma_y} = -z] \\ &= \sum_{z \geq 1} P[\max_{k \geq \sigma_y} \sum_{i=\sigma_y+1}^{k} X_i > y + z, S_{\sigma_y} = -z] \\ &= \sum_{z \geq 1} P[\max_{k \geq 0} \sum_{i=1}^{k} X_i > y + z] \cdot P[S_{\sigma_y} = -z] \\ &= \sum_{z \geq 1} P[M > y + z] \cdot P[S_{\sigma_y} = -z] \\ &= \sum_{z \geq 1} P[M > y + z] \cdot P[S_{\sigma_y} = -z, Q \leq y] \ . \end{aligned}$$

Thus

$$P[Q > y] = P[M > y] - \sum_{z \geq 1} [1 - G(y + z)] \cdot P[S_{\sigma_y} = -z, Q \leq y] \ . \tag{13}$$

It follows from Proposition 1 that

$$P[M > y] = \sum_i \alpha_i R_i^y \text{ with } \alpha_i = \delta_i \cdot R_i/(1 - R_i) \ .$$

So

$$P[Q > y] = \sum_i \alpha_i R_i^y - \sum_{z \geq 1} \left( \sum_i \alpha_i R_i^{z+y} \right) P[S_{\sigma_y} = -z, Q \leq y] \ .$$

Using the following equality

$$P[S_{\sigma_y} = -z, Q \leq y] = P[S^- = -z, Q \leq y] \ ,$$

we get that

$$\begin{aligned} P[Q>y] &= \sum_i \alpha_i R_i^y \cdot \left(1 - \sum_{z\geq 1} R_i^z P[S^- = -z, Q \leq y]\right) \\ &= \sum_i \alpha_i R_i^y \cdot \left(1 - \sum_{z\geq 1} R_i^z (P[S^- = -z] - P[S^- = -z, Q > y])\right) \\ &= \sum_i \alpha_i R_i^y (1 - E[R_i^{-S^-}]) \\ &\quad + \sum_i \alpha_i R_i^y \left(\sum_{z\geq 1} R_i^z P[S^- = -z, Q > y]\right) \ . \end{aligned} \tag{14}$$

Let

$$\mathcal{T}(y) = \sum_{i=1}^{m} \alpha_i R_i^y \left(\sum_{z\geq 1} R_i^z P[S^- = -z, Q > y]\right) \tag{15}$$

and $\alpha = \max_{1\leq i\leq m} |\alpha_i|$. We get

$$|\ \frac{\mathcal{T}(y)}{P[Q>y]}\ | \leq m\alpha R^y \cdot \sum_{z\geq 1} R^z = m\alpha \cdot \frac{R^{y+1}}{(1-R)}\ .$$

So

$$\lim_{y\to+\infty} \frac{\mathcal{T}(y)}{P[Q>y]} = 0\ .$$

Then

$$P[Q>y] \overset{y\to\infty}{\sim} \sum_i K_i R_i^y\ , \tag{16}$$

$$\text{with } K_i = \alpha_i \cdot (1 - E[R_i^{-S^-}]) = \frac{\delta_i \cdot R_i}{1-R_i} \cdot (1 - E[R_i^{-S^-}])\ .$$

Lemma 1 is proved. □

3.2.2. *Distribution of* $\mathbf{H_n}$ In this subsection, we will control $P[Q \leq \lfloor \frac{\ln n}{\lambda} + x \rfloor]$. Let $a_n(x) = \lfloor \frac{\ln n}{\lambda} + x \rfloor$. From (14), we have

$$P[Q > \lfloor \frac{\ln n}{\lambda} + x \rfloor] = \sum_{i=1}^{m} K_i R_i^{a_n(x)+1} + \tilde{\mathcal{T}}(n) \tag{17}$$

with

$$\tilde{\mathcal{T}}(n) = \mathcal{T}(\lfloor \frac{\ln n}{\lambda} + x \rfloor))\ , \tag{18}$$

where $\mathcal{T}$ is defined in (15).

