Combinatorics and biology

Pierre Nicodème

Laboratory Statistics and Génomes
CNRS - La Génopole - Évry - France

nicodeme@genopole.cnrs.fr

and

Algorithmes Project
INRIA - Rocquencourt - France

http://algo.inria.fr/nicodeme

Part I

Experimental Mathematics

30 random texts - 1

word1 = 100 word2 = 111

1 1100101011000010010011010000011011011010110100010011001101010101010
2 11000111
0101101111100
0111110101100
3 100110111
00011100
4 000110001101010000111
5 000010110010100000110111
6 110000110010110111
7 10000000111
0111110100
8 10110000100111
9 11010010011010110010110101100111
1111100
10 10100100110111
11101111011100
11 10010111
11011100
011111011111111100
12 01010000111
13 00100111
14 000000000100110111
15 100111
16 1001010110000100111
17 010010000111
18 00101100000101010110111
19 10001000100111000111
20 11011010011011010000000111
21 01010001101000010110011 1
1101111111010110111100

30 random texts - 2

word1 = 100 word2 = 111

1 00011001011
111011101100
10111010
11101010
010111110100
1110111011111111110101100
000101011100
00011110
2 10000110110111
0111100
3 1101100001101000010111
4 10000101010011011
01111100
5 100111
11011111100
6 0101001001100010110010000101100011001010010111
7 00000100111
10111100
8 1001010111
9 1100011010110110010111
10 1101100001011010111
11 10011000011010011001110111
12 1001000111
01101111100
13 011000111
01010111100
14 10010001000010110101111
11011100
00111100
15 01100001001111

30 random texts - 3

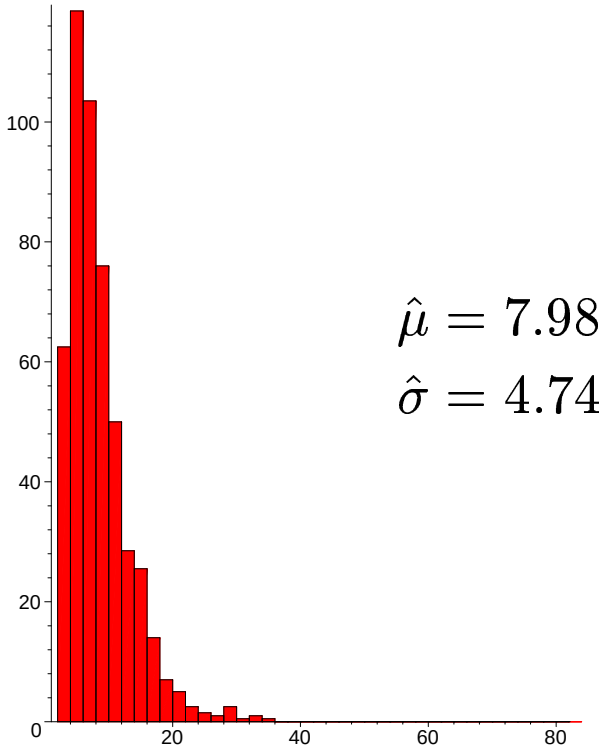
word1 = 100 word2 = 111

111101011010101010100
1 101100010110101100111
2 000001000111
000011110100
3 00100010001101100001100100000111
11100
4 11010000000000101100011010110011111
5 10100110100000100100011011001101111
011100
111011110101101110110100
011111101011010100
6 1100011011000100110101010100000010111
7 0100110001000101100001101010100000111
8 1100111
101111110100
9 100000011000111
10 0100101100110011011010111
010111110111101010100
11 10101010100100100111
001101111101010100
12 0000000100001101100111
13 1001000010111
14 100000010111
15 10000111
011111110111010100
111100
16 00110100110001101011010111
17 1000111
18 0001101001000111
19 10000010111

waiting for the first occurrence

1000 random texts

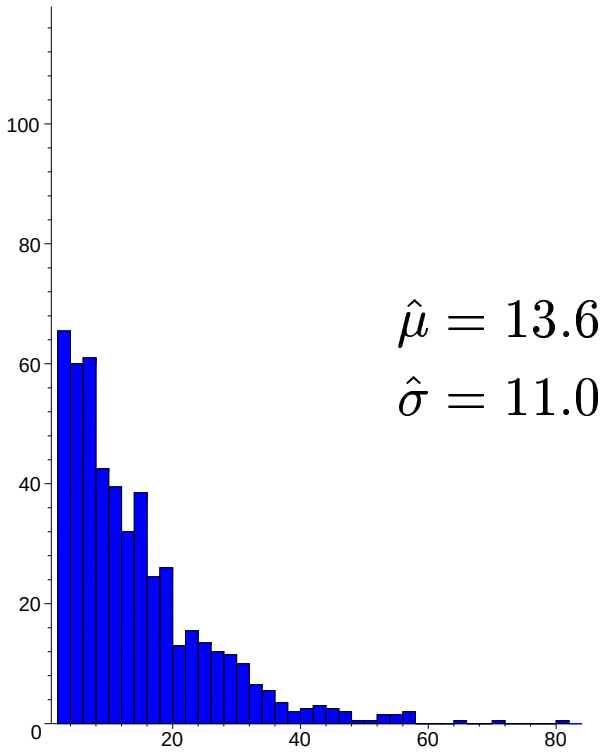
100



$\hat{\mu} = 7.98$

$\hat{\sigma} = 4.74$

111



$\hat{\mu} = 13.67$

$\hat{\sigma} = 11.00$

1 text of size 1000

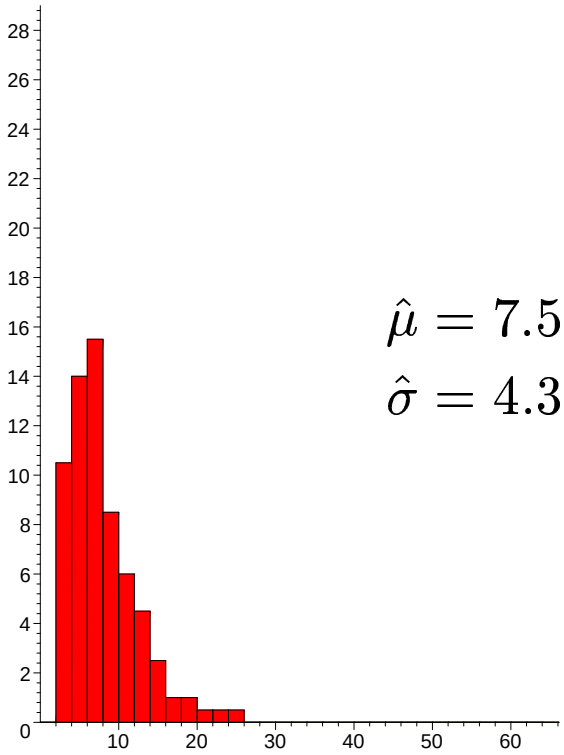
1000

1111

Distance between occurrences

text of size 1000

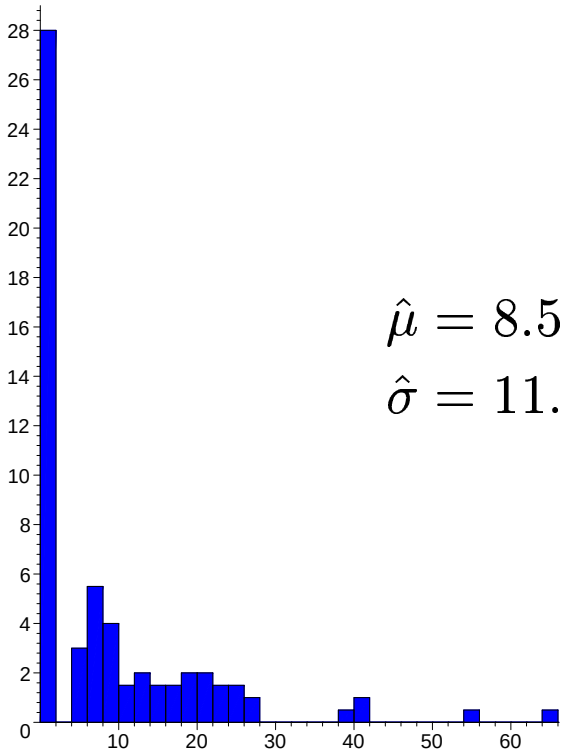
100



$\hat{\mu} = 7.56$

$\hat{\sigma} = 4.31$

111



$\hat{\mu} = 8.53$

$\hat{\sigma} = 11.34$

What is going on?

