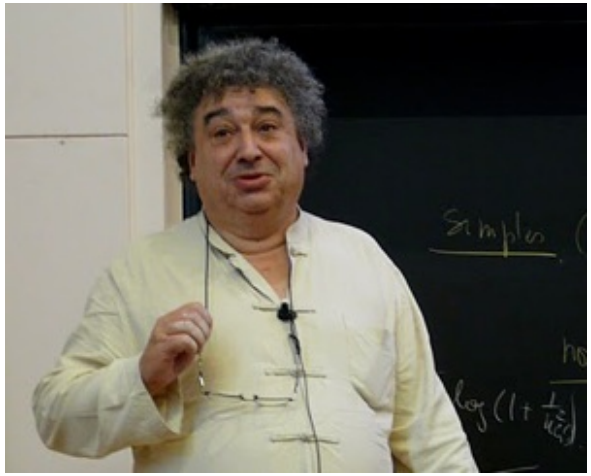
Text Analysis in Philippe Flajolet's work

Pierre Nicodème

CNRS, CALIN team, LIPN, University Paris 13

18/12/2011



Text Analysis Articles

- **Discrepancy of Sequences in Discrete Spaces** (1989), P.Flajolet, P. Kirschenhofer and R.F. Tichy (#76)
- **Deviations from Uniformity in Random Strings** (1988), P.Flajolet, P. Kirschenhofer and R.F. Tichy (#74)
- **Motif statistics** (1999-2002), P. N., B. Salvy and P. Flajolet (#151/#174)
- **Hidden Pattern Statistics** (2001), P. Flajolet, Y. Guivarc'h, W. Szpankowski and B. Vallée (#164)
- **Hidden word Statistics**, (2006), P. Flajolet, W. Szpankowski and B. Vallée (#191)

Text Analysis Articles

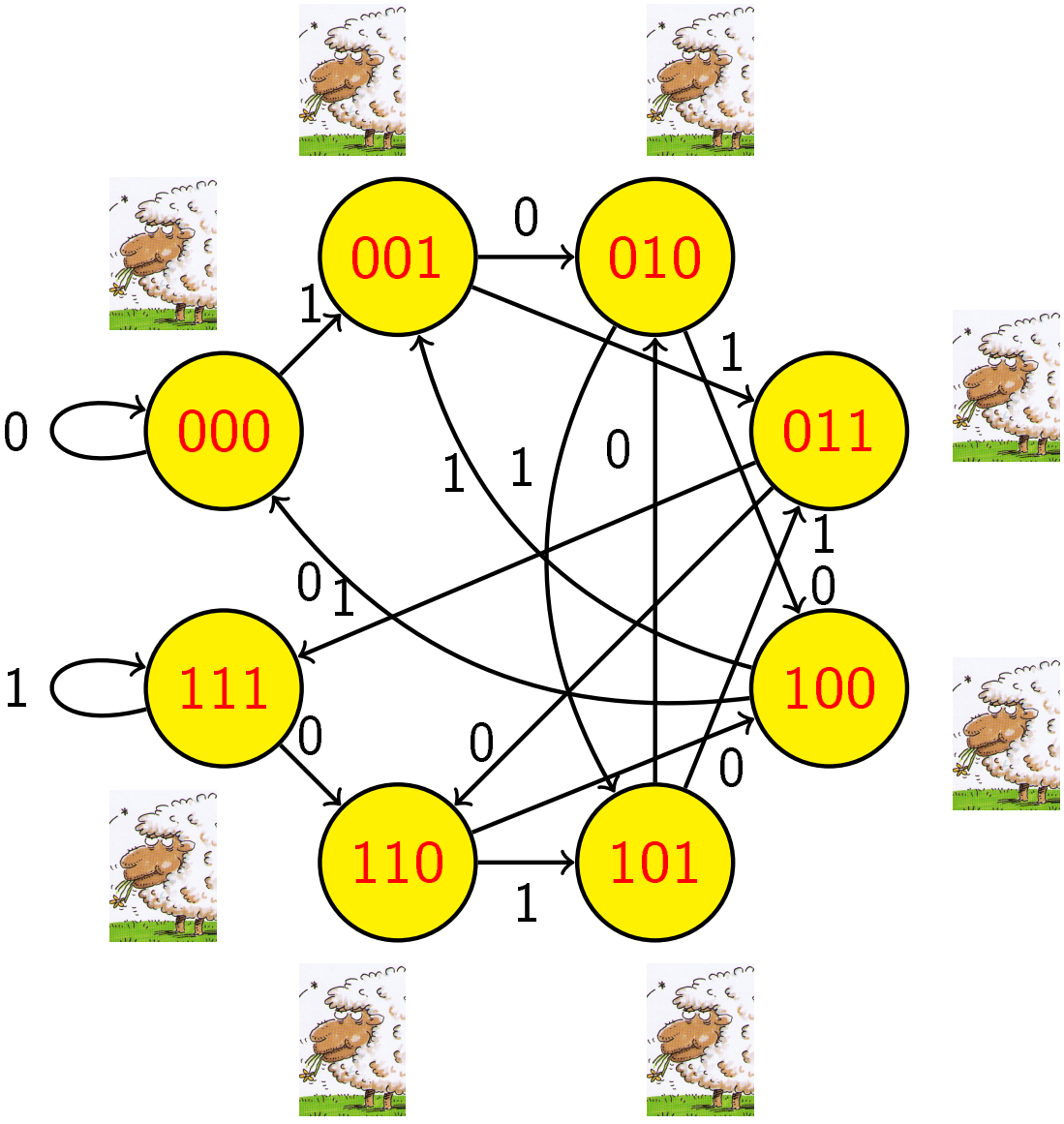
- **Discrepancy of Sequences in Discrete Spaces** (1989), P.Flajolet, P. Kirschenhofer and R.F. Tichy (#76)
- **Deviations from Uniformity in Random Strings** (1988), P.Flajolet, P. Kirschenhofer and R.F. Tichy (#74)

Normal numbers

Blocks of 10000 first bits of π		
Length(k)	Least frequent	Most frequent
2	$(11)^{2480}$	$(00)^{2509}$
3	$(111)^{1226}$	$(000)^{1275}$
4	$(1100)^{603}$	$(0000)^{652}$
5	$(11100)^{296}$	$(00000)^{341}$
6	$(0101100)^{57}$	$(0110110)^{97}$

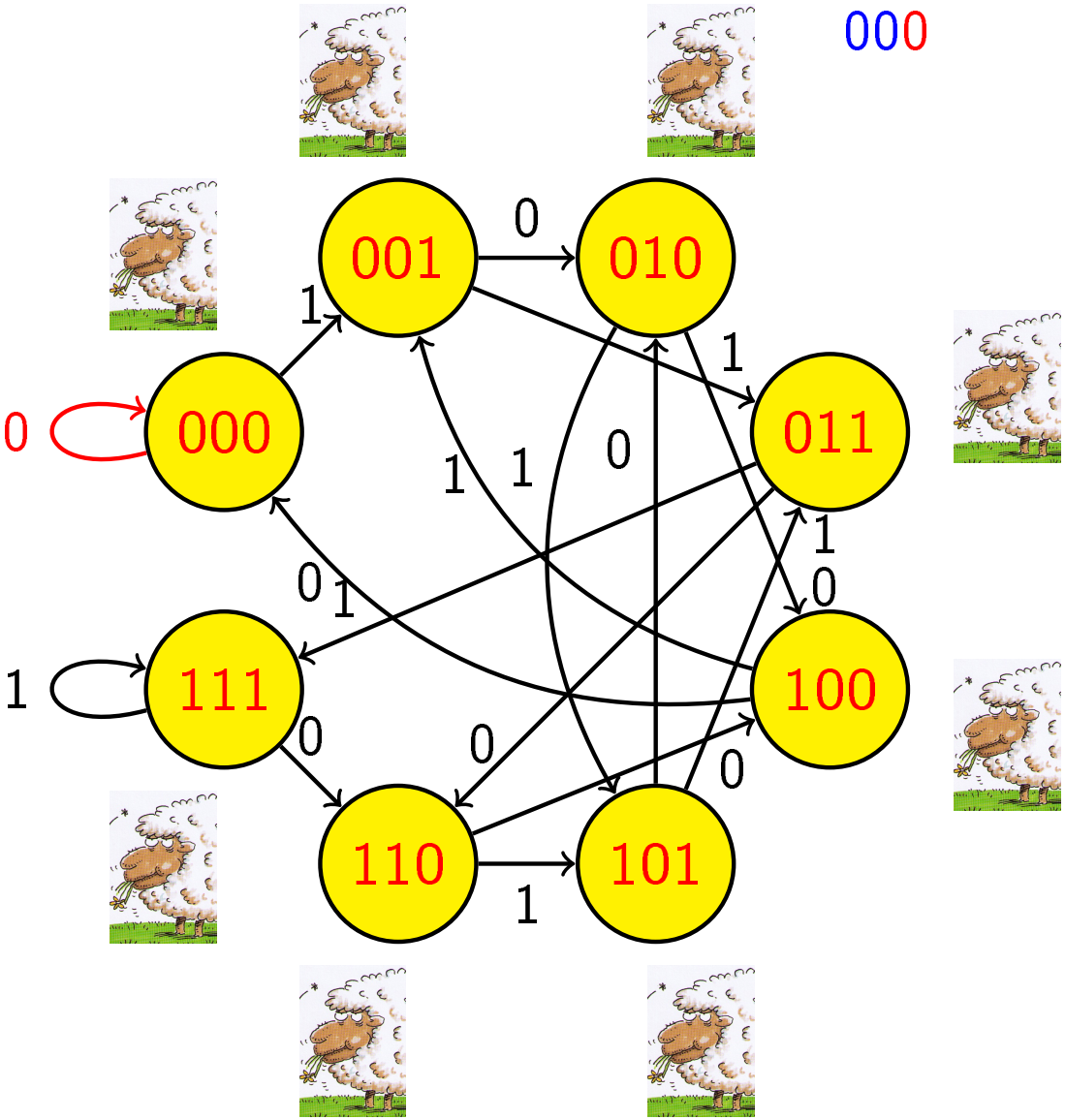
- A number is **normal** if, in its binary representation, all the **blocks of consecutive letters** of length k appear with probability $\frac{1}{2^k}$
- E. Borel proved that **almost all real number** are **normal** (1909)
- It is **conjectured** that e, π or $\sqrt{2}$ are **normal**

The de Bruijn Graph of order 3 over a binary alphabet



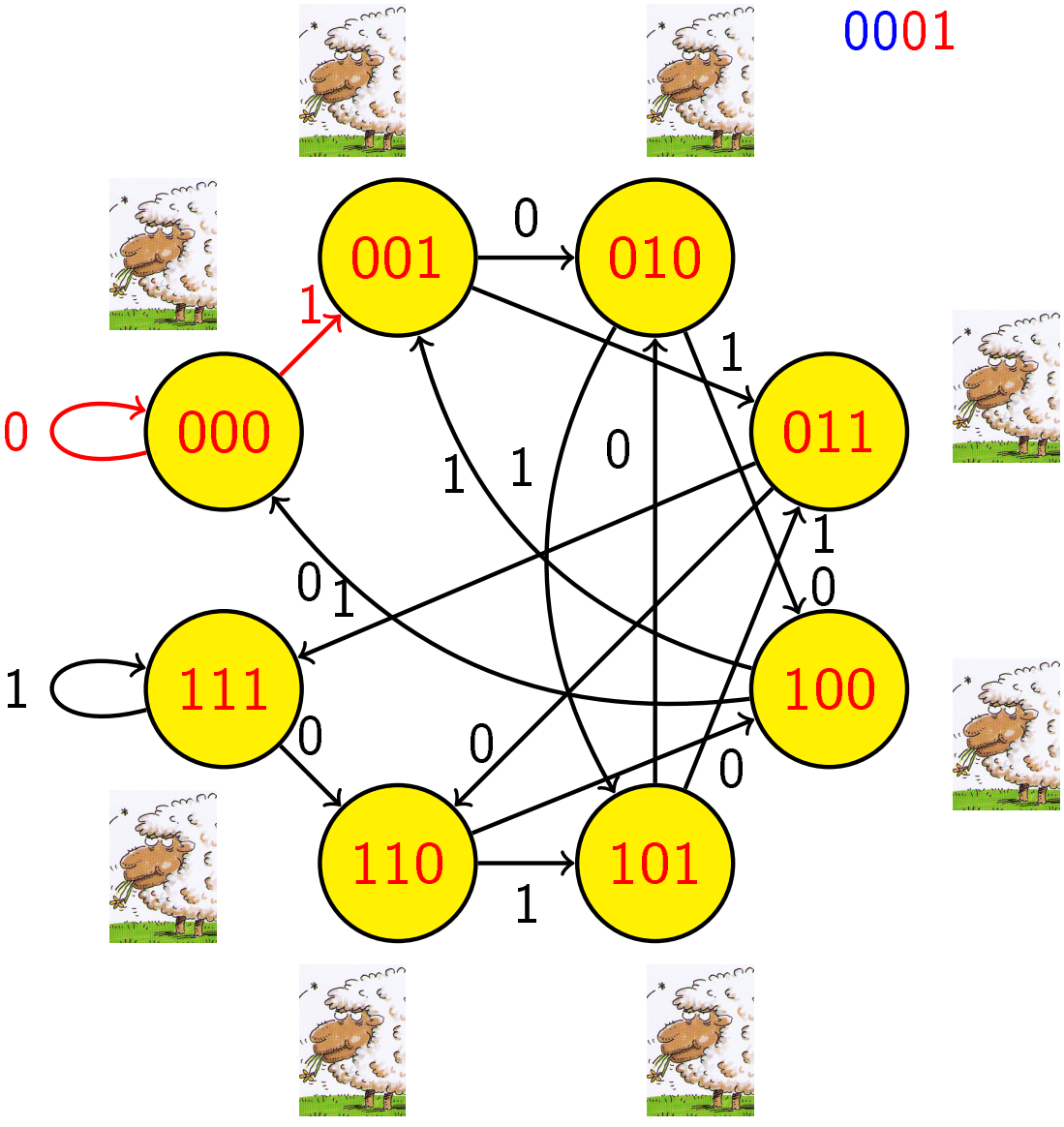
De Bruijn Graph of order $k = 3$; binary alphabet: $f(x) = 2 \times x \pmod{2^k} + \begin{cases} 1 \\ 0 \end{cases}$

