AISM (2007) 59: 741–755
DOI 10.1007/s10463-006-0064-6

**Claudie Hassenforder · Sabine Mercier**

# Exact Distribution of the Local Score for Markovian sequences

Received: 20 July 2005 /
Published online: 1 June 2006


**Abstract** Let $\mathbb{A} = (A_i)_{1\leq i\leq n}$ be a sequence of letters taken in a finite alphabet $\Theta$. Let $s : \Theta \to \mathbb{Z}$ be a scoring function and $\mathbb{X} = (X_i)_{1\leq i\leq n}$ the corresponding score sequence where $X_i = s(A_i)$. The local score is defined as follows: $H_n = \max_{1\leq i\leq j\leq n} \sum_{k=i}^{j} X_k$. We provide the exact distribution of the local score in random sequences in several models. We will first consider a Markov model on the score sequence $\mathbb{X}$, and then on the letter sequence $\mathbb{A}$. The exact $P$-value of the local score obtained with both models are compared thanks to several datasets. They are also compared with previous results using the independent model.

**Keywords** Markov chain · Local score · $P$-value · Sequence analysis

## 1 Introduction

Biostatistics is becoming a very large discipline improving its tools as the biological sequence databases are growing. One of the principal goals of the Human Genome Project started in 1990 consists in developing and improving the tools of sequence analysis. A lot of software exists for providing an analysis of the biological sequences. Some of them focus on the primary structure (succession of the nucleotides, or residues, of the sequence). For example, Antheprot (*Analyse The Protein*, http://antheprot-pbil.ibcp.fr/ie_sommaire.html), Protscale (http://us.expasy.org/cgi-bin/protscale.pl), or Emboss Octanol (http://www.

C. Hassenforder · S. Mercier (✉)
Université de Toulouse II, Equipe GRIMM,
Département Mathématiques et Informatique, UFR SES,
31058 Toulouse cedex 9, France
E-mail: chabriac@univ-tlse2.fr
E-mail: mercier@univ-tlse2.fr
Tel.: +33-561-504131
Fax: +33-561-504173

hgmp.mrc.ac.uk/Software/EMBOSS/Apps/octanol.html), determine protein or nucleic profiles using score scales. A score scale assigns to each component a numerical value, called score, reflecting physico-chemical properties. The two scales most often used are the hydrophobic scale and that corresponding to the parameters of secondary structure conformation (a first step to the spatial configuration of the proteins). Let $s(i)$ be the score of the $i$-th component of the sequence and $H(i)$ the score of the segment of a given length $L$ defined as follows:

$$H(i) = \sum_{k=0}^{L-1} s(i+k).$$

$H(i)$ is calculated onto a sliding window of length $L$ and plotted as a function of the amino acid number. These profiles highlight the maximal score and also the related region of interest. The fixed length can correspond for example to the length of the cellular membrane, converted into a number of amino acids, if one is studying the most hydrophobic regions of transmembrane proteins. But the length of the region of interest is not always known.

The local score is defined as

$$H_n = \max_{1 \le i \le j \le n} \sum_{k=i}^{j} s(k),$$

and the segment of maximal score does not have a determined length.

In order to distinguish common events from events of interest, we need to establish the distribution of the local score. Thus we need to choose a model for the biological sequences.

Let $\mathbb{A}$:$A_1 A_2 ... A_n$ be a biological sequence and $\Theta$ the alphabet corresponding to the biological sequence (for example $\Theta = \{A, C, G, T\}$ if $\mathbb{A}$ is a DNA sequence) and let $s : \Theta \to \{s_{\min}, ..., 0, ..., s_{\max}\}$ be the scoring function, with $-s_{\min}$ and $s_{\max}$ two non-negative integers. Let us define $X_i$ by $X_i = s(A_i)$ and $\mathbb{X} : X_1 X_2 ... X_n$ the score sequence, deduced from $\mathbb{A}$.

Until now the models for local score studies have always been built for the scoring sequence $\mathbb{X}$. The usual model considers $\mathbb{X}$ as a sequence of independent and identically distributed variables, and is called $M_0$ model. Arratia and Waterman (1994) proved the existence of a transition phase, with a linear growth of $H_n$ in $n : H_n = \mathcal{O}(n)$, when the average score is positive, and a logarithmic one: $H_n = \mathcal{O}(\ln(n))$ when the average score is negative. Daudin, Etienne and Valois (2003) prove that $H_n/\sqrt{n}$ converges in distribution to a standard Brownian motion when $E[X_i] = 0$. For an overview of results on the local score, see Waterman (1995), Durbin Eddy, Krogh and Mitchison (1998), Ewens (2002). The most famous result is the approximation of Karlin et al. (see Karlin and Altschul 1990; Karlin and Dembo 1992) implemented in BLAST for the sequence alignment problem

$$P\left[H_n \le \frac{\ln(n)}{\lambda} + x\right] \sim \exp\left(-K\mathrm{e}^{-\lambda x}\right) \quad \text{as } n \to +\infty, \tag{1}$$

where $\lambda$ and $K$ depend only on the parameters of the sequence model. Note that this work deals with the hypothesis of a non-positive average score ($E[X_i] < 0$),

what we call the logarithmic case. The parameter $\lambda$ is the only root in ]0, 1[ of the equation $E[e^{\lambda X_i}] = 1$ and is easy to calculate. The parameter $K$ is more difficult and cannot be calculated easily for the sequence alignment problem. Several recent articles proposed algorithmic methods in order to approximate it accurately and rapidly (see Mott 2000; Bailey and Gribskov 2002 for example). Bacro Daudin, Mercier and Robin (2003) propose a direct and simple proof of (1) and define the parameter $K$ by a new method which is easier to calculate. The result of Karlin et al. is a better approximation when sequences are becoming longer, but must be used with caution for short ones. For small proteins the approximation can be unadapted (see Mercier Cellier, Charlot, and Daudin (2001), for comparison in simple cases).