– Probability of appearance at a given position

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

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

– BUT the 111 occur often by CLUMPS

$$\cdots 0111110 \qquad\qquad \cdots 011110$$

$$111 \qquad\qquad 111\cdots$$

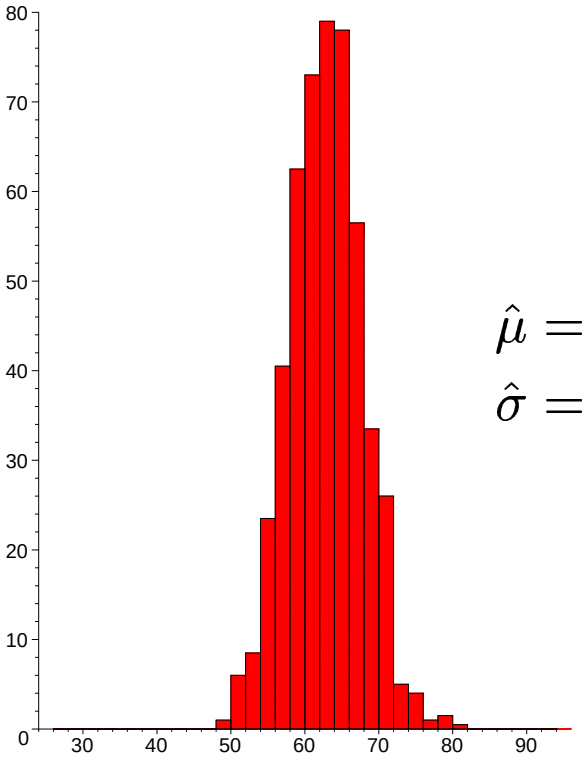
$$111\cdots$$

– while the 100 NEVER OVERLAP

Number of occurrences

1000 texts of size 500

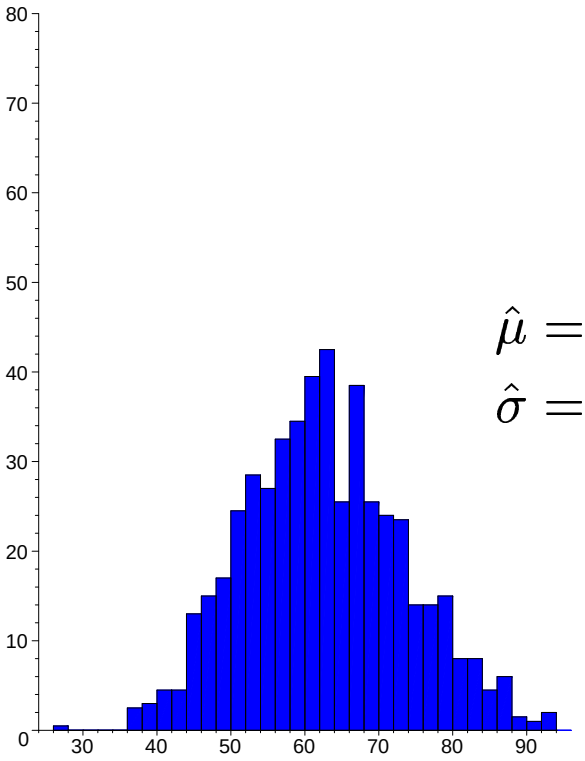
100



$\hat{\mu} = 62.38$

$\hat{\sigma} = 4.90$

111



$\hat{\mu} = 62.07$

$\hat{\sigma} = 10.79$

Part II

Languages and generating functions

The tool of generating functions

sequence	generating function	
$f_0, f_1, \ldots, f_n, \cdots$	$f_0 + f_1 z + \cdots + f_n z^n + \ldots$	$= F(z)$
$1, 1, 1, \ldots, 1, \ldots$	$1 + z + z^2 + \cdots + 1z^n + \ldots$	$= \frac{1}{1-z}$
$1, 2, 3, \ldots, n, \ldots$	$z + 2z^2 + \cdots + nz^n + \ldots$	$= \frac{z}{(1-z)^2}$
$\ldots$	$\ldots$	$\ldots$

Advantage:

1. Manipulation of functions
2. Extraction of the Taylor coefficient of required order
   $[z^n]F(z) = f_n$

Example:

μ_n expectation of number of occurrences in a text of size n

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

Counting generating function

language = set of words

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

$\Sigma^\star = \epsilon + \Sigma + \Sigma^2 + \cdots + \Sigma^n + \ldots$ all the words

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

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

Example

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

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

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

if there are no ambiguities,

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

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

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

Generating function of a language

language = set of words

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

$\Sigma^\star = \epsilon + \Sigma + \Sigma^2 + \cdots + \Sigma^n + \ldots$ all the words

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

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

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

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

if $\mathcal{A}.\mathcal{B}$ non ambiguous,

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

if $\mathcal{A}$ and $\mathcal{B}$ disjoint

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

if $\mathcal{A}^\star$ non ambiguous,

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

Weighted generating function

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

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

Generating function of the language

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

Weighted generating functions

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

p_α proba. of word α

π_n proba. that a word of size n belongs to $\mathcal{L}$

Example

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

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

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

Weighted generating functions

case Bernoulli uniform

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

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

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

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

π_n probability that a word of size n belongs to $\mathcal{L}$

Example

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

$$\to L(z) = z^3 + 1 + z^3 + z^6 + z^9 + \ldots$$

$$= 1 + z^3 + \frac{z^3}{1 - z^3}$$

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

Probability generating function

X random variable (with positive integer values).

Probability generating function $P(z)$

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

Properties

$P(1) = 1$ (sum of probability)

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

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

$$\sigma = \sqrt{m_{(2)} - \mu^2} \quad \text{(standard-deviation)}$$

Bivariate generating function

X_n random variable counting the number of occurrences in a text of size n

Bivariate generating function $F(z,u)$

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

$f_{n,k}$ proba. that a text of size n has k matches

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

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

Part III

Combinatorial approach

1 word - Bernoulli model

h considered word (length m, proba $p = \mathbf{P}(h)$)

Polynomial of autocorrelation

$h = ababa$

$$ababa$$
$$ababa|$$
$$\quad aba\,ba$$
$$\quad\quad a\,baba$$

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

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

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

Languages used

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

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

Minimal $\mathcal{M}$

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

$ababa \qquad\qquad ababa$

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

$ababa$

$ba \subset \mathcal{M}$

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

$ababa \qquad\qquad ababa$

$aabbbabbbbbbb \subset \mathcal{T} \qquad\qquad babbbbbbbbbb \not\subset \mathcal{T}$

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 over the langages

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

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

word beginning with h

- 1 single occurrence of $h \quad \Rightarrow \quad h.\mathcal{T}$
- several occurrences of $h \quad \Rightarrow \quad h.\mathcal{M}.\Sigma^\star$

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

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

word finishing by h

- first occurrence h overlaps last one (or single occurrence) $\quad \Rightarrow \mathcal{F}.\mathcal{A}$
- first occurrence does not overlap last $\Rightarrow \mathcal{F}.\Sigma^\star.h$

Equations over the languages (continued)

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

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

- $h.\mathcal{M}^+$, $\Rightarrow$ words beginning and finishing with h
- $h.\Sigma^\star.h$, $\Rightarrow$ first and last occurrences do not overlap
- $h.(\mathcal{A} - \epsilon)$ $\Rightarrow$ first and last occurrences overlap

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

add a letter to a word of N $\Rightarrow$

- create a match $\Rightarrow \mathcal{F}$
- do not create a match $\Rightarrow \mathcal{N}$

(empty word ϵ forbidden)

Languages $\Rightarrow$ generating functions

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

Add the mark u after each match $\Rightarrow$ marked language $\mathcal{L}$

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

Translation into generating function

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

Solving the system

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

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

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

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

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

Application to words 100 and 111

Model Bernoulli uniform

Word 100

100

100|

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

Word 111

111

111|

11|1

1|11

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

First occurrence and distance between 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	$2\sqrt{6} \approx 4.90$	$2\sqrt{30} \approx 10.95$

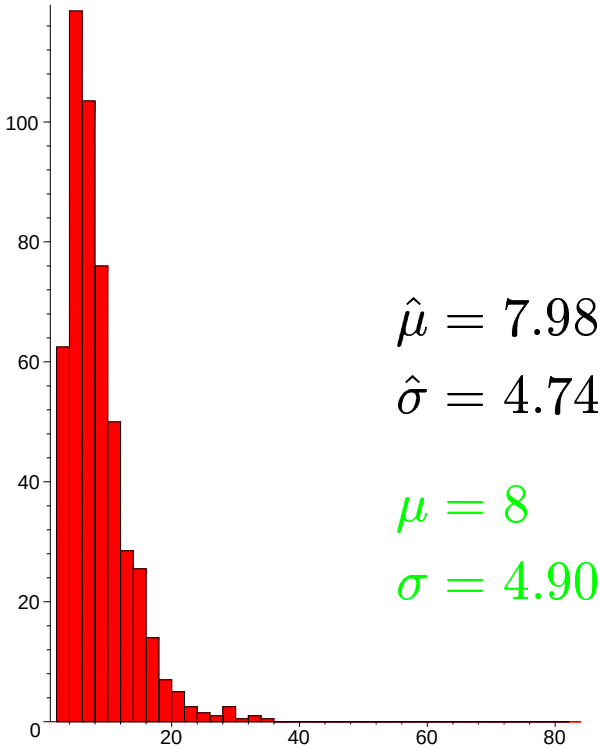
Number of occurrences

	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}	$\dfrac{\sqrt{1500}}{8}\approx 4.84$	$\dfrac{\sqrt{7460}}{8}\approx 10.80$

