Combinatoire et biologie

Pierre Nicodème

Laboratoire Statistiques et Génomes
CNRS - La Génopole - Évry

`nicodeme@genopole.cnrs.fr`

et

Projet Algorithmes
INRIA - Rocquencourt

IHP, le 23 janvier 2002

Partie I

Mathématiques expérimentales

30 textes aléatoires - 1

mot1 = 100   mot2 = 111

1 1100101011000010010011010000011011011010110100010011001101010101011010011000100011000111
2 11000111
010110111100
0111110101100
3 100110111
00011100
4 0001100011010100000111
5 000010110010100000110111
6 11000011001011011 1
7 10000000111
0111110100
8 101100000100111
9 110100100110101100101101011000111
1111100
10 10100100110111
111011110111100
11 10010111
11011100
01111101111111100
12 01010000111
13 00100111
14 0000000001100110111
15 100111
16 100101011000010011 1
17 010010000111
18 00101100000101010110111
19 100010001001100011 1
20 1101101001101101000000111
21 0101000110100001011001 11
11011111111010110111100

30 textes aléatoires - 2

mot1 = 100   mot2 = 111

1 0001100010111
111011101100
101110100
111010100
0101111101100
111011101111111111010101100
000101011100
000111100
2 1000011011010111
0111100
3 1101100001101000010111
4 100001010100110111
01111100
5 100111
11011111100
6 0101001001100010110010000101100011001010001 0111
7 00000100111
101111100
8 1001010111
9 1100011010101100010111
10 1101100001011010111
11 1001100001101001100110111
12 1001000111
0110111100
13 011000111
010101 1100
14 1001000100001011010111
11011100
00111100
15 0110000100111

30 textes aléatoires - 3

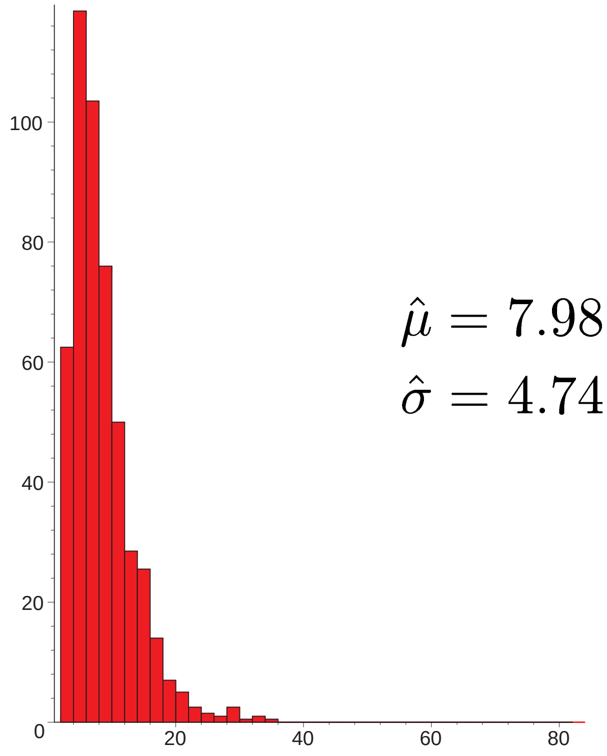
mot1 = 100 mot2 = 111

111101011010101010100
1 101100010110101100111
2 000001000111
00001110100
3 0010001000110110000110010000111
11100
4 110100000000000101100011010110 0111
5 1010011010000010010001101100110111
011100
111011110101101110100
0111111010110101100
6 110001101100010011010101010000010111
7 01001100010001011000011010101000111
8 1100111
1011111110100
9 100000011000111
10 0100101100110011011010111
010111110111101010 0
11 10101010100100101 11
0011011110101 0100
12 0000000100001101100111
13 1001000010111
14 100000010111
15 10000111
0111111110111010100
111100
16 0011010011000110101101 0111
17 1000111
18 0001101001000111
19 10000010111

Attente de la première occurrence

1000 textes aléatoires

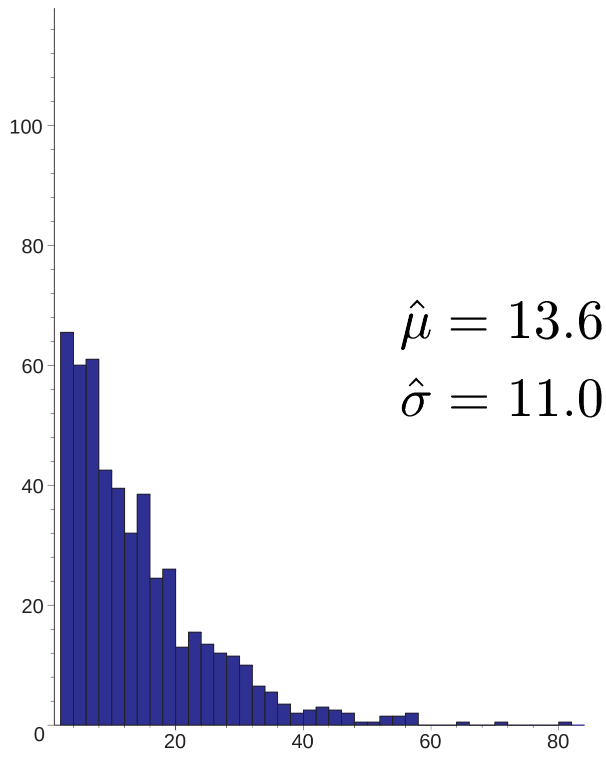
100



$\hat{\mu} = 7.98$

$\hat{\sigma} = 4.74$

111



$\hat{\mu} = 13.67$

$\hat{\sigma} = 11.00$

1 texte de taille 1000

1000

11011000100110110100100011100111101110100101001101
01110100110011110011101000110101011100001110000101
01010110011001111101010111101001110111001000111001
00101000100111000011001111110101110110101011010000
01110000000100001010001010111001001011000000001100
00111101010011111011101110111111011110110001010000
11111010110110010100011110001011001110010100000001
11100001111111111000111000011001101100111010100110
00111110010101100011100110000110010010000110010010
01100101000101100000111000000000010010100101111011
01110100110110010111100101110101011000100100000010
00010001101010111001111101010101000010000011001000
10000010000100011101111011100110000110110000010101
11110111010100110000000111101010001000110011001001
01011010000101111101111111100001010101100111110000
01000011010111110011011011101011001001000000011100
01011001001110110011000000000001000100001001010110
01010100001100001010111110101101101101010011011110
11100100010101110111101110111011101100010001111011
01101101000001010110000100010010101111011101011011

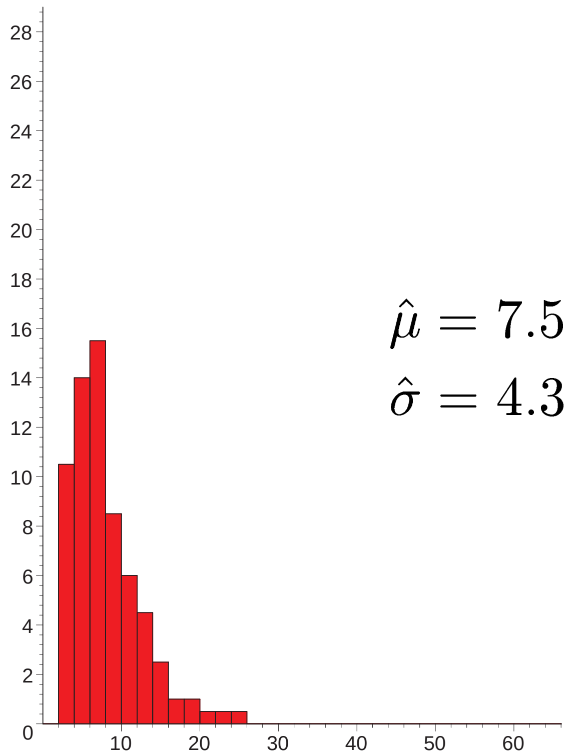
1111

11011000100110110100100011100111101110100101001101
01110100110011110011101000110101011100001110000101
01010110011001111101010111101001110111001000111001
00101000100111000011001111110101110110101011010000
01110000000100001010001010111001001011000000001100
00111101010011111011101110111111011110110001010000
11111010110110010100011110001011001110010100000001
11100001111111111000111000011001101100111010100110
00111110010101100011100110000110010010000110010010
01100101000101100000111000000000010010100101111011
01110100110110010111100101110101011000100100000010
00010001101010111001111101010101000010000011001000
10000010000100011101111011100110000110110000010101
11110111010100110000000111101010001000110011001001
01011010000101111101111111100001010101100111110000
01000011010111110011011011101011001001000000011100
01011001001110110011000000000001000100001001010110
01010100001100001010111110101101101101010011011110
11100100010101110111101110111011101100010001111011
01101101000001010110000100010010101111011101011011

Distance entre les occurrences

texte de taille 1000

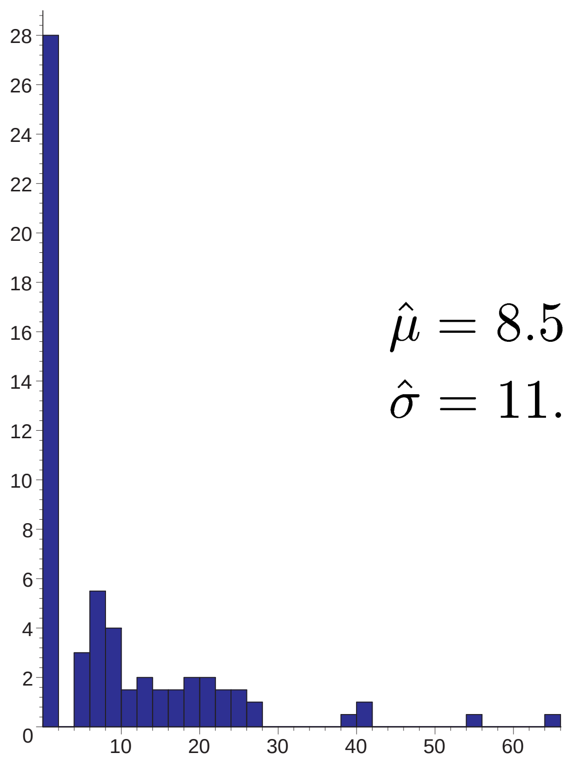
100



$\hat{\mu} = 7.56$

$\hat{\sigma} = 4.31$

111



$\hat{\mu} = 8.53$

$\hat{\sigma} = 11.34$

Que se passe t-il?

