Statistiques de Motifs

Pierre Nicodème
Centre de Recherche sur le Cancer
DKFZ-Heidelberg

Travail en commun avec

Bruno Salvy et Philippe Flajolet

INRIA - Rocquencourt

Février, 1999

Motivations

1) Calcul automatique de la Statistique du nombre d'occurrences d'une Expression Régulière dans un texte aléatoire

2) Application à la calibration probabiliste du nombre d'occurrences des Motifs de la base Prosite dans une base de protéines.

Motifs de Prosite

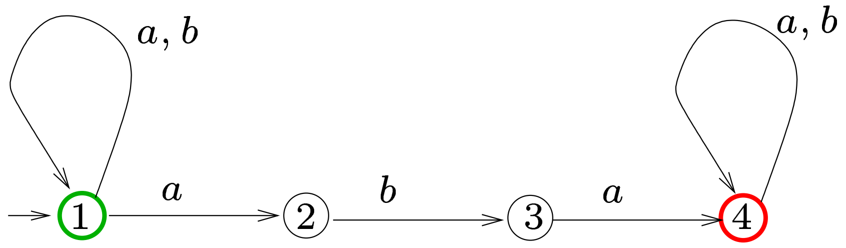
Sous-classe des expressions régulières (pas d' $\star$)

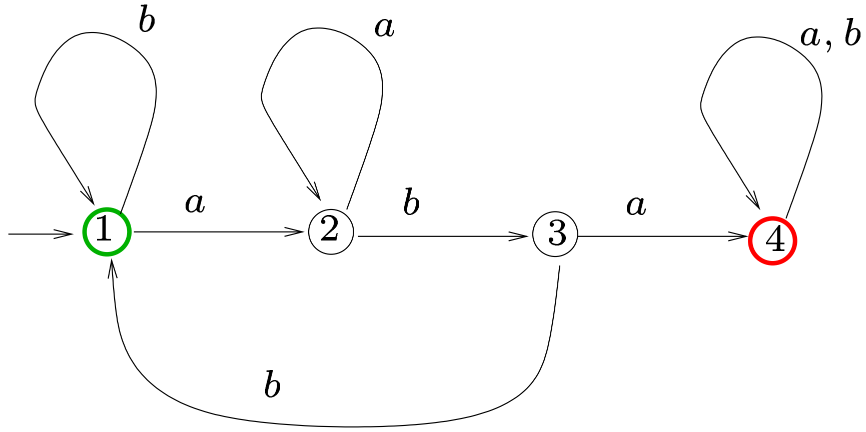
Exemple

```
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;
...
```

Langages réguliers, automates





NFA et DFA reconnaissant

$\mathcal{L} = (a+b)^*aba(a+b)^*$

Exemple: mot $bbbababbbababbb$

Séries génératrices

sér. gén. ordinaire d'un langage

$\mathcal{L} = \{w, w \in \Sigma^*\}$

$\rightarrow \quad F(l_1, l_2, \ldots, l_p) = \sum_{w \in \mathcal{L}} \mathrm{com}(w)$

$\mathcal{L} = (aabaa)^\star \cup ccc \quad \rightarrow F(a, b, c) = \frac{1}{1-a^4 b} + c^3$

Si $A.B$ non ambigu, $F_{A.B}(z) = F_A(z) F_B(z)$

Si A et B disjoints $F_{A \cup B}(z) = F_A(z) + F_B(z)$

Théorèmes

Langages réguliers, automates, séries génératrices

- Kleene - langage régulier si et seulement si reconnu par un NFA
- Rabin & Scott - NFAs et DFAs équivalents (reconnaissent mêmes langages)
- Chomsky & Schützenberger
  lang. régulier $\rightarrow$ série gén. rationnelle

Séries génératrices multivariées

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

$a_{n,k}$ nombre de mots de taille n ayant k occurrences du motif

$F(z/|\Sigma|, u)$ série génératrice de probabilité.