Waiting for the first occurrence

1000 random texts

100



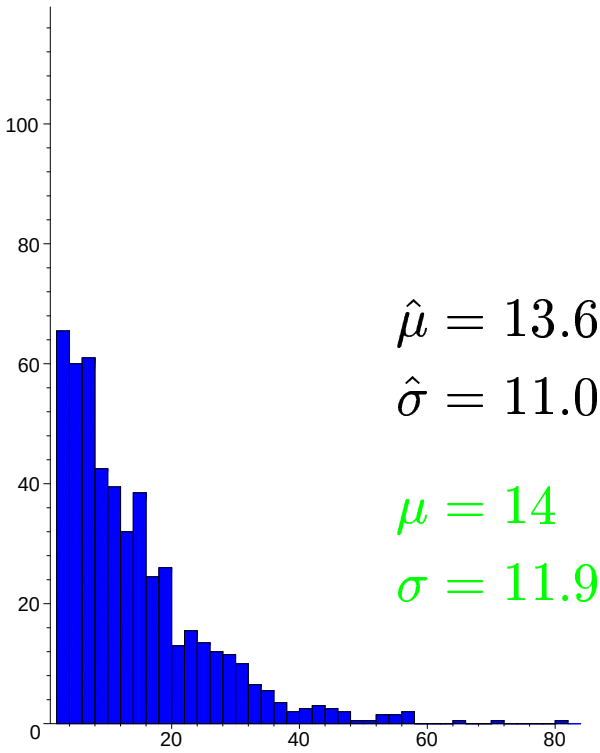
$\hat{\mu} = 7.98$

$\hat{\sigma} = 4.74$

$\mu = 8$

$\sigma = 4.90$

111



$\hat{\mu} = 13.67$

$\hat{\sigma} = 11.00$

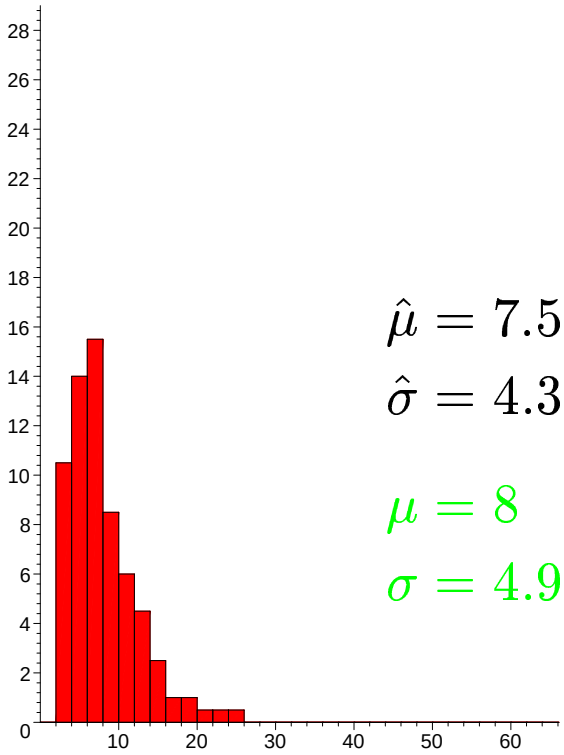
$\mu = 14$

$\sigma = 11.92$

Distance between occurrences

text of size 1000

100



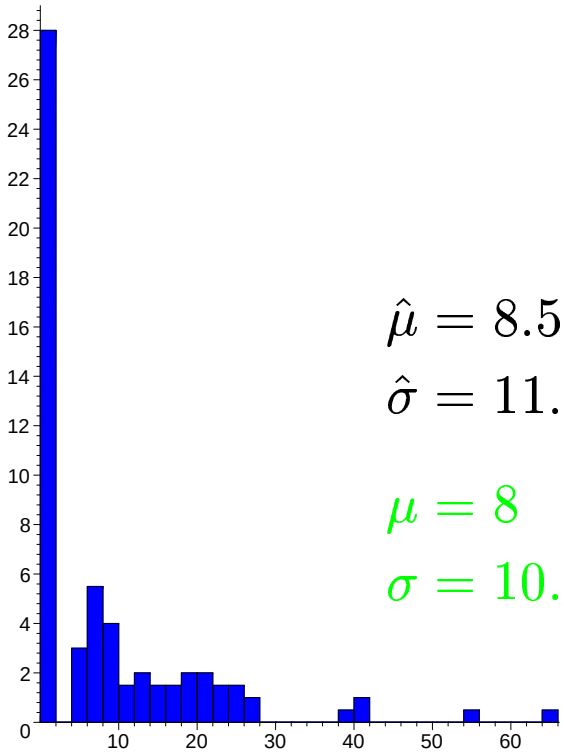
$\hat{\mu} = 7.56$

$\hat{\sigma} = 4.31$

$\mu = 8$

$\sigma = 4.90$

111



$\hat{\mu} = 8.53$

$\hat{\sigma} = 11.34$

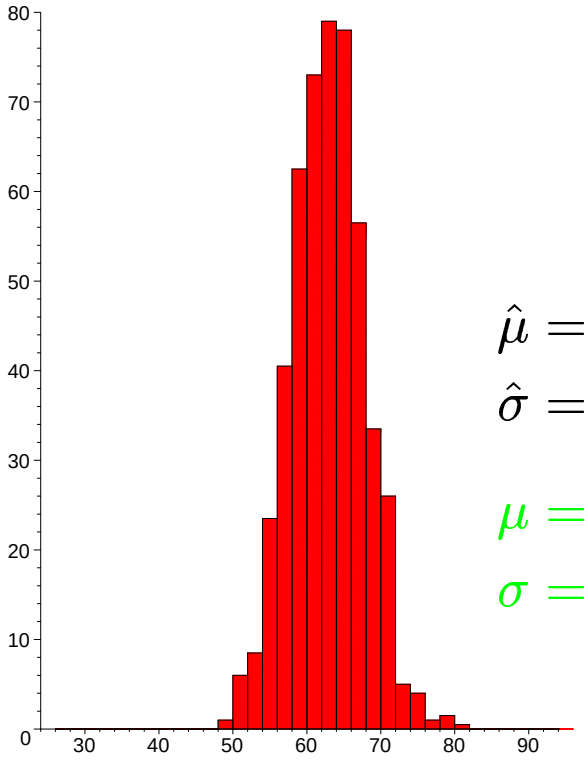
$\mu = 8$

$\sigma = 10.95$

Number of occurrences

1000 texts of size 500

100



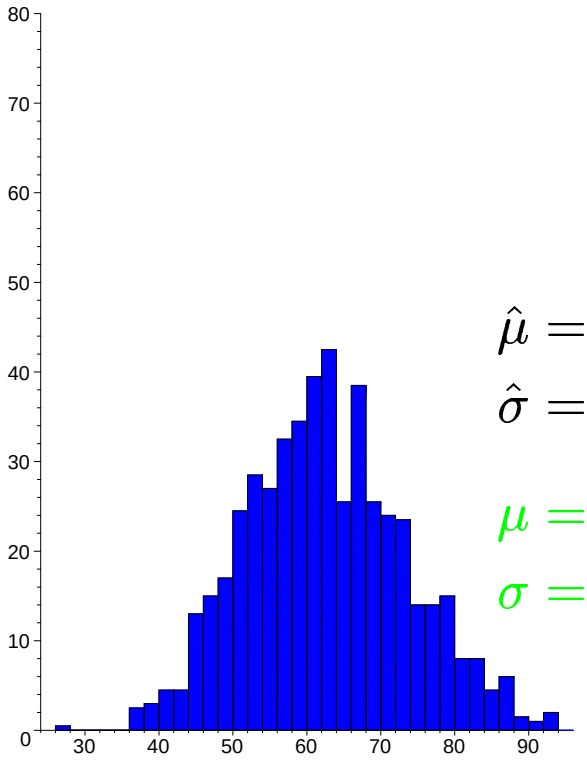
$\hat{\mu} = 62.38$

$\hat{\sigma} = 4.90$

$\mu = 62.25$

$\sigma = 4.84$

111



$\hat{\mu} = 62.07$

$\hat{\sigma} = 10.79$

$\mu = 62.25$

$\sigma = 10.80$

"Multi"- words

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