The de Bruijn Graph of order 3 over a binary alphabet



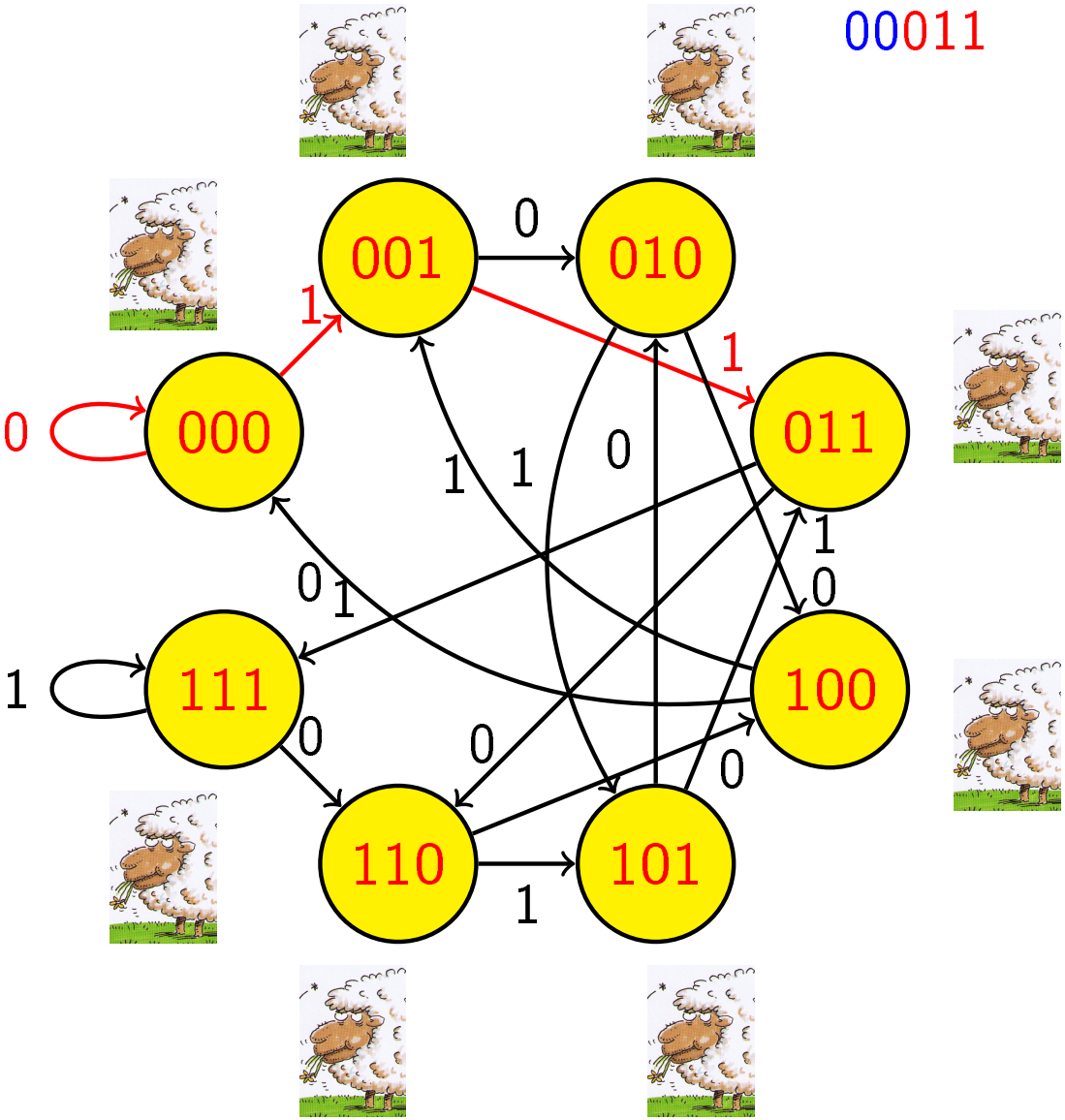
De Bruijn Graph of order $k = 3$; binary alphabet: $f(x) = 2 \times x \pmod{2^k} + \begin{cases} 1 \\ 0 \end{cases}$

The de Bruijn Graph of order 3 over a binary alphabet



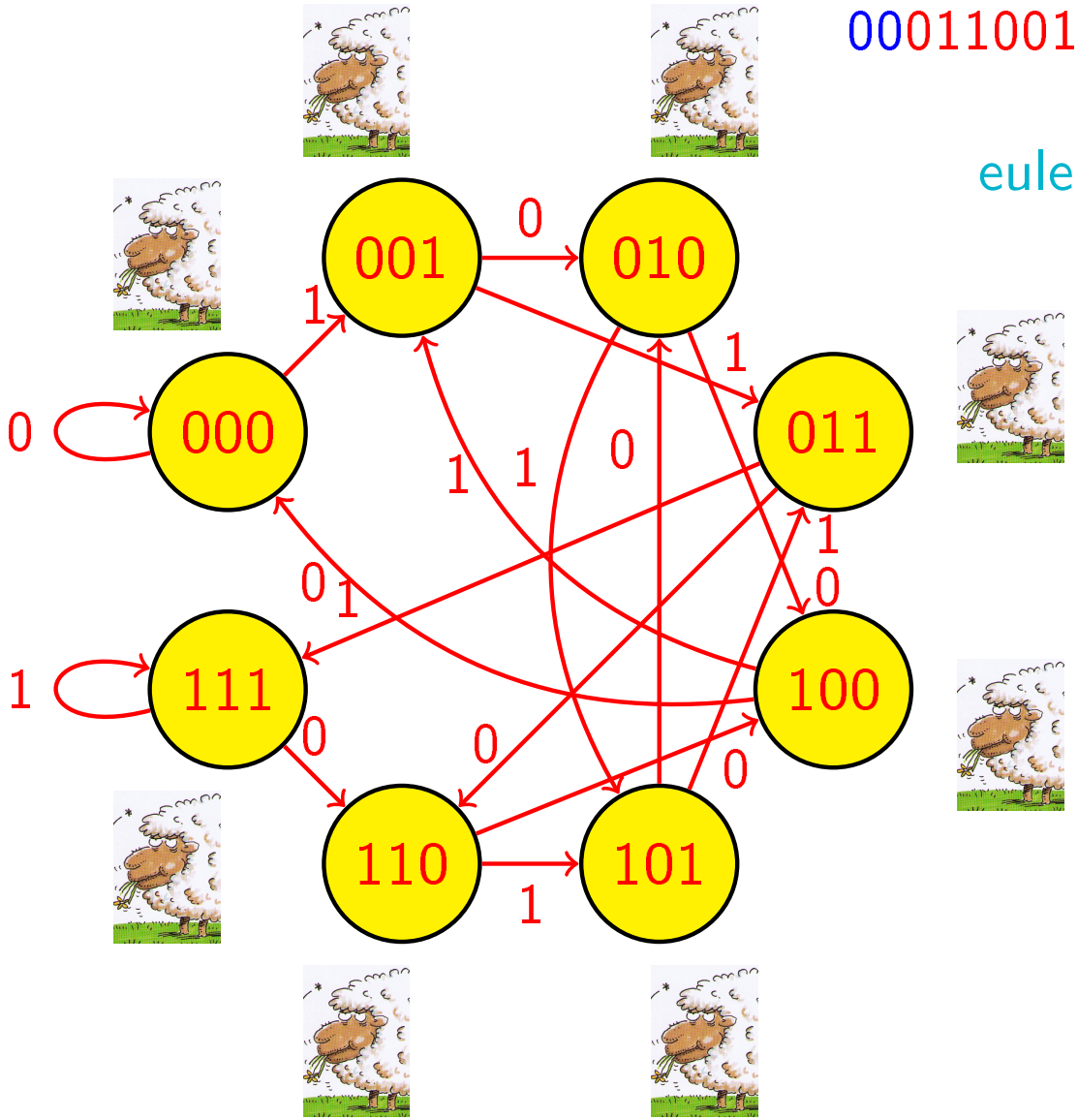
De Bruijn Graph of order $k = 3$; binary alphabet: $f(x) = 2 \times x \pmod{2^k} + \begin{cases} 1 \\ 0 \end{cases}$

The de Bruijn Graph of order 3 over a binary alphabet



De Bruijn Graph of order $k = 3$; binary alphabet: $f(x) = 2 \times x \pmod{2^k} + \begin{cases} 1 \\ 0 \end{cases}$

The de Bruijn Graph of order 3 over a binary alphabet



0001100101111101000

eulerian path

De Bruijn Graph of order $k = 3$; binary alphabet: $f(x) = 2 \times x \pmod{2^k} + \begin{cases} 1 \\ 0 \end{cases}$

k-discrepancy of a binary string S of length n

$$D_k(S) = \max_{|u|=k} \left| \frac{|S|_u}{n-k+1} - \frac{1}{2^k} \right|$$

$|S| = n, \qquad |S|_u =$ number of occurrences of u is S

			0001100101111101000	1001110010111101011	0000000000000000000
000	Fatima		2	0	16
001	Virginie		2	2	0
010	Frederique		2	2	0
011	Carine		2	3	0
100	Marni		2	2	0
101	Claire		2	3	0
110	Elvire		2	2	0
111	Charlotte		2	2	0
$D_k(S)$			$\left\|\frac{2}{16} - \frac{1}{8}\right| = 0$	$\left\|\frac{0}{16} - \frac{1}{8}\right| = \frac{1}{8}$	$\left\|\frac{16}{16} - \frac{1}{8}\right| = \frac{7}{8}$

Uniform distribution of an α-ary string

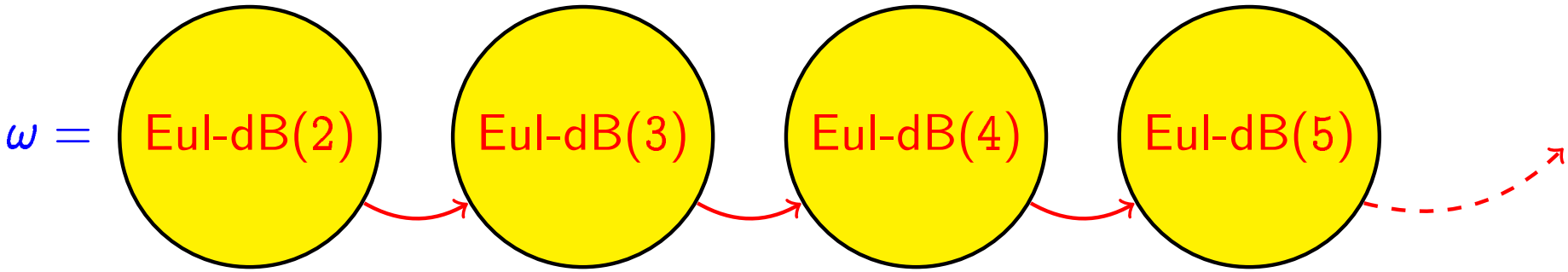
$$D_{k(n)}(S_n) = \max_{|u|=k(n)} \left| \frac{|S_n|_u}{n - k(n) + 1} - \frac{1}{\alpha^{k(n)}} \right|$$

$\alpha = |\mathcal{A}|$, size of alphabet

an infinite sequence $\omega = (x_1, x_2, x_3, \dots)$ is

- $[k(n)]$-**uniformly-distributed**
- if $D_{k(n)}(\omega[1..n]) = o\left(\alpha^{-k(n)}\right) \qquad (n \to \infty)$

A uniformly distributed sequence (1989)



Eul-dB(k) : **Eulerian** path in a **de Bruijn** graph with blocks size $= k$

Thm[Fl-Ki-Ti 89]

- With $k(n) = o(\log\log n)$,
- the **infinite sequence** ω is $[k(n)]$-**uniformly distributed**

A Gaussian theorem for local discrepancy(1989)

Thm[Fl-Ki-Ti 89]
with every u and $k(n)$ such that

- $|u| = k(n)$ and $\omega_n \in \mathcal{A}^n$ (ω_n **random**)
- $\phi(n) = \dfrac{k(n)^2 \alpha^{k(n)}}{n}$ and $\lim_{n\to\infty} \phi(n) = 0$

we have

$$\Pr\left\{\left|\frac{|\omega_n|_u}{n-k(n)+1} - \frac{1}{\alpha^{k(n)}}\right| < y\frac{\sigma_n(u)}{\sqrt{n}}\right\}$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y} e^{-t^2/2}dt + \mathcal{O}\left(\sqrt{\phi(n)}\right) + \mathcal{O}\left(\sqrt{\frac{\log k(n)}{k(n)}}\right)$$

A result with uniform bounds

Thm[Fl-Ki-Ti 89]
With

- $k(n) \leq c \log_\alpha n \qquad (c < 1/3)$,
- $\tau < 1/3 - c$,

for **almost all sequences** $\omega \in \mathcal{A}^\infty$,

we have $D^{[k(n)]}(\omega_n) \leq n^\tau \alpha^{-k(n)} \qquad (n > N_0)$

Example

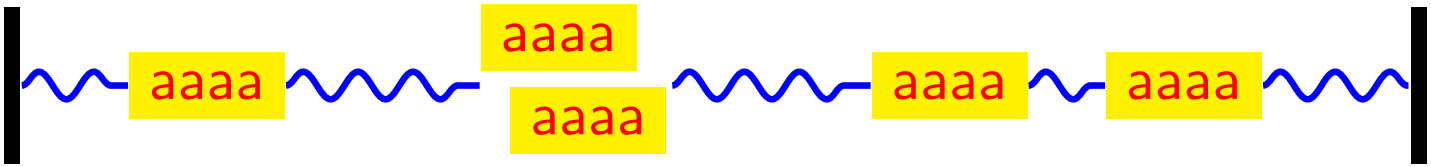
$$c = 1/3 - \epsilon,\ \tau = \epsilon \qquad k(n) < \frac{\log_\alpha(n)}{3+\epsilon} \quad \Longrightarrow \quad D^{[k(n)]}(\omega[1..n]) \leq n^\epsilon \alpha^{-k(n)}$$

Best possible bounds for binary strings - [Fl-Ki-Ti 88]

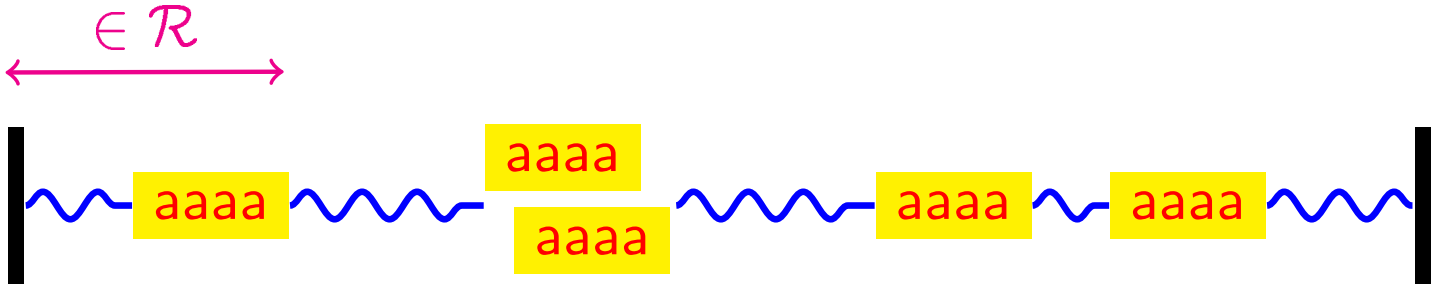
- The **convergence properties** to the **stationary distribution** $\left(\frac{1}{2^k}, \frac{1}{2^k}, \dots\right)$ of the **Markov chain** built upon the **de Bruijn graph** do **not provide** satisfactory results when $k(n)$ tends to **infinity**, in particular when close of $\log_2(n)$.

- Very precise **quasi-uniform estimates** by use of the **Guibas-Odlyzko language decomposition** for **counting occurrences** (1984) of words in strings.