The problem of the length of the sequences combined with that of the parameter $K$ motivates the work of Mercier and Daudin (2001) who establish the exact distribution in the $M_0$ model. This work has several advantages. First, it does not need any hypothesis on the average score. Second, the exact distribution is ideally adapted for small sequences: in order to calculate the $P$-value, $P[H_n \leq a]$, for an observed local score $a$, an $(a+1) \times (a+1)$ matrix corresponding to the transition matrix of a suitable process derived from $\mathbb{X}$ is implemented at the power $n$, with $n$ the length of the sequence. This method is fast for short sequences but becomes more tedious for very long ones ($n > 1,000$). Thus, the two results, the approximation of Karlin et al. and the exact method, can be considered as complementary.

At the present time, Markov chains and their variant, the hidden Markov chains, have an important role in the interaction between biology and mathematics (see Prum 2001). The independent model is not adapted for biological sequences because there exists a dependence between the components, which can be shown in the genetic code for example. The use of a simple model was dictated more by the complexity of the mathematical problem of establishing the distribution of the local score than by a real interest in the model itself. The Markovian model can integrate a certain dependence between the component; it takes into account the different frequencies of words (words of two letters for a Markovian model of order 1) and not only the differences between the frequencies of each component. For example, let us consider the following score scale for amino acids which takes $+2$ for the residues coded as D, E, K, R, H and $-1$ for the others. This example is proposed in Karlin and Altschul (1990), for the research of the most significant amphoteric segments (an amphoteric residue has the property of being charged positively or negatively according to the medium). Let us study the Human protein 67-kDa keratin cytoskeletal type II of length $n = 643$. We deduce from the sequence the matrix of counts, where P (*resp.* N) stands for the residues with a positive (*resp.* negative) score

| | P | N | Total |
|---|---|---|---|
| P | 24 | 110 | 134 |
| N | 110 | 399 | 509 |

The segments of two residues scored $+2$ appear only 24 times, whereas segments of score $+1$ appear 110 times. The probability of the apparition of a segment of high score is influenced by the sparseness of the couple $(+2, +2)$. This observation can be extanded to longer words. We still keep in mind that the length of the segment which realizes the local score is not fixed.

The simplicity of the proof of the exact distribution in $M_0$ model and the importance of the Markovian model for biological sequences encourage us to generalize the exact method to Markovian models.

We first consider in this article a Markov model based on the score sequence $\mathbb{X}$, called the $M_{1-X}$ model, and the exact $P$-value of the local score is given (see Hassenforder and Mercier 2003). Secondly, we consider a Markov model based on the letter sequence $\mathbb{A}$, called the $M_{1-A}$ model and the exact $P$-value is also established. Note that the Markovian dependence on the letter sequence is better justified biologically, and that for $\mathbb{A}$ as a Markov chain and $s$ a scoring function which is not bijective, the sequence $\mathbb{X} = s(\mathbb{A})$ is not a Markov chain, thus the model $M_{1-A}$ is more realistic than the $M_{1-X}$ one. The theoretical results are easy to prove and use classic tools of Markov chain theory.

These new results allow us to compare the $M_0$ model and the Markov chain models for local score significance. We want to see if the improvments of Makovian models are significant enough to encourage us to use them instead of the independent model. These comparisons will be based on exact formulas and thus will focus only on the models. Simulations have been made using different databases. Different scoring functions are also used. Several computational problems appear for the Markovian model based on the letter sequence $\mathbb{A}$.

Section 2 deals with the theoretical $P$-values with proofs in both Markovian models on $\mathbb{X}$ and $\mathbb{A}$. Numerical comparisons are developed in Sect. 3, where some details of the programs are also given. Section 4 provides a conclusion and some perspectives on the study.

## 2 Theorical results and demonstrations

The Markov chains will implicitly be of order 1.

### 2.1 Model for the scoring sequence

Let $\mathbb{X} = (X_k)_{k\geq 1}$ be a Markov chain of probability matrix $\Lambda = (\Lambda_{uv})_{u,v\in\mathbb{Z}}$ and $\gamma$ the initial distribution.

Let $P = \left(P_{(i,u)(j,v)}\right)$ be a matrix such that $(i, u)$ and $(j, v)$ belong to

$$E = \{0, ..., a\} \times \{s_{\min}, ..., 0, ..., s_{\max}\} \quad \text{with } a \in \mathbb{N}, \tag{2}$$

and defined by

$$P_{(a,u)(a,v)} = \Lambda_{uv} \quad \text{and} \quad P_{(a,u)(j,v)} = 0 \text{ for } j \neq a \tag{3}$$

and for $0 \leq i \leq a-1$

$$\begin{cases} P_{(i,u)(0,v)} = \Lambda_{uv} & \text{if } i+u \leq 0 \\ P_{(i,u)(i+u,v)} = \Lambda_{uv} & \text{if } 1 \leq i+u \leq a-1\,, \text{ and} \quad P_{(i,u)(j,v)} = 0 \text{ else.} \\ P_{(i,u)(a,v)} = \Lambda_{uv} & \text{if } i+u \geq a \end{cases} \tag{4}$$

**Theorem 2.1** *The statistical significance of the local score $H_n$ is given by*

$$(\forall a \geq 0) \quad P[H_n \geq a] = \sum_{u,v} \gamma_u \cdot P^n_{(0,u)(a,v)}.$$

Let $S_k$ be the partial sums of the sequence $\mathbb{X}$: $S_0 = 0$ and $S_k = X_1 + \ldots + X_k$. Let $T_k$ be the following stopping times: $T_0 = 0$ and $T_{k+1} = \inf\{i > T_k;\, S_i - S_{T_k} < 0\}$. By definition of the $T_k$, the sequence $(S_{T_k})$ is strictly decreasing, and the $T_k$ are called the successive times of negative records.

Consider the process $\mathbb{U}$ defined by: $U_0 = 0$ and for $T_k \leq j < T_{k+1}$, $U_j = S_j - S_{T_k}$. Fig. 1 illustrates the link between the different processes. We have (see Mercier and Daudin (2001) for the proof of the following lemma):

**Lemma 2.1** $U_j = \max(U_{j-1} + X_j, 0) = (U_{j-1} + X_j)^+$ *and* $H_n = \max_{1 \leq k \leq n} U_k$.

Let $\mathbb{U}^*$ be the process stopped in $a$, with $a \in \mathbb{N}^*$. We get $U_j^* = U_j$ if $j < \tau_a$ and $U_j^* = a$ if $j \geq \tau_a$ with $\tau_a = \inf\{j \geq 1;\, U_j \geq a\}$. And finally, let us define the sequence $\mathbb{Y}$ by: $Y_{n+1} = (U_n^*, X_{n+1})$ for $n \geq 0$. The Markov chain $\mathbb{Y}$ is homogeneous and takes its values in $E$ defined in (2).