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

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

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

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

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

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

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

Part IV

Algorithmic approach by automata

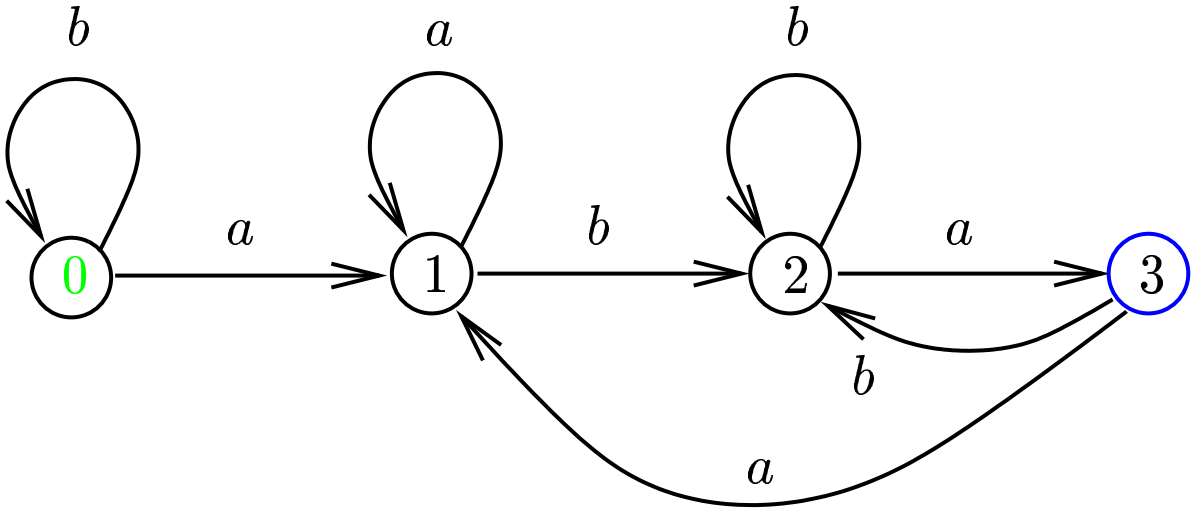
Occurrences of words ab^+a

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

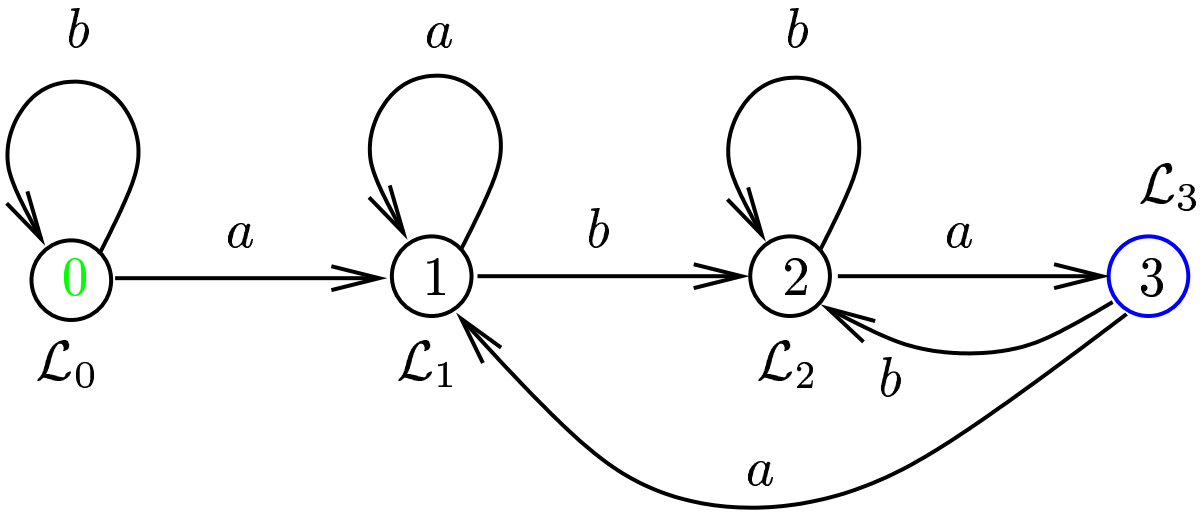
We consider the texts $(a+b)^\star ab^+a$

We build

the automaton recognizing these texts



The algorithm of Chomski and Schützenberger



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

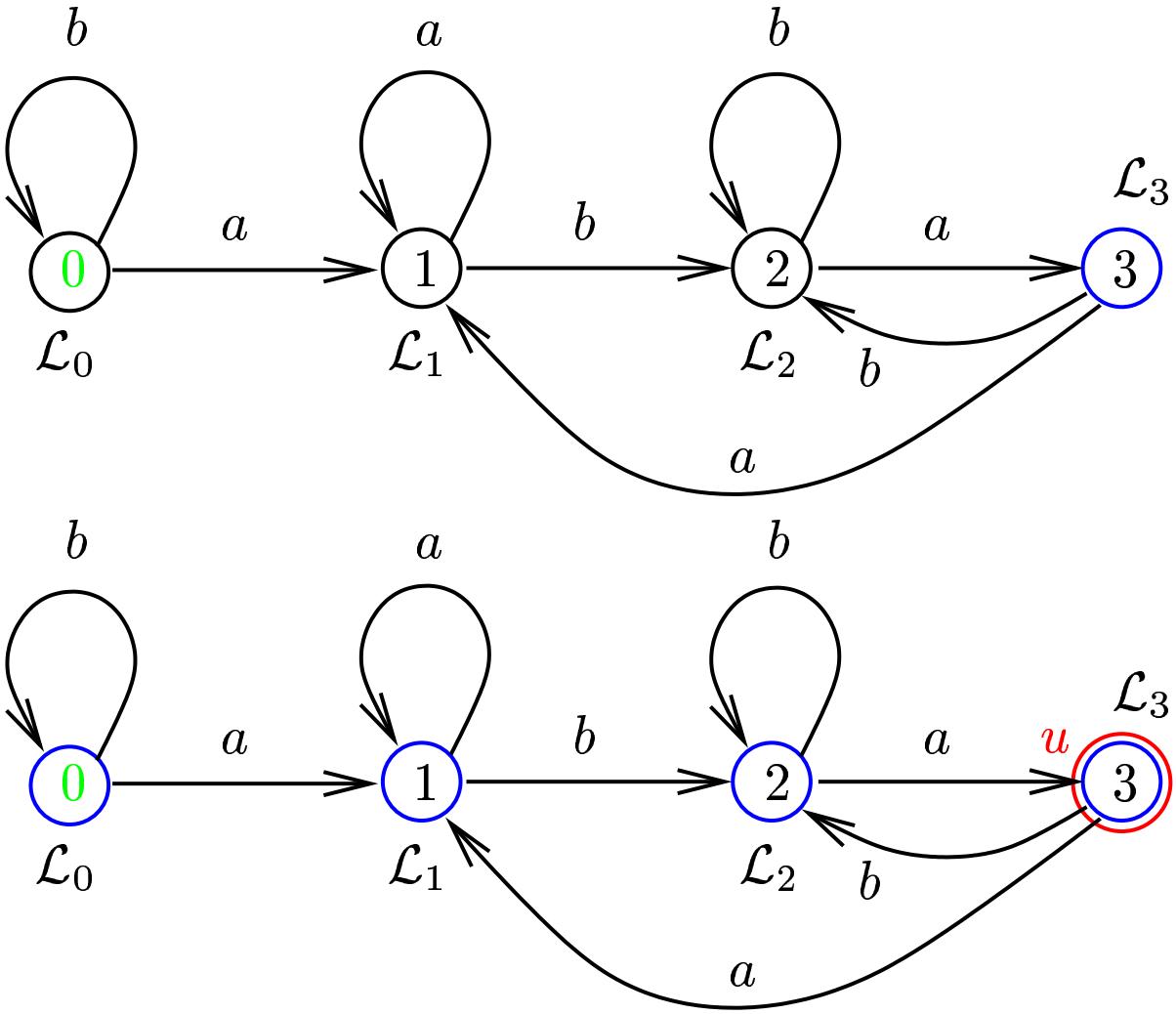
$$L_0(z) = zL_0(z) + zL_1(z) \quad L_1(z) = zL_1(z) + zL_2(z)$$
$$L_2(z) = zL_2(z) + zL_3(z) \quad L_3(z) = zL_1(z) + zL_2(z) + 1$$

