Exact distribution of local score using Finite Markov Chain Imbedding: an effective approach. — extended abstract —

GRÉGORY NUEL *

1 Introduction

Local score statistics are used in a wide range of problems related to biological sequences (homology, hydrophobic segments, genetic markers, G-C content, ...). On i.i.d. sequences, Karlin and Altschul 1990 showed that its distribution is well approximated by a Gumbel one. Recently, Hansen 2005 proved that this result could be extended to Markov dependant sequences. Beside these asymptotic approximations, Mercier and Daudin 2001 (i.i.d. case) and Mercier and Hassenforder 2003 proposed to compute the exact distribution of local score using an elegant Finite Markov Chain Imbedding (FMCI) technique. Unfortunately, the practical computations in the exact case are impracticable except in toy-example cases. In this paper, we propose a new approach to deal with FMCI computation which dramatically outperform the previous ones.

2 Recalls

2.1 Definition

We consider $S = S_1, \ldots, S_n$ a sequence of real scores and we define the local score H_n of this sequence by

$$H_n = \max\left\{0, \max_{i,j}\left(\sum_{\ell=i}^{j} S_\ell\right)\right\}$$

which is exactly the highest partial sum score of a sub-sequence of S.

This local score can be computed in $O(n)$ using the auxiliary process

$$U_0 = 0 \quad \text{and for } 1 \leqslant j \leqslant n \quad U_j = \max\left\{0, \max_{i}\left(\sum_{\ell=i}^{j} S_\ell\right)\right\} = \max\left\{0, U_{j-1} + S_j\right\}$$

because we then have $H_n = \max_j U_j$.

Assuming the sequence S is random (Bernoulli or Markov model), we want to compute p-values relative to the event $\{H_n \geqslant a\}$ where $a > 0$.

2.2 Gumbel approximation

According to Karlin and Altschul 1990, if we are in the Bernoulli case and if $\mathbb{E}[S_1] < 0$ we then can find $\lambda, K > 0$ such as

$$P(H_n \geqslant a) \simeq 1 - \exp\left(-Kne^{-\lambda a}\right)$$

*University of Evry, CNRS (8071), INRA (1142), Laboratoire Statistique et Génome, 523, place des terrasses de l'Agora, 91000 Evry, France (nuel@genopole.cnrs.fr)

2.3 Exact distribution

We first consider the case of integer scores on a Bernoulli sequence. In this case, Mercier and Daudin 2001 introduced the FMCI Z defined by

$$Z_0 = 0 \quad \text{and} \quad Z_j = \begin{cases} U_j & \text{if there is no } a \text{ in } U_0, \dots, U_j \\ a & \text{else} \end{cases}$$

(resulting with a sequence of length $n+1$) with 0 as the only starting state and a as the final absorbing state. The transition matrix Π is given by

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

where

$$p(i) = \mathbb{P}(S_1 = i) \quad f(i) = \mathbb{P}(S_1 \leqslant i) \quad g(i) = \mathbb{P}(S_1 \geqslant i) \quad \forall i \in \mathbb{Z}$$

It is easy to see that

$$\mathbb{P}(H_n \geqslant a) = \Pi^n(0, a)$$

which can be computed with complexities $O\left(\log(n) \times a^2\right)$ in memory and $O\left(\log(n) \times a^3\right)$ in time (using a binary decomposition of n).

At the cost of a larger (sparse) transition matrix, this result can be extended to rational scores as well as Markov cases.

3 New method

We first rewrite the transition matrix as

$$\Pi = \left(\begin{array}{ccc|c} & R & & v \\ \hline 0 & \dots & 0 & 1 \end{array}\right)$$

and it is known that for all $n \geqslant 1$ we have

$$\Pi^n = \left(\begin{array}{ccc|c} & R^n & & y^{n-1} \\ \hline 0 & \dots & 0 & 1 \end{array}\right) \quad \text{with} \quad y^{n-1} = \sum_{i=0}^{n-1} R^i v$$

We hence get

Proposition 1. $\mathbb{P}(H_n \geqslant a) = y_0^{n-1}$ where y^{n-1} is computable through the following recurrence relations:

$$x^0 = y^0 = v \quad \text{and, for all } j \geqslant 0 \quad x^{j+1} = Rx^j \quad \text{and} \quad y^{j+1} = y^j + x^j$$

This proposition leads to a new algorithm which complexities is $O(\zeta)$ in memory and $O(\zeta \times n)$ in time, where ζ is the number of non-zero terms in R (equal to a^2 in the worst case but often far smaller).

An asymptotic development could also be derived (allowing to speed up computations when large n are considered).

