Motif Statistics

Pierre Nicodème

German Cancer Center DKFZ – Heidelberg

Joint work with

Bruno Salvy and *Philippe Flajolet*

Inria - Rocquencourt

ESA'99

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 a target genome

More generally: statistics of the number of occurrences of a regular expression in a random text.

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

Aim & Result

R given regular expression.
X_n number of occurrences in a text of length n.

Aim: $$F(z,u) = \sum_{n,k} \Pr(X_n = k) u^k z^n.$$

Theorem. With or without counting overlap, both in the Bernoulli and Markov model,
($i.$) $F(z,u)$ is rational and can be computed explicitly

($ii.$) $$\begin{cases} \mathrm{E}(X_n) & = \mu n + c_1 + O(A^n), \\ \mathrm{Var}(X_n) & = \sigma^2 n + c_2 + O(A^n). \end{cases}$$

($iii.$) Limit Gaussian law:

$$\Pr\left(\frac{X_n - \mu n}{\sigma\sqrt{n}}\right) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt.$$

Commentaries : the Gaussian law is a new result that generalises previous works (words, finite set of words).

Guibas & Odlyzko, Régnier, Régnier & Szpankowski

Explicit computation is implemantable and fast.

The Right Rational Language

R regular expression over Σ

Key: $\mathcal{L} \subset (\Sigma \cup \{m\})^\star$ contains all the words of $\Sigma^\star$, with a mark (m) after each occurrence of R.

Example: $R = aba$

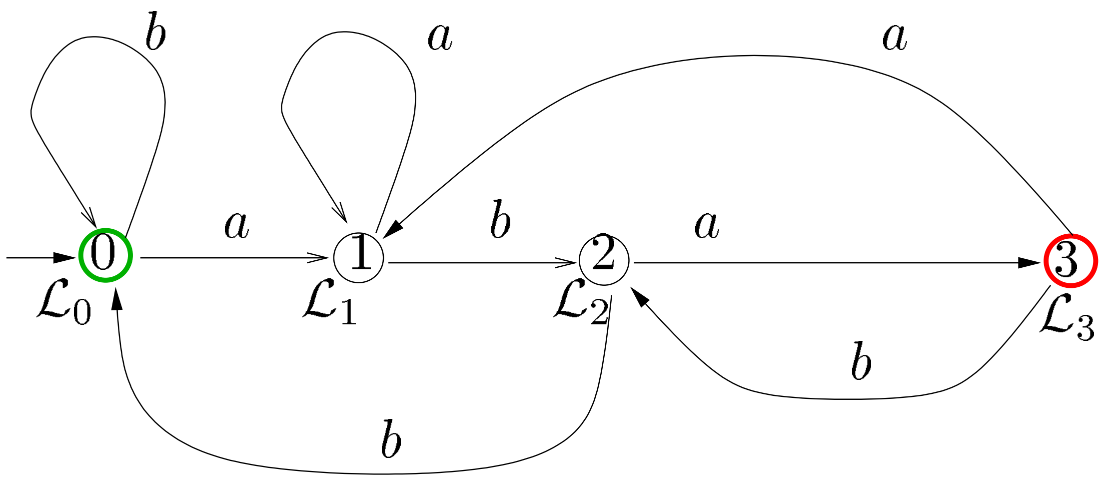
$aaaabambamaaabamaa$ (overlap)

$aaaabamba \quad aaabamaa$ (non-overlap).

Automaton Recognizing $\Sigma^* R$

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

$$aabb\,aba \bullet bbabb\,aba \bullet aa\,aba \bullet bbbb$$



Chomsky-Schützenberger

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

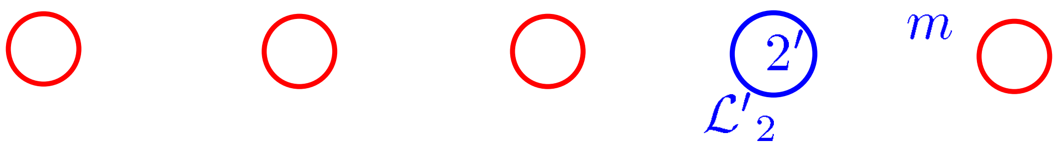
$$\mathcal{L}_2 = a\mathcal{L}_3 + b\mathcal{L}_0, \qquad \mathcal{L}_3 = a\mathcal{L}_1 + b\mathcal{L}_2 + \epsilon,$$

$$L_0 = aL_1 + bL_0, \qquad L_1 = bL_2 + aL_1,$$

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

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

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



$aaaabambamaaabamaa$

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

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

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

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

The algorithmic chain

Input: regular expression R

1. Berry-Sethy $\mapsto$ NFA for $\Sigma^* R$

2. Determinisation $\mapsto$ DFA for $\Sigma^* R$

3. Marking $\mapsto$ marked DFA for $\Sigma^* R$

4. Chomsky-Schützenberger $\mapsto$ $F(z, u)$,

$$F(z, u) = \sum p_{n,k} u^k z^n,$$

$p_{n,k}$: probability that a word of size n contains k occurrences of R.