We solve the system

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

Number of occurrences of $ab^\star a$

$aaaabbaubauaaaaabbbbbbbauaaaaaabbb \quad |u| = 0$



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

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

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

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

Generating function of the number of occurrences of $ab^\star a$

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

We solve the system

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

Model Bernoulli uniform

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

Expectation, standard-deviation

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

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

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

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

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

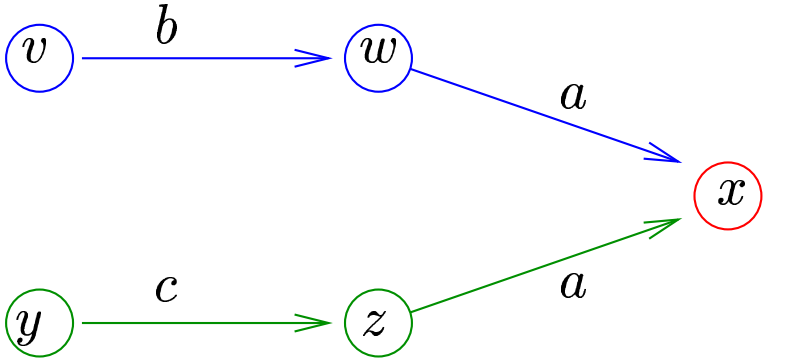
$$= \frac{3}{8} + \frac{z}{8} + \frac{1}{1-z} - \frac{1}{2(1-z)^2} + \frac{1}{8(1-z)^3} - \frac{1}{1-\frac{z}{2}}$$

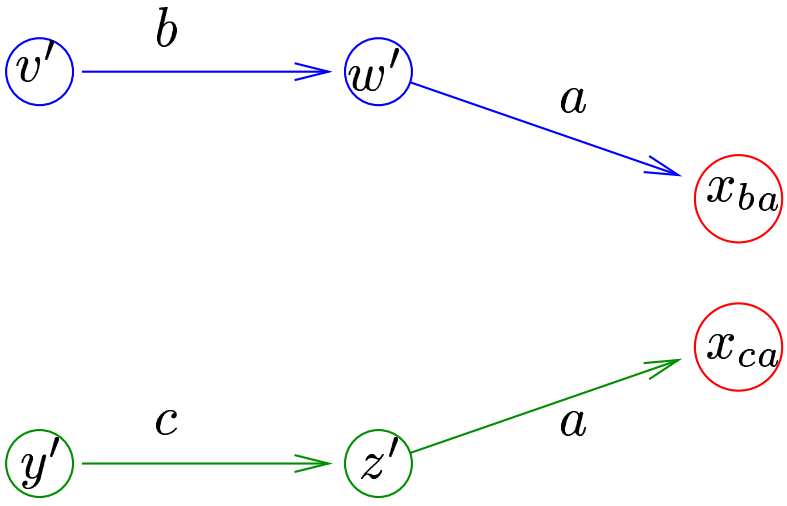
$$\Rightarrow m_{(2)n} = \frac{n^2}{16} - \frac{5n}{16} + \frac{5}{8} - \frac{1}{2^n}$$

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

Algorithm Bernoulli to Markov

Bernoulli automaton



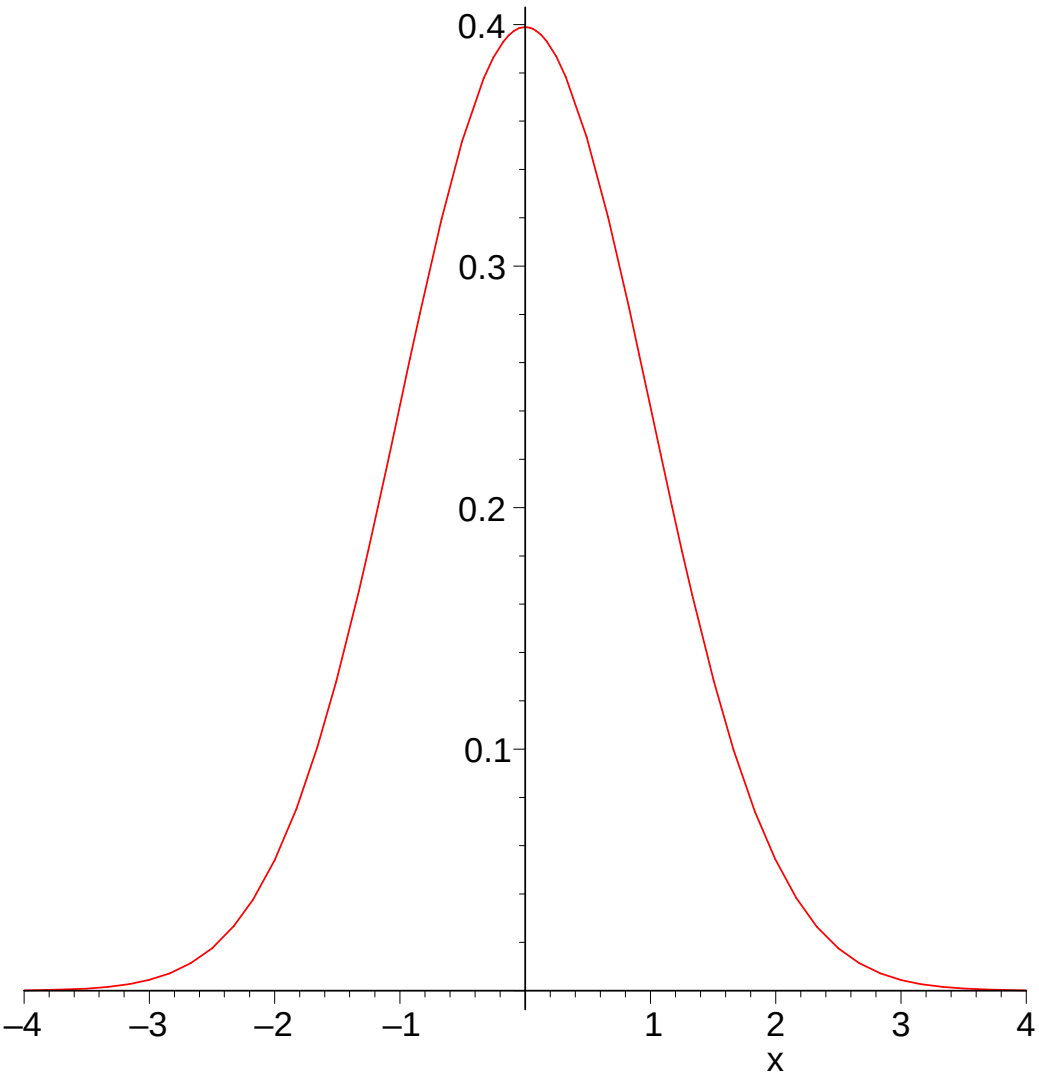


Markov automaton(order 2)

Partie V

Limit distribution of the number of occurrences

Normal law



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

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

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

Laplace Transform

X random variable

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

X discrete law, $\quad \mathbf{P}(X = k) = p_k$

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

$\mathcal{N}$ normal law

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

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x^2 - 2tx)/2} dx = e^{t^2/2}$$

Continuity theorem of Lévy

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

then $X_n \xrightarrow{d} \mathcal{N}$ $\qquad$ (convergence in distribution)

Words $ab^\star a$ - Limit law

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

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

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

We consider the normalized law $\dfrac{X_n - \mu_n}{\sigma_n}$

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

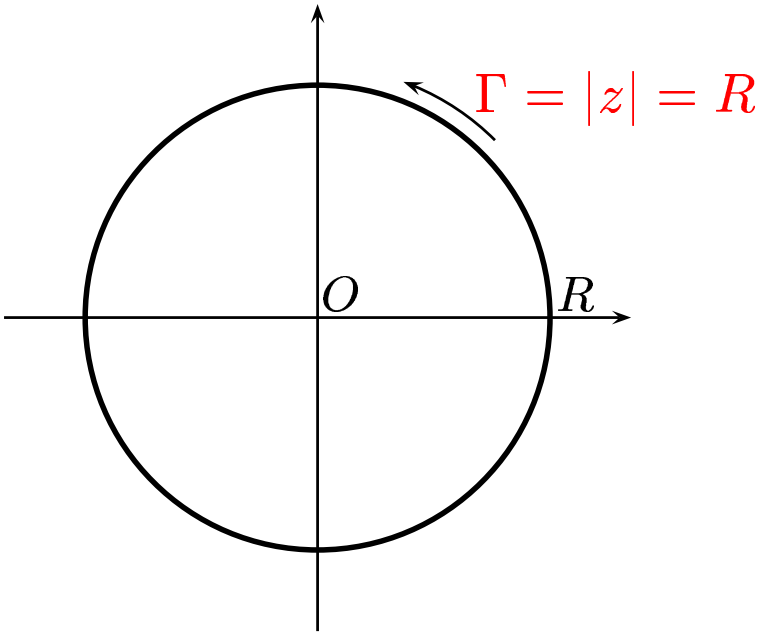
We substitute:

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

We develop at the neighborhood of $t = 0$ $\log(\Phi(t))$

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

Cauchy's integral



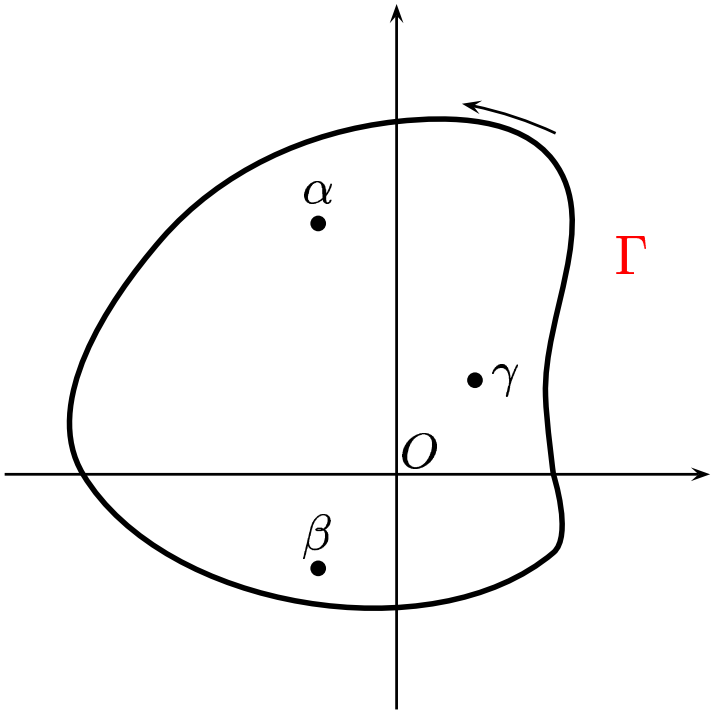
a) $k \in \mathbb{Z} - \{-1\}$

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

b) $k = -1$

change of variable $z = Re^{i\theta} \;\Rightarrow\; dz = Rie^{i\theta}d\theta$

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



$F(z)$ meromorphic inside Γ

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

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

$F(z)$ analytic inside Γ

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

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

Proof of the gaussian law

$$
\begin{aligned}
L_0(z,u) &= zp_aL_1 + zp_bL_0 + 1,\\
L_1 &= zp_bL_2 + zp_aL_1 + 1,\\
L_2 &= zp_auL_3 + zp_bL_2 + 1\\
L_3 &= zp_aL_1 + zp_bL_2 + 1
\end{aligned}
$$

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

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

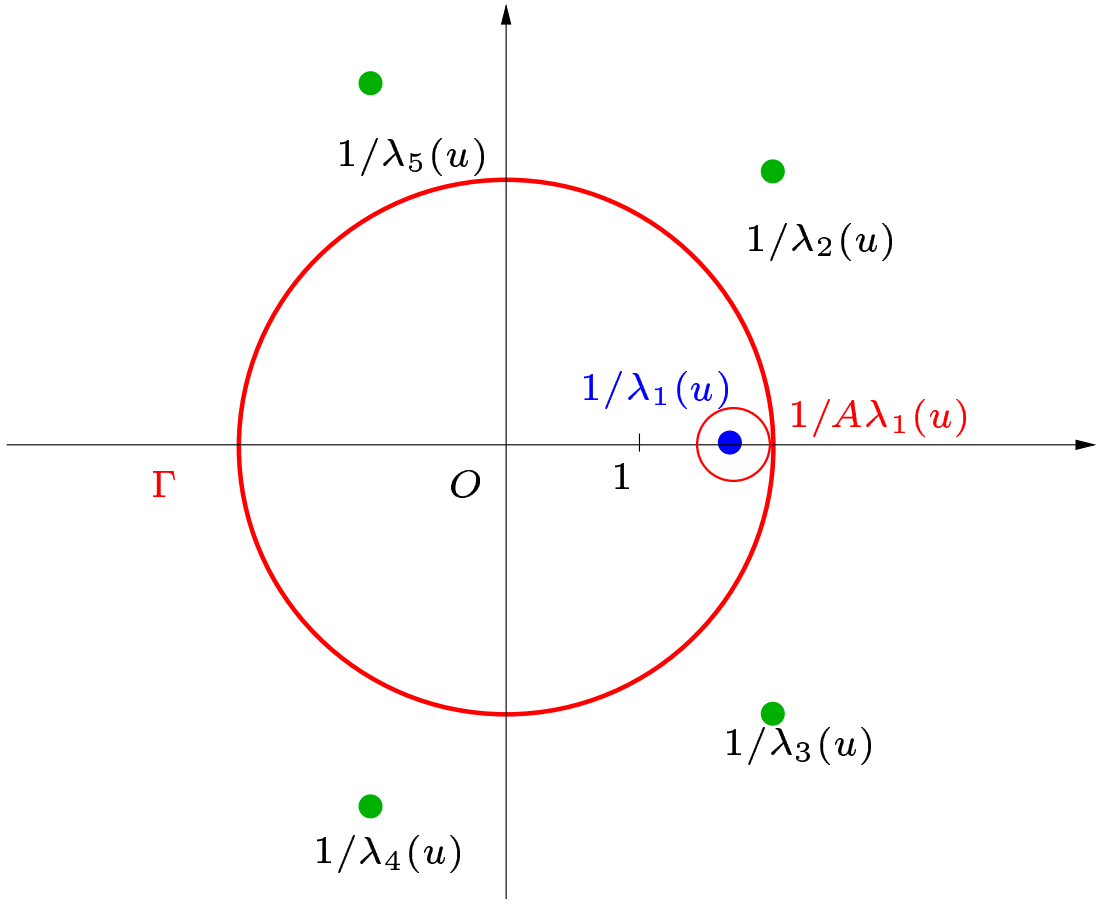
$$
L_0(z,u) = \frac{P(z,u)}{Q(z,u)} = \frac{P(z,u)}{(1-z\lambda_1(u))\cdots(1-z\lambda_n(u))}
$$

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

Perron-Frobenius

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

Property of uniform separation



$$p_n(u) = [z^n]F(z,u) = \frac{1}{2i\pi}\oint_\Gamma \frac{dz}{z^{n+1}}F(z,u),$$

$$= \frac{1}{2i\pi}\oint_\Gamma \frac{c(u)}{z^{n+1}(1-\lambda_1(u)z)} + \frac{1}{z^{n+1}}g(z,u)\,dz,$$

$$= c(u)\lambda_1(u)^n(1+O(A^n)).$$

Hwang's quasi-powers theorem

$\Rightarrow$ gaussian limit law

Part VI

Applications to biology

Over-represented words

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 conditionned by number of occurrences of AAUAAA.

Extract from "A. Denise, M. Régnier, et M. Vandenbogaert. Workshop WABI'01, and INRIA Research Report4132"