- Probabilité d'apparition à une position quelconque

$$\mathbf{P}(100) = \frac{1}{8}$$

$$\mathbf{P}(111) = \frac{1}{8}$$

- MAIS les 111 arrivent souvent GROUPÉS

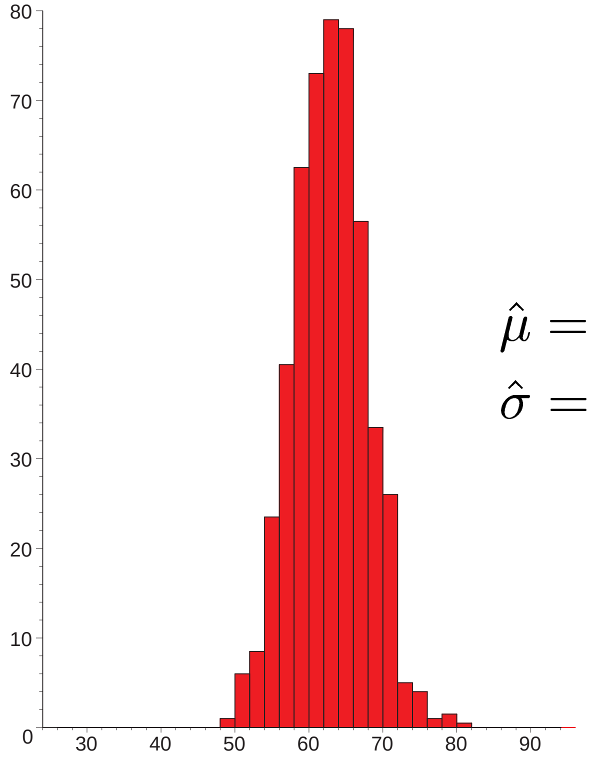
$$\begin{array}{lll} \cdots 0111110 & \quad & \cdots 011110 \\ \quad\;\; 111 & & \quad\;\; 111\cdots \\ \quad\;\;\; 111\cdots & & \end{array}$$

- alors que les 100 ne se recouvrent jamais

Nombre d'occurrences

1000 textes de taille 500

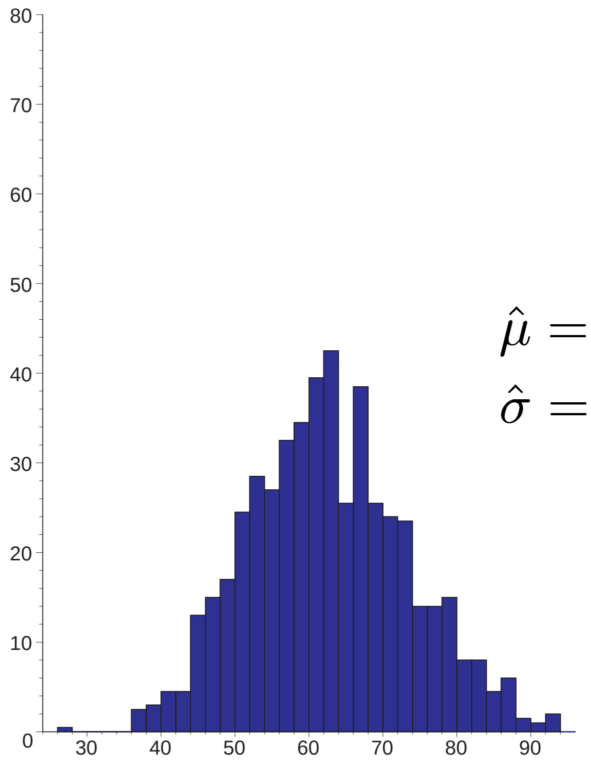
100



$\hat{\mu} = 62.38$

$\hat{\sigma} = 4.90$

111



$\hat{\mu} = 62.07$

$\hat{\sigma} = 10.79$

Partie II

Langages et séries génératrices

L'outil des séries génératrices

suite	série génératrice	
$f_0, f_1, \dots, f_n, \cdots$	$f_0 + f_1 z + \cdots + f_n z^n + \dots$	$= F(z)$
$1, 1, 1, \dots, 1, \dots$	$1 + z + z^2 + \cdots + 1 z^n + \dots$	$= \dfrac{1}{1-z}$
$1, 2, 3, \dots, n, \dots$	$z + 2z^2 + \cdots + n z^n + \dots$	$= \dfrac{z}{(1-z)^2}$
$\dots$	$\dots$	$\dots$

Intérêt:

1. On manipule des fonctions
2. On extrait le coefficient de Taylor d'ordre voulu
   $[z^n]F(z) = f_n$

Exemple:

μ_n espérance du nombre d'occurrences dans un texte de taille n

1. calculer $M(z) = \sum_{n \geq 0} \mu_n z^n$
2. extraire $\mu_n = [z^n]M(z)$

Série génératrice de dénombrement

langage = ensemble de mots

alphabet $\Sigma = \{0, 1\}$

$\Sigma^\star = \epsilon + \Sigma + \Sigma^2 + \cdots + \Sigma^n + \ldots$ tous les mots

$\mathcal{L} \subset \Sigma^\star$

$$\rightarrow \text{série génératrice } L(z) = \sum_{w \in \mathcal{L}} z^{|w|}$$

Exemple

$(001)^\star = \epsilon + 001 + (001)^2 + (001)^3 + \cdots$

$\mathcal{L} = 111 + (001)^\star$

$$\rightarrow L(z) = z^3 + 1 + z^3 + z^6 + \ldots = z^3 + \frac{1}{1 - z^3}$$

s'il n'y a pas d'ambiguïtés,

$$\mathcal{C} = \mathcal{A} \,.\, \mathcal{B} \quad \Rightarrow \quad C(z) = A(z) \times B(z)$$

$$\mathcal{C} = \mathcal{A} + \mathcal{B} \quad \Rightarrow \quad C(z) = A(z) + B(z)$$

$$\mathcal{C} = \mathcal{A}^\star \quad \Rightarrow \quad C(z) = \frac{1}{1 - A(z)}$$

Série génératrice d'un langage

langage = ensemble de mots

alphabet $\Sigma = \{a, b\}$

$\Sigma^\star = \epsilon + \Sigma + \Sigma^2 + \cdots + \Sigma^n + \ldots$ tous les mots

$\mathcal{L} \subset \Sigma^\star$

$F_{\mathcal{L}}(a, b) = \sum_{w \in \mathcal{L}} \text{commute}(w)$

$(aabaa)^\star = \epsilon + aabaa + (aabaa)^2 + (aabaa)^3 + \cdots$

$\mathcal{L} = (aabaa)^\star + bbb \quad \to F_{\mathcal{L}}(a, b) = \frac{1}{1-a^4b} + b^3$

Si $\mathcal{A}.\mathcal{B}$ non-ambigu,

$F_{\mathcal{A}\cdot\mathcal{B}}(a, b) = F_{\mathcal{A}}(a, b) \times F_{\mathcal{B}}(a, b)$

Si $\mathcal{A}$ et $\mathcal{B}$ disjoints

$F_{\mathcal{A}+\mathcal{B}}(a, b) = F_{\mathcal{A}}(a, b) + F_{\mathcal{B}}(a, b)$

Si $\mathcal{A}^\star$ non-ambigu,

$$F_{\mathcal{A}^\star}(a, b) = \frac{1}{1 - F_{\mathcal{A}}(a, b)}$$

Série génératrice sur les probabilités

$\Sigma = \{a, b\} \quad \omega_a = \mathbf{P}(a),\ \omega_b = \mathbf{P}(b)$

$\mathcal{L} \subset \Sigma^\star$

Série génératrice du langage

$M(a,b) = \sum_{\alpha \in \mathcal{L}} \text{commute}(\alpha)$

Série génératrice sur les probabilités

$$\begin{aligned} F(z) &= M(\omega_a z, \omega_b z) \\ &= \sum_{\alpha \in \mathcal{L}} p_\alpha z^{|\alpha|} = \sum \pi_n z^n \end{aligned}$$

p_α proba. du mot α

π_n proba. qu'un mot taille n appartienne à $\mathcal{L}$

Exemple

$\Sigma = \{a, b\} \quad \omega_a = 1/3,\ \omega_b = 2/3 \quad (\epsilon \text{ mot vide})$

$\mathcal{L} = \{\epsilon, aa, ab, ba, aaab\}$

$$\Rightarrow \begin{cases} M(a,b) = 1 + a^2 + 2ab + a^3 b \\ F(z) = M\left(\frac{z}{3}, \frac{2z}{3}\right) = 1 + 0z + \frac{5}{9}z^2 + 0z^3 + \frac{2}{81}z^4 \end{cases}$$

Série génératrice sur les probabilités

cas Bernoulli uniforme

$$\Sigma = \{0, 1\} \quad \mathbf{P}(0) = \mathbf{P}(1) = \frac{1}{2}$$

$$\mathcal{L} \subset \Sigma^\star$$

$$\rightarrow L(z) = \sum_{w \in \mathcal{L}} z^{|w|}$$

$$\rightarrow L_P(z) = \sum_{w \in \mathcal{L}} \left(\frac{1}{2}\right)^{|w|} z^{|w|} = L(z/2) = \sum_{n \geq 0} \pi_n z^n$$

π_n probabilité qu'un mot de taille n appartienne à $\mathcal{L}$

Exemple

$$\mathcal{L} = 111 + (001)^\star$$

$$\begin{aligned} \rightarrow L(z) &= z^3 + 1 + z^3 + z^6 + z^9 + \ldots \\ &= 1 + z^3 + \frac{z^3}{1 - z^3} \end{aligned}$$

$$\rightarrow L_P(z) = 1 + \frac{z^3}{8} + \frac{z^3}{8} + \frac{z^6}{64} + \ldots = 1 + \frac{z^3}{8} + \frac{\frac{z^3}{8}}{1 - \frac{z^3}{8}}$$

Série génératrice de probabilité

X variable aléatoire à valeurs entières positives.

Série génératrice de probabilité $P(z)$

$$P(z) = \sum_{n\geq 0} \mathbf{P}(X = n) z^n = \sum_{n\geq 0} p_n z^n$$

Propriétés

$$P(1) = 1 \qquad \text{(somme des probabilités)}$$

$$\left.\frac{\partial P(z)}{\partial z}\right|_{z=1} = \sum_{n\geq 0} n \times p_n = \mu \qquad \text{(espérance)}$$

$$\left.\frac{\partial}{\partial z} z \frac{\partial P(z)}{\partial z}\right|_{z=1} = \sum_{n\geq 0} n^2 \times p_n = m_{(2)} \qquad \text{(2ème moment)}$$

$$\sigma = \sqrt{m_{(2)} - \mu^2} \qquad \text{(écart-type)}$$

Série génératrice bivariée

X_n variable aléatoire comptant le nombre d'occurrences dans un texte de taille n

Série génératrice bivariée $F(z,u)$

$$F(z,u) = \sum_{n\geq 0}\sum_{k\geq 0}\mathbf{P}\big(X_n = k\big)u^k z^n = \sum_{n,k} f_{n,k}u^k z^n$$

$f_{n,k}$ proba. qu'un texte de taille n ait k matches

$$\left.\frac{\partial F(z,u)}{\partial u}\right|_{u=1} = \sum_{n\geq 0}\mu_n z^n \qquad \text{espérances}$$

$$\left.\frac{\partial}{\partial u}u\frac{\partial F(z,u)}{\partial u}\right|_{u=1} = \sum_{n\geq 0} m_{(2)n} z^n \qquad \text{2ème moments}$$

Partie III

Approche combinatoire

1 mot - modèle Bernoulli

h mot considéré (longueur m, proba $p = \mathbf{P}(h)$)

Polynome d'autocorrélation

$h = ababa$

$$\begin{array}{l} ababa \\ ababa| \\ \quad ababa \\ \qquad ababa \end{array}$$

$\mathcal{A} = \{\epsilon, ba, baba\}$

$A(z) = 1 + \omega_a \omega_b z^2 + \omega_a^2 \omega_b^2 z^4$

$\mathcal{A} = \{\ w, \quad h.w = u.h \quad \text{et} \quad |w| < |h|\ \}$

Langages utilisés

First $\mathcal{F} = \{\ w = u.h \quad \text{et} \quad \nexists r, s,\ w = r.h.s\ \}$

$aaaaaababa \subset \mathcal{F}, \quad bbbbbabababa \not\subset \mathcal{F}$

Minimal $\mathcal{M}$

$= \{\ w, \quad h.w = u.h \ \ \text{et} \ \ \nexists r, s,\ h.w = r.h.s\ \}$

$ababa \qquad\qquad\qquad\qquad ababa$

$aaaaababa \ \subset \mathcal{M} \qquad\qquad babbbbbbbbababa \ \not\subset \mathcal{M}$

$ababa$

$ba \subset \mathcal{M}$

Tail $\mathcal{T} = \{w, \quad \nexists r, s,\ h.w = r.h.s\}$

$ababa \qquad\qquad\qquad\qquad ababa$

$aabbbabbbbbb \subset \mathcal{T} \qquad\qquad babbbbbbbbb \not\subset \mathcal{T}$

Not $\mathcal{N} = \Sigma^\star - \Sigma^\star.h.\Sigma^\star = \{w,\ \nexists r, s,\ w = r.h.s\}$

$$\Sigma^\star = \mathcal{N} + \mathcal{F}.\,(\mathcal{M})^\star.\mathcal{T} \ \Rightarrow\ \mathcal{L}_u = \mathcal{N} + \mathcal{F}u.\,(\mathcal{M}u)^\star.\mathcal{T}$$