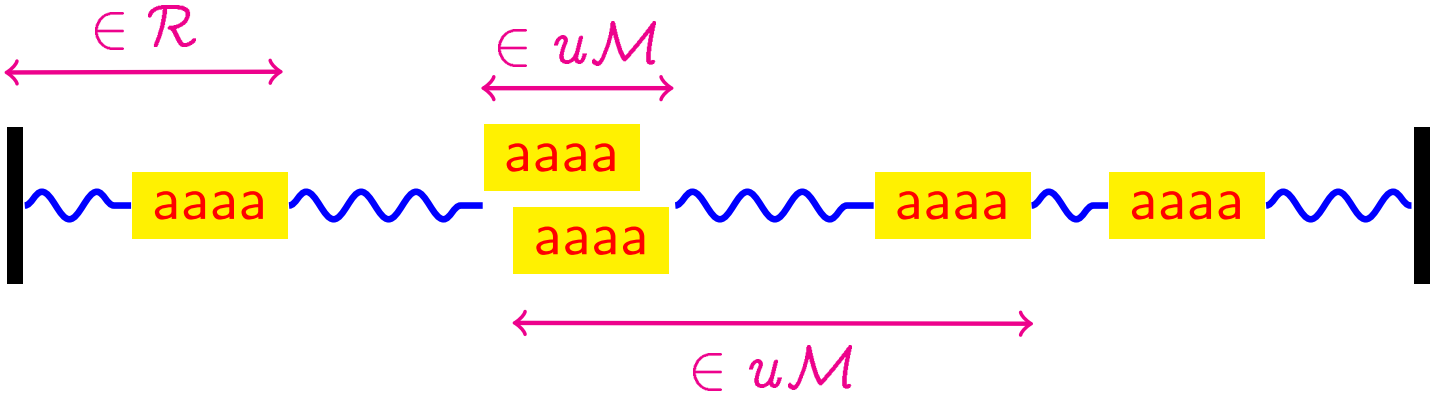
Guibas-Odlyzko decomposition - occurrences of a word u



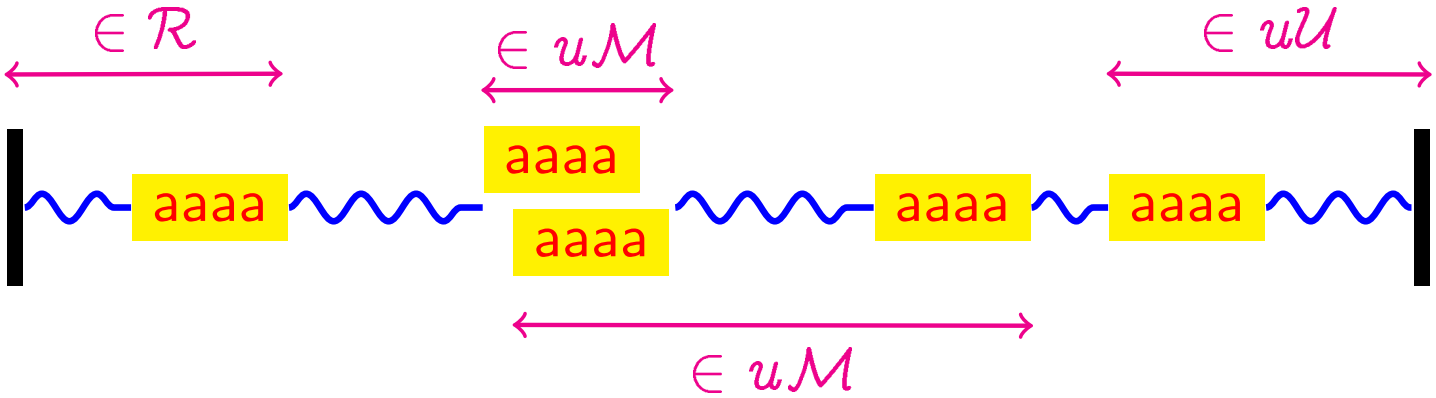
Guibas-Odlyzko decomposition - occurrences of a word u



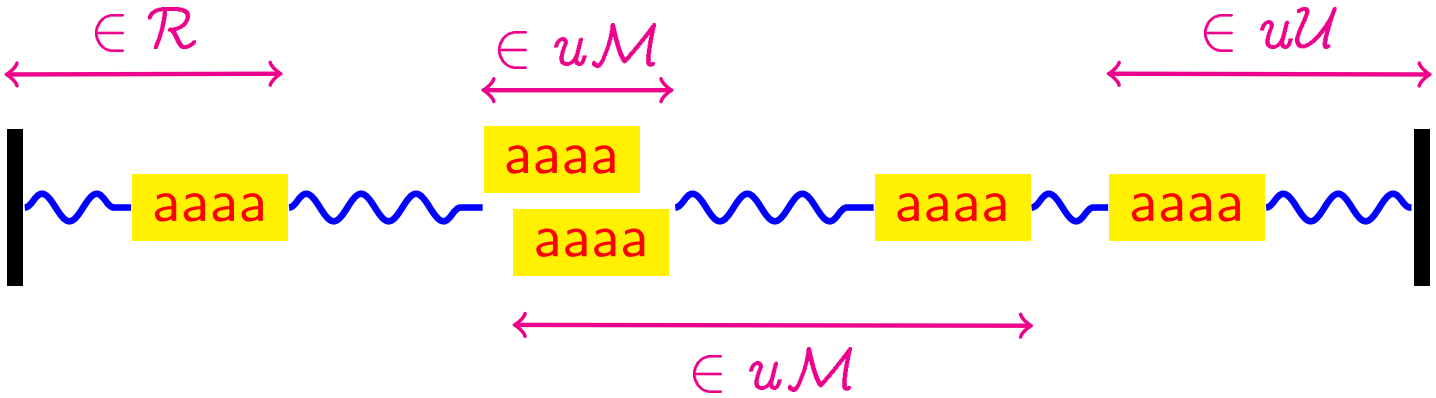
Guibas-Odlyzko decomposition - occurrences of a word u



Guibas-Odlyzko decomposition - occurrences of a word u



Guibas-Odlyzko decomposition - occurrences of a word u



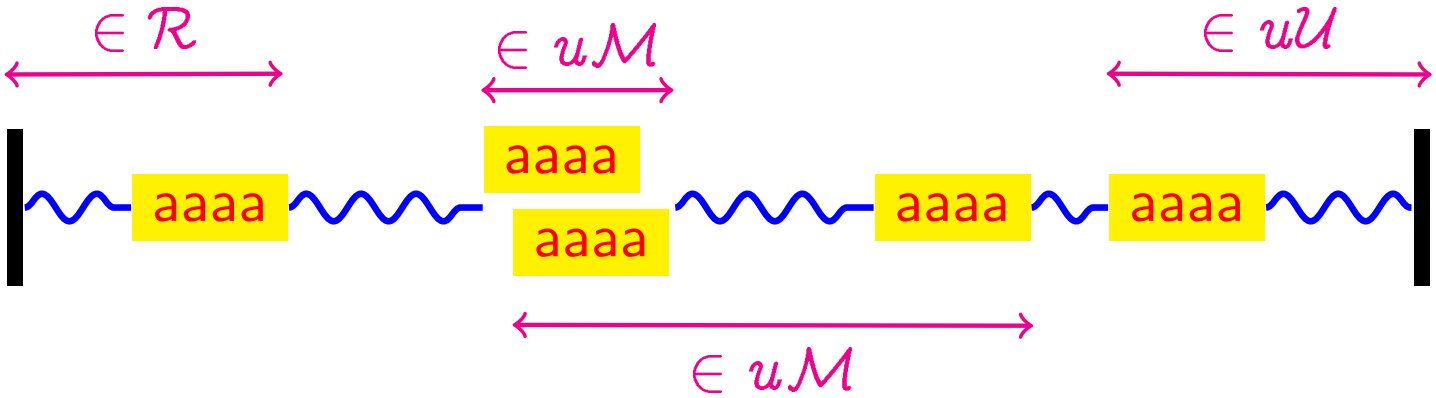
autocorrelation of words

$u = aaaa \qquad \mathcal{C}_u = \{\epsilon, a, aa, aaa\}, \quad C_u(z) = 1 + (z/2) + (z/2)^2 + (z/2)^3$

$v = aabaa \qquad \mathcal{C}_v = \{\epsilon, baa, abaa\}, \quad C_v(z) = 1 + (z/2)^3 + (z/2)^4$

Generating Functions: $L(z) = \sum_{w \in \mathcal{L}} \Pr(w) z^{|w|}$

Guibas-Odlyzko decomposition - occurrences of a word u



autocorrelation of words

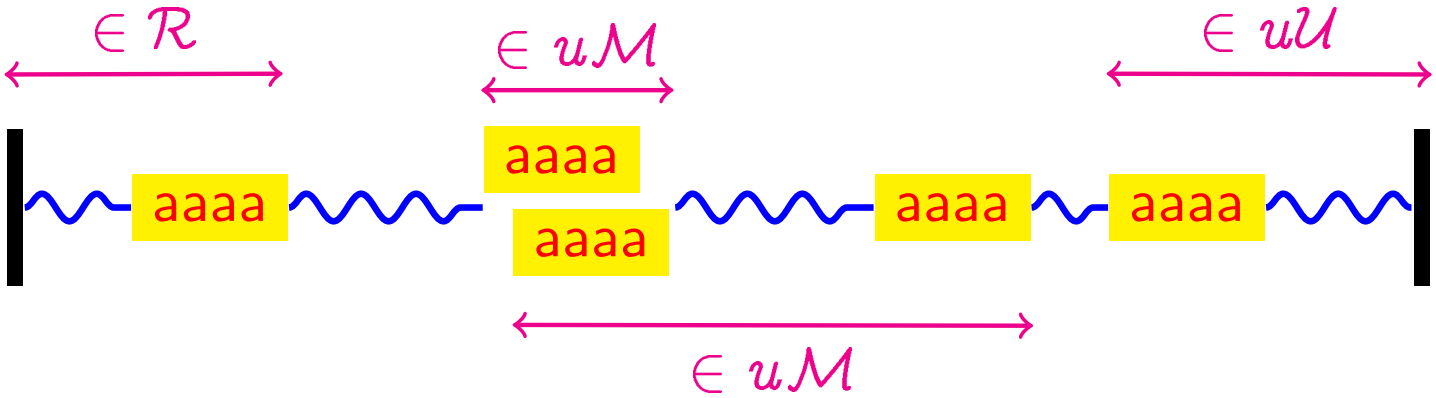
$u = aaaa \qquad \mathcal{C}_u = \{\epsilon, a, aa, aaa\}, \quad C_u(z) = 1 + (z/2) + (z/2)^2 + (z/2)^3$

$v = aabaa \qquad \mathcal{C}_v = \{\epsilon, baa, abaa\}, \quad C_v(z) = 1 + (z/2)^3 + (z/2)^4$

Generating Functions: $L(z) = \sum_{w \in \mathcal{L}} \Pr(w) z^{|w|}$

- $|u| = k$ (blocks of size k)
- $R_u(z) = \dfrac{(z/2)^k}{(z/2)^k + (1-z)\,C_u(z)}, \qquad U_u(z) = \dfrac{1}{(z/2)^k + (1-z)\,C_u(z)}$
- $M_u(z) = \dfrac{(z/2)^k + (1-z)(C_u(z) - 1)}{(z/2)^k + (1-z)\,C_u(z)}$

Guibas-Odlyzko decomposition - occurrences of a word u



autocorrelation of words

$u = aaaa \qquad \mathcal{C}_u = \{\epsilon, a, aa, aaa\}, \quad C_u(z) = 1 + (z/2) + (z/2)^2 + (z/2)^3$

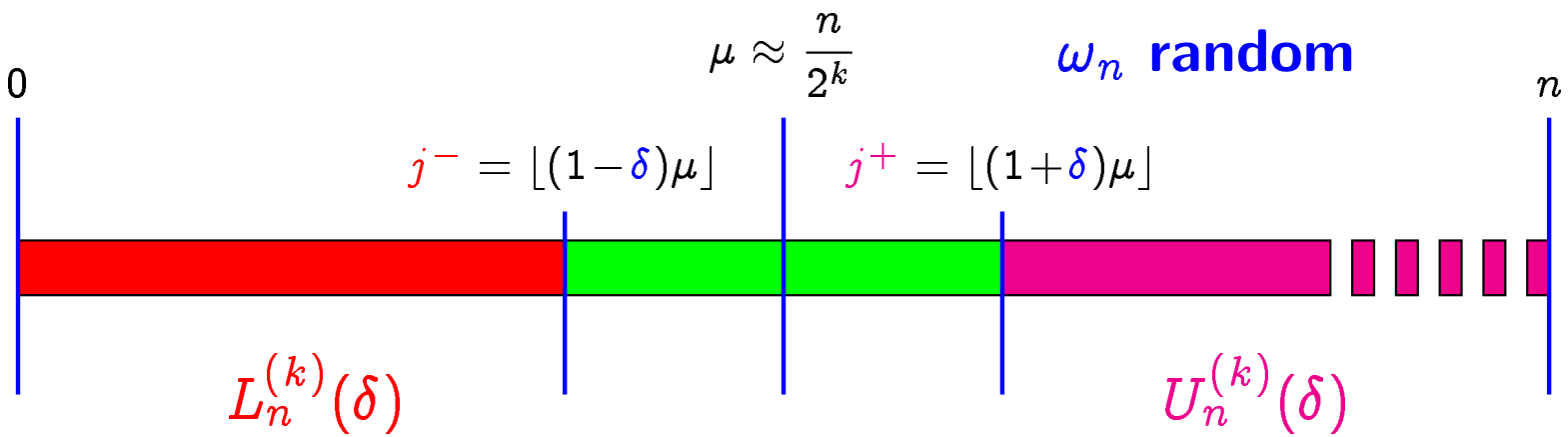
$v = aabaa \qquad \mathcal{C}_v = \{\epsilon, baa, abaa\}, \quad C_v(z) = 1 + (z/2)^3 + (z/2)^4$

Generating Functions: $L(z) = \sum_{w \in \mathcal{L}} \Pr(w) z^{|w|}$

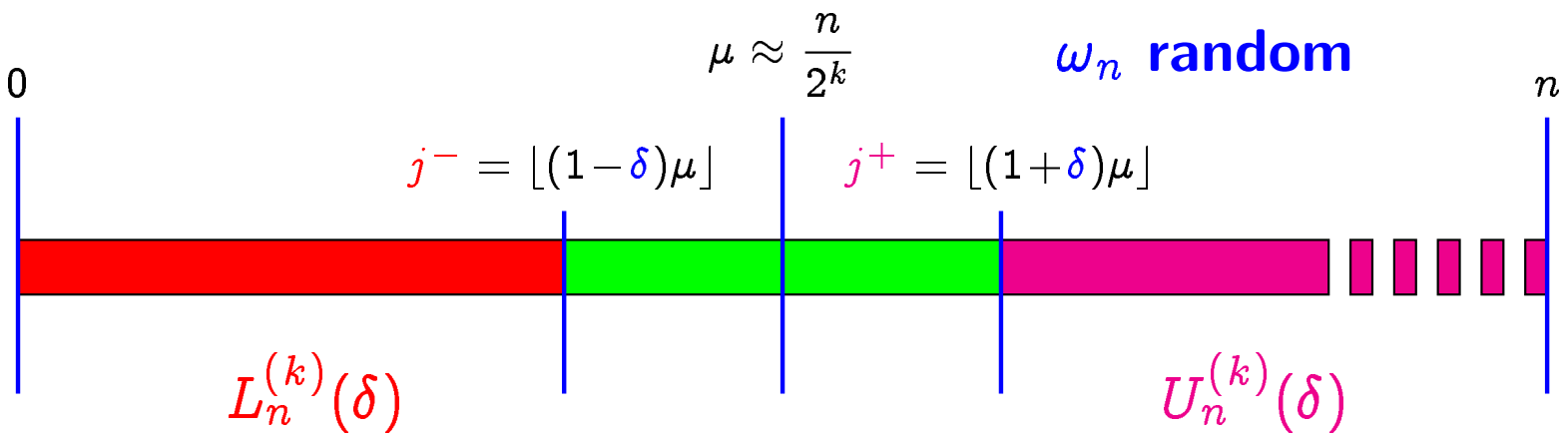
- $|u| = k$ (blocks of size k)
- $R_u(z) = \dfrac{(z/2)^k}{(z/2)^k + (1-z)\,C_u(z)}, \qquad U_u(z) = \dfrac{1}{(z/2)^k + (1-z)\,C_u(z)}$
- $M_u(z) = \dfrac{(z/2)^k + (1-z)(C_u(z) - 1)}{(z/2)^k + (1-z)\,C_u(z)}$

