Fast Approximate Motif Statistics

Pierre Nicodème

Laboratory Statistics and Genome

La Génopole - Evry

To appear in JCB

Motivation : 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;
...
```

Biological pertinence of motifs

with respect to target proteoms

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 $(A^\star = \epsilon + A + A^2 + A^3 + \dots)$

Generating functions on languages

$\mathcal{L} \in \pm^*$ $\quad (\Sigma = \{l_1, l_1, \ldots, l_k\}$ alphabet)

Counting generating function

$F(z) = \sum_{\alpha \in \mathcal{L}} z^{|\alpha|} = \sum f_n z^n$ $\quad$ ($|\alpha|$ taille de α)

Multivariate generating function

$M(l_1, l_2, \ldots, l_p) = \sum_{\alpha \in \mathcal{L}}$ commute(α)

Examples

$\Sigma = \{a, b\}$ $\quad$ (ϵ empty word)

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

$$\Rightarrow \begin{cases} F(z) = 1 + 3z^2 + z^4 \\ M(a,b) = 1 + a^2 + 2ab + a^3b \end{cases}$$

$\mathcal{L} = \{\epsilon, aab, aabaab, \ldots, (aab)^n, \ldots\}$

$$\Rightarrow \begin{cases} F(z) = \frac{1}{1-z^3} \\ M(a,b) = \frac{1}{1-a^2b} \end{cases}$$

Weighted generating function

Bernoulli model, ω_i proba. of letter l_i,

p_α probability of word α

p_α = product of proba. of the letters of the word

Univariate generating function

$F_\omega(z) = \sum_{\alpha \in \mathcal{L}} p_\alpha z^{|\alpha|} = \sum \pi_n z^n$

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

Multivariate generating function

$M_\omega(l_1, l_2, \ldots, l_p) = \sum_{\alpha \in \mathcal{L}} p_\alpha \text{commute}(\alpha)$

Examples

$\Sigma = \{a, b\} \quad \omega_a = 1/3,\ \omega_b = 2/3$