Equations sur les langages

(I) $\Sigma^\star = \mathcal{T} + \mathcal{M}\Sigma^\star$

$$h.\Sigma^\star = h.\mathcal{T} + h.\mathcal{M}\Sigma^\star$$

mot commencant par h

- 1 seule occurrence de $h \quad \Rightarrow \quad h.\mathcal{T}$
- plusieurs occurrences de $h \quad \Rightarrow \quad h.\mathcal{M}.\Sigma^\star$

(II) $\Sigma^\star h = \mathcal{F}.\mathcal{A} + \mathcal{F}.\Sigma^\star.h$

(remarque $\epsilon \subset \mathcal{A}$)

mot finissant par h

- première occurrence h chevauche dernière (ou seule occurrence) $\quad \Rightarrow \mathcal{F}.\mathcal{A}$
- première occurrence ne chevauche pas dernière $\quad \Rightarrow \mathcal{F}.\Sigma^\star.h$

Equations sur les langages (suite)

(III) $\mathcal{M}^+ = \Sigma^\star.h + \mathcal{A} - \epsilon$

$h.\mathcal{M}^+ = h.\Sigma^\star.h + h.(\mathcal{A} - \epsilon)$

- $h.\mathcal{M}^+$, $\Rightarrow$ mots commençant et finissant par h
- $h.\Sigma^\star.h$, $\Rightarrow$ 1ere et dernière occurrence non recouvrantes
- $h.(\mathcal{A} - \epsilon)$ $\Rightarrow$ 1ere et dernière occurrence recouvrantes

(IV) $\mathcal{N}.\sigma = \mathcal{F} + \mathcal{N} - \epsilon$

ajouter une lettre à mot de $N \Rightarrow$

- crée un match $\Rightarrow \mathcal{F}$
- ne crée pas de match $\Rightarrow \mathcal{N}$

(mot vide ϵ interdit)

Langages $\Rightarrow$ séries génératrices

$$
\begin{array}{clcl}
\text{(I)} & \Sigma^\star = \mathcal{T} + \mathcal{M}\Sigma^\star & \Rightarrow & \frac{1}{1-z} = T(z) + \frac{M(z)}{1-z} \\
\text{(II)} & \Sigma^\star h = \mathcal{F}.\mathcal{A} + \mathcal{F}.\Sigma^\star.h & \Rightarrow & \frac{p_h z^m}{1-z} = F(z)\left(A(z) + \frac{p_h z^m}{1-z}\right) \\
\text{(III)} & \mathcal{M}^+ = \Sigma^\star.h + \mathcal{A} - \epsilon & \Rightarrow & \frac{M(z)}{1-M(z)} = \frac{p_h z^m}{1-z} + A(z) - 1 \\
\text{(IV)} & \mathcal{N}.\Sigma = \mathcal{F} + \mathcal{N} - \epsilon & \Rightarrow & zN(z) = F(z) + N(z) - 1
\end{array}
$$

Ajouter la marque u après chaque match $\Rightarrow$
langage marqué $\mathcal{L}$

$$\mathcal{L} = \mathcal{N} + \mathcal{F}.u(\mathcal{M}.u)^\star\mathcal{T}$$

Traduction en série génératrice

$$L(z, u) = N(z) + \frac{uF(z)T(z)}{1 - uM(z)}$$

Résolution du système

$$F(z) = \frac{pz^m}{pz^m + (1-z)A(z)}$$

$$T(z) = \frac{1}{pz^m + (1-z)A(z)}$$

$$N(z) = \frac{A(z)}{pz^m + (1-z)A(z)}$$

$$M(z) = 1 + \frac{z-1}{pz^m + (1-z)A(z)}$$

$$L(z,u) = \frac{1}{1 - z + pz^m \dfrac{1-u}{u + (1-u)A(z)}}$$

Application aux mots 100 et 111

Modèle Bernoulli uniforme

Mot 100

100

100|

$\mathcal{A}_{100} = \{\epsilon\}$ $\qquad A_{100}(z) = 1$

Mot 111

111

111|

 111

  111

$\mathcal{A}_{111} = \{\epsilon, 1, 11\}$ $\qquad A_{111}(z) = 1 + \frac{z}{2} + \frac{z^2}{4}$

Première occurrence
et distance entre occurrences

	100	111
$F(z)$	$\dfrac{z^3}{8-8z+z^3}$	$\dfrac{z^3}{8-4z-2z^2-z^3}$
μ_F	8	14
σ_F	$2\sqrt{6} \approx 4.90$	$\sqrt{142} \approx 11.92$
$M(z)$	$\dfrac{z^3}{8-8z+z^3}$	$\dfrac{4z-2z^2-z^3}{8-4z-2z^2-z^3}$
μ_D	8	8
σ_D	$2\sqrt{6} \approx 4.90$	$2\sqrt{30} \approx 10.95$

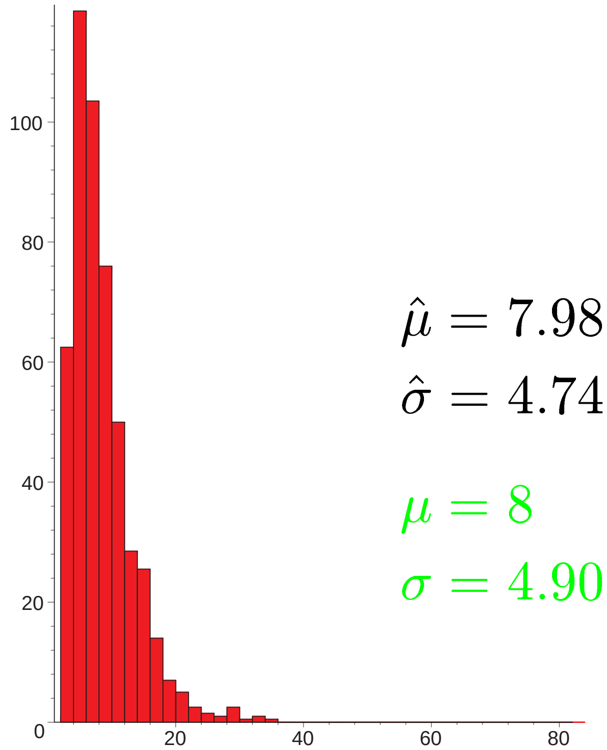
Nombre d'occurrences

	100	111
$L(z,u)$	$\dfrac{1}{1-z+(1-u)\dfrac{z^3}{8}}$	$\dfrac{1+\dfrac{1-u}{8}(4z+2z^2)}{1-z-\dfrac{1-u}{8}(4z-2z^2-z^3)}$
$\mu(z)$	$\dfrac{z^3}{8(1-z)^2}$	$\dfrac{z^3}{8(1-z)^2}$
$m_{(2)}(z)$	$\dfrac{z^3}{8(1-z)^2}-\dfrac{z^3}{32(1-z)^3}$	$\dfrac{8z^3-4z^5-2z^6}{32(1-z)^3}$
μ_n	$\dfrac{n-2}{8}$	$\dfrac{n-2}{8}$
σ_n	$\dfrac{\sqrt{3n}}{8}$	$\dfrac{\sqrt{15n-40}}{8}$
μ_{500}	$\dfrac{249}{4}=62.25$	$\dfrac{249}{4}=62.25$
σ_{500}	$\dfrac{\sqrt{1500}}{8}\approx 4.84$	$\dfrac{\sqrt{7460}}{8}\approx 10.80$

Attente de la première occurrence

1000 textes aléatoires

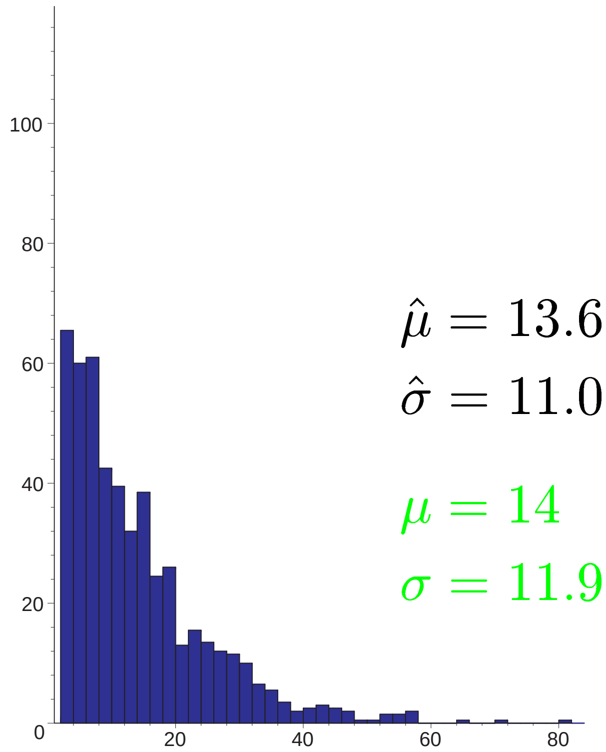
100



$$\hat{\mu} = 7.98$$
$$\hat{\sigma} = 4.74$$

$$\mu = 8$$
$$\sigma = 4.90$$

111



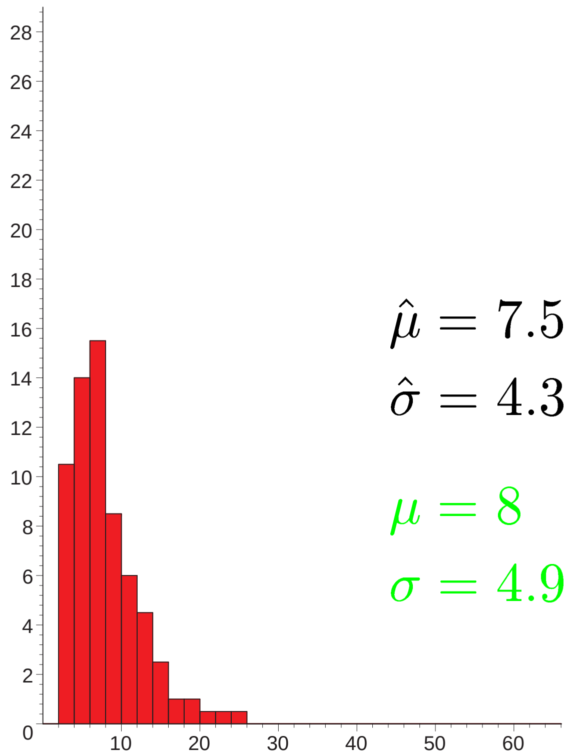
$$\hat{\mu} = 13.67$$
$$\hat{\sigma} = 11.00$$