Using the notations of Theorem 1 and Lemma 1, we get

$$P[Q > \lfloor \frac{\ln n}{\lambda} + x \rfloor] \overset{n\to+\infty}{\sim} \sum_i \frac{K_i \Delta_a^{(i)} \rho_i^{x+1}}{n^{\Gamma_i}} \overset{n\to+\infty}{\sim} \frac{K_1 R^{x+1}}{n} \ . \tag{19}$$

Thus

$$P[Q > \lfloor \frac{\ln n}{\lambda} + x \rfloor] = \mathcal{O}(\frac{1}{n}) \ . \tag{20}$$

We need to control $\tilde{\mathcal{T}}(n)$ as well. We have

$$\tilde{\mathcal{T}}(n) = \sum_i \alpha_i R_i^{a_n(x)} \cdot \left( \sum_{z\geq 1} R_i^z P[Q > a_n(x), S^- = -z] \right) \ .$$

Then, with $\Gamma_i = \ln|R_i|/\ln R = -\ln|R_i|/\lambda$,

$$\begin{aligned} |\tilde{\mathcal{T}}(n)| &\leq \sum_i \left( |\alpha_i R_i^{a_n(x)}| \cdot \sum_{z\geq 1} |R_i^z| P[Q > a_n(x), S^- = -z] \right) \\ &= \sum_i \left( |\alpha_i| \rho_i^{a_n(x)} \cdot \sum_{z\geq 1} \rho_i^z P[Q > a_n(x), S^- = -z] \right) \\ &\leq \sum_i \left( \frac{|\alpha_i| \rho_i^x}{n^{\Gamma_i}} \cdot \sum_{z\geq 1} \rho_i^z P[Q > a_n(x), S^- = -z] \right) \ . \end{aligned}$$

For all $i$, we have

$$\sum_{z\geq 1} {\rho_i}^z \cdot P[Q > a_n(x), S^- = -z] \leq \sum_{z\geq 1} R^z \cdot P[Q > a_n(x)] = P[Q > a_n(x)] \cdot \frac{R}{1-R} \ .$$

Thus, we obtain that

$$\begin{aligned} |\tilde{\mathcal{T}}(n)| &\leq \frac{R}{1-R} \cdot P[Q > \lfloor \frac{\ln n}{\lambda} + x \rfloor] \cdot \sum_{i=1}^{m} \frac{|C_i| \cdot {\rho_i}^x}{n^{\Gamma_i}} \\ &\leq \frac{mCR^{x+1}}{1-R} \cdot \frac{P[Q > \lfloor \frac{\ln n}{\lambda} + x \rfloor]}{n} \ , \end{aligned}$$

using also the fact that the $\Gamma_i > 1$ and $\Gamma_1 = 1$ in relation with $R$.

Let $\beta \in \mathbb{R}^+ \backslash \{0\}$. We have

$$|n^\beta \cdot \tilde{\mathcal{T}}(n)| \leq \frac{mCR^{x+1}}{1-R} \cdot n^{\beta-1} P[Q > \lfloor \frac{\ln n}{\lambda} + x \rfloor] \ .$$

Applying the estimate of $P[Q > \lfloor \frac{\ln n}{\lambda} + x \rfloor]$ which is provided by (19), we get

$$n^{\beta-1} P[Q > \lfloor \frac{\ln n}{\lambda} + x \rfloor] \overset{n\to+\infty}{\sim} K_1 R^{x+1} n^{\beta-2} \ .$$

Then, for $\beta < 2$, we have $\lim_{n\to+\infty} n^\beta \tilde{\mathcal{T}}(n) = 0$. The next result is also proved,

$$(\forall\ 0 < \beta < 2) \qquad \tilde{\mathcal{T}}(n) = o(\frac{1}{n^\beta}) \ . \tag{21}$$

Let $N_n = \sup\{i \in \mathbb{N} : T_i \leq n\}$ denote the number of decreasing excursions in $S_1 ... S_n$. By the law of large number we have

$$\lim_{n\to+\infty} \frac{N_n}{n} = \frac{1}{\mu} \quad \text{as} \ , \tag{22}$$

where $\mu = E[T_1]$. One must not forget that $H_n$ is increasing with $n$ and we get

$$P[H_{T_{N_n+1}} \leq a] \leq P[H_n \leq a] \leq P[H_{T_{N_n}} \leq a] \ .$$

Let $\epsilon > 0$. Using (22), we obtain

$$\begin{aligned}
P[H_{T_{N_n}} \leq a] &= P[H_{T_{N_n}} \leq a;\ n(1/\mu - \epsilon) \leq N_n \leq n(1/\mu + \epsilon)] \\
&\quad + P[H_{T_{N_n}} \leq a;\ |\ N_n/n - 1/\mu\ | > \epsilon] \\
&\leq P[H_{T_{\lfloor n(1/\mu-\epsilon)\rfloor}} \leq a] + P[|\ N_n/n - 1/\mu\ | > \epsilon] \\
&= (P[Q \leq a])^{\lfloor n(1/\mu-\epsilon)\rfloor} + o(1) \ .
\end{aligned}$$