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

$$\Rightarrow \begin{cases} F_\omega(z) = 1 + \frac{5}{9}z^2 + \frac{2}{81}z^4 \\ M_\omega(a, b) = 1 + \frac{1}{9}a^2 + \frac{4}{9}ab + \frac{2}{81}a^3 b \end{cases}$$

Remark

$M(a, b) = 1 + a^2 + 2ab + a^3 b$

$M(\omega_a z, \omega_b z) = F_\omega(z)$

Combinatorial Constructions $\Rightarrow$ Generating functions

Product

If $A_1.A_2 \ldots A_j$ non ambiguous,

$F_{A_1.A_2 \ldots A_j}(z) = F_{A_1}(l_1, \ldots, l_k) \ldots F_{A_j}(l_1, \ldots, l_k)$

Union

If A and B disjoint,

$F_{A \cup B}(l_1, \ldots, l_k) = F_A(l_1, \ldots, l_k) + F_B(l_1, \ldots, l_k)$

Kleene $\star$ operator

If no ambiguity, $F_{A^*}(l_1, \ldots, l_k) = \frac{1}{1 - F_A(l_1, \ldots, l_k)}$

Counter-example

$A_1 = A_2 = \{a, aa\}, \quad \Sigma = \{a\}$

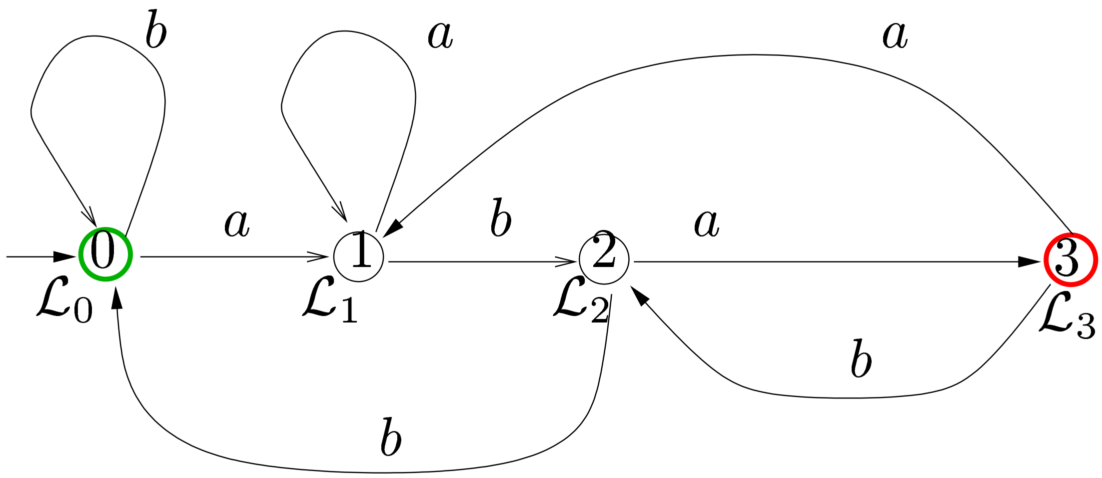
$$\Rightarrow \begin{cases} A_1 A_2 = \{aa, aaa, aaaa\} \\ F_{A_1 A_2}(a) = a^2 + a^3 + a^4 \\ \neq F_{A_1}(a) F_{A_2}(a) = a^2 + 2a^3 + a^4 \end{cases}$$

$$\Rightarrow \begin{cases} A_1 \bigcup A_2 = \{a, aa\} \\ F_{A_1 \bigcup A_2}(a) = a + a^2 \\ \neq F_{A_1}(a) + F_{A_2}(a) = 2a + 2a^2 \end{cases}$$

Automaton Recognizing $\Sigma^* R$

$$\Sigma = \{a, b\} \quad R = aba, \quad E = \Sigma^* R = \Sigma^* aba$$

$$aabbaba\bullet bbabbaba\bullet aaaba\bullet bbbb$$



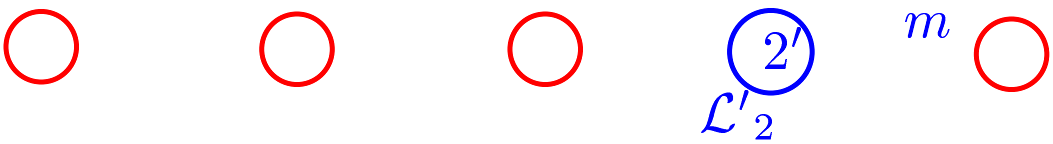
Chomsky-Schützenberger

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

$$L_0 = aL_1 + bL_0, \qquad L_1 = bL_2 + aL_1,$$
$$L_2 = aL_3 + bL_0, \qquad L_3 = aL_1 + bL_2 + 1.$$

$$L_0(a, b) = \frac{a^2 b}{1 - a - b}$$

marking ($\Sigma' = \Sigma \cup \{m\}$)