$$\mu = 14$$
$$\sigma = 11.92$$

Distance entre les occurrences

texte de taille 1000

100



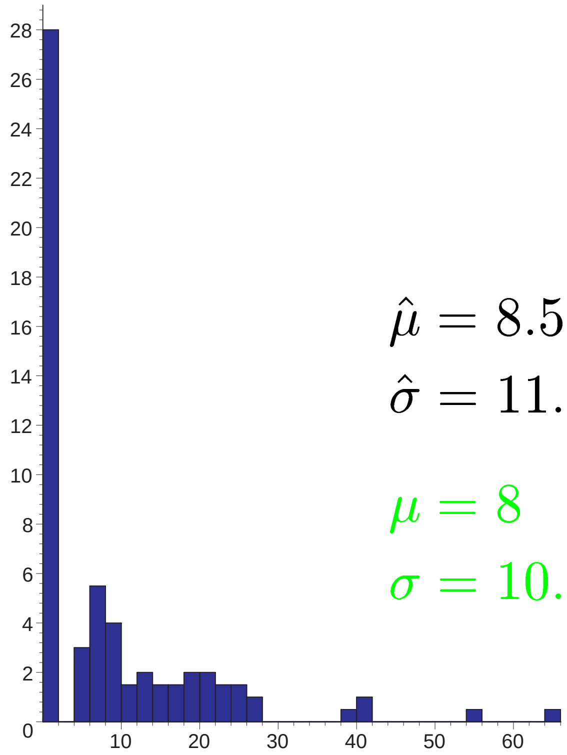
$\hat{\mu} = 7.56$

$\hat{\sigma} = 4.31$

$\mu = 8$

$\sigma = 4.90$

111



$\hat{\mu} = 8.53$

$\hat{\sigma} = 11.34$

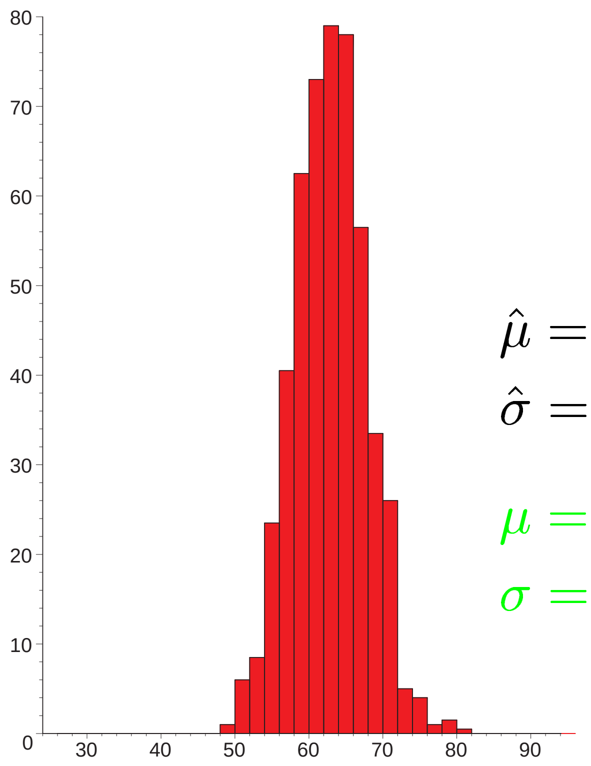
$\mu = 8$

$\sigma = 10.95$

Nombre d'occurrences

1000 textes de taille 500

100



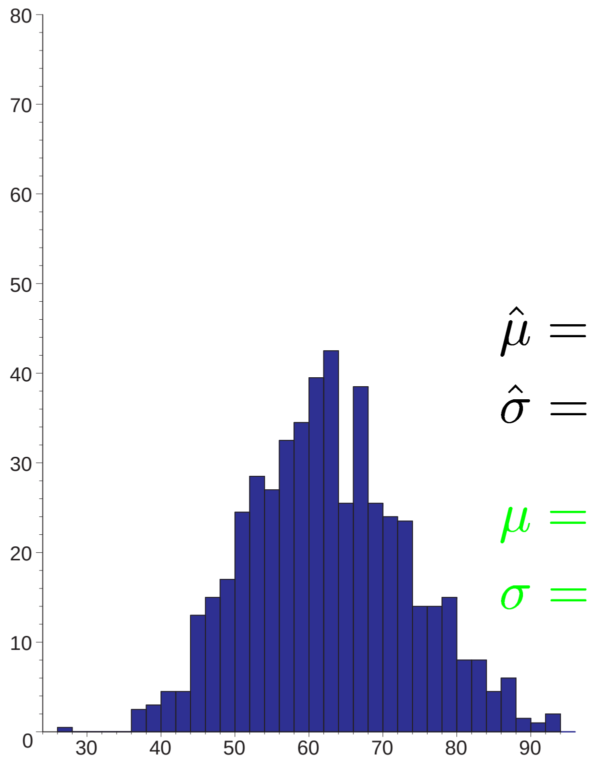
$\hat{\mu} = 62.38$

$\hat{\sigma} = 4.90$

$\mu = 62.25$

$\sigma = 4.84$

111



$\hat{\mu} = 62.07$

$\hat{\sigma} = 10.79$

$\mu = 62.25$

$\sigma = 10.80$

"Multi"- mots

$H = \{h_1, h_2, \ldots, h_m\}$

$h_i = h_{i,1} h_{i,2} \ldots h_{i,l_i}$

$\mathcal{A}_{i,j} = \{w,\ h_i.w = u.h_j,\ \text{et}\ \ \nexists r, s, t \quad h_i.w = s.h_r.t\}$

$\Rightarrow \mathbb{A}(z) = (A_{i,j}(z))$

Exemple $\quad H = \{aab, abaa\}$

$$\mathbb{A}(z) = \begin{pmatrix} 1 & \pi_a^2 z^2 \\ \pi_b z & 1 + \pi_a^2 \pi_b z^3 \end{pmatrix}$$

$\mathcal{F}_i, \mathcal{M}_{i,j}, \mathcal{T}_i \ \Rightarrow\ F_i(z), M_{i,j}(z), T_i(z)$

$\quad \Rightarrow \mathbf{F}(z, u_1, \ldots, u_m), \mathbb{M}(z, u_1, u_2, \ldots, u_m), \mathbf{T}(z)$

Partie IV

Approche algorithmique par automates

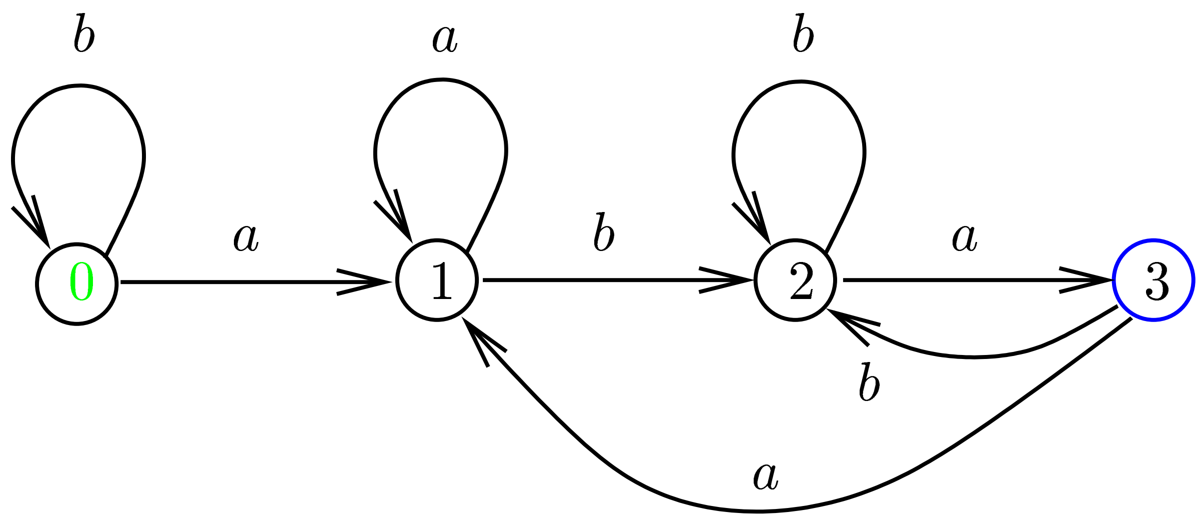
Occurrences des mots ab^+a

$b^+ = b + bb + bbb + bbbb + \ldots$

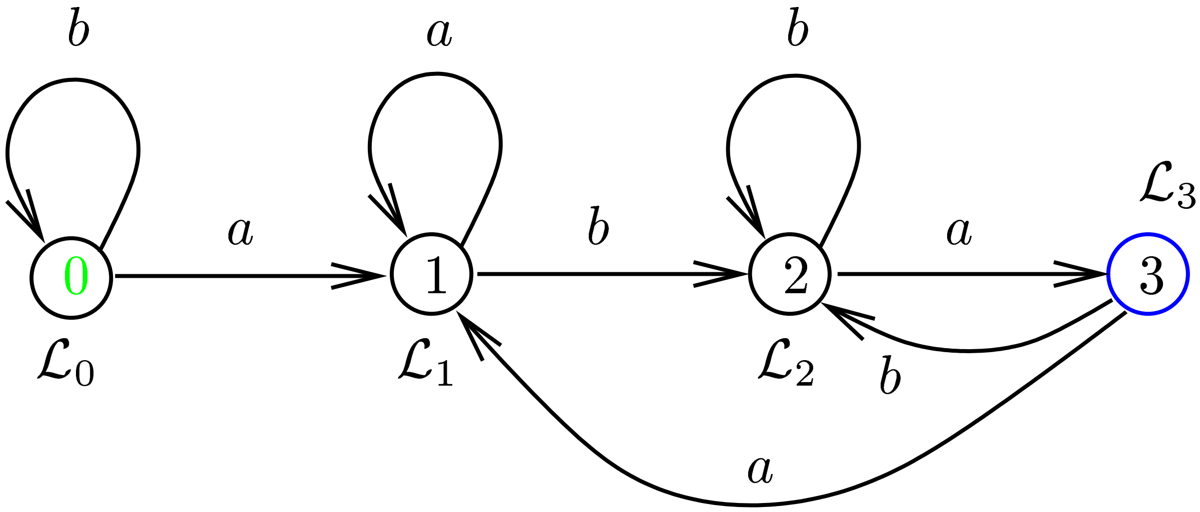
On considére les textes $(a+b)^\star ab^+a$

On construit

l'automate qui reconnaît ces textes



L'algorithme de Chomski et Schützenberger



$$\mathcal{L}_0 = b.\mathcal{L}_0 + a.\mathcal{L}_1 \qquad \mathcal{L}_1 = a.\mathcal{L}_1 + b.\mathcal{L}_2$$
$$\mathcal{L}_2 = b.\mathcal{L}_2 + a.\mathcal{L}_3 \qquad \mathcal{L}_3 = a.\mathcal{L}_1 + b.\mathcal{L}_2 + \epsilon$$

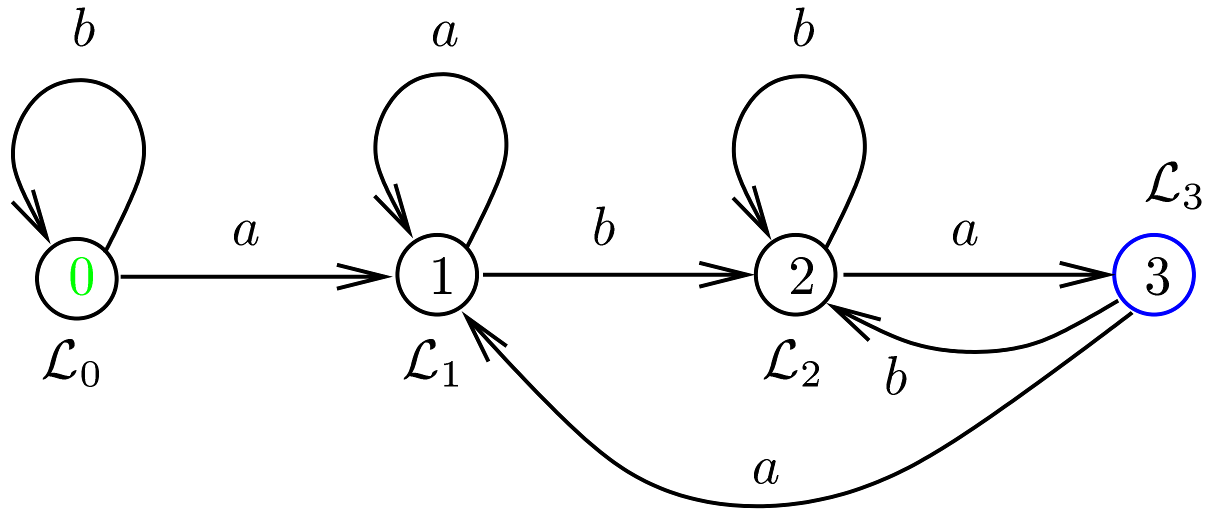
$$L_0(z) = zL_0(z) + zL_1(z) \quad L_1(z) = zL_1(z) + zL_2(z)$$
$$L_2(z) = zL_2(z) + zL_3(z) \quad L_3(z) = zL_1(z) + zL_2(z) + 1$$