r occurrences: $F_u^{(r)}(z) = R_u(z)\, M_u^{r-1}(z)\, U_u(z)$

Occurrences of words - Uniform tail probabilities

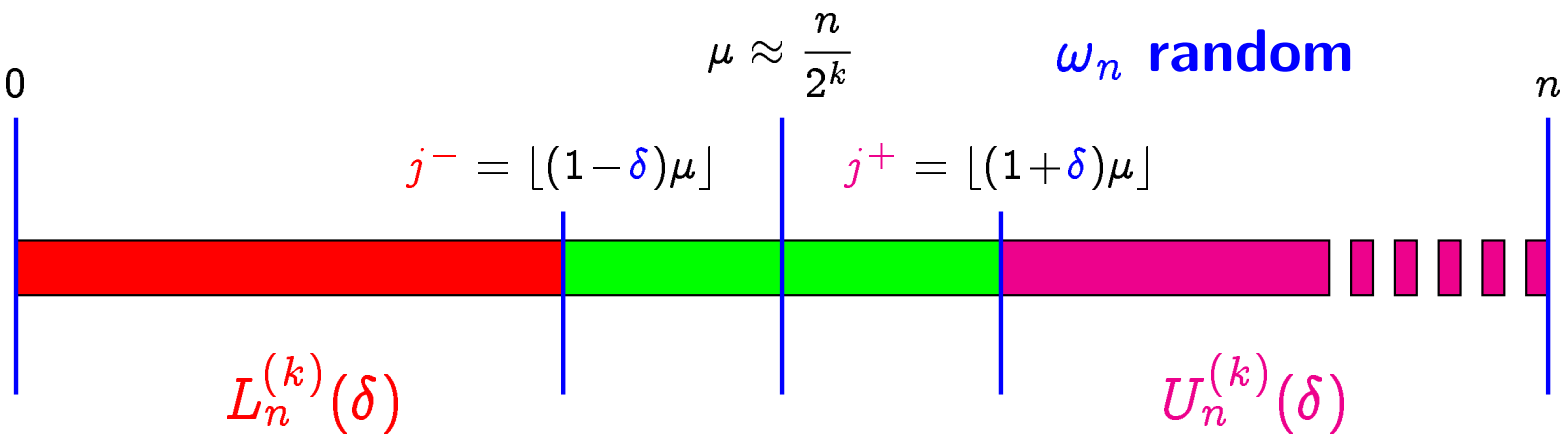


Occurrences of words - Uniform tail probabilities



$$F_u^{(r)}(z) = R_u(z) M_u^{r-1}(z) U_u(z)$$

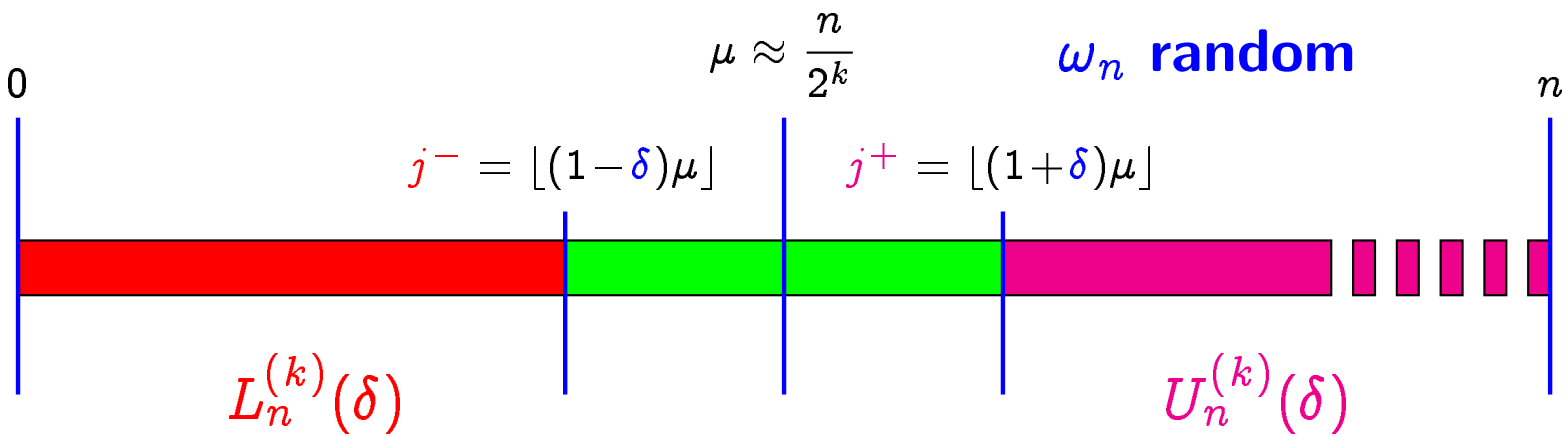
Occurrences of words - Uniform tail probabilities



$$F_u^{(r)}(z) = R_u(z) M_u^{r-1}(z) U_u(z)$$

$$L_{n,u}(\delta) = [z^n] \sum_{i \le j^-} F_u^{(i)}(z) = [z^n] \frac{R_u(z)(1 - M_u(z))^{j^-} U_u(z)}{1 - M_u(z)} < L_n^{(k)}(\delta)$$

Occurrences of words - Uniform tail probabilities

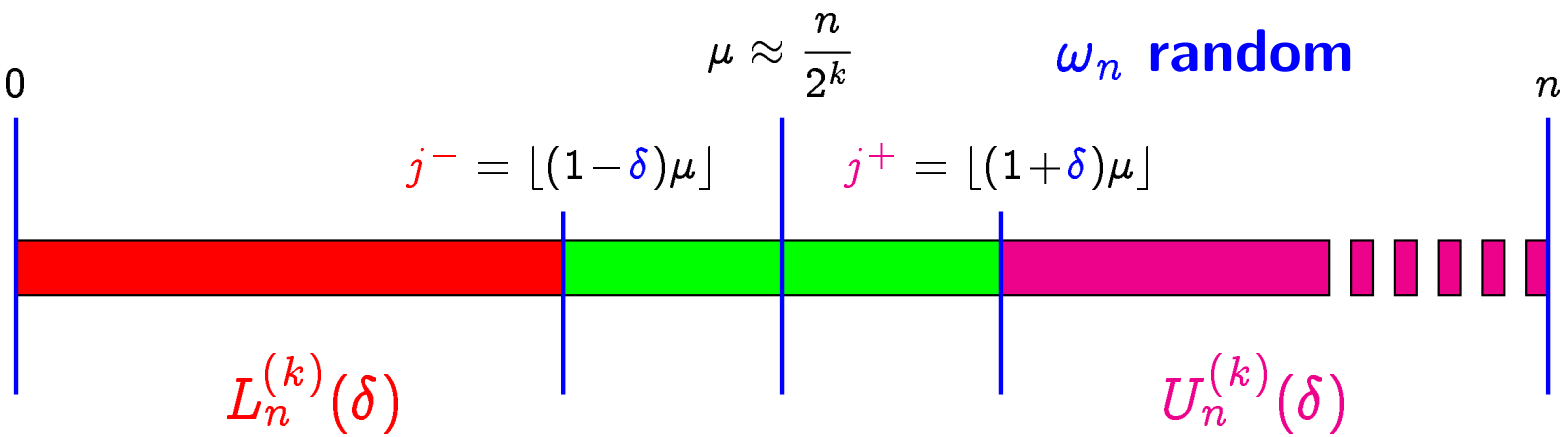


$$F_u^{(r)}(z) = R_u(z) M_u^{r-1}(z) U_u(z)$$

$$L_{n,u}(\delta) = [z^n] \sum_{i \le j^-} F_u^{(i)}(z) = [z^n] \frac{R_u(z)(1-M_u(z))^{j^-} U_u(z)}{1-M_u(z)} < L_n^{(k)}(\delta)$$

$$U_{n,u}(\delta) = [z^n] \sum_{i \ge j^+} F_u^{(i)}(z) = [z^n] \frac{R_u(z)(M_u(z))^{j^+-1} U_u(z)}{1-M_u(z)} < U_n^{(k)}(\delta)$$

Occurrences of words - Uniform tail probabilities



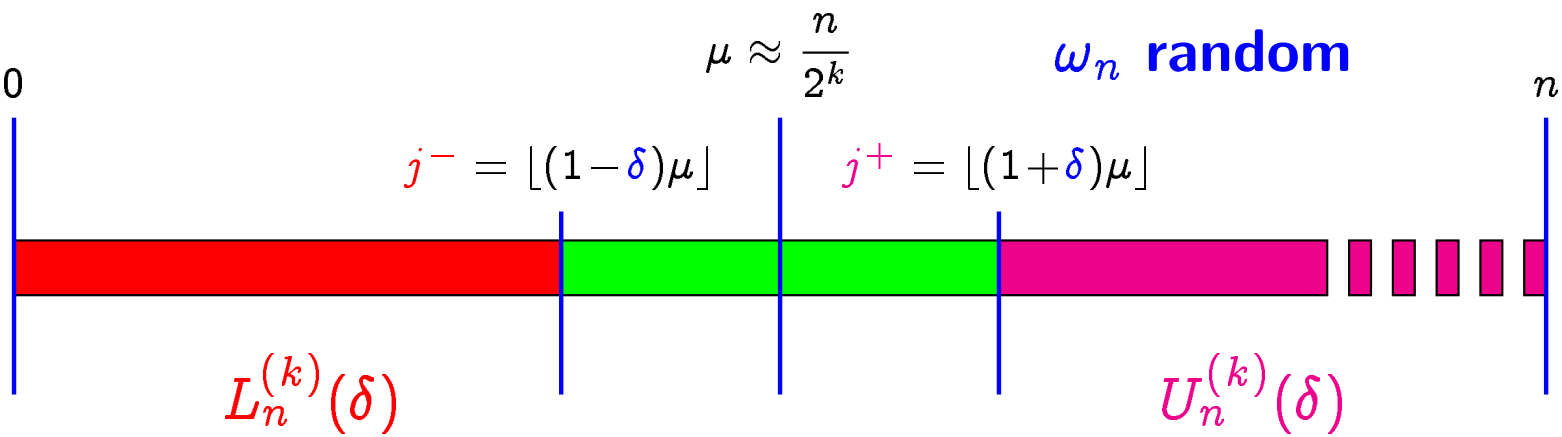
$$F_u^{(r)}(z) = R_u(z) M_u^{r-1}(z) U_u(z)$$

$$L_{n,u}(\delta) = [z^n] \sum_{i \le j^-} F_u^{(i)}(z) = [z^n] \frac{R_u(z)(1-M_u(z))^{j^-} U_u(z)}{1-M_u(z)} < L_n^{(k)}(\delta)$$

$$U_{n,u}(\delta) = [z^n] \sum_{i \ge j^+} F_u^{(i)}(z) = [z^n] \frac{R_u(z)(M_u(z))^{j^+-1} U_u(z)}{1-M_u(z)} < U_n^{(k)}(\delta)$$

Upper bounds for lower and upper tails **independent** of u

Occurrences of words - Uniform tail probabilities



$$F_u^{(r)}(z) = R_u(z) M_u^{r-1}(z) U_u(z)$$

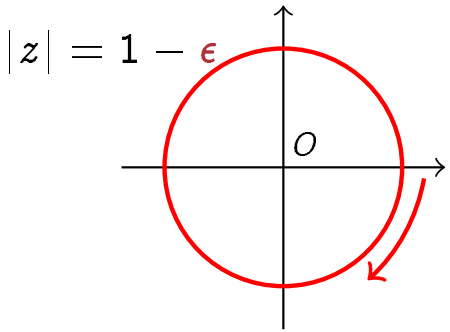
$$L_{n,u}(\delta) = [z^n] \sum_{i \le j^-} F_u^{(i)}(z) = [z^n] \frac{R_u(z)(1-M_u(z))^{j^-} U_u(z)}{1-M_u(z)} < L_n^{(k)}(\delta)$$

$$U_{n,u}(\delta) = [z^n] \sum_{i \ge j^+} F_u^{(i)}(z) = [z^n] \frac{R_u(z)(M_u(z))^{j^+-1} U_u(z)}{1-M_u(z)} < U_n^{(k)}(\delta)$$

Upper bounds for lower and upper tails **independent** of u

$$\Pr\left(D_k(\omega_n) > \frac{\delta}{2^k}\right) \le \sum_{|u|=k} \Pr\left(\left||\omega_n|_u - \frac{n}{2^k}\right| > \frac{\delta n}{2^k}\right) \le 2^k (L_n^{(k)}(\delta) + U_n^{(k)}(\delta))$$

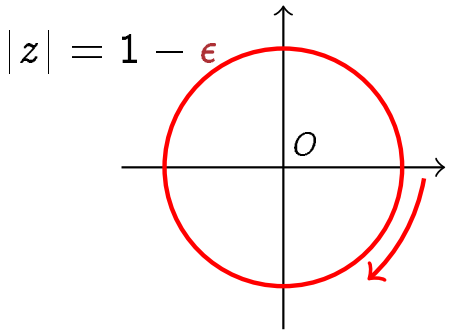
Upper bound - Cauchy integral



$$U_{n,u}(\delta) = \frac{1}{2i\pi} \oint_{|z|=1-\epsilon} 2^k \, a(z) \frac{b^j(z)}{1 - b(z)} \frac{dz}{z^{n+1}}$$

$$a(z) = \frac{z^k}{Q_u^2(z)}, \quad b(z) = 1 + \frac{2^k(z-1)}{Q_u(z)}, \quad Q_u(z) = z^k + 2^k(1-z)\,C_u(z)$$

Upper bound - Cauchy integral

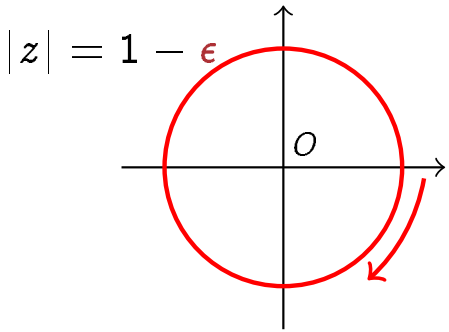


$$U_{n,u}(\delta) = \frac{1}{2i\pi} \oint_{|z|=1-\epsilon} 2^k \, a(z) \frac{b^j(z)}{1 - b(z)} \frac{dz}{z^{n+1}}$$

$$a(z) = \frac{z^k}{Q_u^2(z)}, \quad b(z) = 1 + \frac{2^k(z-1)}{Q_u(z)}, \quad Q_u(z) = z^k + 2^k(1-z)\,C_u(z)$$

Lem:[GuOd78,FlKiTi88] $k \geq k_0$, $\dfrac{1}{Q_u(z)}$ **analytic** in $|z| < 1 + \dfrac{1}{2^{k+1}}$

Upper bound - Cauchy integral



$$U_{n,u}(\delta) = \frac{1}{2i\pi} \oint_{|z|=1-\epsilon} 2^k a(z) \frac{b^j(z)}{1 - b(z)} \frac{dz}{z^{n+1}}$$

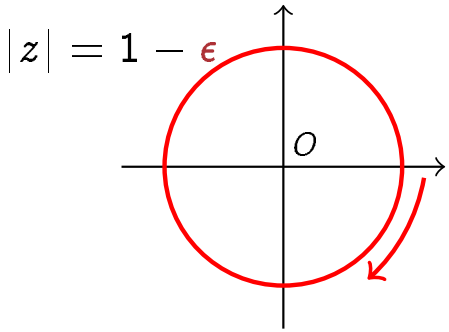
$$a(z) = \frac{z^k}{Q_u^2(z)}, \quad b(z) = 1 + \frac{2^k(z-1)}{Q_u(z)}, \quad Q_u(z) = z^k + 2^k(1-z)C_u(z)$$