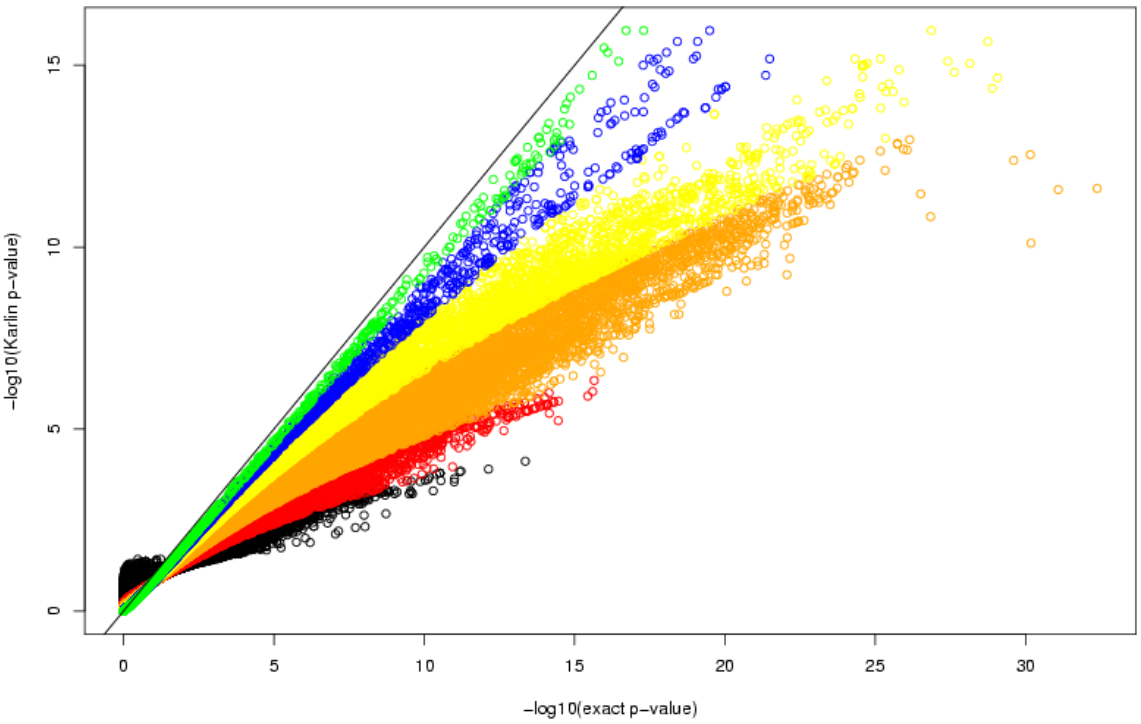

Figure 1: Exact p-value against Karlin ones (in log scale). Color refers to a range of sequence lengths: smaller than 100 in black ($\simeq$ 20 000 sequences), between 100 and 200 in red ($\simeq$ 40 000 sequences), between 200 and 500 in orange ($\simeq$ 90 000 sequences), between 500 and 1 000 in yellow ($\simeq$ 30 000 sequences), between 1 000 and 2 000 in blue ($\simeq$ 6 000 sequences) and greater than 2 000 in green ($\simeq$ 1 000 sequences). The solid line represents $y = x$. Note that 52 sequences for which Karlin's approximations give a p-value of 0.0 have been removed from the data.

4 Results

This new method along with the classical Gumbel approximations have been implemented in a software called pLocalScore. Using this tool, we consider the problem of finding significant hydrophobic segments in the Swissprot database using the Kyte-Doolittle hydrophobic scale (Kyte and Doolittle 1982).

Our program roughly computes 20 exact p-values per seconds and the obtained results are compared to the Gumbel approximations in figure 1. Surprisingly, and in contradiction with the claim of the literature, the Gumbel approximations appear to be very unreliable for the sequences of length smaller than 2,000 (99.5% of the database) both in terms of absolute values as well as ranking (Kendall's tau).

5 Conclusion

Our new method dramatically outperform the previous one allowing for the very first time to compute p-values in a real scale biological study. Using our new algorithms, exact computations could be performed fast enough to be a practical alternative to asymptotic approximations in most cases. Considering the fact that the reliability of the Gumbel approximations seems to be far smaller than expected, we strongly advise to anyone dealing with local score distribution to use these new exact computations whenever it is possible to.

References

Hansen, N. R. 2005. Local alignment of Markov chains. To appear in *Ann. Appl. Prob.*

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

Kyte, J., Doolittle, R. F. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol. 1982* **157(1)**, 105-132.

Mercier, S., Daudin, J.-J. 2001. Exact Distribution for the Local Score of One i.i.d. Random Sequence *J. Comp. Bio.* **8(4)**, 373-380.

Mercier, S., Hassenforder, C. 2003. Exact distribution for the local score of a Markov chain. *C. R. Acad. Sci. Paris* **336 (10)**, 863-868.