Moments d'ordre 1 et 2

$M_1(z) = \frac{\delta F}{\delta u}(z,1)$

$M_2(z) = \frac{\delta}{\delta u} u \frac{\delta F}{\delta u}\Big|_{u=1}$

$\mathcal{L} = (aabaa)^\star \cup ccc \quad \to F(a,b,c) = \frac{1}{1-a^4 b} + c^3$

$\to F(z) = \frac{1}{1-z^5} + z^3$

$\to P(z) = F(z/3) = \frac{1}{1-z^5/243} + z^3/27$

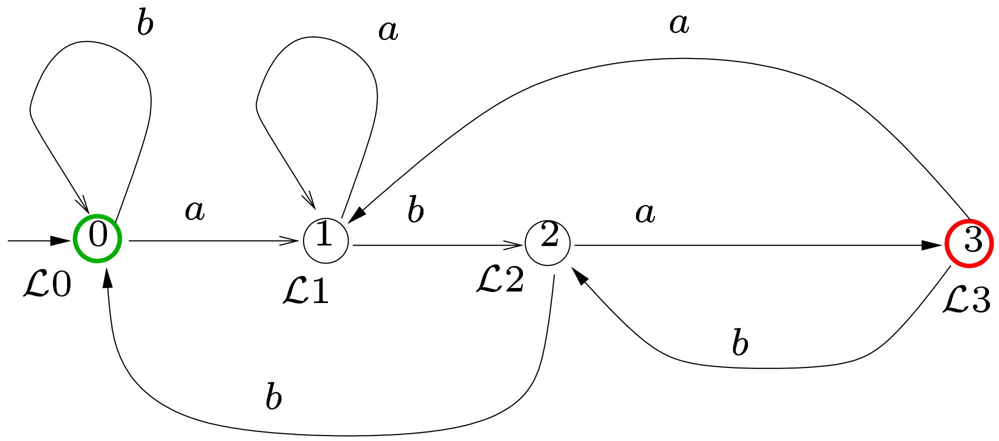
Approche classique

Exp. Rég. $\to$ ϵ-NFA $\to$ DFA $\to$ Func. Gén.

Idée-clé: utiliser automate reconnaissant $\Sigma^* R$ pour compter occ. de R

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

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



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

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

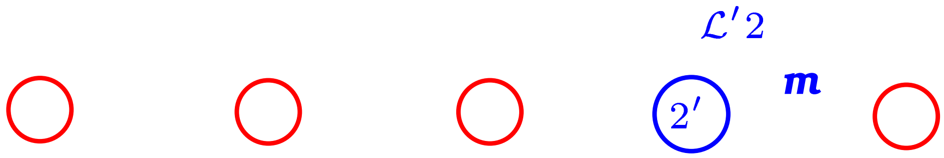
$$L_0 = aL_1 + bL_0, \qquad L_1 = bL_2 + aL_1,$$

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

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

$L(z) = \frac{z^3}{1-2z} = z^3 + 2z^4 + 4z^5 + 8z^6 \ldots$

marquage $(\Sigma' = \Sigma + m)$



Cas avec recouvrement $aaaabambamaaabamaa$

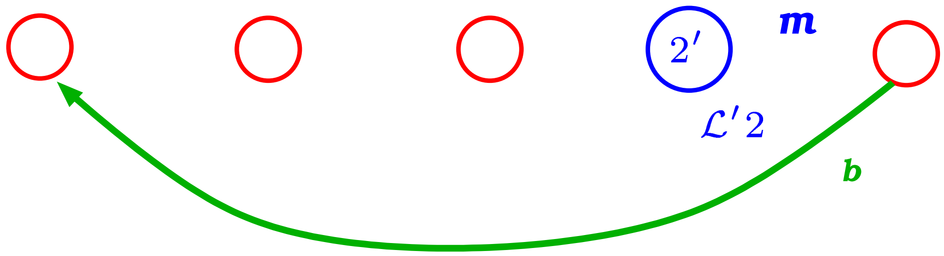
$$\mathcal{L}_0 = a\mathcal{L}_1 + b\mathcal{L}_0 + \epsilon, \qquad \ldots,$$