Lem:[GuOd78,FlKiTi88] $k \geq k_0$, $\frac{1}{Q_u(z)}$ **analytic** in $|z| < 1 + \frac{1}{2^{k+1}}$

- $\epsilon = \epsilon(n) = \frac{\lg n \lg\lg n}{n}$ $\qquad (\lg x = \log_2(x))$
- $k = k(n) = [\lg n - \lg\lg n - 2\lg\lg\lg n]$
- $\eta := 2^k \epsilon \to 0$ $\qquad$ as $n \to \infty$

$$Q_u(1-\epsilon) = (1-\epsilon)^k + \eta C_u\left(\frac{1-\epsilon}{2}\right) = 1 + \mathcal{O}(\eta)$$

Upper bound - Asymptotic Equivalents



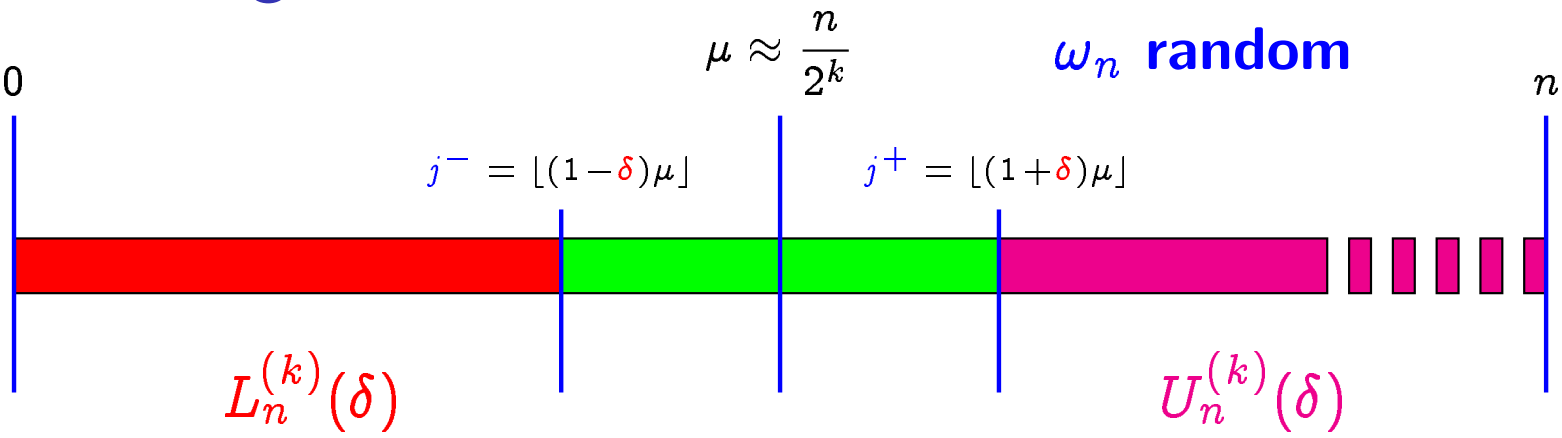
$$U_{n,u}(\delta) = \frac{1}{2i\pi} \oint_{|z|=1-\epsilon} 2^k a(z) \frac{b^j(z)}{1 - b(z)} \frac{dz}{z^{n+1}}$$

$$\begin{cases} \epsilon & = & \dfrac{\lg n \lg\lg n}{n} \\ \eta & = & 2^k \epsilon \to 0 \qquad \text{as } n \to \infty \\ a(1-\epsilon) & = & 1 + \mathcal{O}(\eta) \\ b(1-\epsilon) & = & 1 - \eta + \mathcal{O}(\eta^2) \\ b^j(1-\epsilon) & = & \exp(-\eta j + \mathcal{O}(j\eta^2)) \\ (1+\epsilon)^{-n} & = & \exp(-n\epsilon + \mathcal{O}(n)) \end{cases}$$

$$U_{n,u}(\delta) \leq 2^k \frac{a(1-\epsilon)}{1 - b(1-\epsilon)} b^j(1-\epsilon)(1-\epsilon)^{-n} \quad \text{(trivial bounds)}$$

$$\rightsquigarrow U_{n,u}(\delta) \leq \exp(-\delta \lg n \lg\lg n + c \lg n)$$

Summarizing



- $U_{n,u}(\delta) \leq \exp(-\delta \lg n \lg\lg n + c \lg n)$
- $L_{n,u}(\delta) \leq \exp(-\delta \lg n \lg\lg n + c' \lg n)$

$k = k(n) = [\lg n - \lg\lg n - 2\lg\lg\lg n], \quad 0 < \delta < 1$

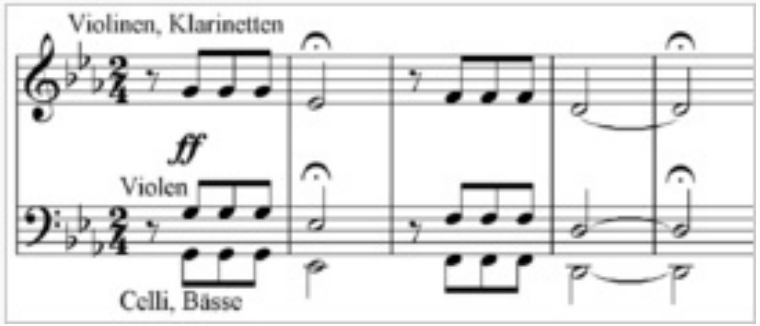
Thm:[Fl-Ki-Ti 88] $\Pr\left(D_k(\omega_n) > \frac{\delta}{2^k} \right) < n^{-\delta \lg\lg n + c''}$

$\delta = \delta(n) = (\lg\lg n)^{-1/2} \implies \sum_n n^{-\delta(n)\lg\lg n + \epsilon} < \infty$ **Borel-Cantelli**

Thm:[Fl-Ki-Ti 88] Almost all infinite strings ω are $k(n)$-uniformly distributed for $k(n) = [(1-\epsilon)\lg n]$ with $\epsilon > 0$

MOTIF STATISTICS

- **Motif Statistics**
  P.N., B. Salvy, P.Flajolet
  - ESA 1999
  - TCS 2002



In Beethoven's Fifth Symphony a four-note figure becomes the most important motif of the work, extended melodically and harmonically to provide the main theme of the first movement.

- **Hidden Pattern Statistics**
  P. Flajolet, Y. Guivar'ch,
  W. Szpankowski, B. Vallée
  ICALP 2001
- **Hidden Word Statistics**
  P. Flajolet, W. Szpankowski, B. Vallée
  J. ACM 2006

Motif Statistics - [Ni-Sa-Fl 99,02]