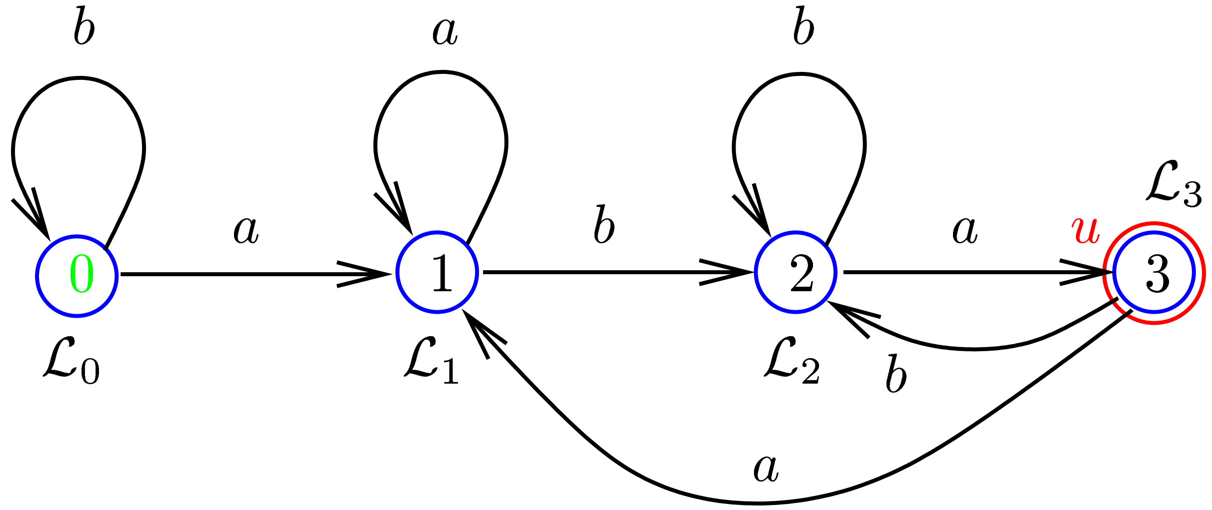
On résout le système

$$L_0(z) = \frac{z^3}{1 - 3z + z^2}$$
$$= z^3 + 3z^4 + 7z^5 + 15z^6 + 31z^7 + 63z^8 + 127z^9 + \dots$$

Nombre d'occurrences de $ab^\star a$

$aaaabba\color{red}{u}\color{black}{ba}\color{red}{u}\color{black}{aaaaabbbbbba}\color{red}{u}\color{black}{aaaaaabbb}$





$$\mathcal{L}_0 = b.\mathcal{L}_0 + a.\mathcal{L}_1 + \epsilon \qquad \mathcal{L}_1 = a.\mathcal{L}_1 + b.\mathcal{L}_2 + \epsilon$$

$$\mathcal{L}_2 = b.\mathcal{L}_2 + a.u.\mathcal{L}_3 + \epsilon \qquad \mathcal{L}_3 = a.\mathcal{L}_1 + b.\mathcal{L}_2 + \epsilon$$

$$L_0(z,u) = zL_0 + zL_1 + 1 \qquad L_1(z,u) = zL_1 + zL_2 + 1$$

$$L_2(z,u) = zL_2 + zuL_3 + 1 \qquad L_3(z,u) = zL_1 + zL_2 + 1$$

Séries génératrices du nombre d'occurrences de $ab^\star a$

$$L_0(z,u) = zL_0 + zL_1 + 1 \qquad L_1(z,u) = zL_1 + zL_2 + 1$$
$$L_2(z,u) = zL_2 + zuL_3 + 1 \qquad L_3(z,u) = zL_1 + zL_2 + 1$$

On résout le système

$$L_0(z,u) = \frac{1 - z + (1-u)z^2}{1 - 3z + 2z^2 + (1-u)(z^2 - z^3)}$$

Modéle Bernoulli uniforme

$$F(z,u) = \sum_{n,k} \mathbf{P}(X_n = k) u^k z^n = L_0\left(\frac{z}{2}, u\right)$$
$$= \frac{1 - \frac{z}{2} + (1-u)\frac{z^2}{4}}{1 - \frac{3z}{2} + \frac{2z^2}{4} + (1-u)\left(\frac{z^2}{4} - \frac{z^3}{8}\right)}$$

Espérance, écart-type

$$F(z,u) = \frac{1 - \frac{z}{2} + (1-u)\frac{z^2}{4}}{1 - \frac{3z}{2} + \frac{2z^2}{4} + (1-u)\left(\frac{z^2}{4} - \frac{z^3}{8}\right)}$$

$$\mu(z) = \sum_{n \geq 0} \mu_n z^n = \left.\frac{\partial F(z,u)}{\partial u}\right|_{u=1} = \frac{z^3}{4(2-z)(1-z)^2}$$

$$= -\frac{1}{4} - \frac{1}{1-z} + \frac{1}{4(1-z)^2} + \frac{1}{1-\frac{z}{2}}$$

$$\Rightarrow \mu_n = \frac{n}{4} - \frac{3}{4} + \frac{1}{2^n}$$

$$m_{(2)}(z) = \left.\frac{\partial}{\partial u} u \frac{\partial F(z,u)}{\partial u}\right|_{u=1} = \frac{z^3(2-2z+z^2)}{8(2-z)(1-z)^3}$$

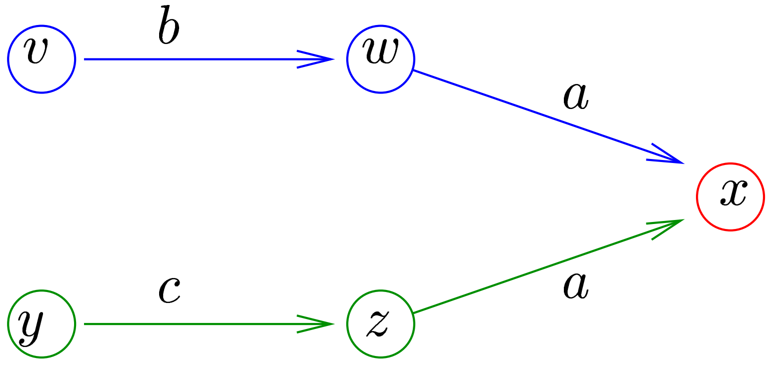
$$= \frac{3}{8} + \frac{z}{8} + \frac{1}{1-z} - \frac{1}{2(1-z)^2} + \frac{1}{8(1-z)^3} - \frac{1}{1-\frac{z}{2}}$$

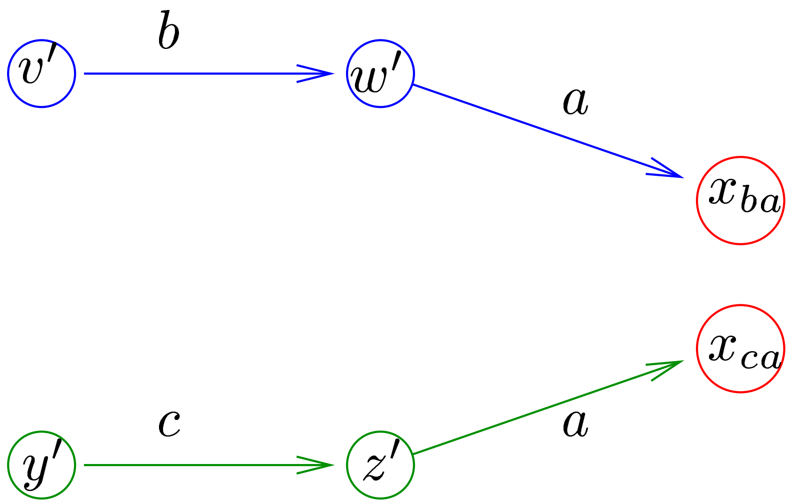
$$\Rightarrow m_{(2)n} = \frac{n^2}{16} - \frac{5n}{16} + \frac{5}{8} - \frac{1}{2^n}$$

$$\Rightarrow \sigma_n = \sqrt{m_{(2)n} - \mu_n^2} = \sqrt{\frac{n}{16} + \frac{1}{16}} + O\left(\frac{1}{2^n}\right)$$

Algorithme Bernoulli vers Markov

automate Bernoulli



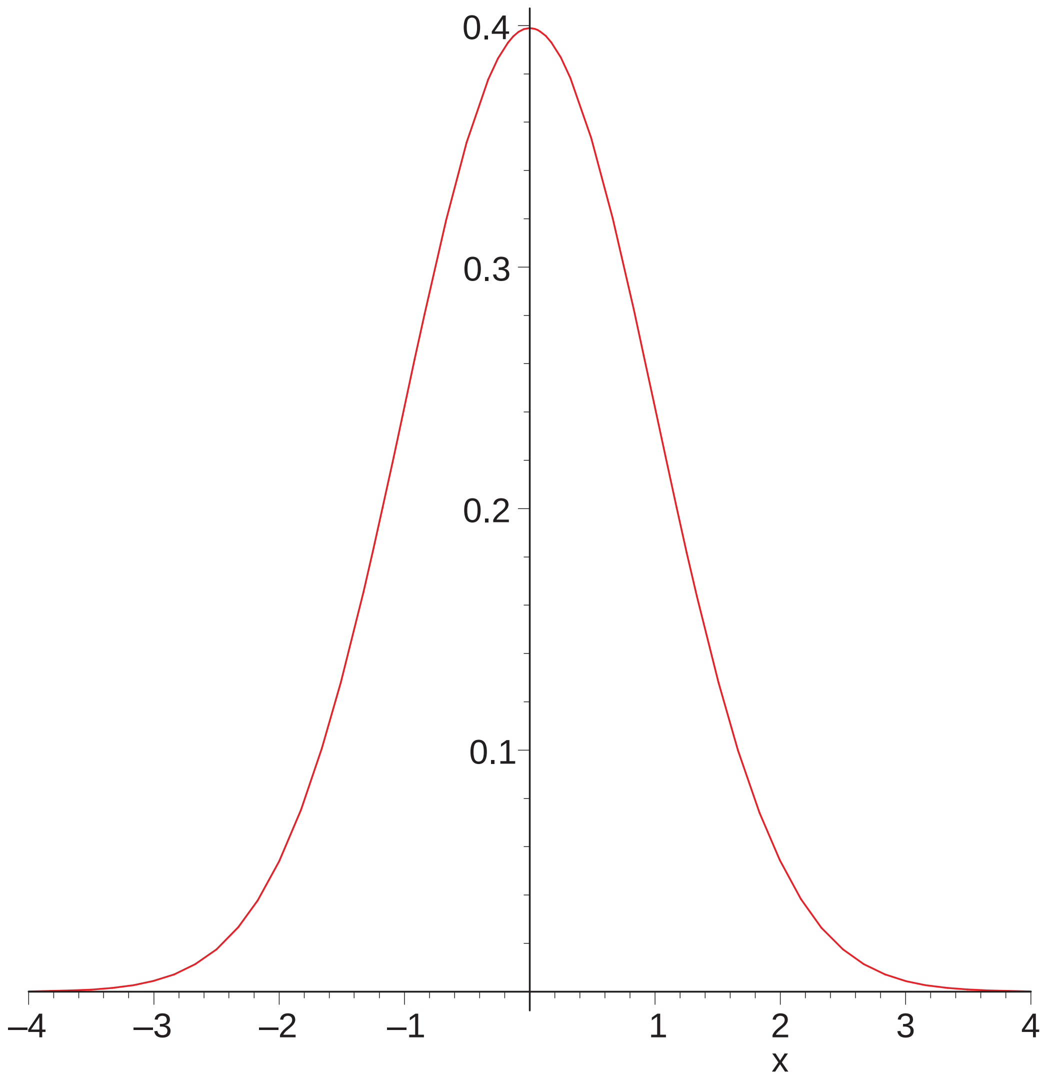


automate Markov (ordre 2)