A similar computation yields

$$\begin{aligned}
P[H_{T_{N_n+1}} \leq a] &= P[H_{T_{N_n+1}} \leq a;\ n(1/\mu - \epsilon) \leq N_n \leq n(1/\mu + \epsilon)] \\
&\quad + P[H_{T_{N_n+1}} \leq a;\ |\ N_n/n - 1/\mu\ | > \epsilon] \\
&\geq P[H_{T_{\lfloor n(1/\mu+\epsilon)\rfloor+1}} \leq a;\ n(1/\mu - \epsilon) \leq N_n \leq n(1/\mu + \epsilon)] \\
&= P[H_{T_{\lfloor n(1/\mu+\epsilon)\rfloor+1}} \leq a] \\
&\quad - P[H_{T_{\lfloor n(1/\mu+\epsilon)\rfloor}} \leq a;\ |\ N_n/n - 1/\mu\ | > \epsilon] \\
&= P[H_{T_{\lfloor n(1/\mu+\epsilon)\rfloor+1}} \leq a] - o(1) \\
&= (P[Q \leq a])^{\lfloor n(1/\mu+\epsilon)\rfloor+1} - o(1) \ .
\end{aligned}$$

So we have

$$(P[Q \leq z])^{\lfloor n(1/\mu+\epsilon)\rfloor+1} - o(1) \leq P[H_n \leq z] \leq (P[Q \leq z])^{\lfloor n(1/\mu-\epsilon)\rfloor} + o(1) \ . \quad (23)$$

Let $z = \lfloor \ln(n)/\lambda + x \rfloor$. We get

$$G_n(\epsilon) - o(1) \leq P[H_n \leq \lfloor \frac{\ln(n)}{\lambda} + x \rfloor] \leq D_n(\epsilon) + o(1) \ ,$$

$$\text{with } G_n(\epsilon) = \left( P[Q \leq \lfloor \frac{\ln(n)}{\lambda} + x \rfloor] \right)^{\lfloor n(1/\mu+\epsilon)\rfloor+1}$$

$$\text{and } D_n(\epsilon) = \left( P[Q \leq \lfloor \frac{\ln(n)}{\lambda} + x \rfloor] \right)^{\lfloor n(1/\mu-\epsilon)\rfloor} \ .$$

Then, for all $\epsilon > 0$

$$\frac{G_n(\epsilon) - o(1)}{\left(1 - \sum_i \frac{K_i \Delta \rho_i^x}{n^{\Gamma_i}}\right)^{\frac{n}{\mu}+1}} \leq \frac{P[H_n \leq \frac{\ln(n)}{\lambda} + x]}{\left(1 - \sum_i \frac{K_i \Delta \rho_i^x}{n^{\Gamma_i}}\right)^{\frac{n}{\mu}+1}} \leq \frac{D_n(\epsilon) + o(1)}{\left(1 - \sum_i \frac{K_i \Delta \rho_i^x}{n^{\Gamma_i}}\right)^{\frac{n}{\mu}+1}} \ ,$$

where $\Delta = \Delta_a^{(i)}$. In order to prove

$$\lim_{n \to +\infty} \frac{P[H_n \leq \frac{\ln(n)}{\lambda} + x]}{\left(1 - \sum_i \frac{K_i \Delta \rho_i^x}{n^{\Gamma_i}}\right)^{\frac{n}{\mu}+1}} = 1 \ ,$$

it suffices to show

$$\lim_{\epsilon \to 0} \lim_{n \to +\infty} \frac{D_n(\epsilon) + o(1)}{\left(1 - \sum_i \frac{K_i \Delta \rho_i^x}{n^{\Gamma_i}}\right)^{\frac{n}{\mu}+1}} = 1 \ , \quad (24)$$

and a similar expression for $G_n(\epsilon)$. (17), (20) and (21) give (24).