$aaaabambamaaabamaa$

$$\mathcal{L}_0 = a\mathcal{L}_1 + b\mathcal{L}_0 + \epsilon, \qquad \ldots,$$

$$\mathcal{L}_2 = m\mathcal{L}'_2 + b\mathcal{L}_0 + \epsilon, \qquad \mathcal{L}'_2 = a\mathcal{L}_3,$$

$$L(a,b,m) = \frac{1 + ab(1-m)}{1 - a - b + ab(1-m) - ab^2(1-m)},$$

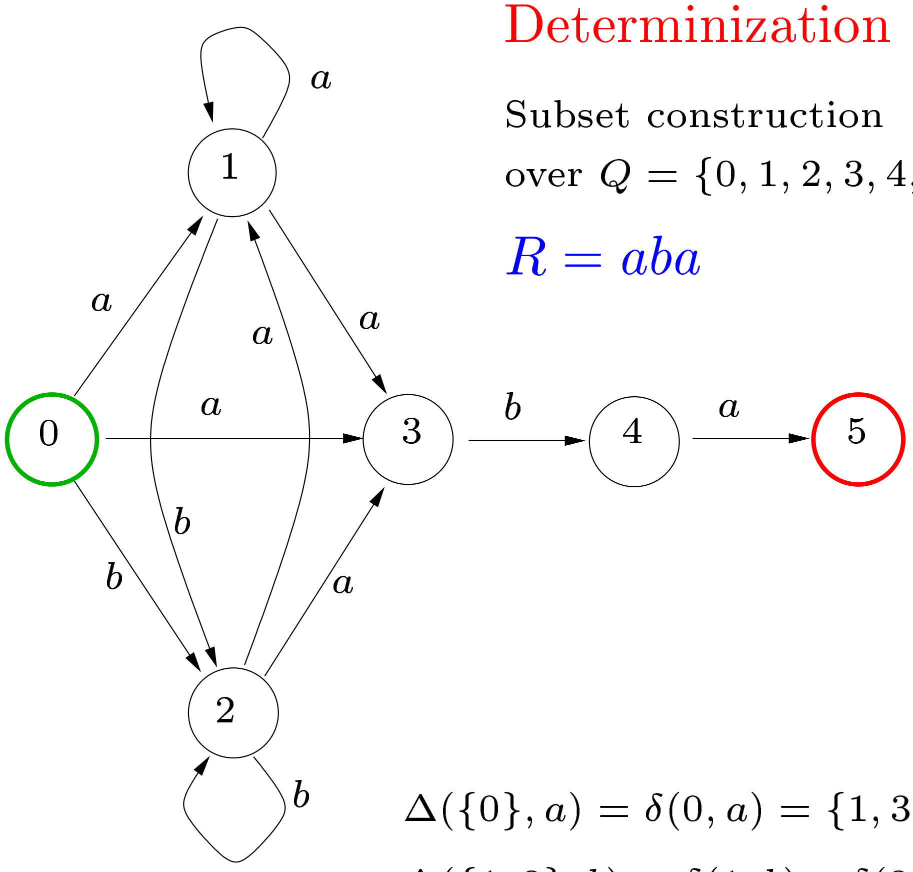
$$F(z,u) = L(p_a z, p_b z, u).$$

Determinization

Subset construction

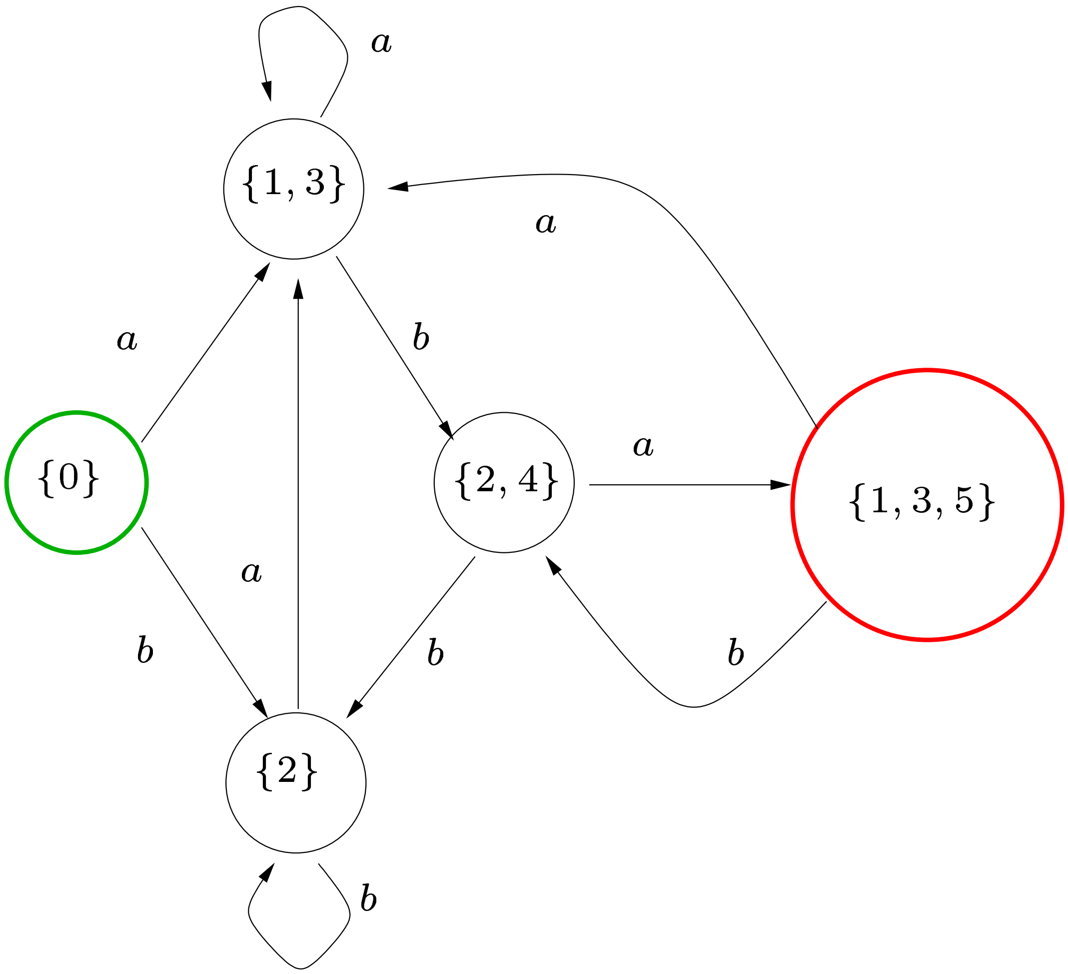
over $Q = \{0, 1, 2, 3, 4, 5\}$

$R = aba$



$$\Delta(\{0\}, a) = \delta(0, a) = \{1, 3\}$$

$$\Delta(\{1, 3\}, b) = \delta(1, b) \cup \delta(3, b) = \{2, 4\}$$



A non-ambiguity lemma

Warning

$A = B = C = \{a, b, ab\}$

$abab = a \cdot b \cdot ab = ab \cdot a \cdot b \in ABC$

$\Rightarrow ABC$ ambiguous

while AB and BC non-ambiguous

R regular expression

$N = \Sigma^* - \Sigma^* R \Sigma^*$ lang. without occ. of R

Lemma:
R non self-overlapping and non nested
$\Rightarrow \quad N(RN)^k$ non-ambiguous $\forall k$

Matches for non-overlapping languages

L non-overlapping, non-nested, O.G.F. $L(z)$

$F(z,u) = \sum p_{n,k} u^k z^n$, O.G.F. nb. matches

μ, σ^2 asympt. expect. and var. for nb. matches

Theorem 1:

(a) $F(z,u) = \frac{1}{1-z+(1-u)L(z)}$

(b) $\mu = L(1) \times n + L(1) - L'(1) + O(A^n)$

(c) $$\sigma^2 = \{L(1) + L(1)^2 - 2L(1)L'(1)\} \times n$$
$$-2L(1)L'(1) + L(1) + 2L(1)L''(1)$$
$$+L(1)^2 - L'(1) + L'(1)^2 + O(A^n)$$

$(A < 1)$

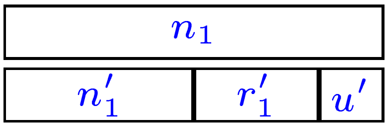
Proof of the lemma

Let us suppose NRN is ambiguous

let $w = n_1 r_1 n_2 = n'_1 r'_1 n'_2 \in NRN$, $\{n_1, n_2, n'_1, n'_2\} \subset N$, $\{r_1, r'_1\} \subset R$

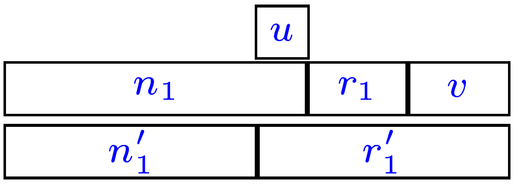