Some biological Motivation

```
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 proteome**

More generally: statistics of the **number** of **matching positions** 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 ($\mathcal{A}^\star = \epsilon + \mathcal{A} + \mathcal{A}^2 + \mathcal{A}^3 + \dots$)

Aim & Result

R given regular expression.
X_n number of occurrences in a text of length n.
Aim: compute $F(z,u) = \sum_{n,k} \Pr(X_n = k) u^k z^n$.

Thm:[Ni-Sa-Fl 99] **With** or **without** counting **overlap**, both in the **Bernoulli** and **Markov** model,

1. $F(z,u)$ is **rational** and **can be computed** explicitly
2. **Moments** $\begin{cases} \mathbb{E}(X_n) & = \mu n + c_1 + O(A^n), \\ \mathrm{Var}(X_n) & = \sigma^2 n + c_2 + O(A^n). \end{cases}$
3. **Limit Gaussian law**:

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

The Right Rational Language

R **regular expression** over $\mathcal{A}$

Key: $\mathcal{L}_m \subset (\mathcal{A} \cup \{m\})^\star$

- contains all the words of $\mathcal{A}^\star$,
- with a **zero-length mark** (m) after each occurrence of R.

Example:

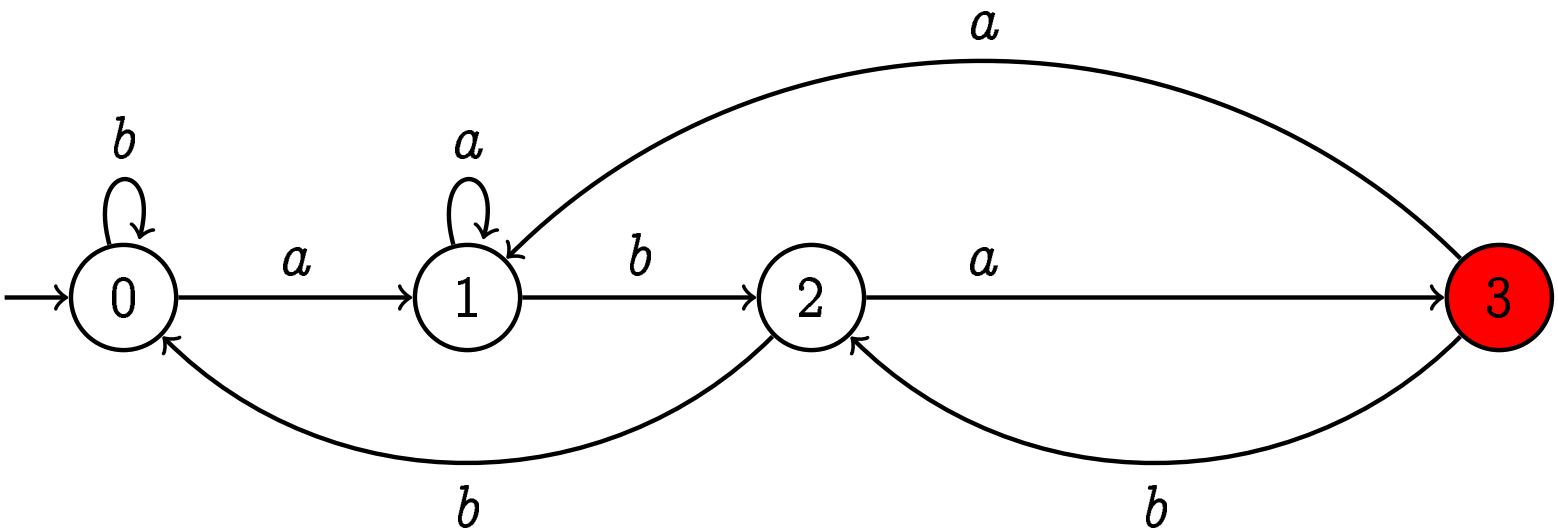
$R = aba$

$aaa\,aba\,m\,ba\,m\,aa\,aba\,m\,aa$ (overlap)
$aaa\,aba\,m\,ba\,aa\,aba\,m\,aa$ (non-overlap).

Automaton recognizing $\mathcal{A}^\star R$ (DFA)

$\mathcal{A} = \{a, b\}$ $R = aba$, $E = \mathcal{A}^\star R = \mathcal{A}^\star aba$

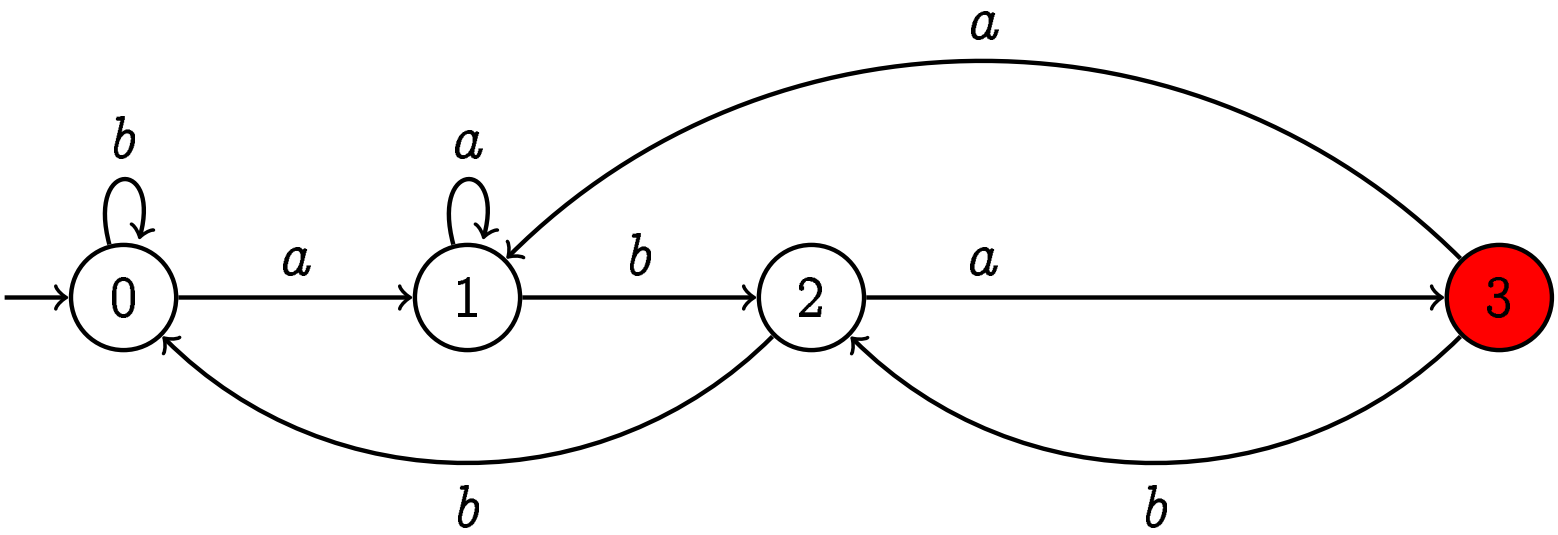
$aabb\,aba \bullet ba \bullet abb\,aba \bullet aa\,aba \bullet bbbb$



Automaton recognizing $\mathcal{A}^\star R$ (DFA)

$\mathcal{A} = \{a, b\}$ $\quad R = aba, \quad E = \mathcal{A}^\star R = \mathcal{A}^\star aba$

$aabb\,aba \bullet ba \bullet abb\,aba \bullet aa\,aba \bullet bbbb$



Chomsky-Schützenberger

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

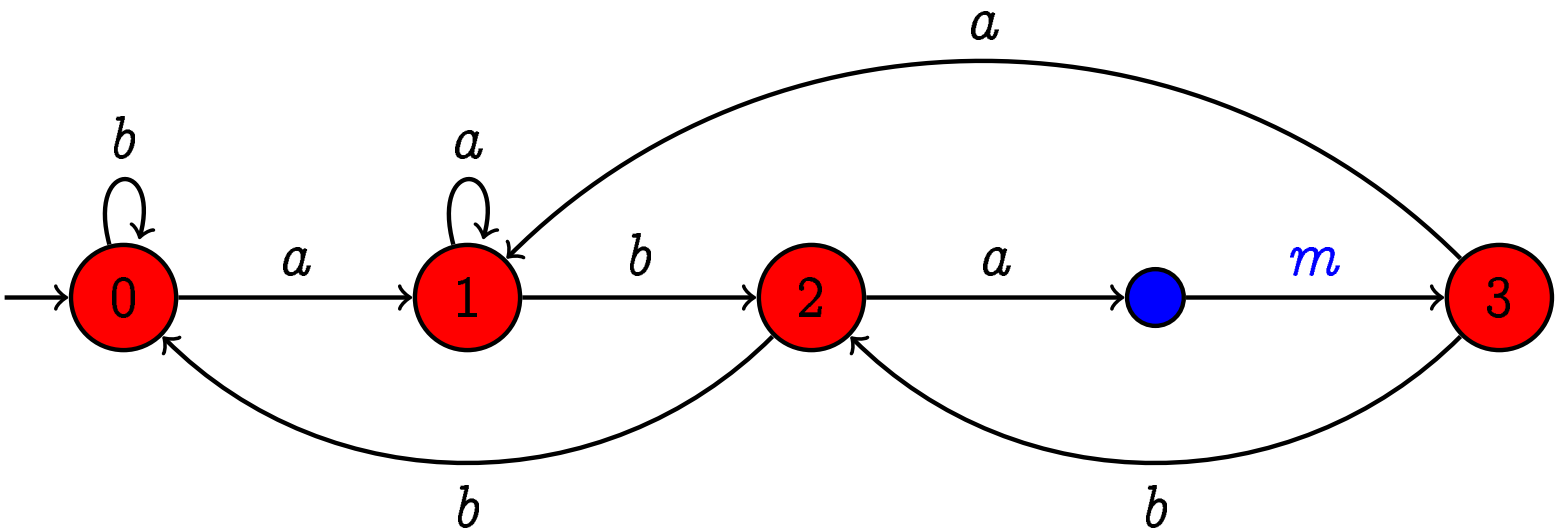
$$L_0 = aL_1 + bL_0, \qquad L_1 = bL_2 + aL_1,$$
$$L_2 = aL_3 + bL_0, \qquad L_3 = aL_1 + bL_2 + 1.$$

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

Automaton recognizing $\mathcal{A}^\star R$ (DFA)

$\mathcal{A} = \{a, b\}$ $R = aba$, $E = \mathcal{A}^\star R = \mathcal{A}^\star aba$

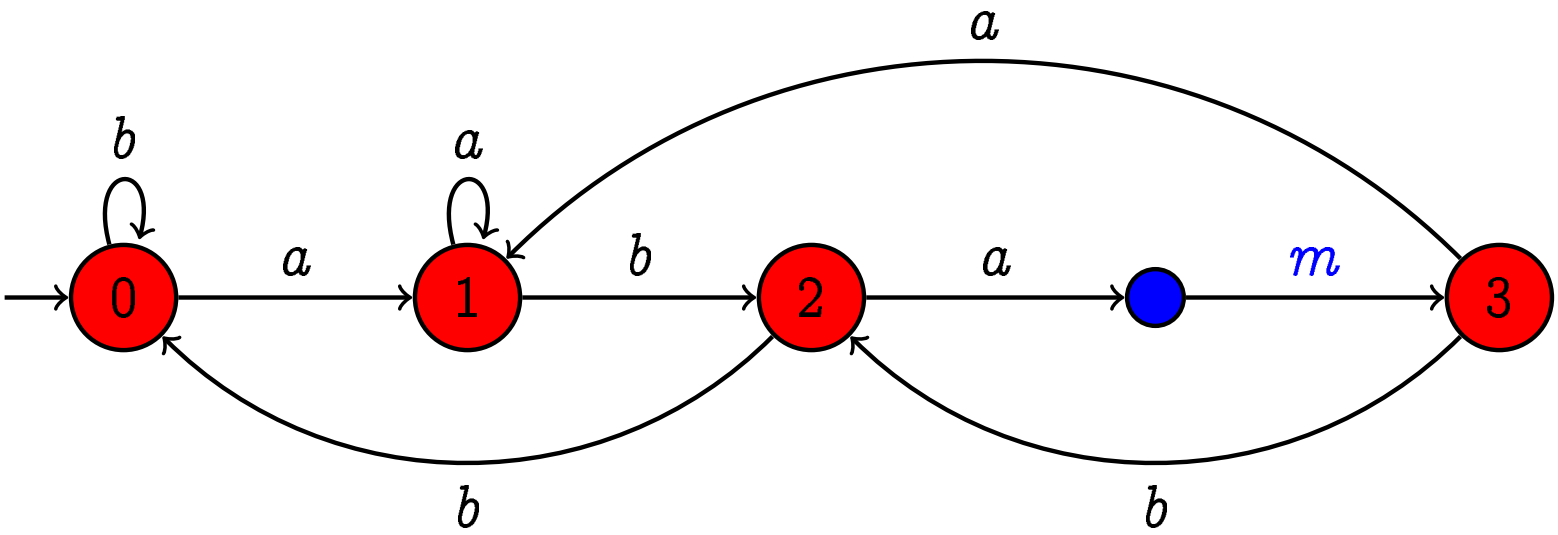
$aabb\,aba \bullet ba \bullet abb\,aba \bullet aa\,aba \bullet bbbb$



Automaton recognizing $\mathcal{A}^\star R$ (DFA)

$\mathcal{A} = \{a, b\}$ $\quad R = aba, \quad E = \mathcal{A}^\star R = \mathcal{A}^\star aba$

$aabb\,aba \bullet ba \bullet abb\,aba \bullet aa\,aba \bullet bbbb$



$$L_0 = aL_1 + bL_0 + 1, \qquad L_1 = bL_2 + aL_1 + 1,$$
$$L_2 = amL_3 + bL_0 + 1, \qquad L_3 = aL_1 + bL_2 + 1.$$

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

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

The algorithmic chain

- **Input**: **regular expression** R
- **Algorithm**
  1. **Berry-Sethy** $\mapsto$ **NFA** for $\mathcal{A}^* R$
  2. **Determinization** $\mapsto$ **DFA** for $\mathcal{A}^* R$
  3. **Marking** $\mapsto$ **marked DFA** for $\mathcal{A}^* R$
  4. **Chomsky-Schützenberger** $\mapsto$ $F(z, u)$,
- **Output** $F(z, u) = \sum p_{n,k} u^k z^n$

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

Exploiting the Output

$$F(z,u) \in \mathbb{Q}(z,u)$$

$$\Longrightarrow \begin{cases} G(z) = \left.\dfrac{\partial F(z,u)}{\partial u}\right|_{u=1} = \sum \mathbf{E}(X_n) z^n \in \mathbb{Q}(z), \\ H(z) = \left.\dfrac{\partial}{\partial u} u \dfrac{\partial F(z,u)}{\partial u}\right|_{u=1} = \sum \mathrm{M}_2(X_n) z^n \in \mathbb{Q}(z), \\ N(z) = F(z,0) = \sum \Pr(X_n = 0) z^n \in \mathbb{Q}(z). \end{cases}$$

- **Fast extraction** of coefficients:
  n-th coefficient in $O(\log n)$ operations
  [implemented in `algolib-gfun`].
- **Exponentially good asymptotics** in **constant time**.

Proof of the Gaussian Law

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

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

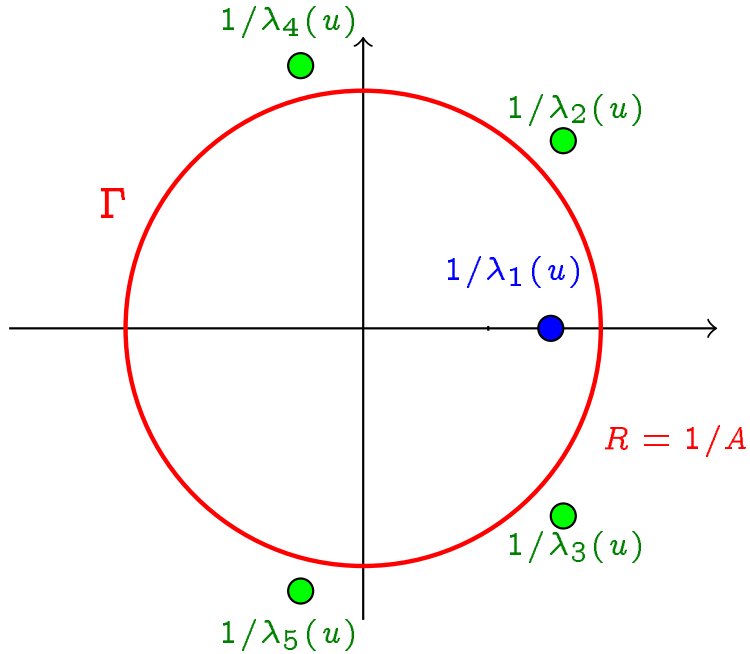
$\mathbb{T}(u)$ **positive** $n \times n$ matrix

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

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

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

Uniform Separation Property

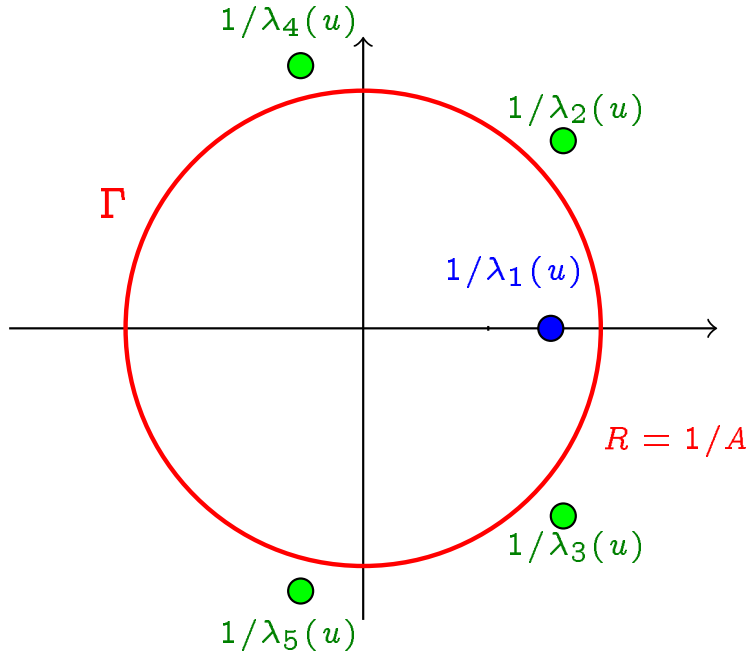


$$p_n(u) = [z^n]F(z,u) = \frac{1}{2i\pi}\oint_\Gamma \frac{dz}{z^{n+1}}F(z,u),$$

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

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

Uniform Separation Property



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

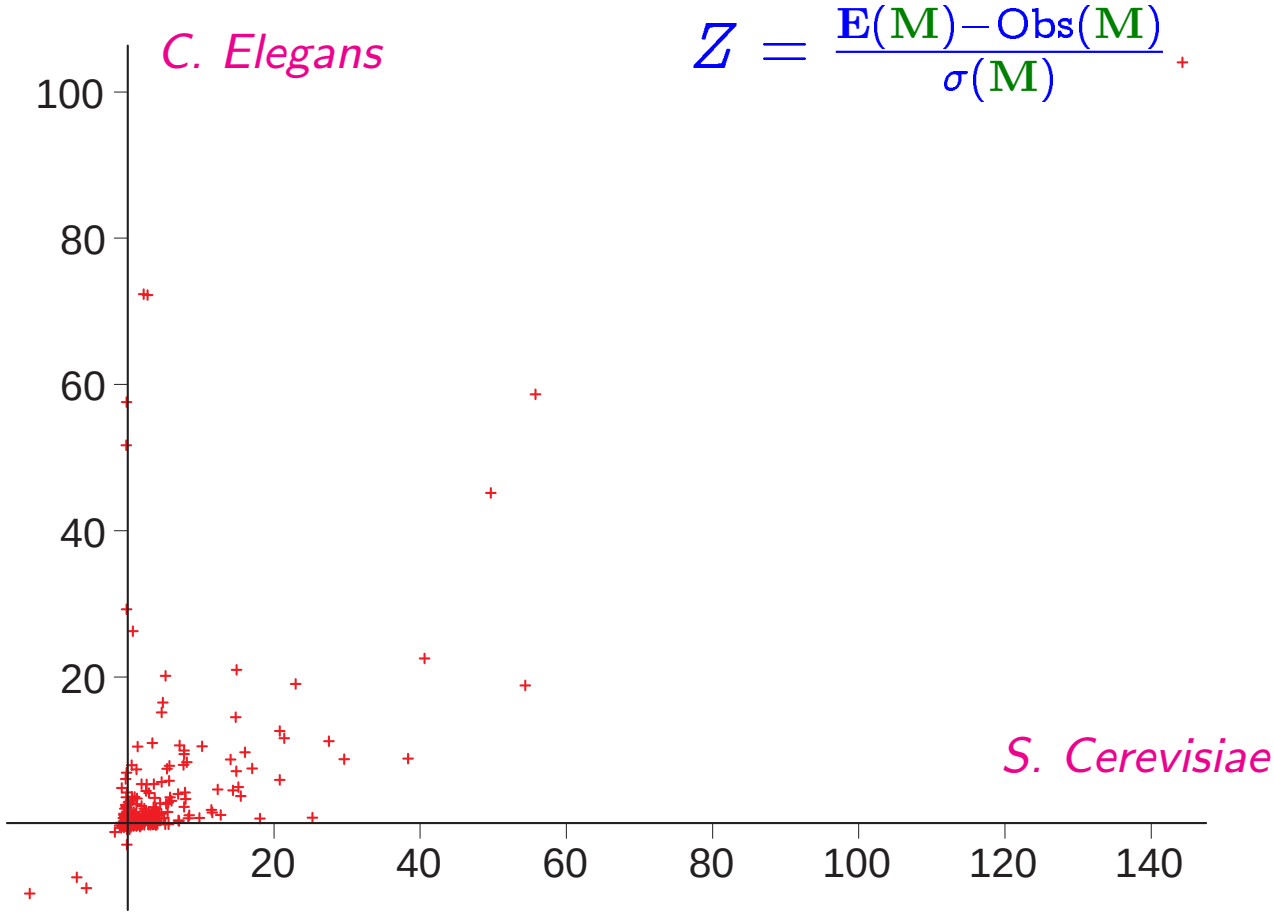
$$= \frac{1}{2i\pi}\oint_\Gamma \frac{c(u)}{z^{n+1}(1-\lambda_1(u)z)} + \frac{1}{z^{n+1}}g(z,u)\,dz,$$

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

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

Variability condition: $\lambda''(1)+\lambda'(1)-\lambda'(1)^2 \neq 0$ $\qquad (\lambda(u)=\lambda_1(u))$

A biological application



Z-scores for *C. Elegans* and *S. Cerevisiae* (305 motifs $\mathbf{M}$)

values normalized for 1,000,000 positions and $\mathbf{E}(\mathbf{M}) \geq 0.05$

Hidden Pattern Statistics - [Fl-Gu-Sz-Va 01]

Counting tool

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

$$\mathcal{A} = \{a, b\}, \qquad p(a) + p(b) = \Big(p(a) + p(b)\Big)^k = 1$$

$$F_{\mathcal{A}^\star}(z) = 1 + z + z^2 + z^3 + \cdots = \frac{1}{1 - z}$$

Hidden Pattern Statistics - [Fl-Gu-Sz-Va 01]

Counting tool

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

$$\mathcal{A} = \{a, b\}, \qquad p(a) + p(b) = \Big(p(a) + p(b)\Big)^k = 1$$

$$F_{\mathcal{A}^\star}(z) = 1 + z + z^2 + z^3 + \cdots = \frac{1}{1-z}$$

Example

- $\mathcal{L} = \mathcal{A}^\star.a.\mathcal{A}^\star.a.\mathcal{A}^\star, \quad \mathcal{A} = \{a, b\}, \quad p(a) = p(b) = \frac{1}{2}$
- $F_{\mathcal{L}}(z) = \frac{1}{1-z} \times \frac{z}{2} \times \frac{1}{1-z} \times \frac{z}{2} \times \frac{1}{1-z}$

Hidden Pattern Statistics - [Fl-Gu-Sz-Va 01]

$\mathcal{W} = a\#_3 a$ - count of number of **occurrences** of **two** a's separated by

- **at most any two** letters,
- or by a **gap** with $|\text{gap}| < 3$

Hidden Pattern Statistics - [Fl-Gu-Sz-Va 01]

$\mathcal{W} = a\#_3 a$ - count of number of **occurrences** of **two** a's separated by

- **at most any two** letters,
- or by a **gap** with $|\text{gap}| < 3$

$\mathcal{W} = a\#_3 a$

$abaaa$

$a\ \ a$

$\ \ aa$

$a\ \ a$

$\ \ aa$

hidden occurrences

$\mathcal{R} = a(\epsilon + x + x^2)a \quad (x \in \mathcal{A})$

$abaaa$

aba

$abaa$

$\ \ aaa$

matching positions

Hidden Pattern Statistics - [Fl-Gu-Sz-Va 01]

$\mathcal{W} = a\#_3 a$ - count of number of **occurrences** of **two** a's separated by