## 4. Applications

### 4.1. The most amphoteric segments

Amphoteric amino acids (mixed charge) can be positively or negatively charged depending on the medium they are. For high-scoring mixed charge segment, we use the scoring scheme given by Karlin *et al.*, [13]: $\sigma = 2$ for Aspartate (D), Glutamate (E), Lysine (K), Arginine (R) and l'Histidine (H); $\sigma = -1$ for the others. Let $p = P[X_0 = 2]$ be the probability to get the five amino acids cited previously. Xe have $E[X_0] = 3p - 1$. The probability transition of the Markov chain $\mathcal{W}$ is given by

$$\begin{aligned} (\forall k \geq 0) \quad \pi_{k,k+2} &= p \ , \\ (0 \leq k \leq 1) \quad \pi_{k,0} &= 1 - p \ , \\ (\forall k > 1) \quad \pi_{k,k-1} &= 1 - p \ . \end{aligned}$$

So the invariant measure satisfies

$$\gamma_0 = \frac{1-p}{p} \cdot \gamma_1 \ , \qquad \gamma_1 = (1-p) \cdot \gamma_2 \ ,$$

$$\text{and for } (k \geq 2) \quad \gamma_k = (1-p) \cdot \gamma_{k+1} + p \cdot \gamma_{k-2} \ .$$

The polynomial $P$ is $P[x] = (1-p) \cdot x^3 - x^2 + p$, and the two roots different from 1 are real.

$$0 < R = \frac{p + \sqrt{p(4-3p)}}{2(1-p)} < 1 \ , \tag{25}$$

$$\text{and } -1 < R' = \frac{p - \sqrt{p(4-3p)}}{2(1-p)} < 0 \ . \tag{26}$$

We have

$$(\forall a \in \mathbb{N}) \quad P[M = a] = \delta \cdot \left[ R^{a+1} - R'^{a+1} \right] \ , \tag{27}$$

$$\text{with } \delta = \frac{(1-3p)}{\sqrt{p(4-3p)}} \ . \tag{28}$$

The scoring function is such that $S^- = -1$. So Lemma 1 brings

$$P[Q > a] \overset{a \to \infty}{\sim} K_1 \cdot R^a + K_2 \cdot R'^a \ , \tag{29}$$

$$\text{with } K_1 = \delta R^2 > 0 \quad \text{ and } \quad K_2 = -\delta R'^2 > 0 \ . \tag{30}$$

**Example 4.1.** The distribution of the local score is approximated by the formula

$$P\left[ H_n \leq \frac{\ln n}{\lambda} + x \right] \overset{n \to +\infty}{\sim} \left[ 1 - \frac{K_1 R^x}{n} - (-1)^{a_n(x)} \cdot \frac{K_2 \rho^x}{n^\Gamma} \right]^{\frac{n}{\mu}+1} \ , \tag{31}$$

where $\rho = |R'|$, $\Gamma = \ln(\rho)/ln(R)$, $a_n(x) = \lfloor \frac{\ln n}{\lambda} + x \rfloor$, $\mu = 1/(1-3p)$, and with $R$, $R'$, $K_1$, $K_2$, $\delta$ given in (25), (26), (28) and (30).

Let's deduce Karlin's formula from this result. We deduce of Example 4.1 with $a = \frac{\ln n}{\lambda} + x$

$$\begin{aligned} P\left[ H_n \leq \frac{\ln n}{\lambda} + x \right] &\overset{n \to +\infty}{\sim} [1 - \frac{K_1 R^x}{n}]^{\frac{n}{\mu}+1} \\ &\overset{n \to +\infty}{\sim} \exp\left( -\frac{K_1}{\mu} R^x \right) \ , \end{aligned}$$

and we want to compare the two constants, $K_1/\mu$ and $K^*$, (cf (1)), in order to verify that the two results are compatible. We have, see [13] and [14]

$$P[H_n \leq \frac{\ln n}{\lambda} + x] \overset{n\to+\infty}{\sim} \exp\left(-K^* \cdot e^{-\lambda x}\right) , \tag{32}$$

with

$$K^* = \frac{(1-e^{-\lambda}) \cdot e^{-\lambda} \cdot E[X_0]^2}{E[X_0 \cdot e^{\lambda X_0}]} .$$

Using $\lambda = \ln(1/R)$ and the fact that $R$ is a root of $P$, we get

$$K^* = \frac{(1-R)R(1-3p)^2}{2p(1/R^2) - (1-p)R} = \frac{(1-R)R^3(1-3p)^2}{2p-(1-p)R^3} = \frac{(1-R)R^3(1-3p)^2}{3p-R^2} ,$$

Our method gives

$$K_1/\mu = \frac{\delta R^2}{\mu} = -\delta R^2 E[X_0] = \delta R^2(1-3p) = \frac{(1-3p)R^2}{\sqrt{p(4-3p)}} .$$

Using (25), It is now easy to show that $K^* = K_1/\mu$.