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

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

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

marquage $\quad (\Sigma' = \Sigma + m)$



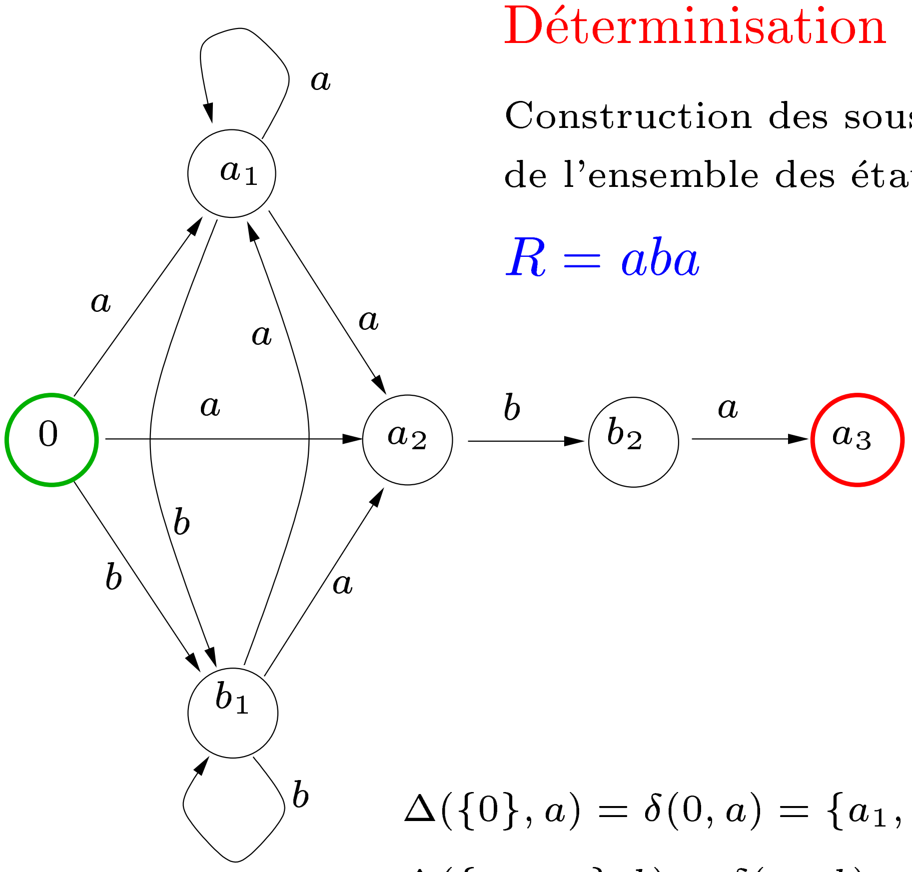
Cas sans recouvrement $aaaabambaaaabamaa$