**Lemma 2.2** $\mathbb{Y}$ *is a Markov chain with probability matrix* $P = \left(P_{(i,u)(j,v)}\right)_{(i,u)\,(j,v)\in E}$, *and* $P_{(i,u)(j,v)} = P[(U_n^* = j) \cap (X_{n+1} = v) \mid (U_{n-1}^* = i) \cap (X_n = u)]$, *determined in (3) and (4).*

*Proof* For $i = a$, we have $P_{(a,u)(j,v)} = 0$ if $j \leq a - 1$ because $U^*$ is stopped in $a$, and $P_{(a,u)(j,v)} = \Lambda_{uv}$ for $j = a$.
For $i \neq a$, we have $P_{(i,u)(j,v)} = P\left[(U_n^* = j) \cap (X_{n+1} = v) \mid (U_{n-1} = i) \cap (X_n = u)\right]$.

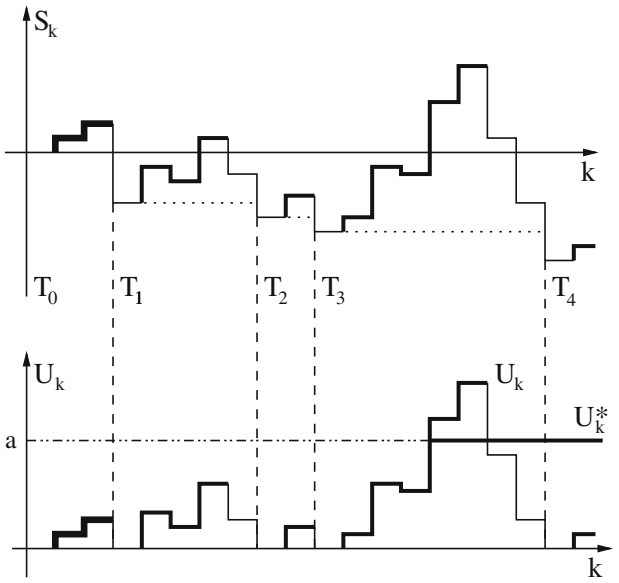


**Fig. 1** Link between processes $S_k$, $U_k$ and $U_k^*$, with $S_k$ the partial sums of the sequence $\mathbb{X}$, $T_k$ the successive times of negative records, $U_0 = 0$ and $U_j = S_j - S_{T_k}$ for $T_k \leq j < T_{k+1}$, and $\mathbb{U}^*$ the process stopped in $a$ for $a$ an observed local score

- If $j = 0$, as $U_{n-1}$ only depends on $X_1$, ..., $X_{n-1}$ and $X_n$ is a Markov chain of order 1 we have

$$\begin{aligned} P_{(i,u)(0,v)} &= P\left[(X_n \leq -U_{n-1}) \cap (X_{n+1} = v) \mid (U_{n-1} = i) \cap (X_n = u)\right] \\ &= P\left[(u \leq -i) \cap (X_{n+1} = v) \mid (U_{n-1} = i) \cap (X_n = u)\right] \\ &= P\left[(u \leq -i) \cap (X_{n+1} = v) \mid (X_n = u)\right], \end{aligned}$$

  Thus $P_{(i,u)(0,v)} = \Lambda_{uv}$ if $u \leq -i$ and 0 else.

- If $1 \leq j \leq a-1$, then

$$\begin{aligned} P_{(i,u)(j,v)} &= P\left[(U_{n-1} + X_n = j) \cap (X_{n+1} = v) \mid (U_{n-1} = i) \cap (X_n = u)\right] \\ &= P\left[(i + u = j) \cap (X_{n+1} = v) \mid (U_{n-1} = i) \cap (X_n = u)\right] \\ &= \Lambda_{uv} \text{ si } j = i + u \quad \text{and 0 else.} \end{aligned}$$

- If $j = a$, we have

$$\begin{aligned} P_{(i,u)(a,v)} &= P\left[(X_n \geq a - U_{n-1}) \cap (X_{n+1} = v) \mid (U_{n-1} = i) \cap (X_n = u)\right] \\ &= P\left[(u \geq a - i) \cap (X_{n+1} = v) \mid (U_{n-1} = i) \cap (X_n = u)\right] \\ &= \Lambda_{uv} \text{ si } i + u \geq a \quad \text{and 0 else.} \end{aligned}$$

**Lemma 2.3** *The distribution of $U_n^*$ is given by*

$$P[U_n^* = j] = \sum_{u,v} \gamma_u \cdot P^n_{(0,u)(j,v)}.$$

From Lemma 2.1, we deduce $P[H_n \geq a] = P[U_n^* = a]$ and using Lemma 2.3 and the explicitation of the $P_{(i,u)(j,v)}$, Theorem 2.1 is proved.

