

Exact distribution of local score using Finite Markov Chain Imbedding an effective approach

Grégory Nuel ¹

¹Laboratoire Statistique et Génome
CNRS (8071), INRA (1152), University of Evry
France

ICAM 2006

Outline

- 1 Introduction
 - Local score
 - Karlin's approximations
- 2 Results
 - Finite Markov Chain Imbedding (FMCI)
 - New recurrence method
 - Hydrophobic domains in proteins

Definition of local score

Definition

$S = S_1, \dots, S_n$ a sequence of (real) scores

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

is the **local score** of S .

Proposition

If $U_0 = 0$ and for all $1 \leq j \leq n$

$$U_j = \max \left\{ 0, \max_i \left(\sum_{\ell=i}^j S_\ell \right) \right\} = \max\{0, U_{j-1} + S_j\}$$

then $H_n = \max_j U_j$

What to do with local score ?

- finding DNA region with **high G-C content**
($S(g) = S(c) = +1$ and $S(a) = S(t) = -1$)
- finding **hydrophobic domains** in proteins (Kyte-Doolittle hydrophobic scale)
- finding high-scoring regions in **ungapped alignments**
($S(\text{match}) = +1$ and $S(\text{mismatch}) = -1$)
- **efficient replacement** for any "sliding window" method (no window size parameter)

⇒ need of **statistical significance**

Gumble approximations

Theorem (Karlin & Altschul, 1990)

If S is i.i.d. with $\mathbb{E}[S_1] < 0$ it exists $\lambda, K > 0$ such as

$$\mathbb{P}(H_n \geq a) \simeq 1 - \exp(-nKe^{-a\lambda})$$

Example

$$\mathbb{P}(S_1 = +1) = 0.5$$

$$\mathbb{P}(S_1 = -2) = 0.5$$

$$\Rightarrow \mathbb{E}(S_1) = -0.5$$

we get:

$$\lambda = 0.481212$$

$$K = 0.163323$$

FMCI: main result

Theorem (Mercier & Daudin 2001)

In the case of *integer score* on *i.i.d. sequence*, for all $a \geq 0$ we define the *FMCI* Z by $Z_0 = 0$ and

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

Z is a *order 1 Markov chain* with transition matrix Π and

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

Corollary

Using a binary decomposition of n , $\mathbb{P}(H_n \geq a)$ is computable with complexities $O(\log(n) \times a^2)$ in *memory* and $O(\log(n) \times a^3)$ in *time*.

FMCI: transition matrix

Proposition

The **transition matrix** of the FMCI 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

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

A recurrence

Remark

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

Theorem

$\mathbb{P}(H_n \geq a) = [y_n]_0$ where y_n is computable through

$x_0 = y_0 = v$ and, for all $j \geq 0$ $x_{j+1} = R x_j$ and $y_{j+1} = y_j + x_j$
with complexities $O(\zeta)$ in memory and $O(\zeta \times n)$ in time (ζ is the number of **non zero terms in R**).

Asymptotic development

Proposition

If R admits a **diagonal form** we have $\forall n \geq \alpha$

$$\mathbb{P}(H_n \geq a) = \left[\sum_{i=0}^{\alpha-1} R^i v \right]_0 + \lambda^\alpha \frac{(1 - \lambda^{n-\alpha})}{(1 - \lambda)} [R^\infty v]_0 + O(\nu^\alpha)$$

with $R^\infty = \lim_{i \rightarrow \infty} R^i / \lambda^i$, where $0 < \lambda < 1$ is the largest eigenvalue of R and ν is the magnitude of the second largest eigenvalue.

Proposition

If R admits a **diagonal form** we have

$$\lim_{n \rightarrow \infty} \frac{\mathbb{P}(H_{n+2} \geq a) - \mathbb{P}(H_{n+1} \geq a)}{\mathbb{P}(H_{n+1} \geq a) - \mathbb{P}(H_n \geq a)} = \lambda$$

Algorithm to compute local score p-value

x size a real vector, $(p_i)_{i \geq 1}$ and $(\lambda_i)_{i \geq 3}$ real, and i integer

initialization $x = v$, $p_1 = [v]_0$, and $i = 0$

main loop while ($i < n$ and (λ_i) not yet converged $\rightarrow \lambda$)

- $i = i + 1$
- $x = R \times x$ (**sparse product**)
- $p_i = p_{i-1} + x_1$
- $\lambda_i = (p_i - p_{i-1}) / (p_{i-1} - p_{i-2})$ (if defined)

end

- $p = p_i$
- if ($i < n$) then $p = p + (p_i - p_{i-1}) \frac{(1 - \lambda^{n-i})}{(1 - \lambda)}$
- return p

Extension to rational scores

Proposition

We denote by $\mathcal{S} \subset \mathbb{Q}$ is the support of a **rational score** then

$$\mathbb{P}(H_n \geq a) = \mathbb{P}(MH_n \geq Ma)$$

with $M = \min_{i \in \mathbb{N}} \{i\mathcal{S} \subset \mathbb{Z}\}$ and we are back to the **integer case**.

Comparing complexities

We have

method	memory	time
classical	$\log(n) \times (M \times a)^2$	$\log(n) \times (M \times a)^3$
recurrence	$M \times a \times \eta$	$n \times M \times a \times \eta$

with η denoting the cardinal of \mathcal{S} .