Partie V

Distribution limite du nombre d'occurrences

Loi normale



$$\mathcal{N} \quad \text{densité } \mathbf{n}(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

$$\int_{-\infty}^{\infty} e^{-x^2/2} dy \int_{-\infty}^{\infty} e^{-y^2/2} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}} dx dy$$

$$= \int_{r=0}^{\infty} \int_{\theta=0}^{2\pi} r e^{-r^2/2} dr d\theta = 2\pi$$

Transformée de Laplace

X variable aléatoire

Transformée de Laplace $\quad \mathbf{L}(X) = \mathbf{E}(e^{tX})$

X loi discrête, $\quad \mathbf{P}(X = k) = p_k$

$$P(z) = \sum_{k \geq 0} p_k z^k \quad \Rightarrow \quad \mathbf{L}(X) = \sum_{k \geq 0} p_k e^{tk} = P(e^t)$$

$\mathcal{N}$ loi normale

$$\begin{aligned}\mathbf{L}(\mathcal{N}) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x^2-2tx)/2} dx = e^{t^2/2}\end{aligned}$$

Théorème de continuité de Lévy

si pour $t \in [-\alpha, +\alpha] \quad \lim_{n\to\infty} \mathbf{E}(e^{tX_n}) = \mathbf{L}(\mathcal{N}) = e^{t^2/2}$

alors $X_n \xrightarrow{d} \mathcal{N}$ (convergence en distribution)

Mots $ab^\star a$ - Loi limite

$$F(z,u) = \frac{1 - \frac{z}{2} + (1-u)\frac{z^2}{4}}{1 - \frac{3z}{2} + \frac{2z^2}{4} + (1-u)\left(\frac{z^2}{4} - \frac{z^3}{8}\right)}$$

$$= \frac{1-u}{u\left(1-\frac{z}{2}\right)} + \frac{1+\sqrt{u}}{u\left(1 - z\frac{1+\sqrt{u}}{2}\right)} + \frac{1-\sqrt{u}}{u\left(1 - z\frac{1-\sqrt{u}}{2}\right)}$$

$$[z^n]F(z,u) = \frac{1}{u}\left(\frac{1+\sqrt{u}}{2}\right)^{n+1} + O\left(\frac{1}{2^n}\right)$$

On considére la loi normalisée $\dfrac{X_n - \mu_n}{\sigma_n}$

$$\Phi(t) = \mathbf{E}\left(e^{\dfrac{t(X_n-\mu_n)}{\sigma_n}}\right) = \exp\left(-\frac{\mu_n t}{\sigma_n}\right)\mathbf{E}\left(\exp\left(\frac{tX_n}{\sigma^n}\right)\right)$$

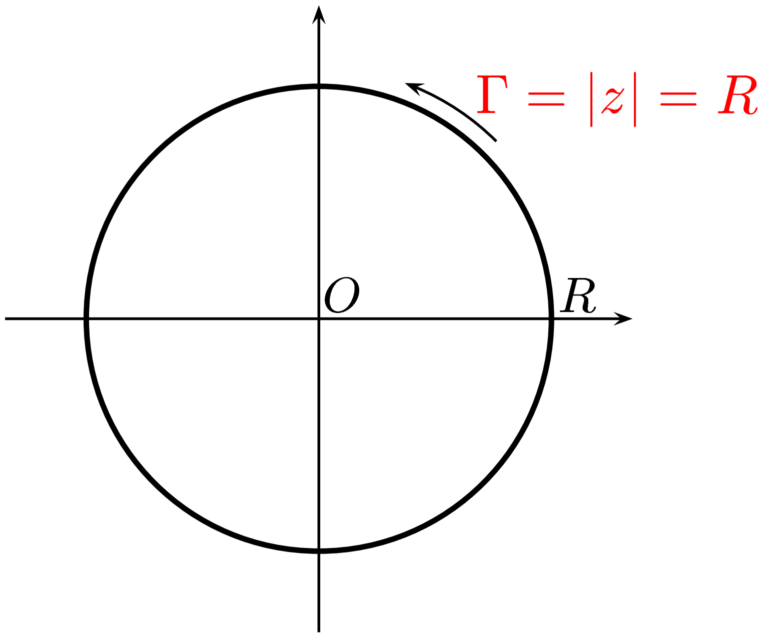
On substitue:

$$u = e^{t/\sigma_n}, \quad \mu_n = \frac{n+1}{4} - 1, \quad \sigma_n = \frac{\sqrt{n+1}}{4}$$

On développe au voisinage de $t = 0$ $\log(\Phi(t))$

$$\log(\Phi(t)) = \frac{t^2}{2} - \frac{t^4}{12(n+1)} + o\left(\frac{t^4}{n^2}\right) \longrightarrow \frac{t^2}{2}$$

Intégrale de Cauchy



$a)\ k \in \mathbb{Z} - \{-1\}$

$$\int_\Gamma z^k dz = \left[\frac{z^{k+1}}{k+1}\right]_{z=R}^{z=R} = 0$$

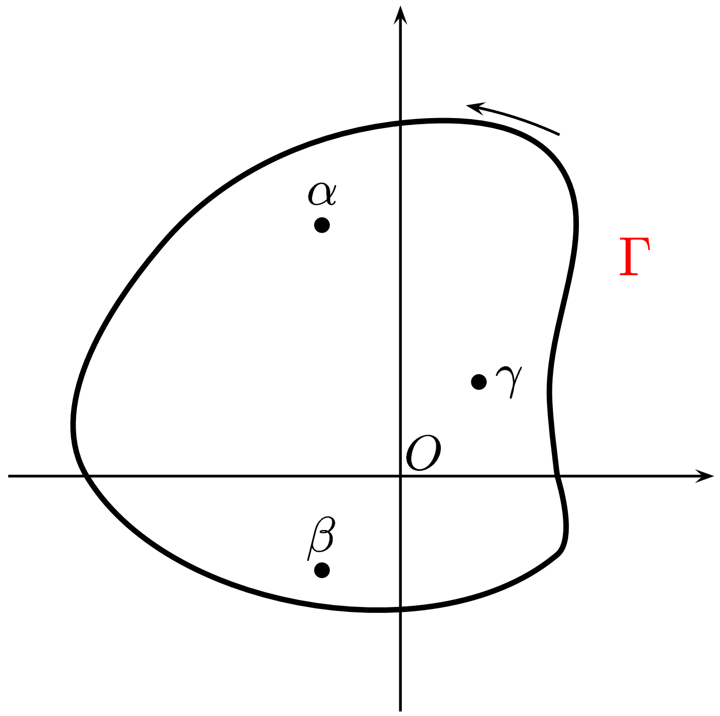
$b)\ k = -1$

changement de variable

$z = Re^{i\theta} \ \Rightarrow \ dz = Rie^{i\theta} d\theta$

$$\int_\Gamma \frac{1}{z} dz = \int_0^{2\pi} i d\theta = 2i\pi$$

Théorème des résidus (Cauchy)



$F(z)$ méromorphe à l'intérieur de Γ

$$\Rightarrow F(z) = \frac{H_{(-r)}}{(z-\alpha)^r} + \frac{H_{-(r-1)}}{(z-\alpha)^{(r-1)}} + \dots$$
$$+ \frac{\text{Résidu}(\alpha)}{z-\alpha}$$
$$+ H_0 + (z-\alpha)H_1 + \dots$$

$$\frac{1}{2i\pi}\int_\Gamma F(z)dz = \text{Résidu}(\alpha) + \text{Résidu}(\beta) + \text{Résidu}(\gamma)$$

$F(z)$ analytique à l'intérieur de Γ

$F(z) = f_0 + f_1 z + \dots + f_n z^n + \dots$

$$\frac{1}{2i\pi}\int_\Gamma \frac{F(z)}{z^{n+1}}dz = 0 + \frac{1}{2i\pi}\int_\Gamma \frac{f_n}{z}dz + 0 = f_n$$

Preuve de la loi Gaussienne

$$
\begin{aligned}
L_0(z,u) &= zp_aL_1 + zp_bL_0 + 1,\\
L_1 &= zp_bL_2 + zp_aL_1 + 1,\\
L_2 &= zp_a u L_3 + zp_bL_2 + 1\\
L_3 &= zp_aL_1 + zp_bL_2 + 1
\end{aligned}
$$

$$L = \begin{pmatrix} L_0 \\ \vdots \\ L_n \end{pmatrix} = zT(u)L + 1$$

$T(u)$ matrice $n \times n$ positive

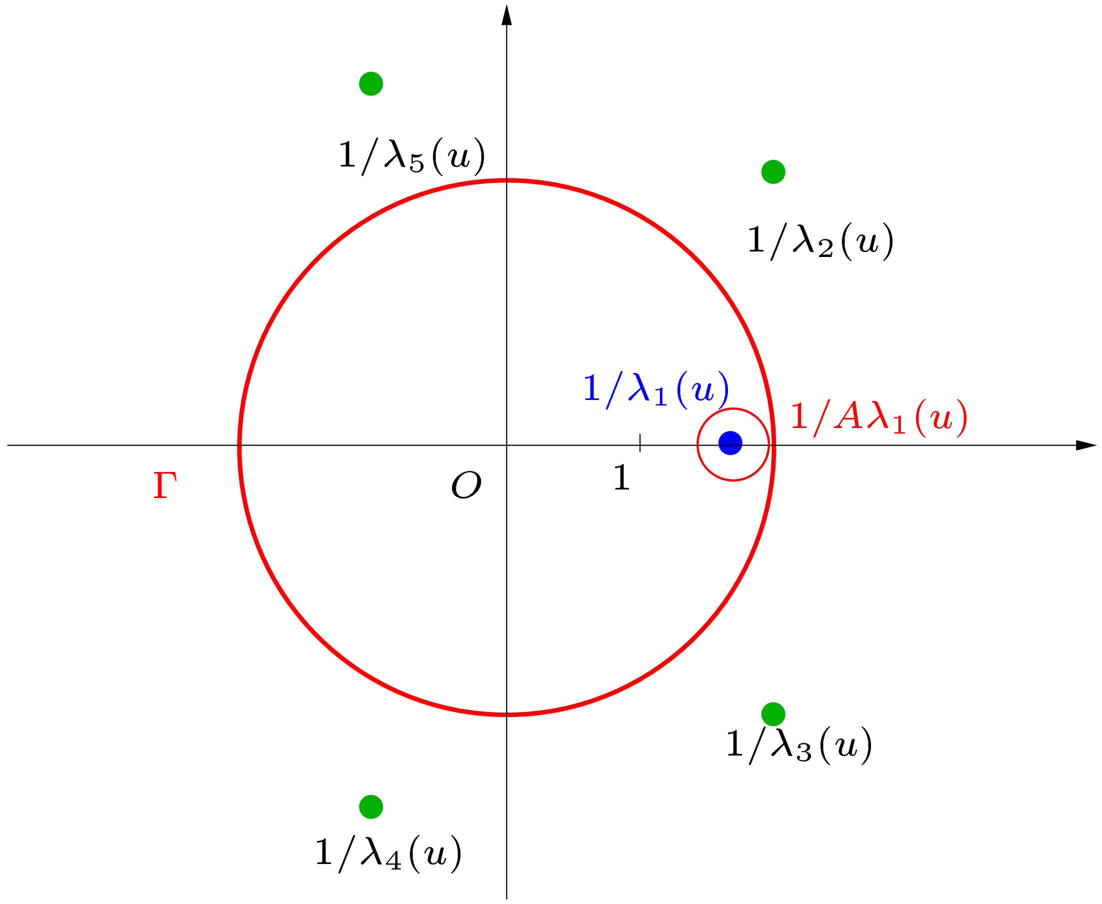
$$L_0(z,u) = \frac{P(z,u)}{Q(z,u)} = \frac{P(z,u)}{(1 - z\lambda_1(u)) \cdots (1 - z\lambda_n(u))}$$

$$\frac{1}{|\lambda_1(u)|} < \frac{1}{|\lambda_2(u)|} \leq \dots$$

Perron-Frobenius

$\lambda_1(u)$ unique, réelle, positive.

Propriété de séparation uniforme



$$
\begin{aligned}
p_n(u) &= [z^n]F(z,u) = \frac{1}{2i\pi}\oint_\Gamma \frac{dz}{z^{n+1}}F(z,u),\\
&= \frac{1}{2i\pi}\oint_\Gamma \frac{c(u)}{z^{n+1}(1-\lambda_1(u)z)} + \frac{1}{z^{n+1}}g(z,u)\,dz,\\
&= c(u)\lambda_1(u)^n(1+O(A^n)).
\end{aligned}
$$

Théorème des quasi-puissances de Hwang

$\Rightarrow$ distribution limite Gaussienne.