## 2.2 Model for the letters sequence

Let $\Theta$ be the set of letters. We suppose that the sequence $\mathbb{A}$ of these letters is a 1-order Markov chain, with transition matrix $\Lambda = (\Lambda_{\alpha,\beta})_{\alpha,\beta\in\Theta}$ and initial distribution $\mu$. Let:

$$E = \{0, ..., a\} \times \Theta^2 \text{ with } a \in \mathbb{N}. \tag{5}$$

Let us introduce the matrix $Q = \left(Q_{(i,\alpha,\beta),(j,\gamma,\delta)}\right)$, where $(i, \alpha, \beta)$ and $(j, \gamma, \delta)$ are in $E$, defined by

$$\left\{ \begin{array}{l} Q_{(a,\alpha,\beta)(a,\beta,\delta)} = \Lambda_{\beta,\delta} \\ \left\{ \begin{array}{ll} Q_{(i,\alpha,\beta)(0,\beta,\delta)} = \Lambda_{\beta,\delta} & \text{if } s(\beta)+i \leq 0 \\ Q_{(i,\alpha,\beta)(s(\beta)+i,\beta,\delta)} = \Lambda_{\beta,\delta} & \text{if } 1 \leq s(\beta)+i \leq a-1 \\ Q_{(i,\alpha,\beta)(a,\beta,\delta)} = \Lambda_{\beta,\delta} & \text{if } s(\beta)+i \geq a \end{array} \right\} \text{ for } 0 \leq i \leq a-1 \\ Q_{(i,\alpha,\beta),(j,\gamma,\delta)} = 0 \text{ else.} \end{array} \right. \tag{6}$$

We have the following result:

**Theorem 2.2** *The statistic significance of the local score $H_n$ is given by the following formula:*

$$(\forall a \geq 0) \quad P[H_n \geq a] = \sum_{\alpha,\beta,\gamma,\delta} \mu_\alpha \cdot Q^n_{((s(\alpha),0)^+,\alpha,\beta)(a,\gamma,\delta)}.$$

*Proof* Let us denote $S_k$ the partial sums associated with the sequence of the $s(A_k)$: $S_0 = 0$ and $S_k = s(A_1) + ... + s(A_k)$. Consider the sequence of stopping times $T_k$ defined by: $T_0 = 0$ and $T_{k+1} = \inf\{i > T_k;\, S_i - S_{T_k} < 0\}$.

Let $\mathbb{U}$ be the sequence defined by $U_0 = 0$ and for $T_k \leq j < T_{k+1}$, $U_j = S_j - S_{T_k} = s(A_{T_k+1}) + ... + s(A_j)$. We have in particular $U_{T_k} = 0$ for all $k \geq 0$. The sequence $\mathbb{U}$ is positive but not necessarily bounded. As proved in Mercier and Daudin (2001), we have got the following results:

**Lemma 2.4**

$$U_j = \max(U_{j-1} + s(A_j), 0) = (U_{j-1} + s(A_j))^+ \quad \textit{and} \quad H_n = \max_{1 \leq j \leq n} U_j.$$

Consider $\mathbb{U}^*$ the process from $U$ stopped in $a$, where $a$ is in $\mathbb{N}^*$.

$$U^*_j = U_j \text{ if } j < \tau_a \text{ and } U^*_j = a \text{ if } j \geq \tau_a \text{ with } \tau_a = \inf\{j \geq 1\,;\, U_j \geq a\}.$$

In the case of an i.i.d. sequence $\mathbb{A}$ (see Mercier and Daudin 2001), $U^*$ is a Markov chain of order 1 and it is therefore easy to establish the distribution of $U^*_n$, but this is no longer true in the case of a Markovian sequence $\mathbb{A}$. In order to establish the distribution of $U^*_n$, consider the chain $\mathbb{Z} = (Z_n)$ defined by:

$$(\forall\, j \geq 0) \quad Z_{j+1} = (U^*_j, A_j, A_{j+1}),$$

which is of order 1 and for which the set of states is $E$ as defined in (5).

**Lemma 2.5 (Transition matrix of $\mathbb{Z}$)** *$(Z_k)_{k\geq 1}$ is a Markov chain with transition matrix $Q = \left(Q_{(i,\alpha,\beta)(j,\gamma,\delta)}\right)$, with $(i, \alpha, \beta)$ and $(j, \gamma, \delta)$ in $E$, where the $Q_{(i,\alpha,\beta)(j,\gamma,\delta)}$ are given by* (6)*. We have:*

$$Q_{(i,\alpha,\beta)(j,\gamma,\delta)} = P\left[(U^*_n = j) \cap (A_n = \gamma) \cap (A_{n+1} = \delta) \mid (U^*_{n-1} = i) \cap (A_{n-1} = \alpha) \cap (A_n = \beta)\right].$$

**Lemma 2.6 (Distribution of $U^*_n$)**

$$P[U^*_n = k] = \sum_{\alpha,\beta,\gamma,\delta} \mu_\alpha \cdot Q^n_{((s(\alpha),0)^+,\alpha,\beta)(k,\gamma,\delta)}$$

From Lemma 2.4, we deduce $P[H_n \geq a] = P[U^*_n = a]$. Theorem 2.2 is deduced from Lemma 2.6 and the explanation of the $Q_{(i,\alpha,\beta)(j,\gamma,\delta)}$ given by (6).

## 3 Numerical comparisons

### 3.1 Empirical and theorical $P$-values

We simulate 10,000 letter sequences of a given length $n$ on the amino-acid alphabet, using two different models: the independent model where letters are independently and identically distributed, model noted IID, and a Markovian one, noted MC. Parameters of the simulated sequences are derived from a real protein (Human protein 67-kDa keratin cytoskeletal type II). For each sequence of the dataset, the local score is calculated using a given scoring function $s$. The parameters of the different models are also derived from the dataset, model $M_{1-A}$, and from both the dataset and the scoring function $s$ for the models standing on the scoring sequences, model $M_0$ and $M_{1-X}$.

For each observed local score $a$, an empirical $P$-value, noted $p_{\text{emp}}$ is calculated as followed

$$p_{\text{emp}}(a) = P_{\text{emp}}[H_n \geq a] = \frac{N_a}{N}$$

where $N$ is the number of sequences of the dataset ($N = 10,000$) and $N_a$ is the number of sequences of the dataset with a local score equal or up to $a$. The different theorical $P$-values, noted $p_{\text{theo}}$ when the method is not specified, are also derived.

$$\begin{aligned} p_{\text{theo}} &= p_K \text{ for the approximated } P\text{-value of Karlin et al.,} \\ &= p_{M_0} \text{ for the exact } P\text{-value with } M_0 \text{ model,} \\ &= p_{M_{1-X}} \text{ for the exact } P\text{-value with Markovian model on } \mathbb{X}. \end{aligned}$$

Simulating letter sequences assume us to be under the null hypothesis "sequences are ordinary", or "common", and to get every sequence of same length. This last point allows us to estimate an empirical $P$-value: we need to observe realisations of $H_n$, for a fixed length $n$.