Let us suppose $|n_1| > |n'_1|$ $\quad$ ($|n_1| < |n'_1|$ is symmetrical)

$|n_1| > |n'_1 r'_1|$



$\Rightarrow n_1 \notin N$

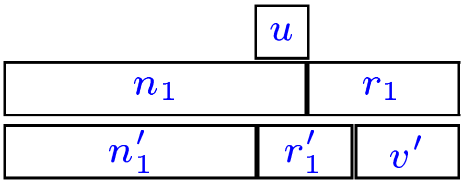
$|n_1 r_1| < |n'_1 r'_1|$



$\Rightarrow$ $r'_1 = u r_1 v$

R nested

$|n_1 r_1| > |n'_1 r'_1|$



$\Rightarrow$ $u r_1 = r'_1 v$

R overlapping

$\Rightarrow \quad n_1 = n'_1 \quad \Rightarrow \quad r_1 = r'_1 \quad \Rightarrow \quad n_2 = n'_2$

(elsewhere R is overlapping)

$\Rightarrow$ NRN non-ambiguous

$\Rightarrow$ $N(RN)^k$ $\quad (k > 1)$ non-ambiguous (same proof)

Proof of Theorem 1

$\mathcal{U}_L$ mark. process - add u after each match

$\mathcal{U}_L(t) = t_u \qquad (t \in \Sigma^\star, u \notin \Sigma)$

$\mathcal{U}_{abb}(aaabbbabbaa) = aaabbubabbuaa$

$t_u \to \phi_{t_u}(z, u) = \pi(t) u^k z^{|t|}$

($\pi(t)$ probability of word t, $\qquad k$ nb. matches)

$$F(z,u) = \sum_{t_u \in \Sigma_u^\star} \phi_{t_u}(z,u) = \sum_{t \in \Sigma^\star} \pi(t) u^k z^{|t|}$$

$$= \sum_{n=|t|\geq 0} z^n \sum_{u \geq 0} p_{n,k} u^k = \sum p_{n,k} u^k z^n$$

$p_{n,k}$ proba k matches in text size n

$\Sigma_u^\star = \bigcup_{t \in \Sigma^\star} = (NLu)^\star N$

$\Rightarrow F(z,u) = \frac{N(z)}{1-uN(z)L(z)}$

$\Rightarrow \quad F(z,1) = \frac{1}{1-z} = \frac{N(z)}{1-N(z)L(z)}$

elimin. $N(z) \quad \Rightarrow \quad F(z,u) = \frac{1}{1-z+(1-u)L(z)}$

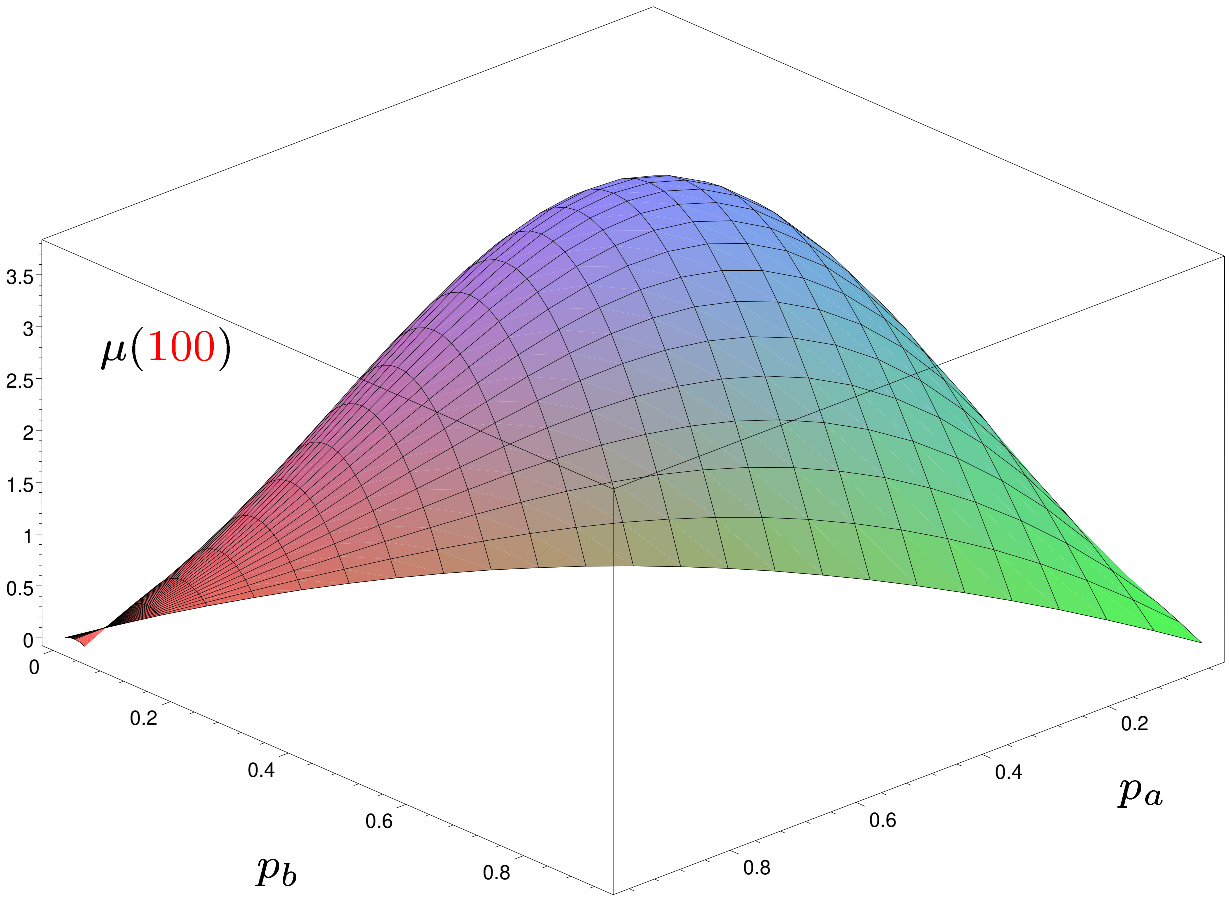
Application

$$L = \bigcup_{k\geq 1} a^k b^k c^k$$

$$L(z) = \frac{p_a p_b p_c z^3}{1 - p_a p_b p_c z^3} \qquad\qquad (p_a + p_b + p_c = 1)$$

$$F(z,u) := \frac{1}{1 - z + \frac{(1-u)\, p_a\, p_b\, p_c\, z^3}{1 - p_a\, p_b\, p_c\, z^3}}$$