we see now how to exploit the results and to obtain the theorem

Exploiting the Output

$$F(z,u) \in \mathbb{Q}(z,u) \Rightarrow \begin{cases} G(z) = \sum \mathrm{E}(X_n) z^n \in \mathbb{Q}(z), \\ H(z) = \sum M_2(X_n) z^n \in \mathbb{Q}(z), \\ N(z) = \sum \Pr(X_n \geq 1) z^n \in \mathbb{Q}(z). \end{cases}$$

– Fast extraction of coefficients: nth coefficient in $O(\log n)$ operations [implemented in gfun].

– Exponentially good asymptotics in constant time.

($\lambda_{\max} = 1$). To the proof of the theorem and asymptotic results

Proof of the Gaussian Law

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

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

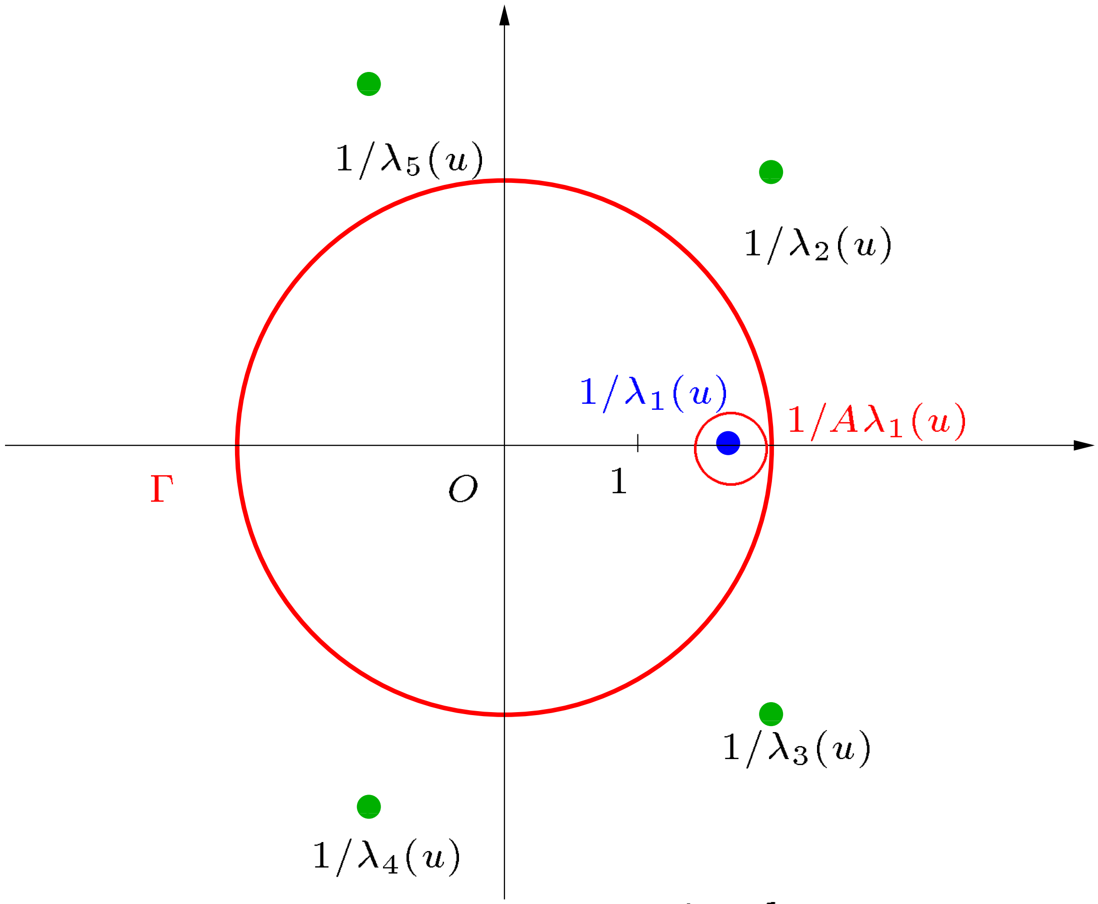
$T(u)$ positive $n \times n$ matrix

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

$$