Partie VI

Applications
à la biologie

Mots sur-représentés

Hexamer	Obs.	Rk	Exp.	Z-sc.	Rk	Cd.Exp.	Cd.Z-sc.	Rk
AAUAAA	3456	1	363.16	167.03	1			**1**
AAAUAA	1721	2	363.16	71.25	2	1678.53	1.04	1300
AUAAAA	1530	3	363.16	61.23	3	1311.03	6.05	404
UUUUUU	1105	4	416.36	33.75	8	373.30	37.87	**2**
AUAAAU	1043	5	373.23	34.67	6	1529.15	-12.43	4078
AAAAUA	1019	6	363.16	34.41	7	848.76	5.84	420
UAAAAU	1017	7	373.23	33.32	9	780.18	8.48	211
AUUAAA	1013	8	373.23	33.12	10	385.85	31.93	**3**
AUAAAG	972	9	184.27	58.03	4	593.90	15.51	34
UAAUAA	922	10	373.23	28.41	13	1233.24	-8.86	4034
UAAAAA	922	11	363.16	29.32	12	922.67	9.79	155
UUAAAA	863	12	373.23	25.35	15	374.81	25.21	**4**
CAAUAA	847	13	185.59	48.55	5	613.24	9.44	167
AAAAAA	841	14	353.37	25.94	14	496.38	15.47	36
UAAAUA	805	15	373.23	22.35	21	1143.73	-10.02	4068

Table of the most frequent hexanucleotides. *Obs*: number of observed occurrences. *Rk*: Rank. *Exp.*: (non-conditional) expectation. *Cd.Exp.*: expectation conditioned by number of occurrences of AAUAAA.

Extrait de “A. Denise, M. Régnier, et M. Vandenbogaert. Workshop WABI’01, et Rapport de Recherche INRIA 4132”

Comparaison de protéomes

par étude statistique des motifs de Prosite

```
AC   PS00723;
DE   Polyprenyl synthetases signature 1.
...
PA   [LIVM](2)-x-D-D-x(2,4)-D-x(4)-R-R-[GH].

...
DR   P14324, FPPS_HUMAN, T; ... P49353, FPPS_MAIZE, T;
DR   P08524, FPPS_YEAST, T; ... P08836, FPPS_CHICK, P;
...
```

Question?

pertinence biologique des motifs

respectivement à des protéomes cibles

Expressions régulières

$$R = (a \cdot b \cdot a + (c^4 \cdot e)^\star \cdot b \cdot b)^\star$$

Opérateurs

$+$	Union
$\cdot$	Concaténation
$\star$	Opérateur-étoile ($A^\star = \epsilon + A + A^2 + A^3 + \dots$)

Motifs de Prosite

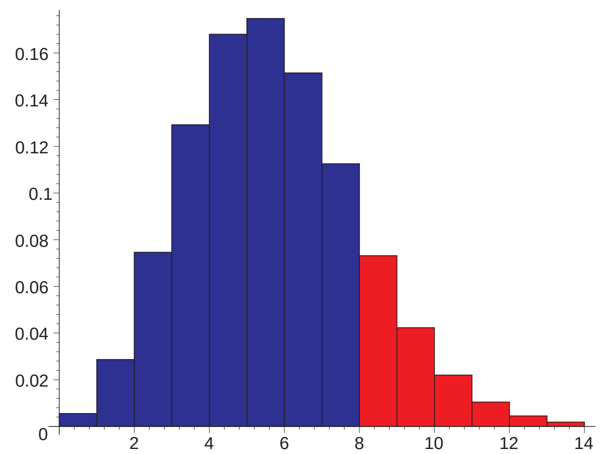
PA = [LIVM](2)-x-D-D-x(2,4)-D-x(4)-R-R-[GH].

$$PA = (L+I+V+M)^2 \cdot \Sigma \cdot D \cdot D \cdot (\Sigma^2 + \Sigma^3 + \Sigma^4) \cdot D \cdot \Sigma^4 \cdot R \cdot R \cdot (G+H)$$

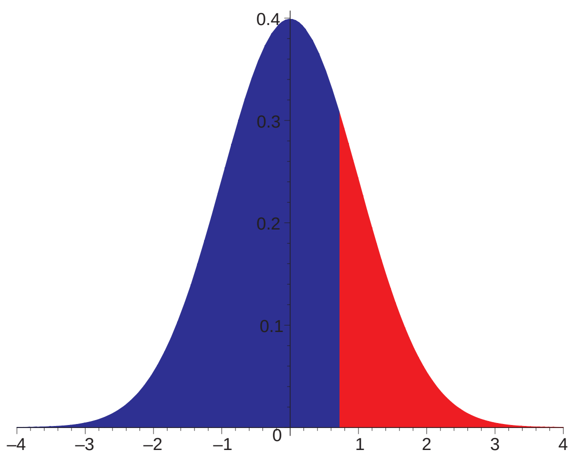
$$\text{with } \Sigma = \begin{array}{l} A+R+N+D+C+Q+E+G+H+I \\ +L+K+M+F+P+S+T+W+Y+V \end{array}$$

$$\Sigma^k = \Sigma \cdot \Sigma \cdot \quad \cdots \quad \cdot \Sigma \qquad (k \text{ times})$$

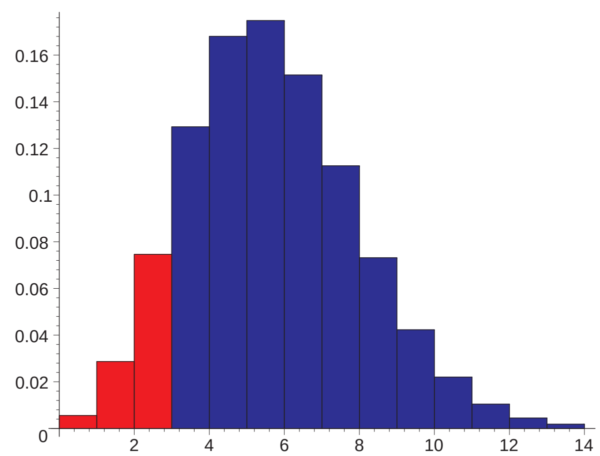
Normalisation Gaussienne



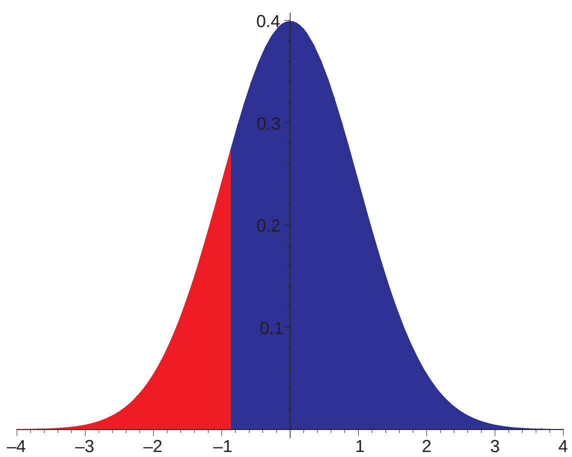
$\omega = 8$



$\gamma = +0.72$



$\omega = 2$



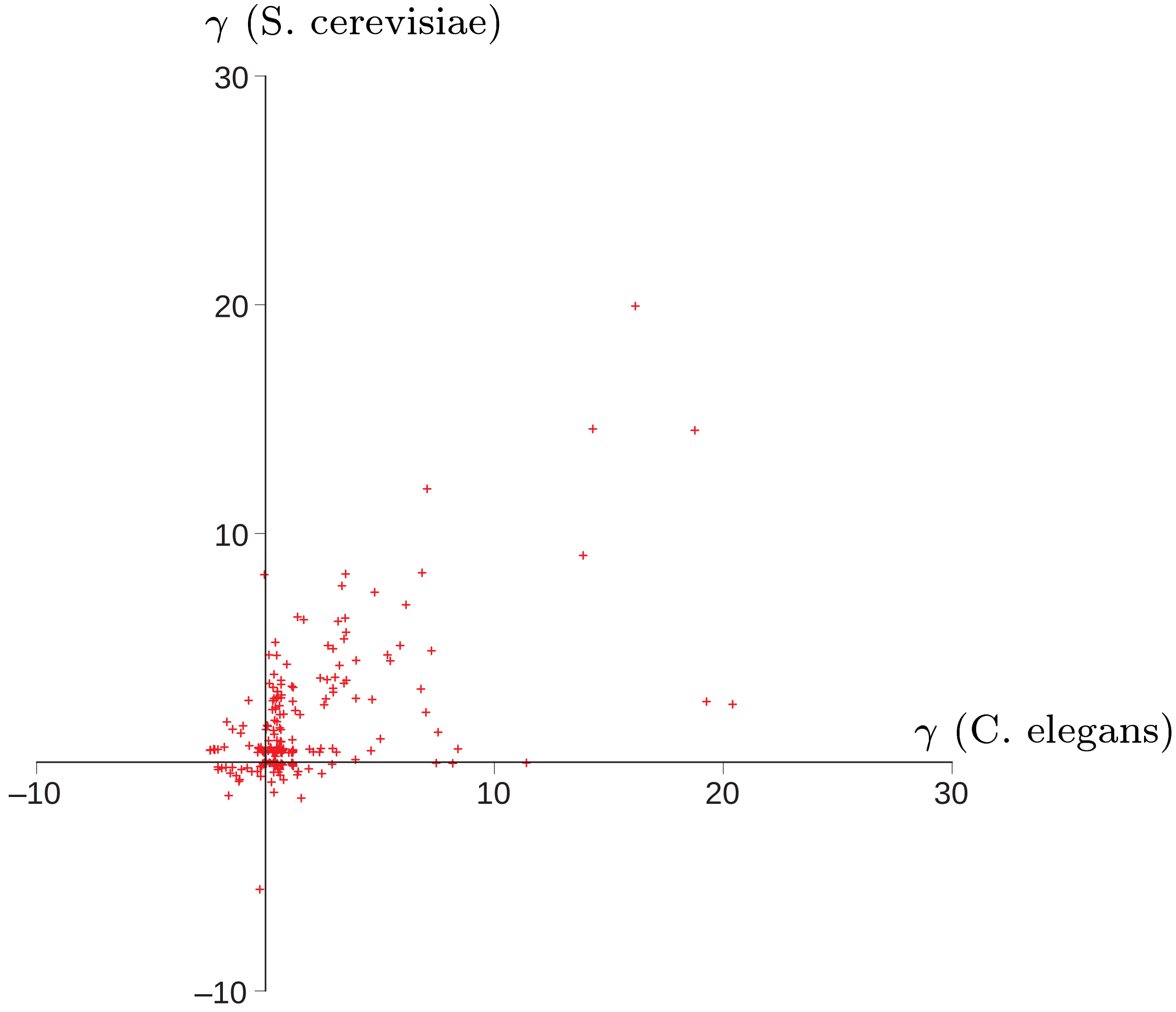
$\gamma = -0.81$

Espérance Poisson $\mu = 5.2$

Distribution Gaussienne

ω nombre d'occurrences observés

Comparaison du ver et de la levure



Comparaison des motifs de Prosite
pour *C. elegans* et *S. cerevisiae*

- 273 motifs avec $\mu \geq 0.5$ sont représentés -

```
PS00022 GC   20.41 GS    2.54 OC    430 EC       3.666 OS      4 ES        .321
PS00022
PS00022 C-x-C-x(5)-G-x(2)-C
PS00022 EGF_1;PATTERN.   EGF-like domain signature 1.

PS01186 GC   19.27 GS    2.66 OC    384 EC       3.260 OS      3 ES        .270
PS01186
PS01186 C-x-C-x(2)-[GP]-[FYW]-x(4,8)-C
PS01186 EGF_2;PATTERN.   EGF-like domain signature 2.

PS00010 GC   11.40 GS    -.03 OC    113 EC        .576 OS      0 ES        .061
PS00010
PS00010 C-x-[DN]-x(4)-[FY]-x-C-x-C
PS00010 ASX_HYDROXYL;PATTERN.   Aspartic acid and asparagine hydroxylation site.

PS00109 GC    8.42 GS     .58 OC     81 EC       1.306 OS      2 ES        .570
PS00109
PS00109 [LIVMFYC]-x-[HY]-x-D-[LIVMFY]-[RSTAC]-x(2)-N-[LIVMFYC](3)
PS00109 PROTEIN_KINASE_TYR;PATTERN.   Tyrosine protein kinases specific active-site sign..

PS00232 GC    8.18 GS    -.04 OC     53 EC        .168 OS      0 ES        .103
PS00232
PS00232 [LIV]-x-[LIV]-x-D-x-N-D-[NH]-x-P
PS00232 CADHERIN;PATTERN.   Cadherins extracellular repeated domain signature.

PS00018 GC    7.55 GS    1.32 OC    184 EC      24.606 OS     30 ES      20.373
PS00018
PS00018 D-x-[DNS]-{ILVFYW}-[DENSTG]-[DNQGHRK]-{GP}-[LIVMC]-[DENQSTAGC]-x(2)-[DE]-[LIVMFYW]
PS00018 EF_HAND;PATTERN.   EF-hand calcium-binding domain.

PS00280 GC    7.47 GS    -.04 OC     62 EC        .721 OS      0 ES        .088
PS00280
PS00280 F-x(3)-G-C-x(6)-[FY]-x(5)-C
PS00280 BPTI_KUNITZ;PATTERN.   Pancreatic trypsin inhibitor (Kunitz) family signature.
```