In order to evaluate the accuracy of the $P$-values using the Markovian model on letters, noted $p_{M_{1-A}}$, we also use SCOP database and more precisely the old parseable file 1.37 of SCOP, used by Bailey and Gribskov (2002), that contains about 10,000 non-redundant sequences.

$$p_{\text{theo}} = p_{M_{1-A}} \text{ for the exact } P\text{-value with Markovian model on } \mathbb{A}.$$

We cut the end of the sequences to obtain the same length.

### 3.2 The scoring functions

The scoring functions, or score scales, which are used by biologists and rely on rational scores are very definite and quite various (see for example Kyte and Doolittle 1982). Results with rational scores can be deduced from the integer case, but the time of computation is increasing as it is a function of the range of the scores (one can see in Table 1 that the time of computation of the theorical $P$-value for Markovian models is directly linked with this range). In order to limit the global time of computations, we prefer to create scoring functions very similar to that

**Table 1** Complexities in the different models with binary decomposition of the length $n$, with $a$ the observed local score calculated, $R$ the range of the scores and $\nu$ the cardinal of the alphabet: $\nu = 20$ for proteins, and 4 for DNA

| $M_0$ | $M_{1-X}$ | $M_{1-A}$ |
|---|---|---|
| $a \times \log(n)$ | $a \times R \times \log(n)$ | $a \times \nu^2 \times \log(n)$ |

**Table 2** Score functions used for the numerical examples

| –5 | –4 | –3 | –2 | –1 | 0 | +1 | +2 | +3 | +4 | +5 |
|---|---|---|---|---|---|---|---|---|---|---|
| Score function 1 | | | | | | | | | | |
| F, K | A, L P, T | D, R | E, N, W | H, Q, Y | S | C | V | G | M | I |
| Score function 2 | | | | | | | | | | |
| –5 | –4 | –3 | –2 | –1 | 0 | +1 | +2 | +3 | +4 | +5 |
| A, I, W | E, K, M, P | D, G, H | N, T | C, R | L | Q | S | V | Y | F |
| Score function 3 | | | | | | | | | | |
| | | –2 | –1 | 0 | +1 | +2 | | | | |
| | | A, D, N, T | C, E, G, I, L, R | P, Q, W | K, M | F, H, S, V, Y | | | | |
| Score function 4 | | | | | | | | | | |
| | | –2 | –1 | 0 | +1 | +2 | | | | |
| | | K, R | D, E, H, N, Q | G, P, S, T, W, Y | A, C, F, M | I, L, V | | | | |
| Score function 5 | | | | | | | | | | |
| | | –2 | –1 | 0 | +1 | +2 | | | | |
| | | I, L, V | A, C, F, M | G, P, S, T, W, Y | D, E, H, N, Q | K, R | | | | |

proposed by biologists, but with integer scores (see Table 2). The score function 4 corresponds to that proposed by Karlin and Altschul (1990) for an hydrophobic example.

### 3.3 Measures for comparison

Three different measures are calculated to evaluate the possible improvments.

Bailey and Gribskov (2002) proposed a new method for evaluating the $P$-values of the local score for sequence alignment: the PSE ($P$-value slope error). Let $m$ be the least-squares estimation of the slope:

$$\log(p_{\text{theo}}) = m \cdot \log(p_{\text{emp}}) + b,$$

where $p_{\text{emp}}$ and $p_{\text{theo}}$ are defined in Sect. 3.1. They defined PSE by $\text{PSE} = 1 - m$, which gives an indication of the direction and magnitude of the errors. Logarithmic plot has the advantage of focusing the measure on the queue of the distribution.

Mean square error, noted MSE, is also calculated using the log($P$-value). Mean square error between $p_{\text{emp}}$ and $p_{M_{1-X}}$, for example, is given by:

$$\text{MSE}(p_{\text{emp}}, p_{M_{1-X}}) = \frac{1}{\#a} \cdot \sum_{a} \left[\log(p_{\text{emp}}(a)) - \log(p_{M_{1-X}}(a))\right]^2,$$

where $\#a$ is the number of different observed local scores.

We also use the Kullback distance, noted $d_{KL}$. Let $p = (p_1, ..., p_\kappa)$ and $q = (q_1, ..., q_\kappa)$ be two discrete distributions, $d_{KL}(p, q)$ is given by:

$$d_{KL}(p, q) = \sum_{i=1}^{\kappa} p_i \cdot \log_2 \left(\frac{p_i}{q_i}\right).$$

We derive the different distributions, empirical and theorical, from the $P$-values using the obvious equality: $P[H_n = a] = P[H_n \geq a] - P[H_n \geq a+1]$. Note first that the Kullback distance is not symetric, and secondly that we need to cluster the extreme values to avoid null probabilities.

As we will see in Sect. 3.5, the different measures give similar conclusion.

### 3.4 About the programs

We use the algorithm $kiss()$ with a period of $2^{25}$ (see Robert 1996) to simulate our data.

For the exact $P$-values in all three models, a matrix at a given power $n$ corresponding to the length of the sequences has to be calculated. Using a binary decomposition, the complexity of the programs should be as indicated in Table 1.

We do not use the same method to compute the model based on $\mathbb{A}$ because the considered matrix $Q$ (see (6)) is too large. We use the fact that it is also particularly sparse in this model: for example, for $a = 9$ and an alphabet of 20 amino acids, we have a $4,000 \times 4,000$ matrix, and there are at most 20 terms different from zero in each horizontal line. The implementation problems come both from large amount of memory required and the slow execution speed.