Comparison of proteomes

by statistical analysis of Prosite motifs

```
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?

biological soundness of motifs

with respect to target proteomes

Regular expressions

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

Operators

- $+$ Union
- $\cdot$ Concatenation
- $\star$ Star-operator ($A^\star = \epsilon + A + A^2 + A^3 + \dots$)

Prosite motifs

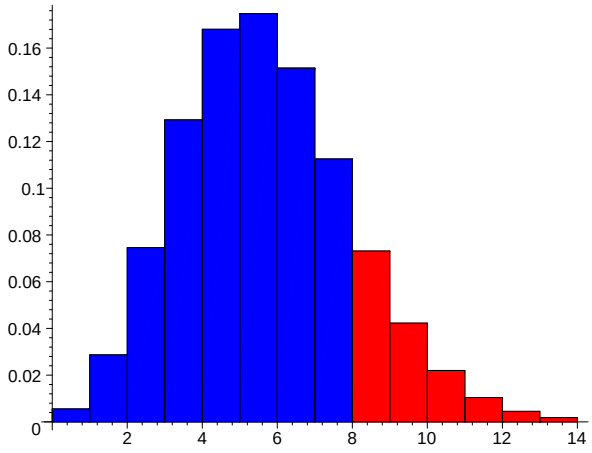
PA = [LIVM](2)-x-D-D-x(2,4)-D-x(4)-R-R-[GH].

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

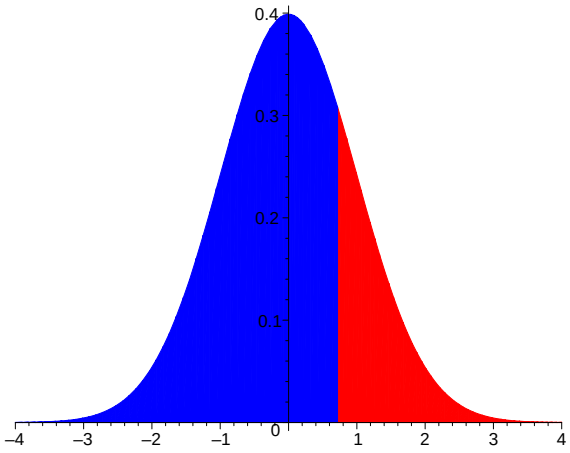
$$\text{with } \Sigma = A+R+N+D+C+Q+E+G+H+I + L+K+M+F+P+S+T+W+Y+V$$

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

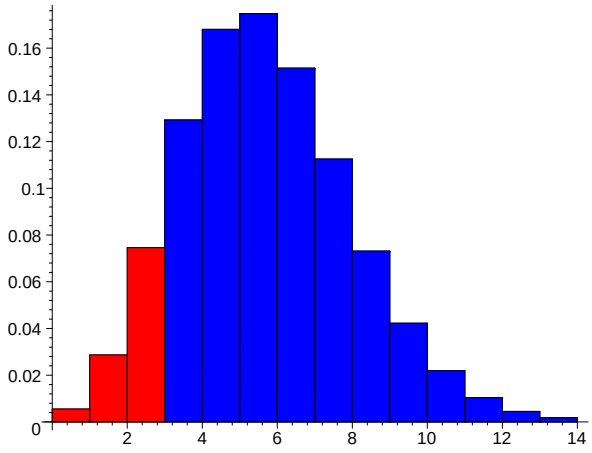
Gaussian normalisation



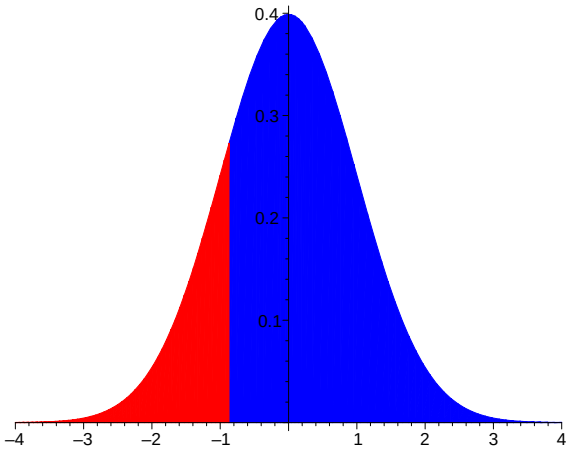
$\omega = 8$



$\gamma = +0.72$



$\omega = 2$



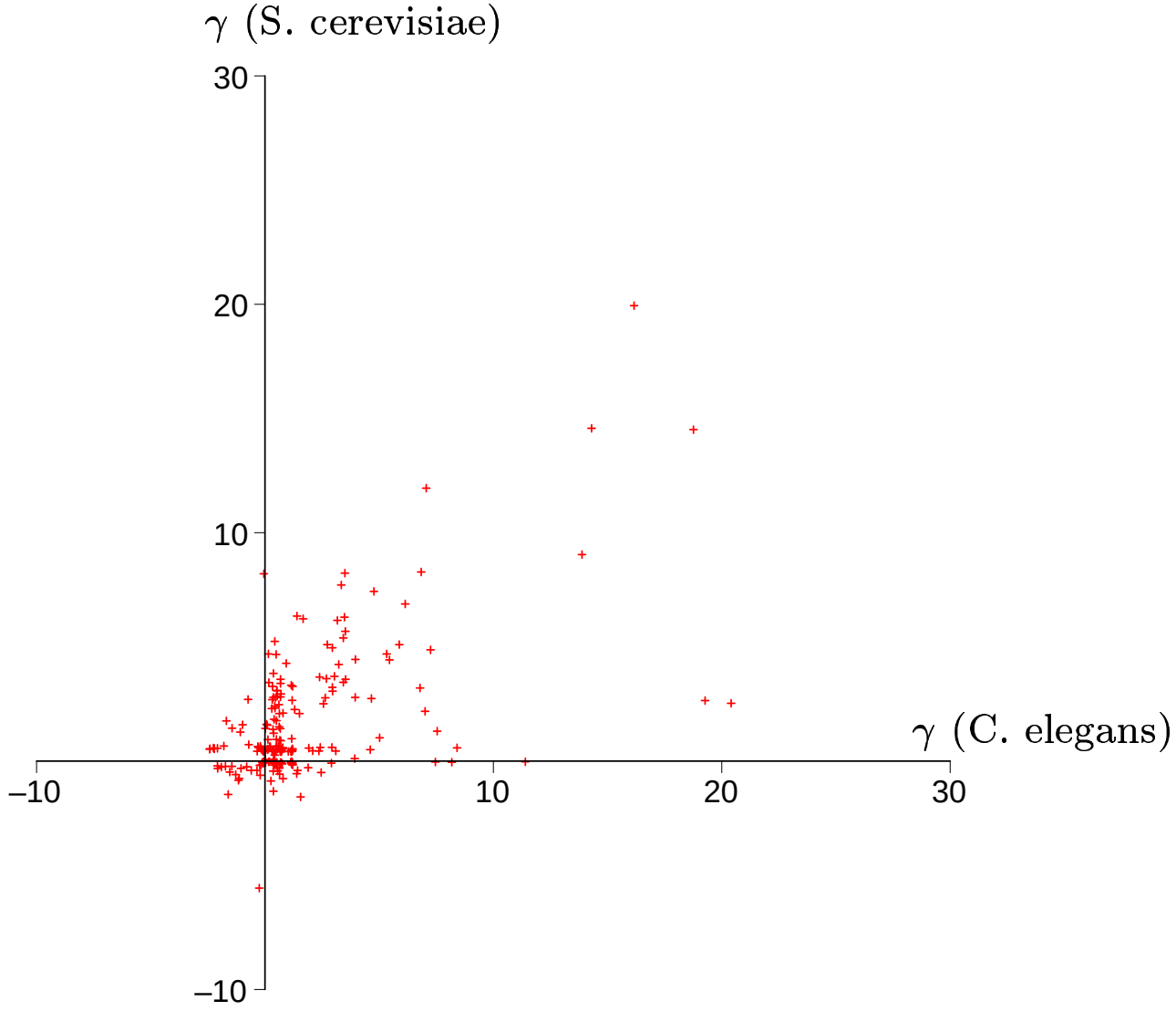
$\gamma = -0.81$

Poisson expectation $\mu = 5.2$

Gaussian distribution

ω number of observed occurrences

Comparison of worm and yeast



Comparison of Prosite motifs
for *C. elegans* and *S. cerevisiae*

- 273 motifs with $\mu \geq 0.5$ are shown -

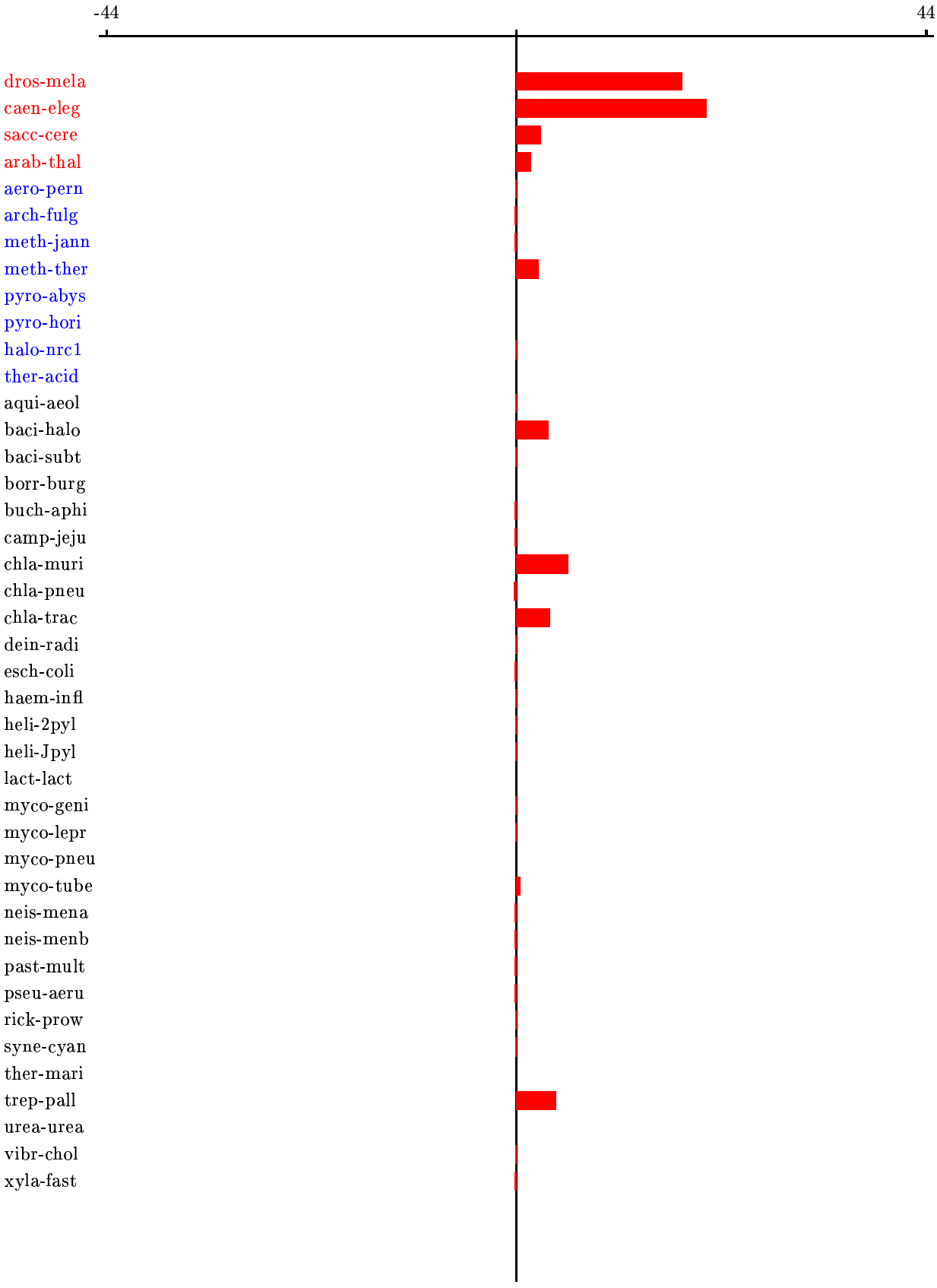
```
PS00022 GC   20.41 GS    2.54 OC    430 EC       3.666 OS      4 ES        .321
PS00022
PS00022 C-x-C-x(5)-G-x(2)-C
PS00022 EGF_1;PATTERN.   EGF-like domain signature 1.