- **at most any two** letters,
- or by a **gap** with $|\text{gap}| < 3$

$\mathcal{W} = a\#_3 a$

$abaaa$

$a\ \ a$

aa

$a\ \ a$

aa

hidden occurrences

$\mathcal{R} = a(\epsilon + x + x^2)a \quad (x \in \mathcal{A})$

$abaaa$

aba

$abaa$

aaa

matching positions

$$\mathcal{W} = a\#_\infty a \rightsquigarrow F(z) = \frac{1}{1-z} \times \frac{z}{2} \times \frac{1}{1-z} \times \frac{z}{2} \times \frac{1}{1-z}$$

$$\Longrightarrow \quad \mathbf{E}(O_n(\mathcal{W})) = [z^n]\,\frac{z^2}{4} \times \frac{1}{(1-z)^3} = \frac{1}{2\times 4} \times n^2 + O(n)$$

Motivations

- **Intrusion detection** in **computer security** repeated **signatures** as **subsequences**
  - find an **indicator** to set up **alarms**
  - **avoid false alarms**
- **Bioinformatics**: introns versus exons, tandem repeats in DNA

Motivations

- **Intrusion detection** in **computer security** repeated **signatures** as **subsequences**
  - find an **indicator** to set up **alarms**
  - **avoid false alarms**
- **Bioinformatics**: introns versus exons, tandem repeats in DNA
- **Fun**: **secret codes**:
  - **Moby Dick predicted in details Lady Di's accidental death**!!
  - The **Bible** and the **Koran** predict the date at which **baby Giulia Sarkozy** will get **her first tooth**

Some definitions

$w = abracadabra$

$\mathcal{W} = a|b\#_2 r\#a|c\#a\#d\#_4 a\#b|r\#a, \qquad (| = \#_1)$

Some definitions

$w = abracadabra$

$\mathcal{W} = a|b\#_2 r\#a|c\#a\#d\#_4 a\#b|r\#a, \qquad (| = \#_1)$

Constraint

$\mathcal{D} = (d_1, d_2, \ldots, d_{m-1}) \in (\mathbb{N}^+ \cup \{\infty\})^{m-1} \quad (m = |w|)$

Some definitions

$w = abracadabra$

$\mathcal{W} = a|b\#_2 r\#a|c\#a\#d\#_4 a\#b|r\#a, \qquad (| = \#_1)$

Constraint

$\mathcal{D} = (d_1, d_2, \ldots, d_{m-1}) \in (\mathbb{N}^+ \cup \{\infty\})^{m-1} \qquad (m = |w|)$

$\mathcal{D}_{\mathcal{W}} = (1, 2, \infty, 1, \infty, \infty, 4, \infty, 1, \infty)$

Some definitions

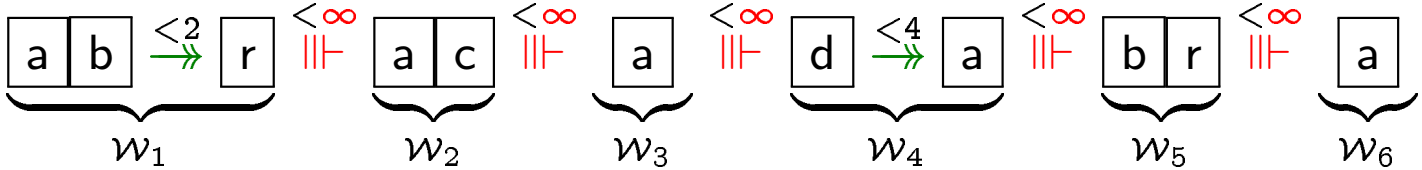
$w = abracadabra$

$\mathcal{W} = a|b\#_2 r\#a|c\#a\#d\#_4 a\#b|r\#a, \qquad (| = \#_1)$

Constraint

$\mathcal{D} = (d_1, d_2, \ldots, d_{m-1}) \in (\mathbb{N}^+ \cup \{\infty\})^{m-1} \qquad (m = |w|)$

$\mathcal{D}_{\mathcal{W}} = (1, 2, \infty, 1, \infty, \infty, 4, \infty, 1, \infty)$



Some definitions

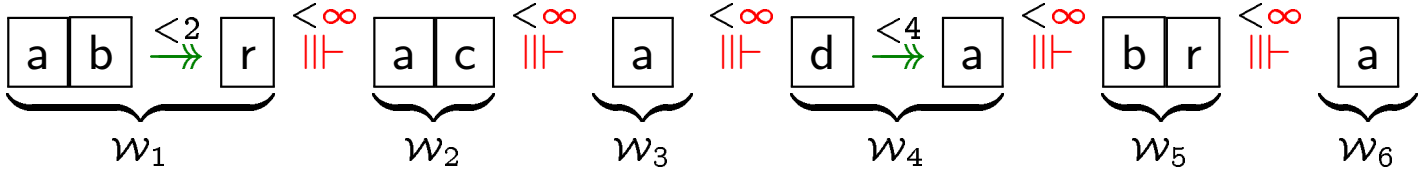
$w = abracadabra$

$\mathcal{W} = a|b\#_2 r\#a|c\#a\#d\#_4 a\#b|r\#a, \qquad (| = \#_1)$

Constraint

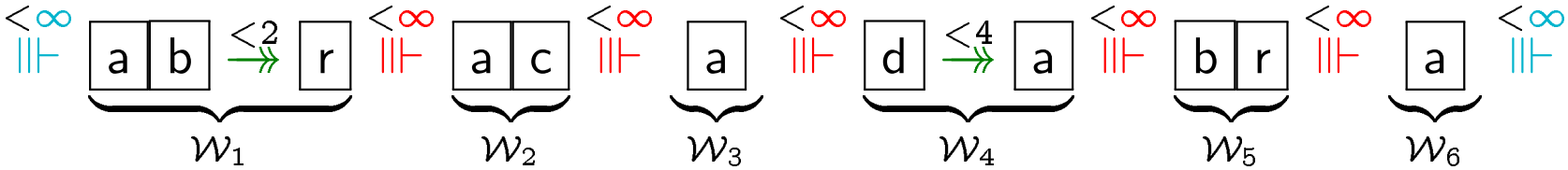
$\mathcal{D} = (d_1, d_2, \ldots, d_{m-1}) \in (\mathbb{N}^+ \cup \{\infty\})^{m-1} \quad (m = |w|)$

$\mathcal{D}_{\mathcal{W}} = (1, 2, \infty, 1, \infty, \infty, 4, \infty, 1, \infty)$

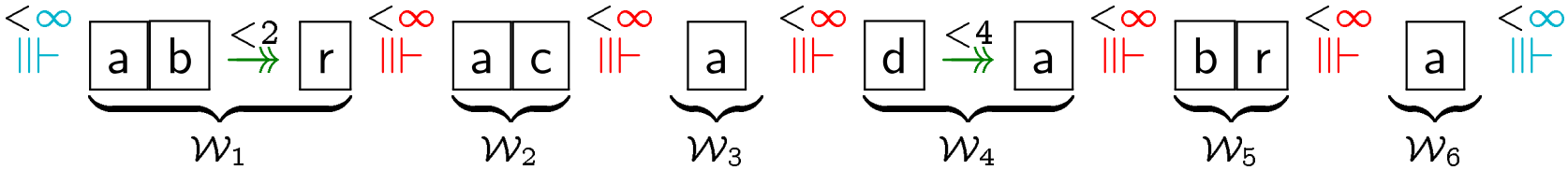


- **Valid position**: $I = (i_1, i_2, \ldots, i_m)$ with $i_{j+1} - i_j \leq d_j$
- **Occurrence position** in text T: (I, T) with $\begin{cases} I \text{ valid} \\ T[i_j] = w[j] \end{cases}$

Blocks and subpositions

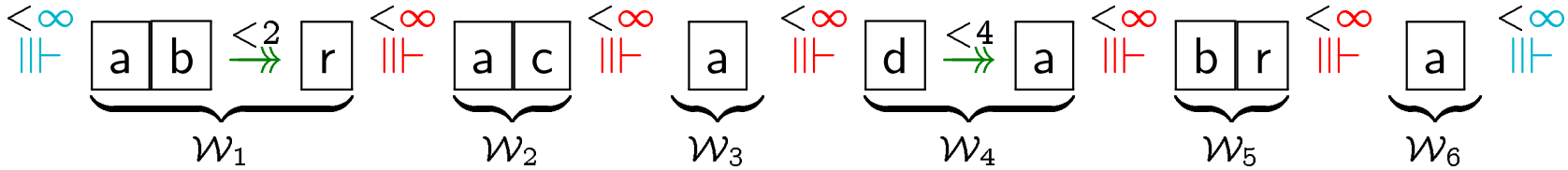


Blocks and subpositions



position: $I = (6, 7, 9, 18, 19, 22, 30, 33, 50, 51, 60)$

Blocks and subpositions

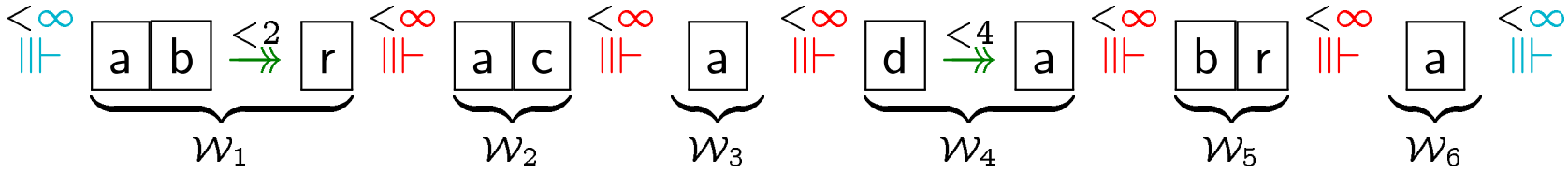


position: $I = (6, 7, 9, 18, 19, 22, 30, 33, 50, 51, 60)$

$\rightsquigarrow$ 6 **subpositions**:

$$\overbrace{(6,7,9)}^{I^{[1]}},\quad \overbrace{(18,19)}^{I^{[2]}},\quad \overbrace{(22)}^{I^{[3]}},\quad \overbrace{(30,33)}^{I^{[4]}},\quad \overbrace{(50,51)}^{I^{[5]}},\quad \overbrace{(60)}^{I^{[6]}}$$

Blocks and subpositions



position: $I = (6, 7, 9, 18, 19, 22, 30, 33, 50, 51, 60)$

$\rightsquigarrow$ 6 **subpositions**:

$$\overbrace{(6,7,9)}^{I^{[1]}},\quad \overbrace{(18,19)}^{I^{[2]}},\quad \overbrace{(22)}^{I^{[3]}},\quad \overbrace{(30,33)}^{I^{[4]}},\quad \overbrace{(50,51)}^{I^{[5]}},\quad \overbrace{(60)}^{I^{[6]}}$$

$\rightsquigarrow$ 6 **blocks**:

$$\underbrace{[6,9]}_{B^{[1]}},\quad \underbrace{[18,19]}_{B^{[2]}},\quad \underbrace{[22]}_{B^{[3]}},\quad \underbrace{[30,33]}_{B^{[4]}},\quad \underbrace{[50,51]}_{B^{[5]}},\quad \underbrace{[60]}_{B^{[6]}}$$

number of **blocks**: $b = |\mathcal{W}|_\infty + 1$ Example: $b = 5 + 1 = 6$

Mean value analysis

$$\Omega(T) = \sum_{I \in \mathcal{P}_{[T]}} X_I(T), \quad \begin{cases} X_I(T) = [\![w \text{ occurs at position } I \text{ in } T]\!] \\ \mathcal{P}_{[T]} \text{ set of valid positions for T} \end{cases}$$

Mean value analysis

$$\Omega(T) = \sum_{I \in \mathcal{P}_{[T]}} X_I(T), \quad \begin{cases} X_I(T) = [\![w \text{ occurs at position } I \text{ in } T]\!] \\ \mathcal{P}_{[T]} \text{ set of valid positions for T} \end{cases}$$

$\mathcal{O}$ **set** of **texts** with **hidden occurrences**

$$\mathcal{O} = \mathcal{A}^\star \times \{w_1\} \times \mathcal{A}^{< d_1} \times \{w_2\} \times \mathcal{A}^{< d_2} \times \ldots \times \{w_{m-1}\} \times \mathcal{A}^{< d_{m-1}} \times \{w_m\} \times \mathcal{A}^\star$$

Mean value analysis

$$\Omega(T) = \sum_{I \in \mathcal{P}_{[T]}} X_I(T), \quad \begin{cases} X_I(T) = [\![w \text{ occurs at position } I \text{ in } T]\!] \\ \mathcal{P}_{[T]} \text{ set of valid positions for T} \end{cases}$$

$\mathcal{O}$ **set** of **texts** with **hidden occurrences**

$$\mathcal{O} = \mathcal{A}^\star \times \{w_1\} \times \mathcal{A}^{<d_1} \times \{w_2\} \times \mathcal{A}^{<d_2} \times \ldots \times \{w_{m-1}\} \times \mathcal{A}^{<d_{m-1}} \times \{w_m\} \times \mathcal{A}^\star$$

set of **indices** of **finite constraints** $\mathcal{F} = \{ j;\ d_j < \infty \}$

number of **unbounded constraints**: $|\mathcal{D} \setminus \mathcal{F}| + 2 = b - 1 + 2$

$$O(z) = \sum_{n \geq 0} \mathbb{E}_n[\Omega] z^n = \left(\frac{1}{1-z} \right)^{b+1} \times \left(\prod_{i=1}^{m} p_{w_i} z \right) \times \left(\prod_{j \in \mathcal{F}} \frac{1-z^j}{1-z} \right)$$

Mean value analysis

$$\Omega(T) = \sum_{I \in \mathcal{P}_{[T]}} X_I(T), \quad \begin{cases} X_I(T) = [\![w \text{ occurs at position } I \text{ in } T]\!] \\ \mathcal{P}_{[T]} \text{ set of valid positions for T} \end{cases}$$

$\mathcal{O}$ **set** of **texts** with **hidden occurrences**

$$\mathcal{O} = \mathcal{A}^\star \times \{w_1\} \times \mathcal{A}^{<d_1} \times \{w_2\} \times \mathcal{A}^{<d_2} \times \ldots \times \{w_{m-1}\} \times \mathcal{A}^{<d_{m-1}} \times \{w_m\} \times \mathcal{A}^\star$$

set of **indices** of **finite constraints** $\mathcal{F} = \{ j;\ d_j < \infty \}$

number of **unbounded constraints**: $|\mathcal{D} \setminus \mathcal{F}| + 2 = b - 1 + 2$

$$O(z) = \sum_{n \geq 0} \mathbb{E}_n[\Omega] z^n = \left(\frac{1}{1-z} \right)^{b+1} \times \left(\prod_{i=1}^{m} p_{w_i} z \right) \times \left(\prod_{j \in \mathcal{F}} \frac{1 - z^j}{1 - z} \right)$$