Our approximation deals with two terms. Taking into account only the main one allows us to obtain by a different way the result of Karlin *et al.* Numerical investigations indicate that our approximation is better, see [18].

### 4.2. The most hydrophobic segments

A scoring scheme quite consistent with the Kyte-Doolittle scale of hydrophobicity, (see [17]), takes $\sigma = 2$ for I, L, V; $\sigma = 1$ for F, M, A, C; $\sigma = 0$ for G, S, Y, W, T, P; $\sigma = -1$ for N, Q, H, D, E; $\sigma = -2$ for K, R.

This example of scoring function is proposed in [15] and [14] (Example 3). The polynomial is

$$P(x) = q_2x^4 + q_1x^3 + (p_0-1)x^2 + p_2x + p_1 .$$

Let $p = P[\sigma \geq 0]$. Only two roots are in $]-1,+1[$, noted $R$ and $R'$, such that $0 < |R'| < R < 1$. The distribution of the maximum $M$ is given by

$$(\forall k \leq 2) \quad p_2 \cdot \gamma_{k-2} + p_1 \cdot \gamma_{k-1} + p_0 \cdot \gamma_k + q_1 \cdot \gamma_{k+1} + q_2 \cdot \gamma_{k+2} = \gamma_k \tag{33}$$

$$[p_0 + (1-p)] \cdot \gamma_0 + (1-p) \cdot \gamma_1 + q_2 \cdot \gamma_2 = \gamma_0 , \tag{34}$$

$$p_1 \cdot \gamma_0 + p_0 \cdot \gamma_1 + q_1 \cdot \gamma_2 + q_2 \cdot \gamma_3 = \gamma_1 \ , \tag{35}$$

The recurrent equation (33) brings us

$$P[M = k] = \gamma_k = \delta_1 R^k + (-1)^k \delta_2 R'^k \ , \tag{36}$$

with $R_1 = R$ and $R_2 = -R'$ the two roots of $P(x)$ satisfying $|R_i| < 1$ for $i = 1, 2$.
Injecting (36) in (34) and (35) brings the same relation:

$$\delta_1 = \frac{R}{R'} \cdot \delta_2 \ .$$

Using the fact that $\sum_{k \geq 0} \gamma_k = 1$, we have also

$$\frac{\delta_1}{1 - R} + \frac{\delta_2}{1 + R'} = 1 \ .$$

Then we get with

$$\tau = \frac{(1 - R)(1 + R')}{R(1 + R') + R'(1 - R)} \ ,$$

$$\delta_1 = R'\tau, \ \delta_2 = R\tau \ .$$

We define $\Pi_i = P[S^- = -i]$, $i = 1, 2$. Considering the work of Asmussen [5] (Chapter 3), there is a much more elegant way to get the distribution of $S^-$ using the $v$ roots noted $z_i$ of $P(x)$ with modulus superior or equal to 1:

$$(P(z_i) = 0, \ |z_i| \geq 1) \Rightarrow E[(z_i)^{-S^-}] = 1 \ . \tag{37}$$

Using Asmussen (see Prop. 3.15, [5]), we have exactly 2 roots of $P(x)$ with $|R_i| < 1$ and two roots $z_i$ with $|z_i| \geq 1$. Let's note $R_3 = 1$ and $R_4$ the last root of $P(x)$ with $|R_4| \geq 1$. The equation (37) gives

$$\begin{cases} 1 = & \Pi_1 + \Pi_2 \\ 1 = & \Pi_1 \cdot R_4 + \Pi_2 \cdot R_4^2 \ . \end{cases}$$

So we get $\Pi_1 = (1 + R_4)/R_4 \in ]0, 1[$, and $E[S^-] = (1 - R_4)/R_4$.
This method has the advantage to be generalized to any scoring function taking value in $\{-v, \ ...,0, \ 1, \ 2, \ ...,+u\}$, whith $\{-1, \ -2, \ ...,-v\}$ the possible value of $S^-$. The equation (37) gives a linear system of $v$ equations to get the $(\Pi_i = P[S^- = -i])_{1 \leq i \leq v}$.