1/|\lambda_1| \le 1/|\lambda_2| \le \dots
$$

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

Uniform Separation Property

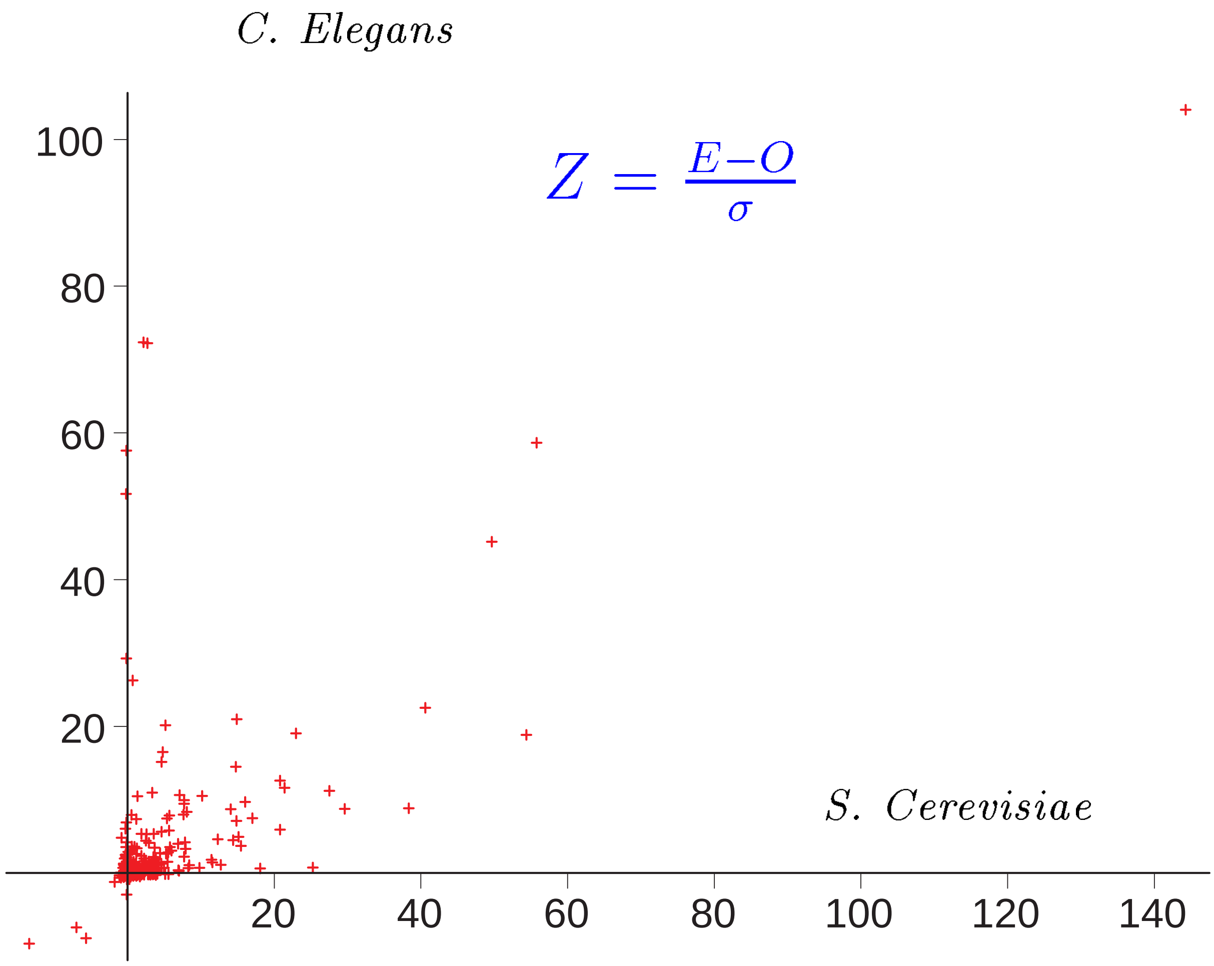


$$
\begin{aligned}
p_n(u) &= [z^n]F(z,u) = \frac{1}{2i\pi}\oint_\Gamma \frac{dz}{z^{n+1}}F(z,u),\\
&= \frac{1}{2i\pi}\oint_\Gamma \frac{c(u)}{z^{n+1}(1-\lambda_1(u)z)} + \frac{1}{z^{n+1}}g(z,u)\,dz,\\
&= c(u)\lambda_1(u)^n(1+O(A^n)).
\end{aligned}
$$

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

- $\lambda_1(u), c(u)$ analytiques dans un voisinage de $u = 1$;
- $c(1) \neq 0$;
- $g(z, u)$ analytique pour $|z| < 1/A\lambda_1(u)$ avec $A < 1$.

Hwang gives also linear mean and variance
Everything is computable
also in the Markovian Case



Z-scores for *C. Elegans* and *S. Cerevisiae*

values normalized for 1,000,000 positions and $E \geq 0.05$

(305 motifs)