Even with such an improved program, the computation is not adapted (exponential growth time with the value of the local score $a$). This results from the fact that the matrices are still large and the implemented structure is not adapted for not so sparse a matrix: the matrix $Q^2$ is not as sparse as $Q$. Critical threshold seems to be about 30% of filling. The main idea for improving the programs is to use the fact that $Q$ is actually built up with blocks which are partially filled with lines of $\Lambda$: the lines of $\Lambda$ are distributed in the different column-blocks defined by the value of $i$ and $j$ for $Q$. Consider the following numerical example with a simple scale $[-1; 0; +1]$, and $a = 2$. The matrix corresponds to an $800 \times 800$ matrix. (The size of which prohibits inclusion in this article.) Thus, as the property can also be seen in the matrix $P$ of the Markovian model on $\mathbb{X}$, we give the numerical example for $P$. With

$$\Lambda = \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.1 & 0.4 & 0.5 \\ 0.33 & 0.33 & 0.34 \end{pmatrix}$$

and $a = 2$, we get:

$$P = \left(\begin{array}{ccc|ccc|ccc} 0.5 & 0.25 & 0.25 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.1 & 0.4 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.33 & 0.33 & 0.34 & 0 & 0 & 0 \\ \hline 0.5 & 0.25 & 0.25 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.1 & 0.4 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.33 & 0.33 & 0.34 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0.25 & 0.25 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0.4 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.33 & 0.33 & 0.34 \end{array}\right)$$

We have few numerical results with the letter model, as it is much more time consuming.

### 3.5 Numerical results

We highlight the real improvment the Markovian model $M_{1-X}$ can achieve compared with the $M_0$ model in Fig. 2. We plot the Kullback distance between the empirical distribution and the exact distribution using model $M_0$, $d_{\mathrm{KL}}(\mathrm{emp}, \mathrm{M}_0)$, versus the Kullback distance between the empirical distribution and that calculated using the Markovian model on the scoring sequence, $d_{\mathrm{KL}}(\mathrm{emp}, M_{1-X})$. The

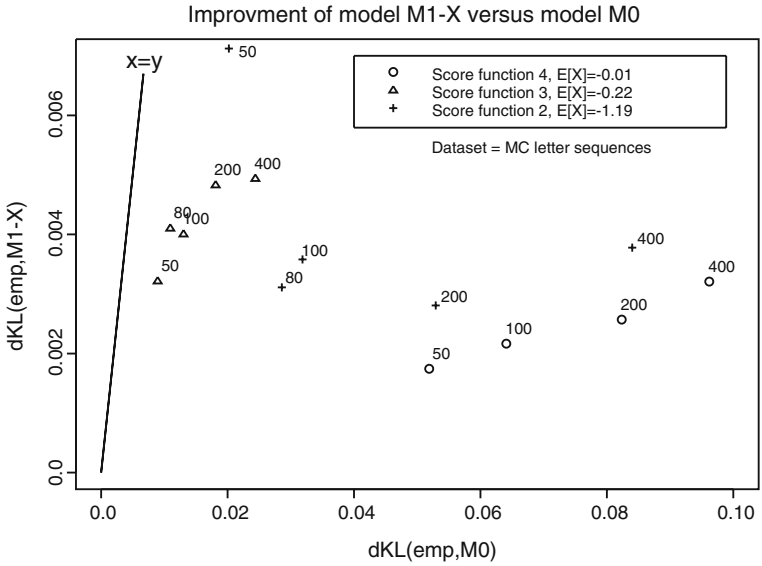


**Fig. 2** Plot–plot of Kullback distances: $d_{\mathrm{KL}}(\mathrm{emp}, \mathrm{M}_{1-\mathrm{X}})$ (resp. $d_{\mathrm{KL}}(\mathrm{emp}, \mathrm{M}_0)$) is the Kullback distance between empirical distribution of $H_n$ and the theorical distribution calculated using the exact method with model $M_{1-X}$ (resp. $M_0$). Letter sequences of the dataset are simulated using the Markovian model. Score functions and parameters of the simulated sequences vary to obtain different mean scores $E[X]$. Numbers close to the points correspond to the length $n$ of the simulated letter sequences. The different scoring functions are given in Table 2

advantage of using the $M_{1-X}$ model for establishing the statistical significance of an observed local score is considerable and clear in this figure: the line $'x = y'$ is close to the vertical axis due to the different scales of the two axes. It seems that there is no particular influence of the mean score $E[X]$. Simulation with $n = 50$ and $E[X] = -1.19$ apart, one can observe that the Kullback distance seems to increase with the length $n$.

The measures (MSE, PSE, $d_{KL}$ defined in Sect. 3.3) used on the different examples which correspond to the logarithmic case ($E[X] < 0$) are summarized in Table 3. We give the average of each measurements between empirical values and the theorical values, using the approximation of Karlin et al. and the exact methods for the $M_0$ and $M_{1-X}$ models. The averages are calculated on 14 values both for IID sequences and MC sequences. These 14 values correspond to the different cases studied for $E[X] < 0$, making length and scoring function used vary. For the IID case, the scoring sequences are also independent and identically distributed, thus we expect to get measures close to zero for the exact method using both the $M_0$ and $M_{1-X}$ models. The corresponding averages allow us to appreciate the accuracy of our method of comparison, in respect to the problem of parameter estimation and to the precision of the measuring. For the Markovian dataset, we can see that even if the scoring sequences are not Markovian, the model $M_{1-X}$ gives very good average measurements, on the same order that of the IID case, and that the improvment of model $M_{1-X}$ over model $M_0$ is of real interest (more than a factor $10^{-1}$ for MSE and $d_{\mathrm{KL}}$ measures). For the linear case, with $E[X] > 0$, we obtain similar results: $\mathrm{MSE}(p_{\mathrm{emp}}, p_{M_0}) = 8.79 \cdot 10^{-2}$ and $\mathrm{MSE}(p_{\mathrm{emp}}, p_{M_{1-X}}) = 5.70 \times 10^{-3}$ for $n = 100$, scoring function number 5 of Table 2, with $E[X] = +0.02$.

The parameters of the Markovian model on the letter sequence, model $M_{1-A}$, are estimated on truncated sequences ($n = 100$) of a non-redundant database (SCOP, old parseable file 1.37). Due to a considerable time calculation, only one case is presented. The scoring function used is the second one given in Table 2 and corresponds to a mean score $E[X]$ equal to $-1.5$. The numerical results are given in Table 4 (see also Fig. 3). Note that time calculation is too excessive for the model $M_{1-A}$ for observed local score $a$ up to 10. Thus the comparison between the different models, including the model $M_{1-A}$, is done only for the small values of $a$ ($a \leq 9$). Both Markovian models achieve a real improvment on real sequences, especially for model $M_{1-A}$.