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.
```

Exceptional motifs for the worm and yeast

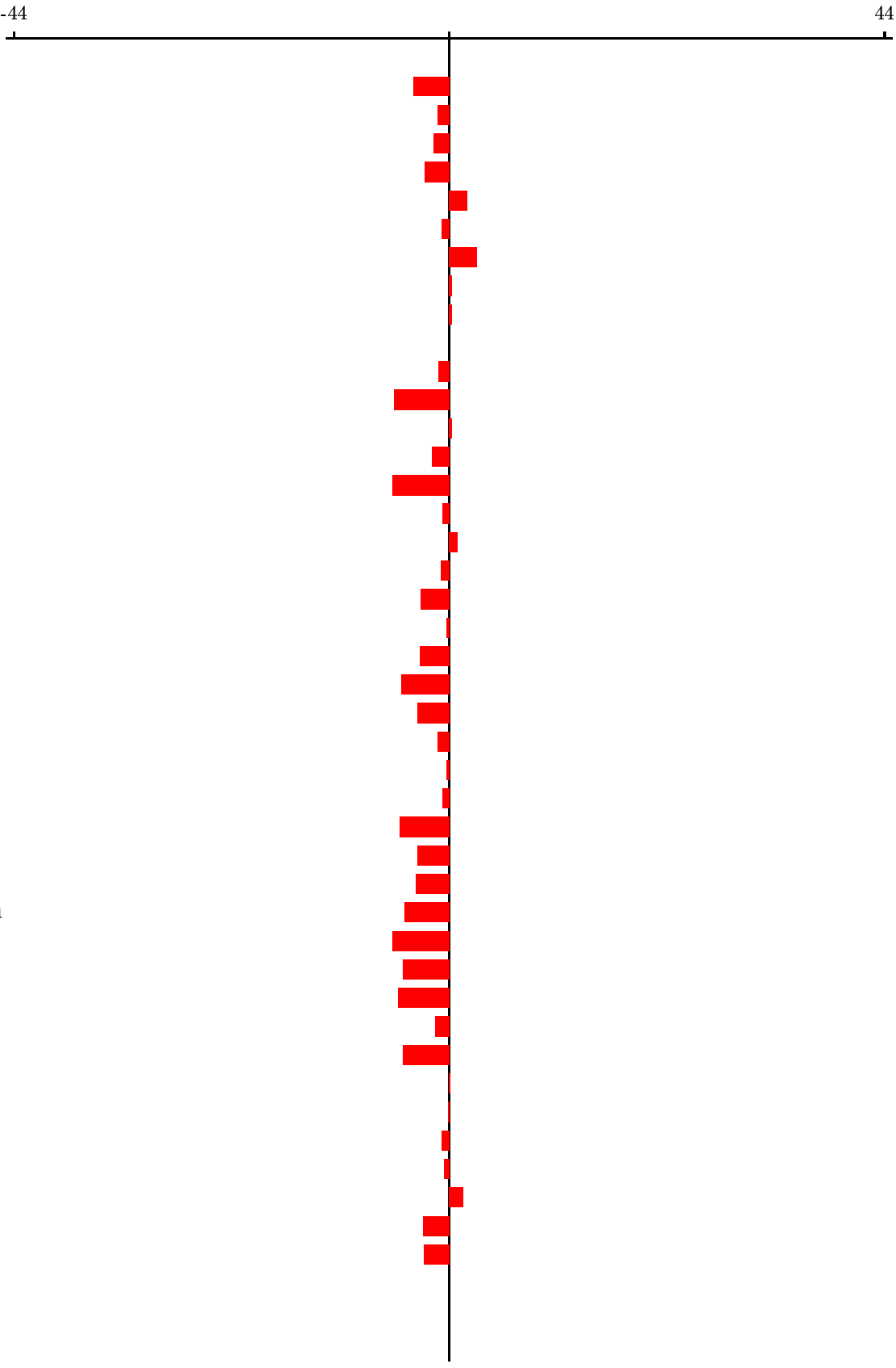
GC, OC, EC et GS, OS, ES respectively are the gaussian deviations γ, the observations ω and the expectations μ of the number of occurrences for *C. elegans* and *S. cerevisiae*.

Comparison of 42 proteomes

	Species	Length of proteome
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



dros-mela	2.71	280
caen-eleg	3.67	430
sacc-cere	.32	4
arab-thal	4.48	13
aero-pern	.05	0
arch-fulg	.08	0
meth-jann	.07	0
meth-ther	.07	1
pyro-abys	.01	0
pyro-hori	.01	0
halo-nrc1	.02	0
ther-acid	.01	0
aqui-aeol	.02	0
baci-halo	.03	2
baci-subt	.04	0
borr-burg	.01	0
buch-aphi	.02	0
camp-jeju	.05	0
chla-muri	.09	2
chla-pneu	.09	0
chla-trac	.09	1
dein-radi	.03	0
esch-coli	.16	0
haem-infl	.04	0
heli-2pyl	.04	0
heli-Jpyl	.04	0
lact-lact	.00	0
myco-geni	.00	0
myco-lepr	.04	0
myco-pneu	.01	0
myco-tube	.09	1
neis-mena	.05	0
neis-menb	.05	0
past-mult	.06	0
pseu-aeru	.16	0
rick-prow	.02	0
syne-cyan	.08	0
ther-mari	.01	0
trep-pall	.17	2
urea-urea	.00	0
vibr-chol	.09	0
xyla-fast	.10	0



dros-mela	1187.45	883
caen-eleg	1124.24	987
sacc-cere	375.28	316
arab-thal	2104.73	1738
aero-pern	163.23	184
arch-fulg	134.95	125
meth-jann	67.31	84
meth-ther	168.11	173
pyro-abys	102.87	105
pyro-hori	92.28	96
halo-nrc1	309.74	292
ther-acid	103.86	68
aqui-aeol	69.83	74
baci-halo	200.05	170
baci-subt	180.22	102
borr-burg	28.23	26
buch-aphi	16.72	18
camp-jeju	44.69	39
chla-muri	44.33	34
chla-pneu	46.59	45
chla-trac	43.40	33
dein-radi	327.35	247
esch-coli	288.99	228
haem-infl	78.23	69
heli-2pyl	46.81	45
heli-Jpyl	47.36	47
lact-lact	82.98	49
myco-geni	12.43	8
myco-lepr	180.51	149
myco-pneu	22.81	13
myco-tube	557.93	409
neis-mena	129.41	91
neis-menb	129.13	87
past-mult	95.17	82
pseu-aeru	634.32	482
rick-prow	24.79	25
syne-cyan	192.26	190
ther-mari	110.88	102
trep-pall	81.88	77
urea-urea	14.81	18
vibr-chol	191.57	153
xyla-fast	200.56	170

- Part II
  * P.Flajolet and R.Sedgewick. *An introduction to the analysis of algorithms.* Addison-Wesley, 1996, (undergraduate to graduate level)
  * W.Szpankowski. *Average case analysis of algorithms on sequences.* Wiley-Interscience, 2001, (masters level)
- Part III
  * M.Régnier and W.Szpankowski. *On pattern frequency occurrences in a markovian sequence?* Algorithmica, vol. 22, n°4, 1998, pages 631–649.
  * M.Régnier. *A unified approach to word occurrences probabilities.* Discrete Applied Mathematics, vol. 104, n°1, 2000, pages 259–280.
- Part IV
  * P.N., B.Salvy and P.Flajolet. *Motif statistics.* Theoretical Computer Science - To appear.
    Presented at ESA'99, Prague, July 1999
  * P.N., T.Doerks and M.Vingron. *Genome comparisons based on motif statistics*
    To be presented at ECCB02, Saarbrücken, October 2002
- Part V
  * R.Durrett. *Probability: Theory and examples.* Duxbury Press, 1996