$$\mu(n) := \frac{p_a\, p_b\, p_c}{1 - p_a\, p_b\, p_c} n - 2\,\frac{p_a\, p_b\, p_c}{1 - p_a\, p_b\, p_c} - 3\,\frac{{p_a}^2\, {p_b}^2\, {p_c}^2}{(1 - p_a\, p_b\, p_c)^2}$$



Poisson law for rare occurrences

L non overlapping, non nested, O.G.F. $L(z)$

$p_{n,k}$ prob. k matches in text size n

Theorem 2:

(a) $\exists \rho \in \mathbb{R}^+$, $\rho < 1$, and $1 - \rho + L(\rho) = 0$

(b) for $k = O(1)$, $\exists \kappa > \rho$

$p_{n,k} = \sum_{j=1}^{k+1} (-1)^j a_j \binom{n}{j-1} \rho^{-(n+j)} + O(\kappa^{-n})$

and

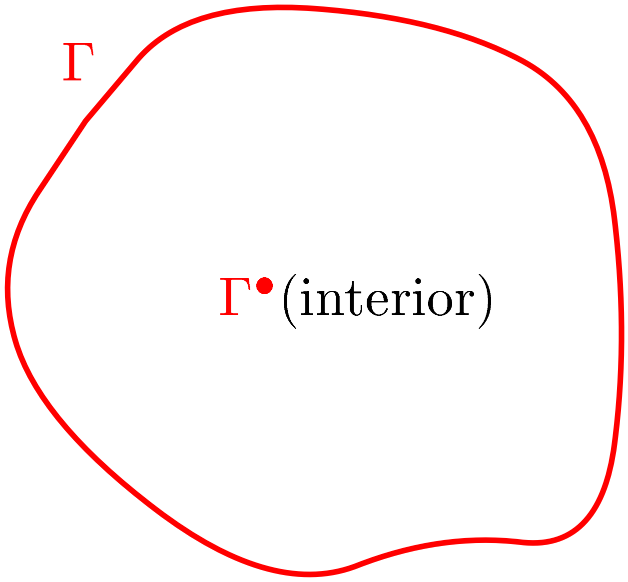
$a_{k+1} = \frac{L(\rho)^k}{\rho^{k+1}(L'(\rho)-1)^{k+1}}$

(c) asymptotically, for $k = O(1)$

$$p_{n,k} = \frac{\rho^{-n-1}}{1 - L'(\rho)} \frac{(\alpha n)^k}{k!} \left(1 + O\left(\frac{1}{n}\right)\right),$$

$$\alpha = \left(\frac{\rho^{-1} L(\rho)}{1 - L'(\rho)}\right)$$

Rouché's Theorem



1) f and g differentiable in $\{\Gamma^\bullet$ - finite nb. poles$\}$

2) $0 < |f(z)| < \infty \quad$ for $z \in \Gamma$

3) $|g(z)| < |f(z)| \quad$ for $z \in \Gamma$

$\Rightarrow \quad f + g \quad$ and $\quad f$

have same number of zeroes inside $\Gamma^\bullet$

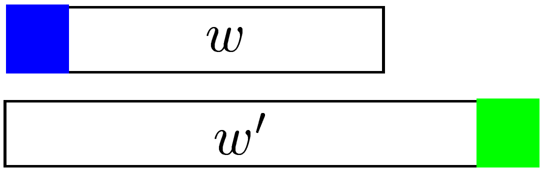
Demonstration: homotopy on

$$\frac{1}{2\pi i}\int_\Gamma \frac{f' + tg'}{f + tg}, \qquad t \in [0,1]$$

Proof of Theorem 2

$$F(z,u) = \frac{1}{1-z+(1-u)L(z)} = \sum_{k\geq 0} \frac{L(z)^k}{(1-z+L(z))^{k+1}} u^k$$

asymptotics come from zeroes of $1 - z + L(z)$



$B = \{b, \quad b \cdot w \in L\}$

$T = \{t, \quad w' \cdot t \in L\}$

L non overlapping $\quad \Rightarrow \quad T \subseteq \Sigma - B$