We have then

$$\mu = \frac{E[S^-]}{E[X_0]} = \frac{R_4 - 1}{R_4 \cdot [2(p_2 - q_2) + p_1 - q_1]} \ .$$

**Example 4.2.** An approximation of the distribution of $H_n$ is

$$P[H_n \leq \frac{\ln n}{\lambda} + x] \overset{n \to +\infty}{\sim} \left[1 - \frac{K_1 R^x}{n} - (-1)^{a_n(x)} \frac{K_2 R'^x}{n^\Gamma}\right]^{\frac{n}{\mu}+1} \ , \tag{38}$$

with $K_1 = \delta_1 R$, $K_2 = \delta_2 R'$, $R$, $R'$, $\delta_i$, $\mu$ given above, and $\lambda = -\ln R$, $\Gamma = \ln R' / \ln R$ and $a_n(x) = \lfloor \frac{\ln n}{\lambda} + x \rfloor$.

Karlin *et al.* in [14] (example 3) give

$$K^* = \frac{(1 - \Pi_1 e^{-\lambda} - \Pi_2 e^{-2\lambda})^2 E[X_0]^2}{(2 - \Pi_1)^2 (e^\lambda - 1) E[X_0 \exp(\lambda X_0)]} \ .$$

The two different approximations for the hydrophobic example have been numerically tested and compared with empirical distribution of the local score or with the distribution given by Daudin and Mercier [19], see Figure 1.

## 5. Conclusion

Proposition 1 deals with the maximal partial sum of a random walk. An approximation of its distribution has been obtained previously using the renewal theory, see [14]. We give explicitly in this paper the exact distribution using classical results of Markov Chain Theory.

The application of this result to the distribution of the local score allows us to get a better approximation of this distribution. Theoretical formulas seems not so easy to manipulate, but applications are practical and simple. This approximation allows us to confirm and improve numerically the result of Karlin *et al.* [13].
Some work remains to deal with the alignment scoring problem of two sequences but we think that our result can be easily generalized to the case of Markovian dependent sites.

## References

[1] Arratia, R., Golstein, L. and Gordon, L. (1989). Two moments suffice for Poisson approximations : the Chen-Stein method. *Ann. Prob.*, 17:9–25.