Extension to Markov sequences

Theorem (Mercier & Hassendorf 2003)

For *integer score* on *order m Markov sequence*, we define the *FMCI Z* by

$$Z_j = \begin{cases} (S_{j-m}, \dots, S_{j-1}, U_j) & \text{if there is no } a \text{ in } U_0, \dots, U_j \\ f & \text{else} \end{cases}$$

and get the same result than before.

Comparing complexities

method	memory	time
classical	$\log(n) (M \times a \times \eta^m)^2$	$\log(n) (M \times a \times \eta^m)^3$
recurrence	$M \times a \times \eta^{m+1}$	$n \times M \times a \times \eta^{m+1}$

Kyte-Doolittle hydrophobic scale on SwissProt (47.8)

Score distribution

a. a.	F	M	I	L	V	C	W	A	T	G
\mathbb{P} in %	4.0	2.4	5.9	9.6	6.7	1.5	1.2	7.9	5.4	6.9
score	2.8	1.9	4.5	3.8	4.2	2.5	-0.9	1.8	-0.7	-0.4
a. a.	S	P	Y	H	Q	N	E	K	D	R
\mathbb{P} in %	6.9	4.8	3.1	2.3	3.9	4.2	6.6	5.9	5.3	5.4
score	-0.8	-1.6	-1.3	-3.2	-3.5	-3.5	-3.5	-3.9	-3.5	-4.5

Numerical parameters

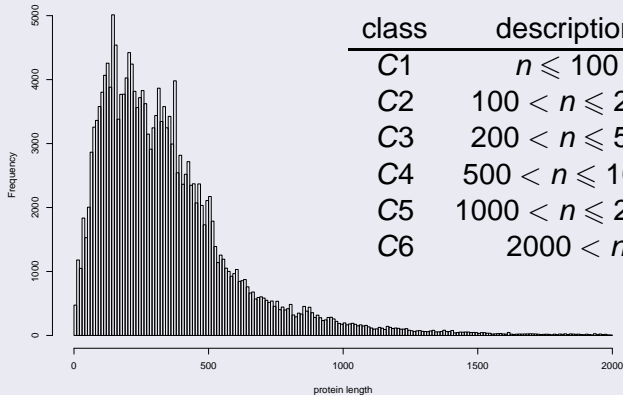
We have $M = 10$, $\eta = 20$, $n \leq 8000$ and

$$\mathbb{E}[S_1] = -0.244$$

so we can use **Karlin's approximations** with

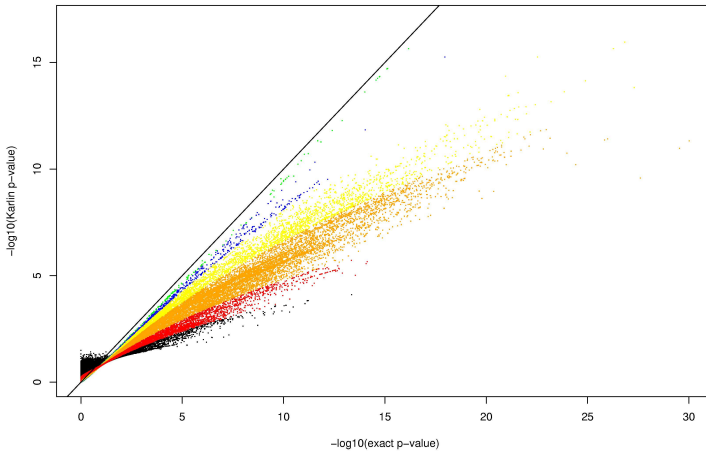
$$\lambda = 5.144775 \times 10^{-3} \quad \text{and} \quad K = 1.614858 \times 10^{-2}$$

Protein length distribution



class	description	freq.
C1	$n \leq 100$	20 000
C2	$100 < n \leq 200$	40 000
C3	$200 < n \leq 500$	90 000
C4	$500 < n \leq 1000$	30 000
C5	$1000 < n \leq 2000$	2 000
C6	$2000 < n$	1 000

Comparison with approximations (1)



Comparison with approximations (2)

predictive accuracy

e-value	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
exact	9 473	7 772	6 271	4 563	3 232	2 348
Karlin	3 417	2 047	1 056	439	195	96
accuracy	34%	26%	17%	10%	6%	4%

rank accordance (Kendall's tau)

n. p-values	all	C1	C2	C3	C4	C5	C6
10	0.30	0.64	0.24	-0.20	0.58	0.64	0.97
50	0.14	0.73	0.50	0.46	0.56	0.78	0.97
100	0.37	0.70	0.67	0.62	0.61	0.80	0.98

Summary

What have we done ?

- **dramatic improvement** of FMCI exact method
- Karlin's approximations **unreliable for 99.5%** of the data
- exact computations are quite fast (**20 p-values per s.**)
- a GPL software available: **pLocalScore**
<http://stat.genopole.cnrs.fr/plocalscore/>

Outlook

- add **Markov case support** in pLocalScore
- add support for **more sophisticated approximations**
- **one reference**: Nuel, 2006, AMB (in revision)