$b \in B, t \in T \quad \Rightarrow$ *bbt* and *btt* $\quad$ (of $B\Sigma T$) overlap

$\Rightarrow \quad L \subset B\Sigma^\star T \subseteq \Phi = B\Sigma^\star(\Sigma - B)$

$L(z) < \Phi(z) = \pi_B z \frac{1}{1-z}(1-\pi_B)z \leq \frac{z^2}{4(1-z)}$

On $\Gamma = \{|z| = 2\} \qquad |L(z)| < 1 \leq |1-z|$

Apply Rouché's Theorem with
$f = 1 - z, \quad f + g = 1 - z + L(z)$

$\Rightarrow 1 - z + L(z)$ has only one root ρ inside Γ

Perron-Frobenius $\quad \Rightarrow \quad \rho \in \mathbb{R}^+$

$L \quad$ prefix-free $\quad \Rightarrow \quad 1 < \rho < \tau$ (radius conv $(f+g)$)

Computation of $R(z), R'(z), R(1), R'(1)$

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

$R(z) = (p_L + p_I + p_V + p_M)^2 z^2 \times z \times p_D z \times p_D z \times (z^2 + z^3 + z^4) \times p_D z \times z^4 \times p_D z \times p_D z \times (p_G + p_H) z$

$= (p_L + p_I + p_V + p_M)^2 p_D^5 (p_G + p_H) z^{13} (z^2 + z^3 + z^4)$

$\rightarrow R'(z), R(1), R'(1)$

Validation of the approximation

Comparison of expectations computed

1. by the exact method (automata) - μ_e
2. by the non-overlapping approximation - μ_a

rate of error r

$$r = \frac{|\mu_a - \mu_e|}{\mu_e}$$

for 96% motifs $r < 1\%$

largest error $r = 34\%$

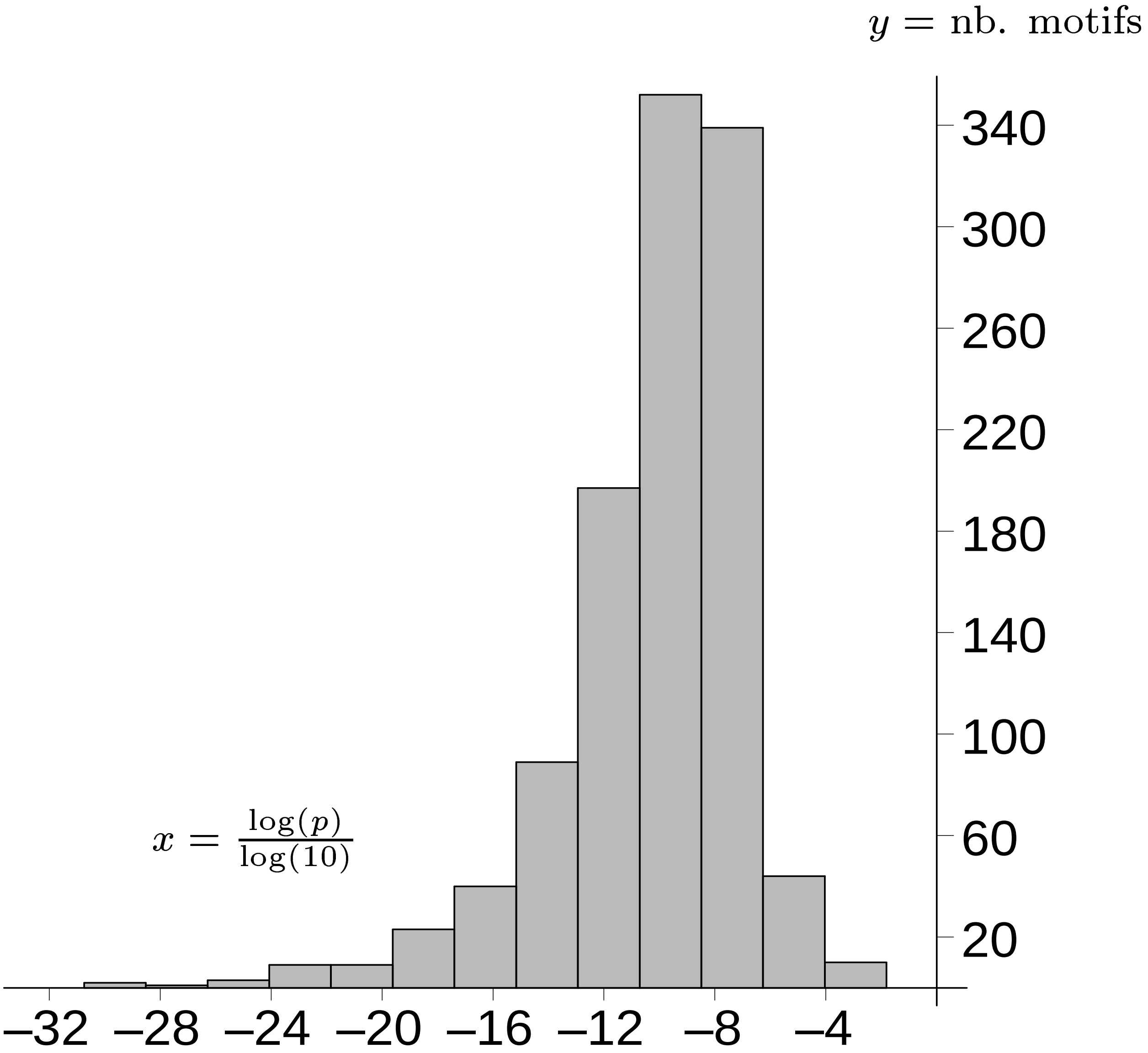


Figure 1: Histogram of probability of occurrence p of a motif at a given position for the 1118 motifs exactly computed. (ProDom distribution for amino-acids)

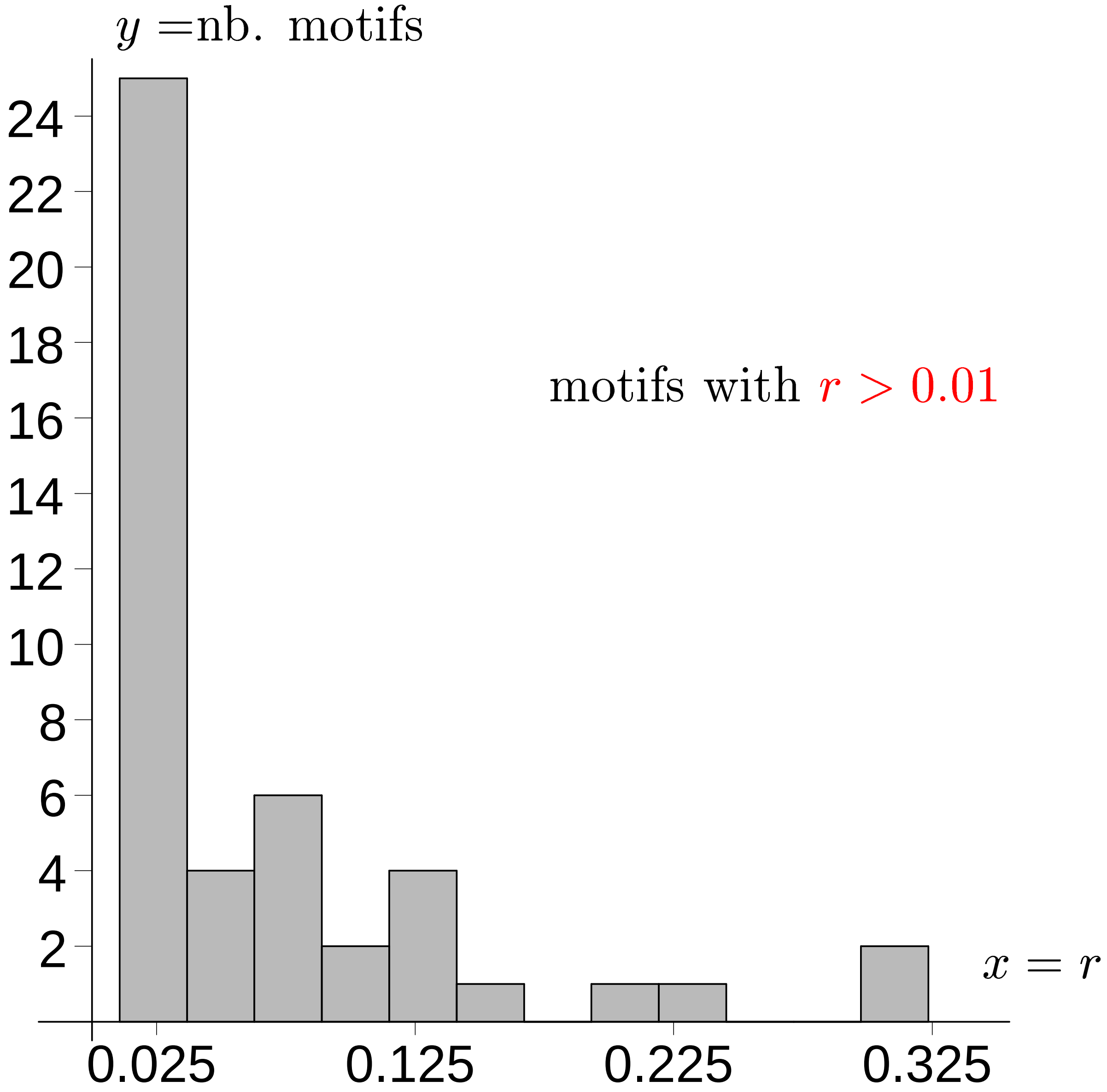


Figure 2: Rate of error r for the 1118 motifs exactly computed. E_e is the exact expectation, E_a is the approximate expectation on a sequence of length 6,756,720 positions with the ProDom amino-acids distribution