We also want to compare the exact method with the $M_{1-X}$ model and the approximation of Karlin et al. (see Fig. 4) to indicate a possible length threshold

**Table 3** Mean of the three different measures defined in Sect. 3.3 between the empirical values and the theorical values for independent and identically distributed sequences (IID) and for Markovian sequences (MC)

| | IID generated sequences | | | MC generated sequences | | |
|---|---|---|---|---|---|---|
| | MSE | PSE | $d_{\mathrm{KL}}$ | MSE | PSE | $d_{\mathrm{KL}}$ |
| $K$ | $9.47 \times 10^{-2}$ | 0.182 | $8.32 \times 10^{-1}$ | $3.85 \times 10^{-1}$ | 0.391 | 1.76 |
| $M_0$ | $5.88 \times 10^{-3}$ | 0.026 | $2.66 \times 10^{-3}$ | $9.87 \times 10^{-2}$ | 0.153 | $4.20 \times 10^{-2}$ |
| $M_{1-X}$ | $5.07 \times 10^{-3}$ | 0.022 | $2.74 \times 10^{-3}$ | $7.98 \times 10^{-3}$ | 0.038 | $3.66 \times 10^{-3}$ |

*PSE* $P$-value slope error, *MSE* mean square error, $d_{KL}$ to Kullback distance, $K$ for the approximation of Karlin *et al.*, $M_0$, resp., $M_{1-X}$, exact method using model $M_0$, resp. $M_{1-X}$

**Table 4** *SCOP:* PSE corresponds to $P$-value slope error, $MSE$ to mean square error, $d_{KL}$ to Kullback distance

| | SCOP database ($n = 100$, Scoring function 2) | | | | | |
|---|---|---|---|---|---|---|
| | $a \leq 35$ | | | $a \leq 9$ | | |
| | MSE | PSE | $d_{\mathrm{KL}}$ | MSE | PSE | $d_{\mathrm{KL}}$ |
| $K$ | $2.91 \times 10^{-2}$ | 0.102 | $6.44 \times 10^{-2}$ | $5.16 \times 10^{-4}$ | $-0.169$ | $3.84 \times 10^{-2}$ |
| $M_0$ | $2.89 \times 10^{-2}$ | 0.095 | $6.91 \times 10^{-2}$ | $3.18 \times 10^{-4}$ | 0.267 | $1.94 \times 10^{-2}$ |
| $M_{1-X}$ | $1.58 \times 10^{-2}$ | 0.061 | $6.14 \times 10^{-2}$ | $1.60 \times 10^{-4}$ | 0.138 | $1.19 \times 10^{-2}$ |
| $M_{1-A}$ | – | – | – | $1.18 \times 10^{-4}$ | 0.105 | $9.91 \times 10^{-3}$ |

The numerical results are calculated with the empirical $P$-values and the different theorical values: $K$ for the approximation of Karlin et al., $M_0$ (resp. $M_{1-X}$, $M_{1-A}$) the exact values with model $M_0$ (resp. $M_{1-X}$, $M_{1-A}$) and with the corresponding distributions for Kullback distance. For the $M_{1-A}$ model, the measurements are given for only small observed local score $a$ ($a \geq 9$) and we also give the different measurements for comparison

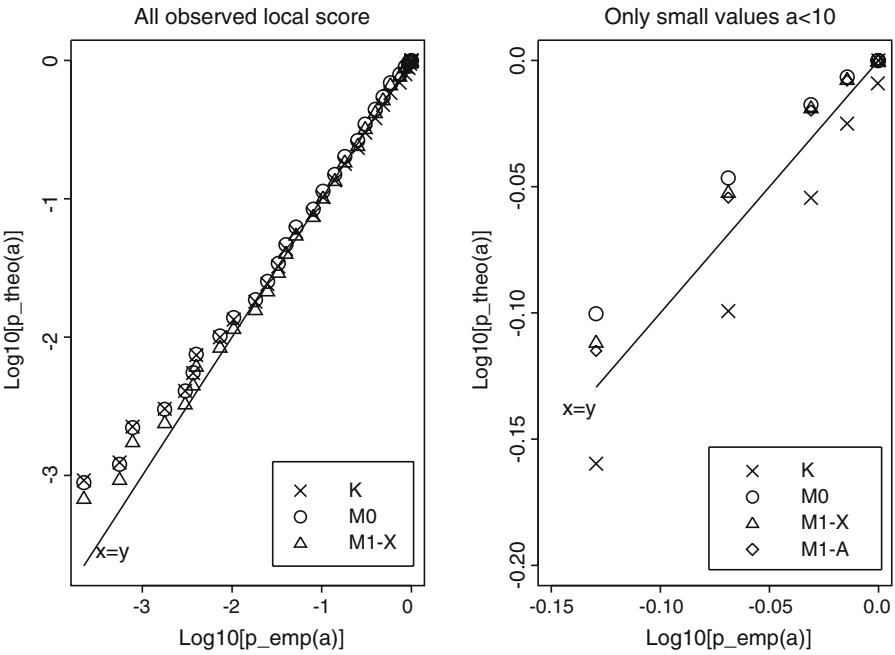


**Fig. 3** SCOP: Plot–plot of logarithm of the $P$-values calculated for all observed local scores $a$ on SCOP database and only for small observed ones. The length $n$ is 100, and the score function used is number 1 in Table 2. Measurements corresponding to this example are given in Table 4

for which the asymptotic approximation is just as accurate as the exact method. This figure clearly shows the asymptotic property of the approximation, but we cannot determine any threshold because the accuracy also greatly depends on the mean score $E[X]$. (For $E[X]$ close to zero, the approximation is not good at all even for sequences with length up to the mean length of real sequences, $\simeq 350$ residues, whereas for strongly negative mean score, see $E[X] = -3.3$ in Fig. 4, Karlin's approximation gives not-so-bad results even for length equal or less than