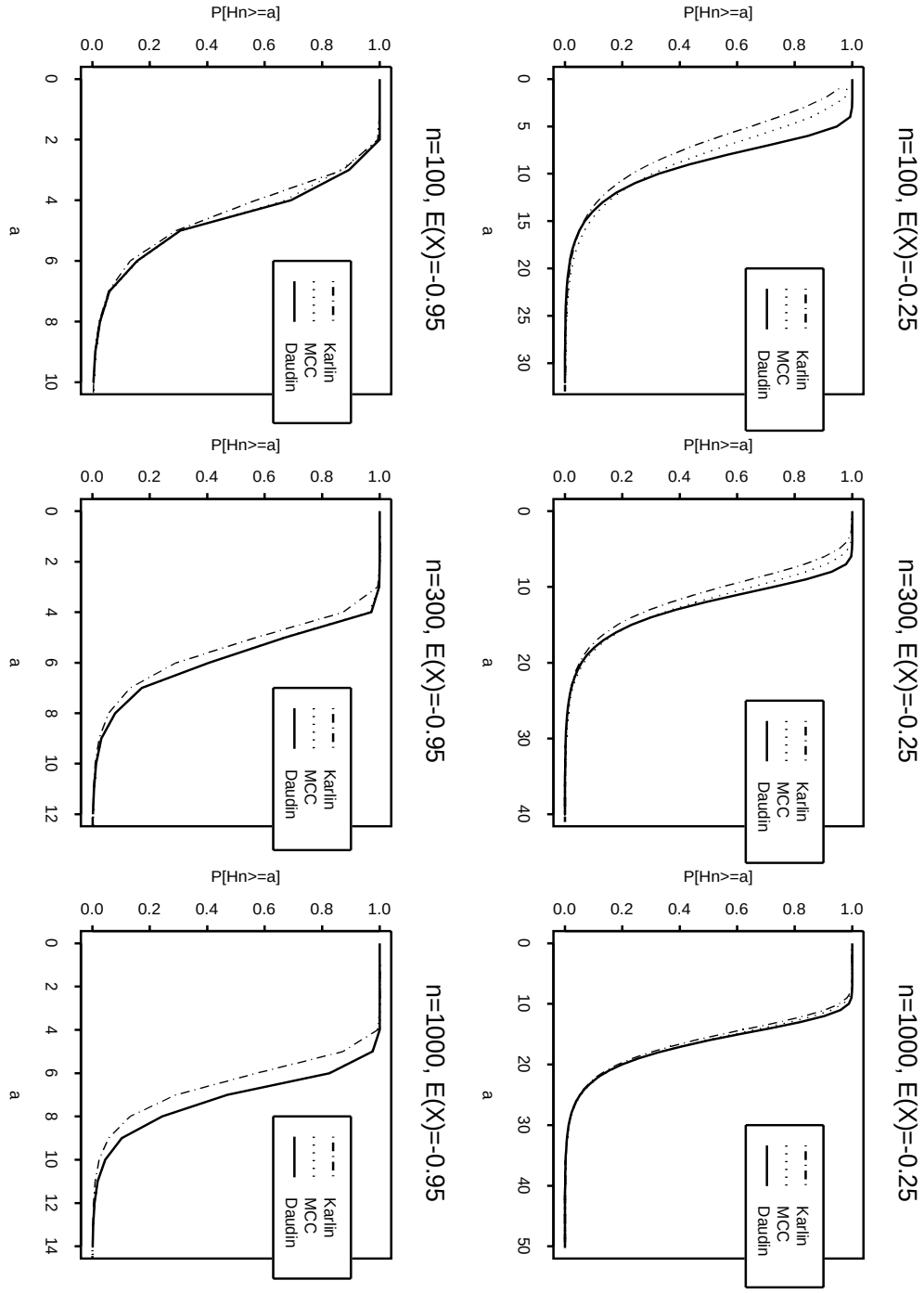


FIGURE 1: Hydrophobic case: comparison between Karlin approximation and the approximation given in this article. The real distribution is calculated with the exact method of Daudin and Mercier [19] instead of empirical distribution that spent more time.

[2] Arratia, R., Gordon, L. and Waterman, M.-S. (1990). The Erdos-Rényi law in distribution for coin tossing and sequence matching. *Ann. Statist.*, 18:539–570.

[3] Arratia, R.and Waterman, M.-S. (1994). A phase transition for the score in matching random sequences allowing deletions. *Ann. Appl. Prob.*, 4:200–225.

[4] Asmussen, S. (1987). *Applied Probability and Queues*. Wiley.

[5] Asmussen, S. (1995). *Stationary Distributions via First Passage Times, in Advances in Queuing*. J. H. Dshalalow, CRC Press.

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

[7] Daudin, J.-J. and Mercier, S. (1999). Distribution exacte du score local d'une suite de variables indépendantes et identiquement distribuées. *C. R. Acad. Sc. Paris*, 329(1):815–820.

[8] Dembo, A., Karlin, S. (1991). Strong limit theorems of empirical functionnals for large exceedences of partial sums of i.i.d. variables. *Ann. Prob.*, 19(4):1737–1755.

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

[10] 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.

[11] Feller, W. (1966). *An Introduction to Probability Theory and Its Applications, Volume II*. Wiley, New-York.

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

[13] 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.

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

[15] Karlin, S., Dembo, A., Kawabata, T. (1990). Statistical composition of high-scoring segments from molecular sequences. *Ann. Statist.*, 18:571–581.

[16] Karlin, S., and Taylor, H.M. (1981). *A second course in stochastic processes*. Academic Press.

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

[18] Mercier, S., Cellier, D., F. Charlot and Daudin, J.-J. (2001). Exact and asymptotic distribution for the local score of one i.i.d. random sequence. *LNCS*, volume for JOBIM 2000, (2066):74–85.

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

[20] Wald, A. (1947). *Sequential Analysis*, John Wiley, New York.

[21] Waterman, M.S. (1995). *Introduction to Computational Biology*. Chapman and Hall, London.

[22] Zhang, Y. (1995). A limit theorem for matching random sequences allowing deletions. *Ann. Appl. Prob.*, 5:1236–1240.