Motifs exceptionnels pour le ver et la levure

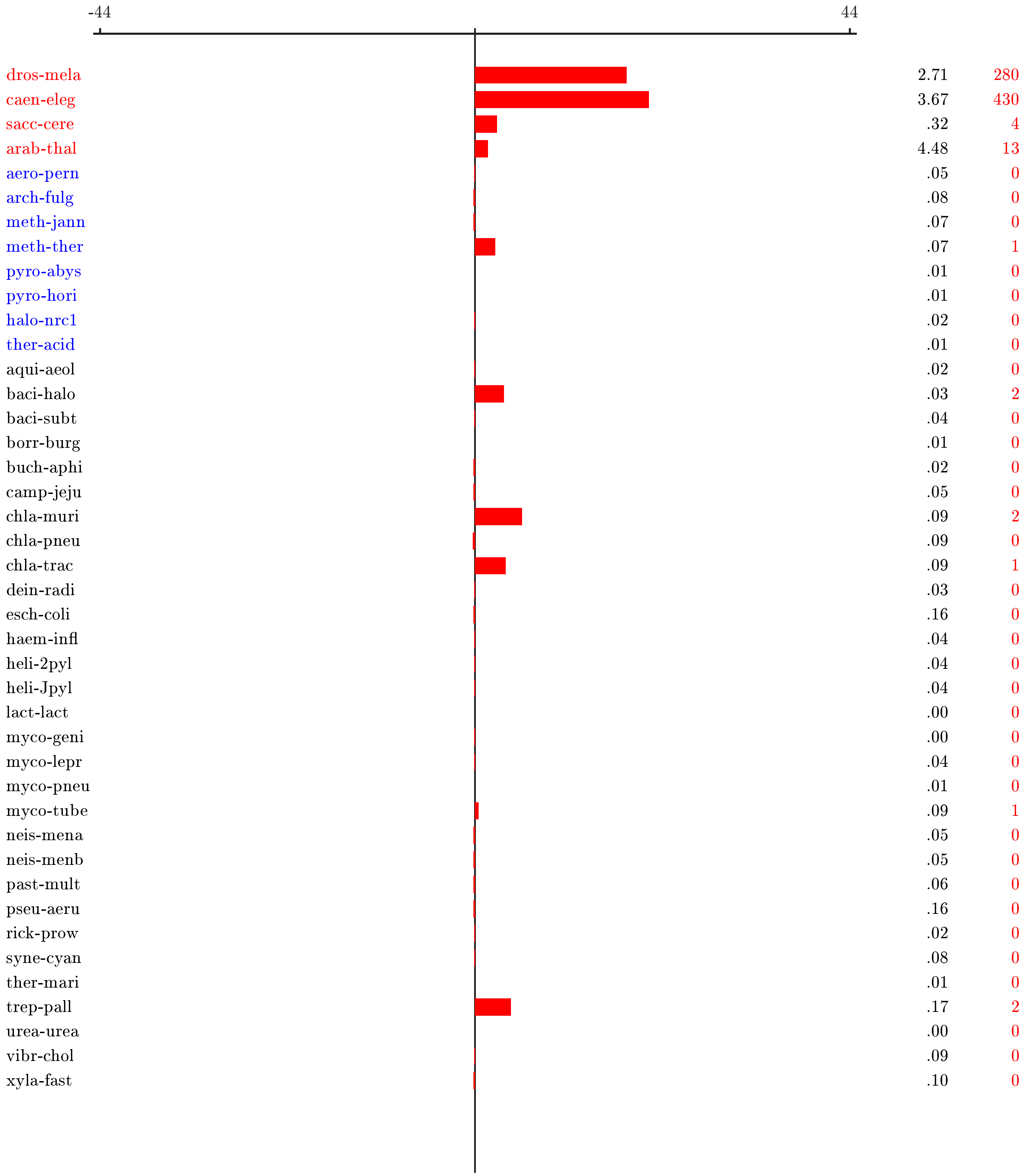
GC, OC, EC et GS, OS, ES sont respectivement les écarts Gaussiens γ, les observations ω et les espérances μ du nombre d'occurrences pour *C. elegans* et *S. cerevisiae*.

Comparaison de 42 protéomes

	Espèce	Longueur protéome
Eucaryotes	dros-mela	6617803
	caen-eleg	7863058
	sacc-cere	2937205
	arab-thal	11304307
Archae-bacteries	aero-pern	638811
	arch-fulg	660832
	meth-jann	502362
	meth-ther	525164
	pyro-abys	535786
	pyro-hori	568584
	halo-nrc1	624005
	ther-acid	453115
Bacteries	aqui-aeol	488484
	baci-halo	1169204
	baci-subt	1218487
	borr-burg	352644
	buch-aphi	187023
	camp-jeju	503501
	chla-muri	324428
	chla-pneu	361707
	chla-trac	312118
	dein-radi	951455
	esch-coli	1373785
	haem-infl	526801
	heli-2pyl	491768
	heli-Jpyl	493049
	lact-lact	655989
	myco-geni	175729
	myco-lepr	520057
	myco-pneu	237930
	myco-tube	1317198
	neis-mena	582084
	neis-menb	573863
	past-mult	667675
	pseu-aeru	1856757
	rick-prow	279044
	syne-cyan	1027015
	ther-mari	582037
	trep-pall	349913
	urea-urea	227717
	vibr-chol	1156096
	xyla-fast	744437

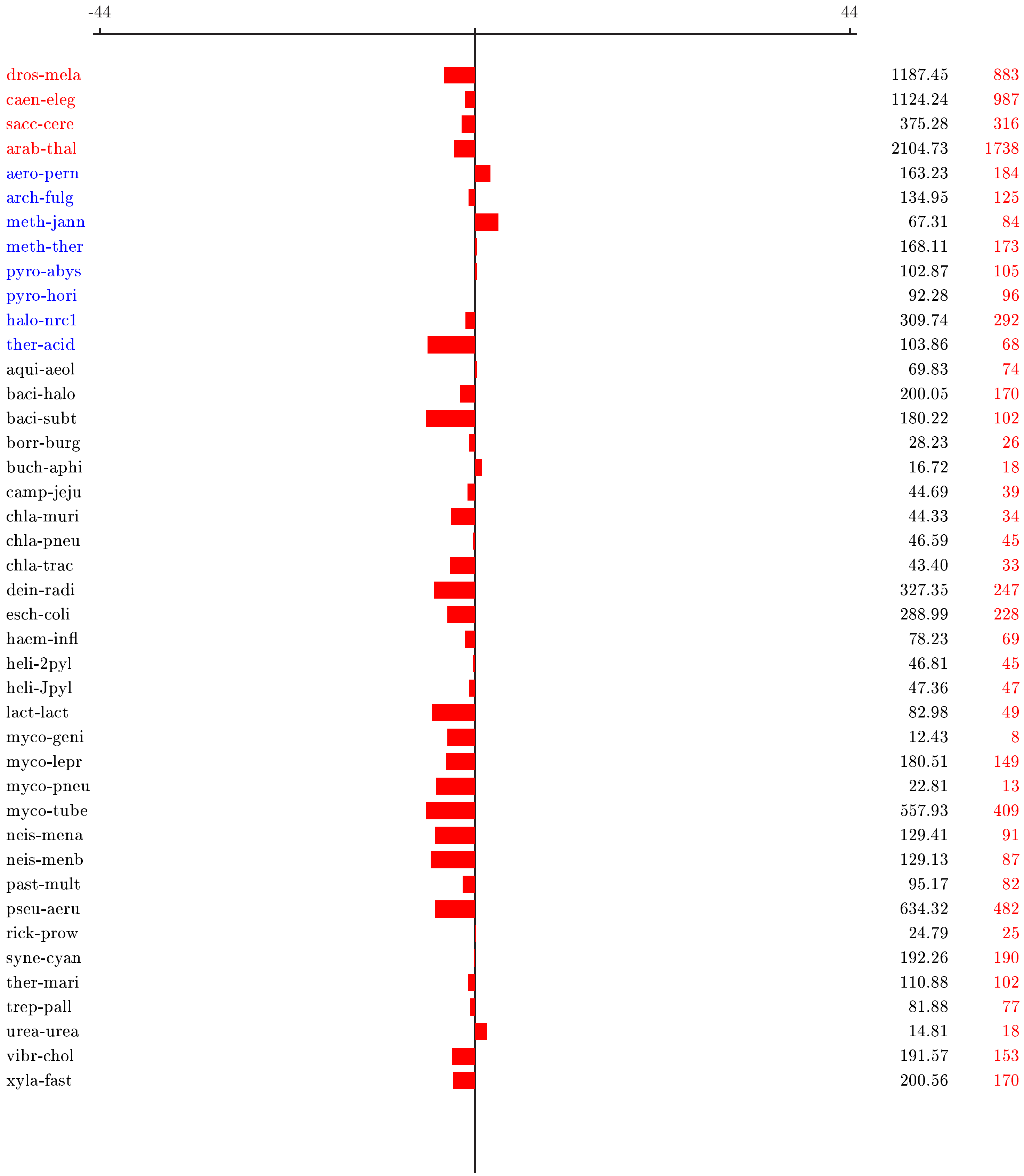
PS00022 EGF-1;PATTERN. EGF-like domain signature 1.

C-x-C-x(5)-G-x(2)-C



PS00016 RGD;PATTERN. Cell attachment sequence.

R-G-D



Bibliographie

– Partie II

  ∗ Flajolet et Sedgewick. *An introduction to the analysis of algorithms.* Addison-Wesley, 1996, (niveau DEUG, licence, v.f. chez Vuibert)

  ∗ Szpankowski. *Average case analysis of algorithms on sequences.* Wiley-Interscience, 2001, (niveau maîtrise, DEA)

– Partie III

  ∗ Régnier et Szpankowski. *On pattern frequency occurrences in a markovian sequence?* Algorithmica, vol. 22, n°4, 1998, pages 631–649.

  ∗ Régnier. *A unified approach to word occurrences probabilities.* Discrete Applied Mathematics, vol. 104, n°1, 2000, pages 259–280.

– Partie IV

  ∗ Nicodème, Salvy et Flajolet. *Motif statistics.* Theoretical Computer Science, 2001. - To appear.
  Presented at ESA'99, Prague, July 1999

– Partie V

  ∗ Durrett. *Probability: Theory and examples.* Duxbury Press, 1996