$$\mathbb{E}_n[\Omega] = [z^n] O(z) = \frac{n^b}{b!} \left(\prod_{j \in \mathcal{F}} d_j \right) p(w) \left(1 + O\left(\frac{1}{n} \right) \right)$$

Variance Analysis

Centered random variables

$$\Xi_n = \Omega_n - \mathbf{E}_n[\Omega] = \sum_{I \in \mathcal{P}_n} Y_I, \quad \text{with} \quad Y_I = X_I - \mathbf{E}[X_I] = X_I - p(w)$$

$$\mathbf{E}[\Xi_n^2] = \sum_{I,J \in \mathcal{P}_n} \mathbf{E}[Y_I Y_J]$$

Variance Analysis

Centered random variables

$$\Xi_n = \Omega_n - \mathbf{E}_n[\Omega] = \sum_{I \in \mathcal{P}_n} Y_I, \quad \text{with} \quad Y_I = X_I - \mathbf{E}[X_I] = X_I - p(w)$$

$$\mathbf{E}[\Xi_n^2] = \sum_{I,J \in \mathcal{P}_n} \mathbf{E}[Y_I\, Y_J]$$

one needs analysing pairs of **positions** (I, J)

$$\mathcal{O}_2 = \{(I, J, T): \quad I, J \in \mathcal{P}_{[T]}\}$$

Variance Analysis

Centered random variables

$$\Xi_n = \Omega_n - \mathbf{E}_n[\Omega] = \sum_{I \in \mathcal{P}_n} Y_I, \quad \text{with} \quad Y_I = X_I - \mathbf{E}[X_I] = X_I - p(w)$$

$$\mathbf{E}[\Xi_n^2] = \sum_{I,J \in \mathcal{P}_n} \mathbf{E}[Y_I\, Y_J]$$

one needs analysing pairs of **positions** (I, J)

$$\mathcal{O}_2 = \{(I, J, T): \quad I, J \in \mathcal{P}_{[T]}\}$$

$$O_2(z) = \sum_{(I,J,T) \in \mathcal{O}_2} Y_I(T)\, Y_J(T) p(T) z^{|T|}$$

$$= \sum_{n \geq 0} \sum_{I,J \in \mathcal{P}_n} \mathbf{E}[Y_I(n)\, Y_J(n)] z^n = \sum_{n \geq 0} \mathbf{E}[\Xi_n^2] z^n$$

Variance Analysis

$$O_2(z) = \sum_{n \geq 0} \sum_{I, J \in \mathcal{P}_n} \mathbf{E}[Y_I Y_J] z^n = \sum_{n \geq 0} \mathbf{E}[\Xi_n^2] z^n$$

$I \cap J = \emptyset$: Y_I and Y_J **independent**, $\mathbf{E}[Y_I Y_J] = 0$

only need to consider **intersecting positions** and **blocks**

$\mathcal{W}_{I \cap J}$: **subpattern** of $\mathcal{W}$ occurring at **positions** $I \cap J$

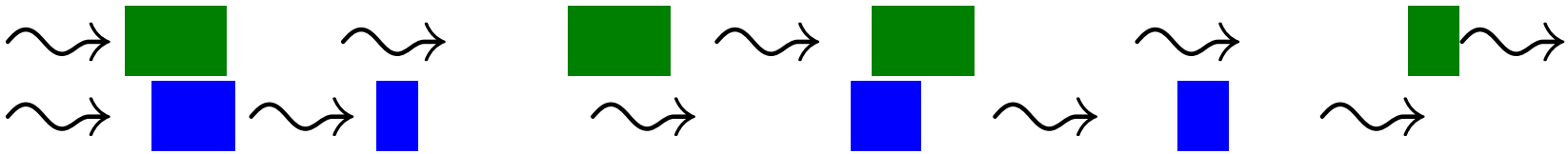
$$\mathbf{E}[X_I X_J] = \frac{p^2(w)}{p\left(\mathcal{W}_{I \cap J}\right)}$$

$$\Longrightarrow \quad \mathbf{E}[Y_I Y_J] = \mathbf{E}[X_I X_J] - p^2(w) = p^2(w) e(I, J)$$

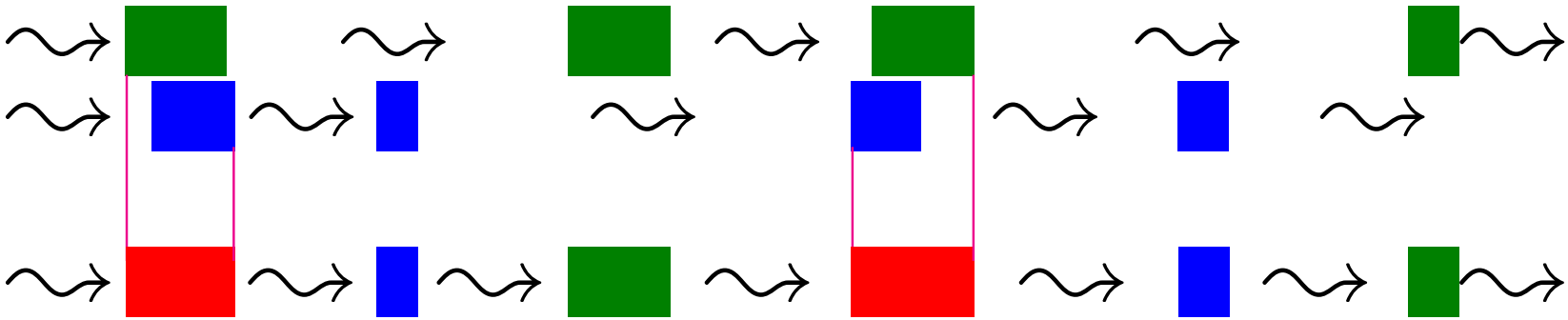
$$e(I, J) = \frac{1}{p\left(\mathcal{W}_{I \cap J}\right)} - 1$$

correlation number

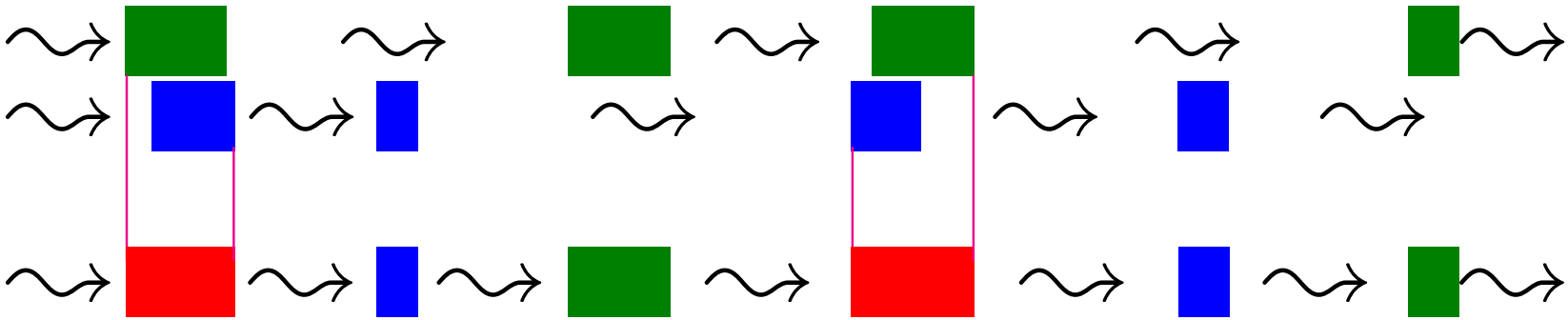
Variance Analysis - Hitting blocks



Variance Analysis - Hitting blocks



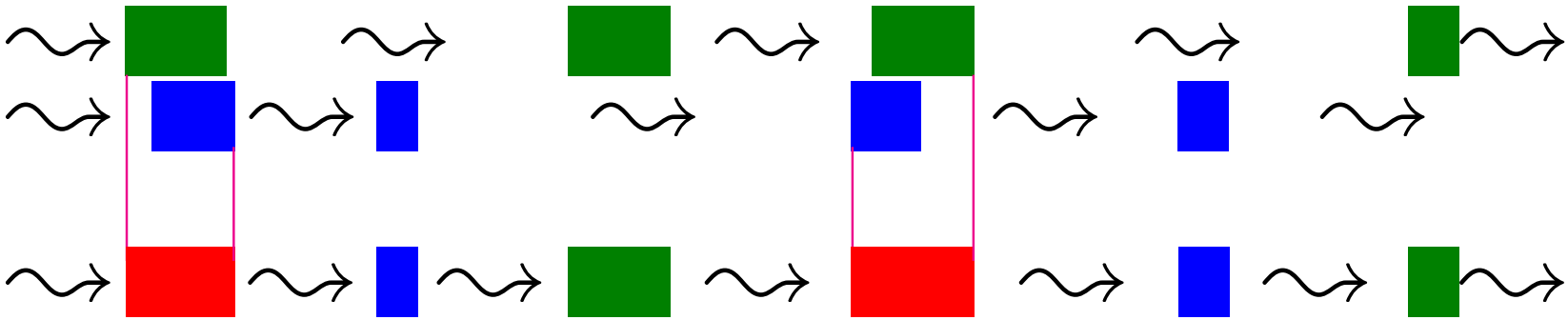
Variance Analysis - Hitting blocks



ϕ: degree of **freedom**, c: number of **collisions**

- **positions** I and J: $b = 4$ **blocks**, $\Phi = 5$
- **position** $I \cup J$: $c = 2$, $b_{I \cup J} = 6$ **blocks**, $\Phi = 2 \times 4 - 2 + 1 = 7$

Variance Analysis - Hitting blocks



ϕ: degree of **freedom**, c: number of **collisions**

- **positions** I and J: $b = 4$ **blocks**, $\Phi = 5$
- **position** $I \cup J$: $c = 2$, $b_{I \cup J} = 6$ **blocks**, $\Phi = 2 \times 4 - 2 + 1 = 7$

General case

$$\Phi = 2b - c + 1 \quad \rightsquigarrow \quad O_2^{[c]}(z) \cong \left(\frac{1}{1-z}\right)^{2b-c+1} \times \mathrm{Poly}(z)$$

corresponding term in the **variance**: $O\left(n^{2b-c}\right)$

dominant terms: $c = 1$, **only one block collision**: $O\left(n^{2b-1}\right)$

Expectation and Variance

Thm[Fl-Gu-Sz-Va 01] For a general constraint $\mathcal{D}$ with b **blocks**, the random variable Ω counting the number of hidden occurrences of a pattern $\mathcal{W}$ in texts verifies

$$\mathbf{E}_n[\Omega] = \frac{p(w)}{b!}\left(\prod_{j;d_j<\infty} d_j\right) n^b \left(1+O\left(\frac{1}{n}\right)\right),$$

$$\mathbf{Var}_n[\Omega] = \sigma^2(\mathcal{W}) \times n^{2b-1}\left(1+O\left(\frac{1}{n}\right)\right),$$

where the "**variance coefficient**" $\sigma^2(\mathcal{W})$ verifies

$$\sigma^2(\mathcal{W}) = \frac{p^2(\mathcal{W})}{(2b-1)!}\kappa^2(\mathcal{W}) \quad \text{with} \quad \kappa^2(\mathcal{W}) := \sum_{(I,J)\in\mathcal{B}_2^{[1]}}\left(\frac{1}{p(\mathcal{W}_{I\cap J})}-1\right)$$

$\sigma^2(\mathcal{W})$ { - connections with the **autocorrelation polynomial**
- computable by **dynamic programming**

Higher moments - Central Limit Laws

$$\widetilde{\Xi}_n := \frac{\Xi_n}{n^{b-1/2}} = \frac{\Omega_n - \mathbf{E}_n[\Omega]}{n^{b-1/2}}$$

$$\mathbf{E}[\Xi^r] = \sum_{(I_1,\ldots,I_r)\in\mathcal{P}_n^r} \mathbf{E}[\,Y_1 \ldots Y_r]$$

- same line of proof as in variance analysis
- **minimum** number of **collisions**: **maximum freedom degree**

Higher moments - Central Limit Laws

$$\widetilde{\Xi}_n := \frac{\Xi_n}{n^{b-1/2}} = \frac{\Omega_n - \mathbf{E}_n[\Omega]}{n^{b-1/2}}$$

$$\mathbf{E}[\Xi^r] = \sum_{(I_1,\ldots,I_r)\in\mathcal{P}_n^r} \mathbf{E}[\,Y_1 \ldots Y_r]$$

- same line of proof as in variance analysis
- **minimum** number of **collisions**: **maximum freedom degree**

Odd moments:

$$\mathbf{E}_n[\Xi^{2s+1}] = o\left(n^{(2s+1)(b-1/2)}\right) \quad \Longrightarrow \quad \mathbf{E}_n[\widetilde{\Xi}^{2s+1}] = o(1)$$

Even moments: **symmetries**, **involutions** and **shuffling**

$$\mathbf{E}_n[\Xi^{2s}] \simeq 1.3.5.\ldots.(2s-1)n^{(2b-1)s}\sigma^{2s}$$

$$\Longrightarrow \mathbf{E}_n[\widetilde{\Xi}^{2s}] = 1.3.5.\ldots.(2s-1)\sigma^{2s}$$

Thm[Fl-Gu-Sz-Va 01]

Convergence by moments to the **Gaussian distribution**

Experiments

Hamlet text stripped of non-alphabetic characters
n=120,057 alphabetic characters
w = thelawisgaussian "*The Law is Gaussian*"
$\widetilde{w}$ = naissuagsiwaleth (mirror image of w)
d: maximum distance authorized between letters

		thelawisgaussian		naissuagsiwaleth	
d	Expected (E)	Occurred (Ω)	Ω/E	Occurred (Ω)	Ω/E
13	9.195E+01	0	0.00	18	0.19
14	2.794E+02	693	2.47	371	1.32
20	5.886E+04	124,499	2.11	41,066	0.69
50	5.482E+10	76,146,232,395	1.38	48,386,404,680	0.88
∞	1.330E+48	1.36554E+48	1.03	1.38807E+48	1.04

Fully Constrained Case - [Fl-Sz-Va 06]

Use a de Bruijn graph

$$\left| \text{  } \right| = \max(\mathcal{W}) - 1 = |w| + \left(\sum_{d_j \in \mathcal{D}} (d_j - 1) \right) - 1$$

Example: $\mathcal{W} = a\#_2 b$

$$\mathbb{T}(u) = \begin{array}{c} \\ aa \\ ab \\ ba \\ bb \end{array} \begin{array}{c} \begin{array}{cccc} aa & ab & ba & bb \end{array} \\ \begin{pmatrix} p_a & p_b u^2 & 0 & 0 \\ 0 & 0 & p_a & p_b u \\ p_a & p_b u & 0 & 0 \\ 0 & 0 & p_a & p_b \end{pmatrix} \end{array}$$

Perron-Frobenius, properties of **cycles** in **graphs**, **quasi-powers**

Thm[Fl-Sz-Va 06] **Precise distribution results**

- **Speed of convergence** to the **Gaussian** law: $O(1/n)$
- **Large deviation** result
- **Local limit law** for **primitive** patterns (all $d_j = 1$, or one letter of the alphabet is missing, in constrast to "*The quick brown foxes jump over lazy dogs*")

A humoristic nod to Philippe

A humoristic nod to Philippe