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

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

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

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



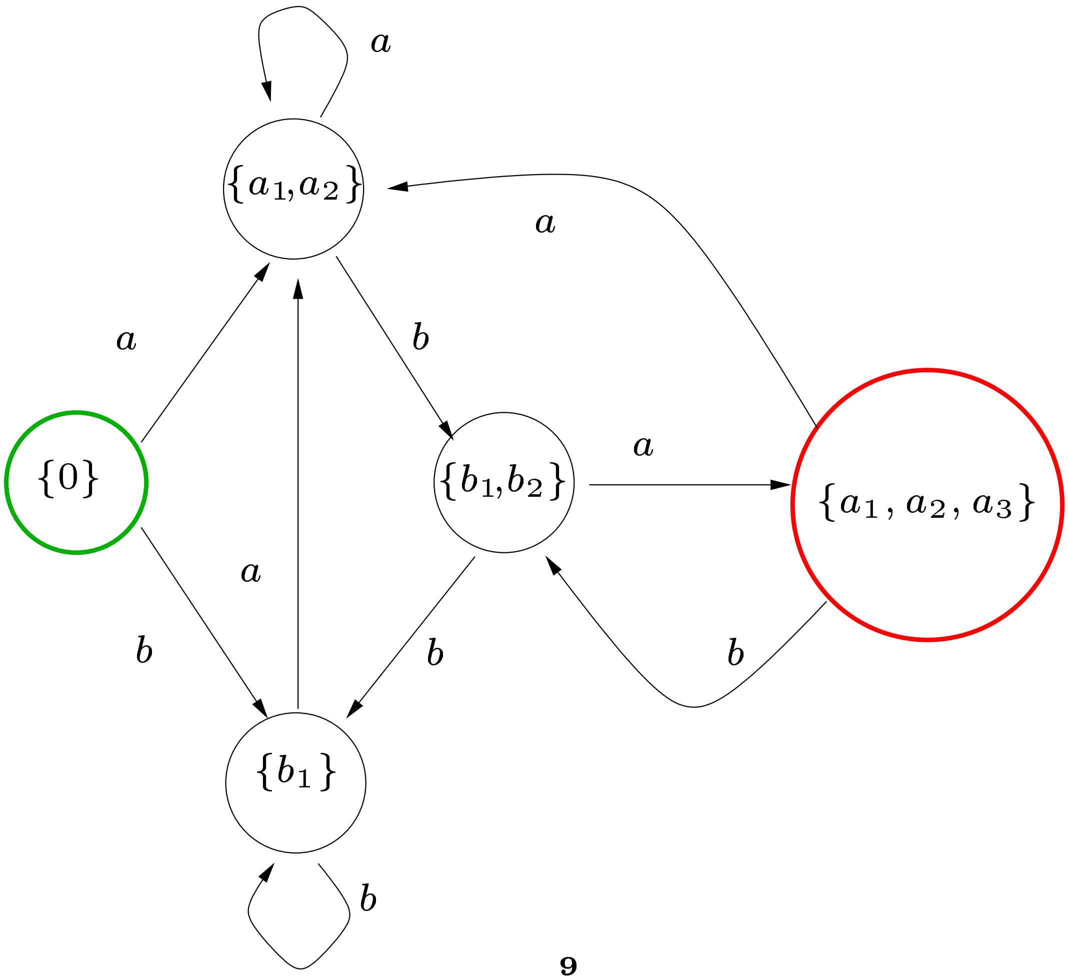
Déterminisation

Construction des sous-ensembles
de l'ensemble des états Q

$R = aba$

$$\Delta(\{0\}, a) = \delta(0, a) = \{a_1, a_2\}$$

$$\Delta(\{a_1, a_2\}, b) = \delta(a_1, b) \cup \delta(a_2, b) = \{b_1, b_2\}$$



Exp. Rég. vers NFA par Berry-Sethy

$E = (a + b)^* aba$

1) marquer les occurrences des lettres
$E' = (a_1 + b_1)^* a_2 b_2 a_3$

2) utiliser les constructeurs first, last, follow

$\mathrm{first}(R) = \{a_1, b_1, a_2\}$

$\mathrm{last}(R) = \{a_3\}$

$\mathrm{follow}(R, b_1) = \{a_1, b_1, a_2\}$

3) automate

lettres marquées $\rightarrow$ états,

suppression des indices $\rightarrow$ transitions

$\delta(b_1, a) = \{a_1, a_2\}, \quad \delta(b_1, b) = \{b_1\}$

Algorithme Berry-Sethy

déf. récursive de first, last, follow et nullable

$\text{nullable}(R) = \mathit{true}$ si $\epsilon \in$ langage de R

$\text{first}(R_1 R_2) =$

$$\begin{cases} \text{first}(R_1) \cup \text{first}(R_2) & \text{si} \quad \text{nullable}(R_1), \\ \text{first}(R_1) & \text{sinon} \end{cases}$$

$\text{follow}(R_1 R_2, x) =$

$$\begin{cases} \text{follow}(R_2, x) & \text{si} \quad x \in R_2, \\ \text{follow}(R_1, x) \cup \text{first}(R_2) & \text{si} \quad x \in \text{last}(R_1) \\ \text{follow}(R_1, x) & \text{sinon} \end{cases}$$

$\text{follow}(R^*, x) =$

$$\begin{cases} \text{follow}(R, x) \cup \text{first}(R) & \text{si} \quad x \in \text{last}(R), \\ \text{follow}(R, x) & \text{sinon} \end{cases}$$

Condition tech. $\Rightarrow$ complexité quadratique

Récupitalif de la chaîne algorithmique

Donnée: expression régulière R

1) Berry-Sethy pour $\Sigma^* R$ $\quad\rightarrow\quad$ NFA pour $\Sigma^* R$

2) Déterminisation $\quad\rightarrow\quad$ DFA pour $\Sigma^* R$

3) Marquage $\quad\rightarrow\quad$ DFA marqué pour $\Sigma^* R$

4) Chomsky-Schützenberger

$$\rightarrow \quad BGF(z,u) = \sum a_{n,k} u^k z^n$$

$a_{n,k} =$ nombre mots taille n avec k occ. de R

ou

$$\rightarrow \quad PGF(z,u) = \sum p_{n,k} u^k z^n$$

$p_{n,k} =$ proba. mot de taille n ait k occ. de R

Statistique du nombre d'occurrences

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

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

$$\mathcal{L}'_2 = m\mathcal{L}_3$$

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

$$L = \begin{pmatrix} L_0 \\ L_1 \\ \dots \\ L_n \end{pmatrix} = zT(u)L + \mathbf{1}$$