Algorithmic chain

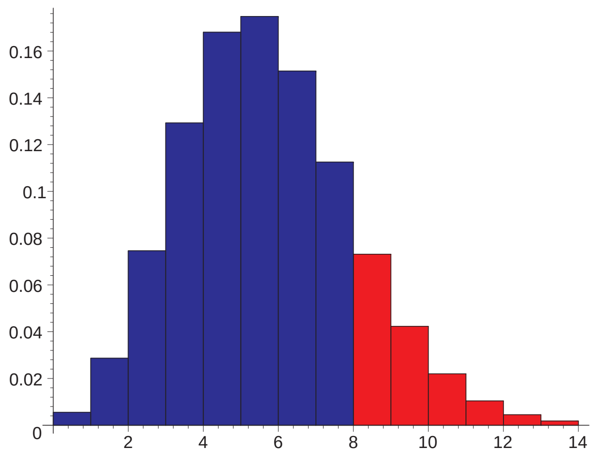
Input

- a proteom (text, lenght, a.a. distribution)
- a (set) of motif(s)

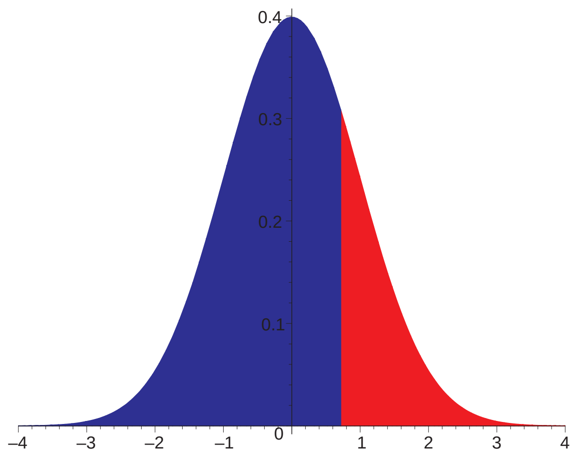
Steps of the probabilistic calibration

1. compute $\mu = \mu_a$ (approximate)
2. if $\mu > 10$ compute $\mu = \mu_e$ (exact)
3. compute the observed ω in proteom
4. compute $\Pr\{\mathrm{Poisson}_\mu > \omega\}$ (resp. $<$)
5. evaluate the gaussian quantile

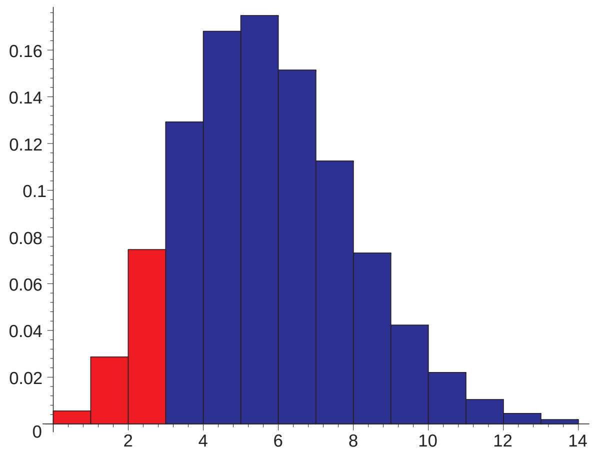
Gaussian quantile equivalence



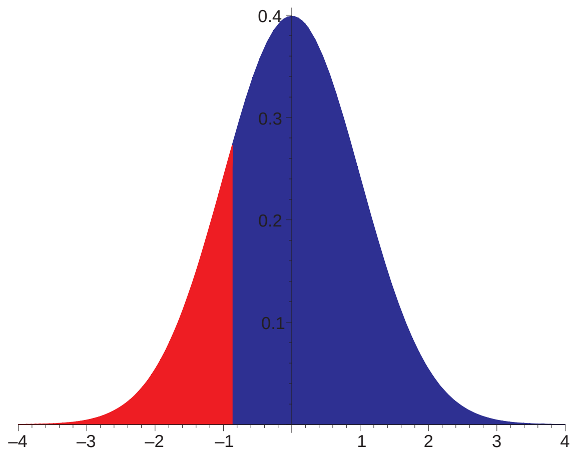
$\omega = 8$



$\gamma = +0.72$



$\omega = 2$

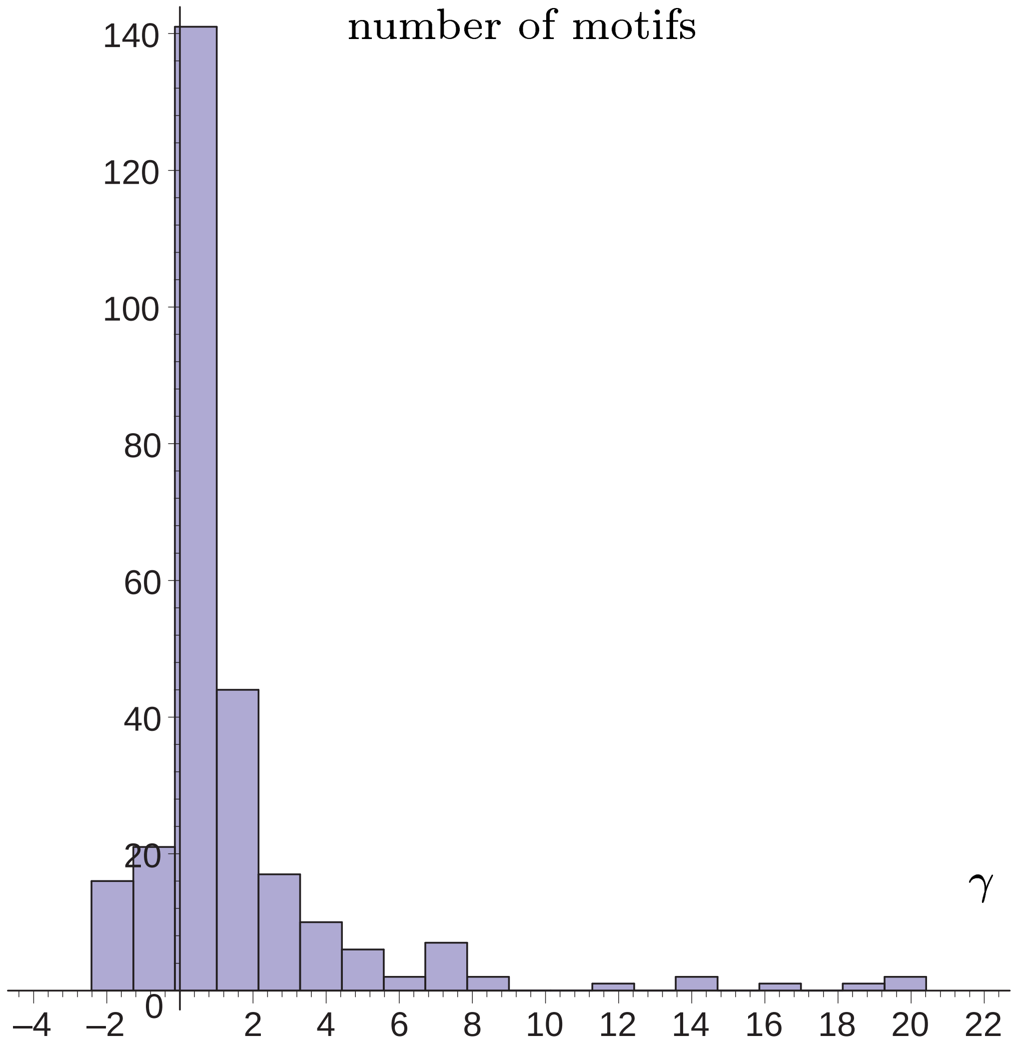


$\gamma = -0.81$

Poisson expectation $\mu = 5.2$

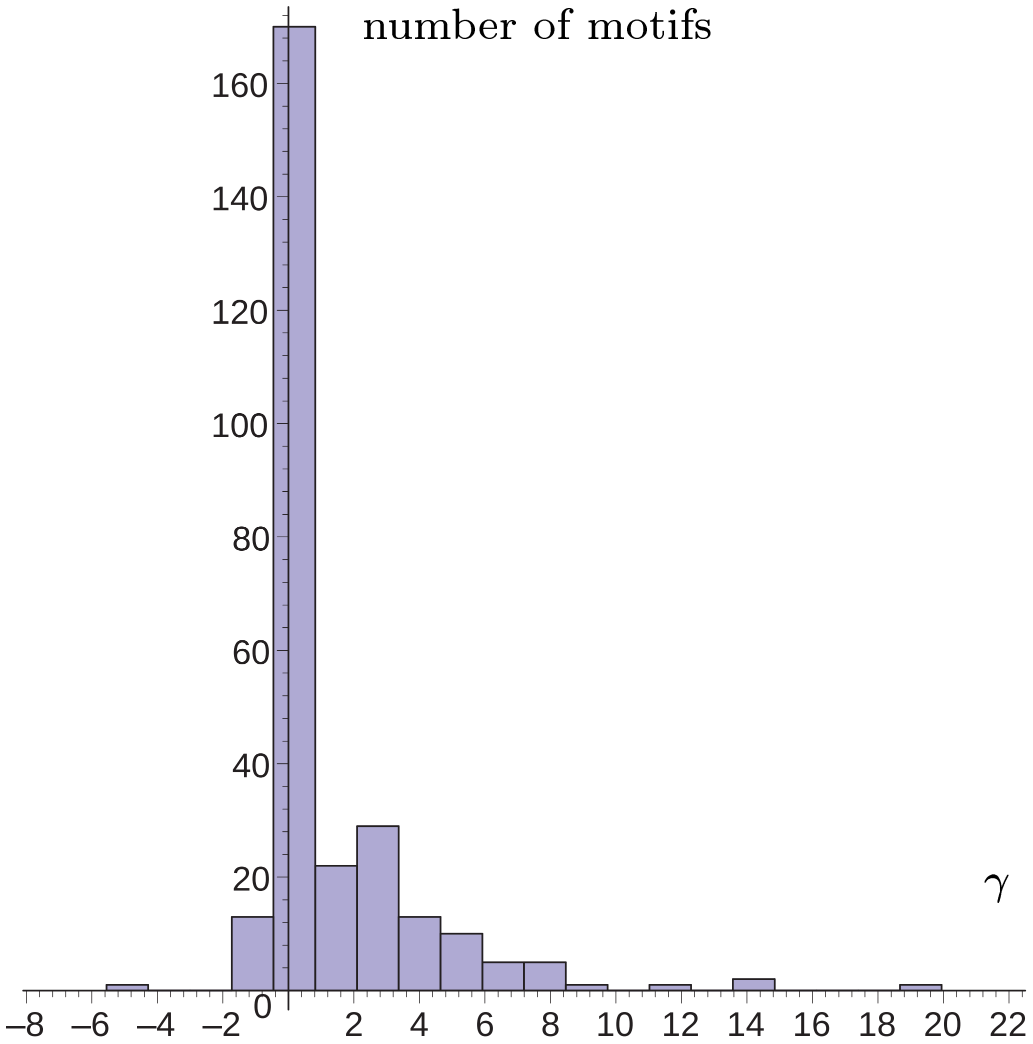
Normal distribution

ω number of matches observed



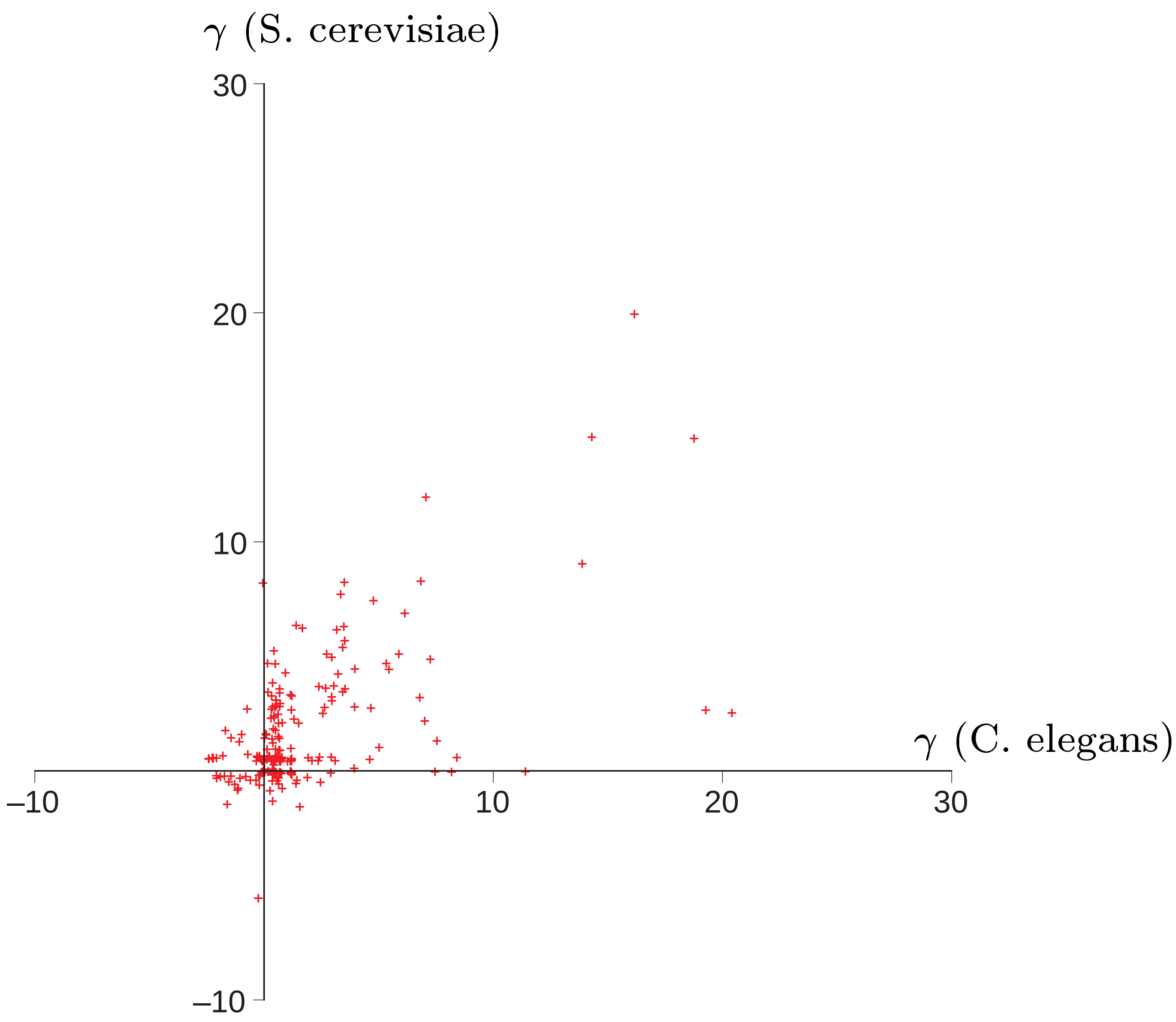
Histogram of gaussian quantiles γ of Prosite motifs for *C. elegans*

- 273 motifs plotted ($E \geq 0.5$) -



Histogram of gaussian quantiles γ of Prosite motifs for *S. cerevisiae*

- 273 motifs plotted ($E \geq 0.5$) -



Gaussian quantiles of the Prosite motifs for *C. elegans* and *S. cerevisiae*

- 273 motifs with $E \geq 0.5$ are plotted -

	A	R	N	D	C	Q	E	G	H	I
C. El.	.0622	.0513	.0496	.0523	.0207	.0408	.0641	.0531	.0232	.0624
S. Ce.	.0557	.0444	.0612	.0580	.0129	.0392	.0648	.0499	.0217	.0656
	L	K	M	F	P	S	T	W	Y	V
C. El.	.0870	.0644	.0262	.0497	.0484	.0804	.0586	.0111	.0322	.0623
S. Ce.	.0956	.0729	.0209	.0448	.0438	.0898	.0589	.0104	.0336	.0561

Probability of amino-acids occurrence in *C. elegans* and *S. cerevisiae.*