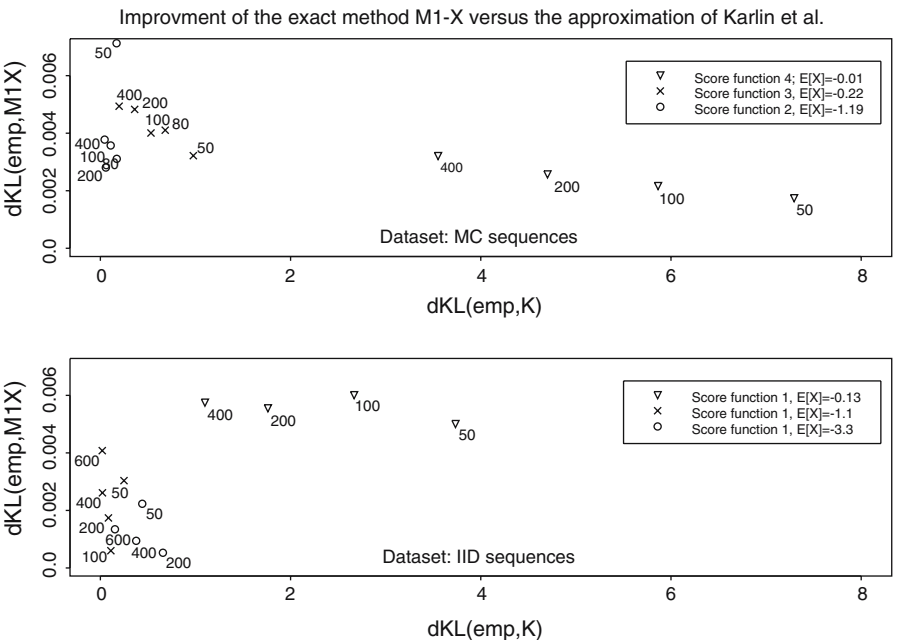


**Fig. 4** Plot–plot of Kullback distances: $d_{\mathrm{KL}}(\mathrm{emp}, \mathrm{K})$ (*resp.* $d_{\mathrm{KL}}(\mathrm{emp}, \mathrm{M}_{1-\mathrm{X}})$) is the Kullback distance between empirical distribution of $H_n$ and the theorical distribution calculated using the approximation of Karlin et al. (*resp.* the exact method with model $M_{1-\mathrm{X}}$). Score functions and parameters of the simulated sequences vary to obtain different mean scores $E[X]$. Numbers close to the points correspond to the length $n$ of the simulated letter sequences. The different scoring functions are given in Table 2

50.) The line $'x = y'$ does not appear in the figures because it is too close to the vertical axis (see the different scales of the two axes): the improvment is very considerable.

## 4 Conclusion and perspectives

As is already known, the asymptotic approach must be used for long sequences, but we have also shown that the exact methods are preferable in the case of mean score average 0, even for not-so-small sequences. Results in this case (accuracy and speed) should be compared with the Brownian approach of Daudin *et al.* (2003).

The Markovian model is performed on the score sequence for scoring function with reasonable range and the numerical results achieved point out the real advantage of this model.

The computation of the exact method with the Markovian model on the letters requires that significant work be done (before it can be efficently utilized). Easy improvments of computation using mathematical properties could be made which allow the important benefit of such a model to be realized (see Nuel 2006).

The comparisons are done with a "mathematical" approach which focuses on the distribution itself. Biologists' use of $P$-value stands more on the rank of the

most exceptional sequences deduced from the $P$-values. Studies should be completed using this aspect. Accuracy of the different methods should also be measured using sensibility and specificity criteria.

**Acknowledgement** We would like to thank the referees for their helpful remarks in clarifying this article.

## References

Arratia, R., Waterman, M.-S. (1994). A phase transition for the score in matching random sequences allowing deletions. *Annals of Applied Probability 4*, 200–225.

Bacro, J.-N., Daudin, J.-J., Mercier, S., Robin, S. (2003). Back to the local score in the algorithmic case: a direct and simple proof. *Annals of the Institute of Statistical Mathematics*, *54(4)*, 748–757.

Bailey T.L., Gribskov M. (2002). Estimating and evaluating the statistics of gapped local-alignment scores. *Journal of Computational Biology*, *9(3)*, 575–593.

Daudin, J.-J., Etienne, M.-P., Valois P. (2003). Asymptotic behavior of the local score of independent and identically distributed random sequences. *Stochastic Processes and their Applications*, *107*, 1–28.

Durbin, R., Eddy, S., Krogh, A., Mitchison, G. (1998). *Biological sequence analysis. probabilistic models of proteins and nucleic acids*. Cambridge: Cambridge University Press.

Ewens W. (2002) *Statistical methods in bioinformatics*. Berlin Heidelberg New York: Springer

Hassenforder, C., Mercier, S. (2003). Exact Distribution for the local score of a Markov chain. *Comptes rendus de l'Académie des sciences*, *336(10)*, 863–868.

Karlin, S., Altschul, S.F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of National Academy of Sciences, USA*, *87*, 2264–2268.

Karlin, S., Dembo, A. (1992). Limit distributions of maximal segmental score among Markov-dependent partial sums. *Advances in Applied Probability*, *24*, 113–140.

Karlin, S., Taylor, H.M. (1981). *A second course in stochastic processes*. New York: Academic.

Kyte, J., Doolittle, R.F. (1982). A simple method for displaying the hydrophatic character of a protein. *Journal of Molecular Biology*, *157*, 105–132.

Mercier, S., Cellier, D., Charlot, F., Daudin, J.-J. (2001). Exact and asymptotic distribution for the local score of one I.I.D. Random sequence. *Lecture Notes in Computational Science*, volume for JOBIM 2000, *2066*, 74–85.

Mercier, S., 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.

Mott, R.F. (2000). Accurate formula for $P$-values of gapped local score and profile alignments. *Journal of Molecular Biology*, *300*, 649–659.

Prum, B. (2001). Probabilités, statistique et génomes. *Matapli*, *64*.

Nuel, G. (2006). Exact distribution of local score using Finite Markov Chain Imbedding: an effective approach. ICAM 2006, Santiago, Chile.

Robert, C. (1996). *Méthodes de Monte Carlo par Chaînes de Markov* (Economica).

Waterman, M.S. (1995). *Introduction to computational biology* London. Chapman and Hall.