($\mathbf{1}$ vecteur colonne unité, $T(u)$ matrice $n \times n$)

$$L_0(z,u) = \frac{P(z,u)}{Q(z,u)} = \frac{P(z,u)}{(1 - z\lambda_1(u))(1 - z\lambda_2(u)) \dots (1 - z\lambda_n(u))}$$

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

$T(1)$ stochastique $\Rightarrow \quad L_0(z,1) = \frac{1}{1-z}, \quad \lambda_1(1) = 1$

$\lambda(u)$ valeur propre de $T(u) \quad \Rightarrow \quad \frac{1}{1-z\lambda(u)}$ vecteur propre

Perron-Frobenius, T(u) réelle, positive

$\Rightarrow \quad \lambda_1(u) \quad$ unique, réelle, positive, module max.

Statistique du nombre d'occurrences

Théorème

X_n nb d'occ. de R dans un texte aléatoire

$T(u)$ défini par le système $L = zT(u)L + 1$,

$\lambda(u)$ plus grande val. propre positive de $T(u)$.

Alors,

$E(X_n) = n\lambda'(1) + c_1 + O(A^n)$

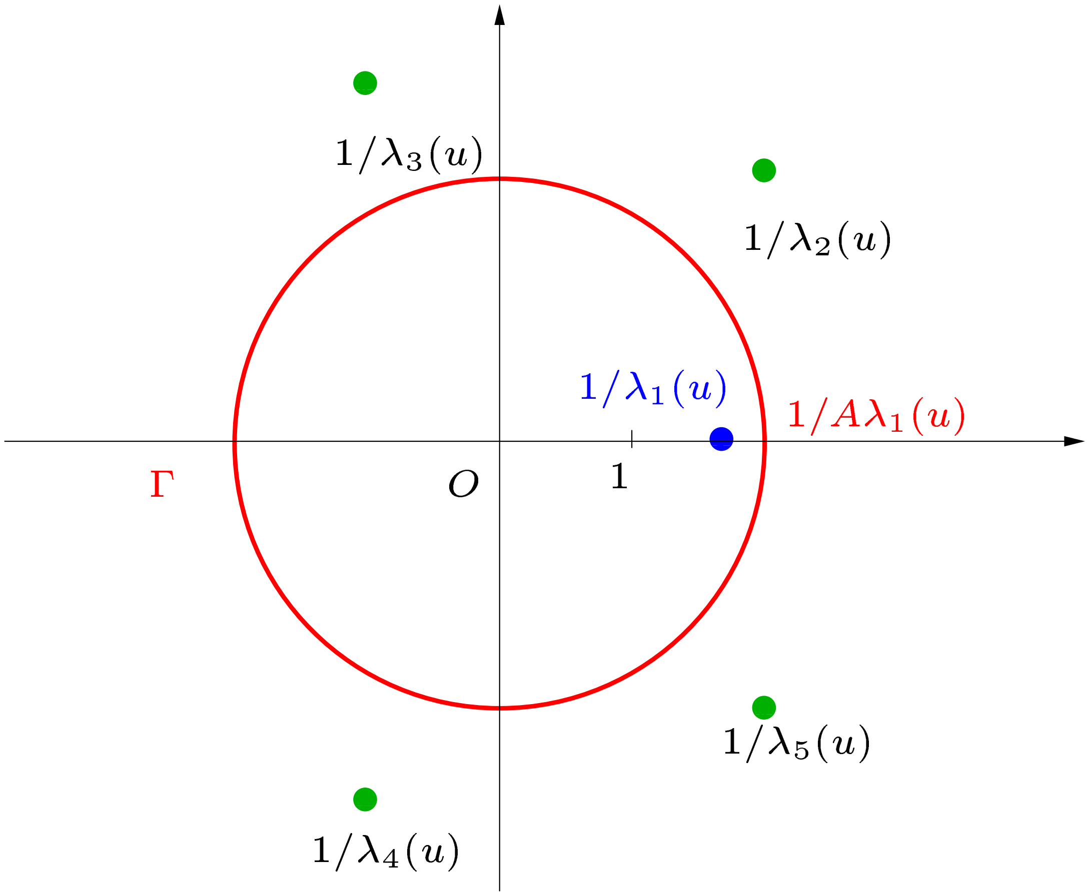
$Var(X_n) = n\sigma^2 + c_2 + O(A^n)$

$\sigma^2 = \lambda''(1) + \lambda'(1) - \lambda'(1)^2$

$A < 1, \quad c_1, c_2 \quad$ constantes

$\frac{X_n - n\lambda'(1)}{\sigma\sqrt{n}} \to \mathcal{N}(0,1)$

Condition nécessaire $\lambda''(1) + \lambda'(1) - \lambda'(1)^2 \neq 0$



$$F(z,u) = \sum p_{n,k} u^k z^n = \frac{c(u)}{1-\lambda_1(u)z} + g(z,u)$$

$\lambda_1(u)$ analytique dans un voisinage de $u = 1$

$c(1) \neq 1$, $g(z,u)$ analytique pour $|z| < 1/A\lambda_1(u)$ pour un $A < 1$

Formule intégrale de Cauchy

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

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

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

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

$p_n(u) = c(u)\lambda(u)^n(1 + O(A^n))$

$(\lambda(u) = \lambda_1(u))$

X_n une v.a. de fonct. de probabilité $p_n(u)$.

$X_n \approx Z + Y_1 + Y_2 + \cdots + Y_n$

Y_i v.a. indépendantes.

Cas voisin du théorème central limite.

Théorème des quasi-puissances de Hwang

$\rightarrow$ distribution asymptotique normale

Exploitation des résultats

Extraction rapide des coefficients