```
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 C. elegans

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

Proof of the Gaussian Law

$$L_0(z,u) = zp_aL_1 + zp_bL_0 + 1,$$
$$L_1 = zp_bL_2 + zp_aL_1 + 1,$$
$$L_2 = zp_a u L_3 + zp_bL_0 + 1$$
$$L_3 = zp_aL_1 + zp_bL_2 + 1$$

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

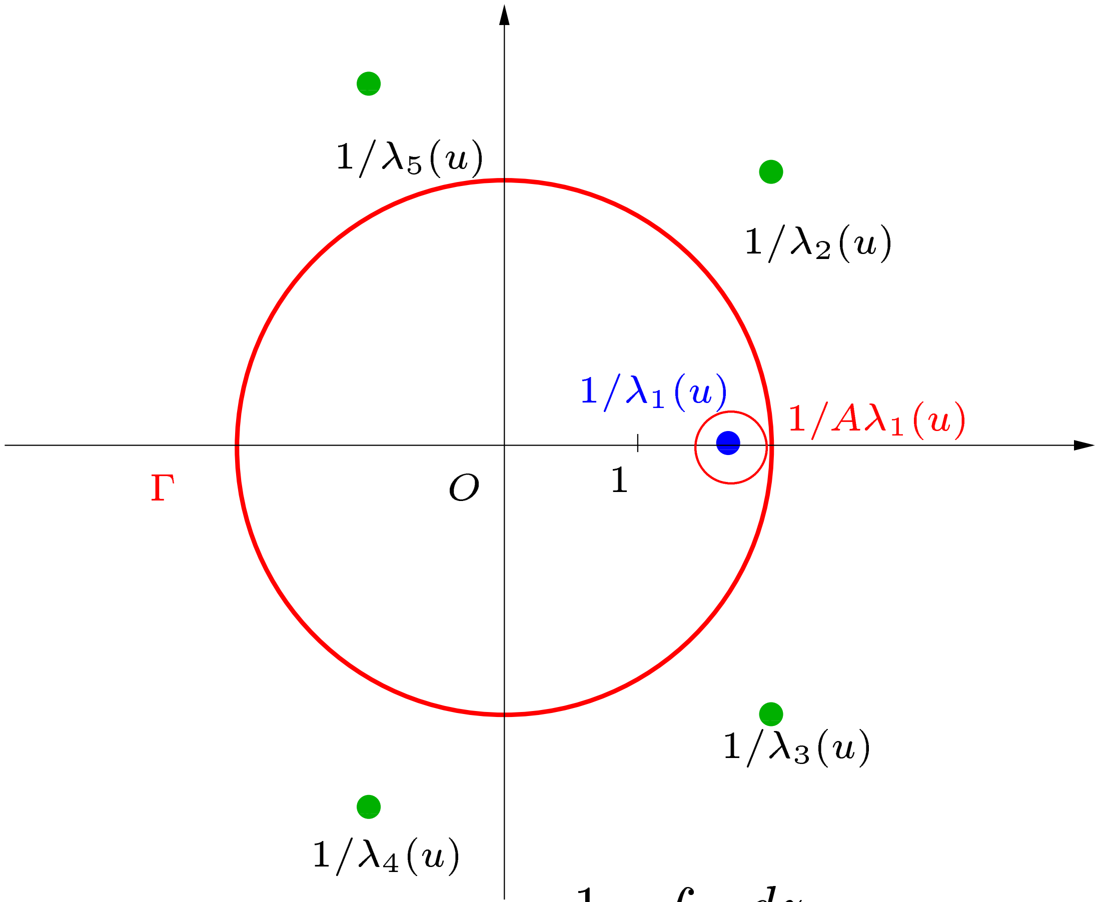
$T(u)$ positive $n \times n$ matrix

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

$$1/|\lambda_1| \leq 1/|\lambda_2| \leq \ldots$$

Perron-Frobenius: $\lambda_1(u)$ unique, real, positive.

Uniform Separation Property



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

Hwang's quasi-power theorem $\rightarrow$
limiting Gaussian distribution.