$$F(z,u) = \frac{P(z,u)}{Q(z,u)} \to \begin{cases} E(z) = \frac{U(z)}{V(z)}, \\ M_2(z) = \frac{H(z)}{K(z)} \end{cases}$$

$$E(X_n) = [z^n]E(z), \quad M_2(X_n) = [z^n]M_2(z)$$

But: extraire rapidement le nième coefficient de Taylor d'une fraction rationnelle

Méthode

$$E(z) = \sum e_n z^n = \frac{\sum_{0 \le i \le j} u_i z^i}{\sum_{0 \le i \le k} v_i z^i} \qquad (u_i, v_i \quad \text{connus})$$

$$\to \sum_{0 \le i \le k} v_i z^i \sum e_n z^n = \sum_{0 \le i \le j} u_i z^i$$

pour $m > j$, $\quad e_m v_0 + e_{m-1} v_1 + \cdots + e_{m-k} v_k = 0$

$$E_m = (e_m, e_{m-1}, \ldots, e_{m-k})$$

$$E_m^t = A E_{m-1}^t = A(e_{m-1}, e_{m-2}, \ldots, e_{m-k-1})^t$$

$$A = \begin{pmatrix} -v_1/v_0 & -v_2/v_0 & \ldots & -v_k/v_0 \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & & & \end{pmatrix} \qquad \text{matrice carrée}$$

$$E_m = A^{m-k} E_k$$

exponentiation binaire pour calculer A^{m-k}

Exemple - Pattern *aba*

$$\frac{\delta F}{\delta u}(z/2,1) = \frac{z^3/2}{4-8z+5z^2-2z^3+z^4}$$

on veut calculer $E(X_{4000})$

$$e_n = 2e_{n-1} - \frac{5}{4}e_{n-2} + \frac{1}{2}e_{n-3} - \frac{1}{4}e_{n-4}$$

$$E^t_{4000} = \begin{pmatrix} 2 & -5/4 & 1/2 & -1/4 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}^{3997} \begin{pmatrix} 1/8 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

3997(base 10) = 111110011101(base 2)   (12 bits)

12 multiplications de matrices

9 multiplications matrice par vecteur (nb. de bits à 1)

$E(X_{4000}) = 399.84$ (évaluation flottante)

calcul effectué en 0.7 sec.

Complexité $O(\log n)$ pour le calcul du $n$ième coefficient.

Démo Maple $R = aba$

```
> restart:read(`/home/nicodeme/soft/regexp/programs/RegExp.maple`):
> read(`/home/nicodeme/soft/regexp/programs/regexptomarkgram.maple`):
> with(combstruct): with(gfun):
> G:={R=Prod(a,b,a),a=Atom,b=Atom}:
> Auto_Mark:=regexptomarkgram(G,R,m);
```

$$\mathit{Auto\_Mark} := \{ w1 = \mathrm{Union}(\mathrm{E}, \mathrm{Union}(\mathrm{Prod}(a, w2), \mathrm{Prod}(b, R))), w3 = \mathrm{Union}(\mathrm{E}, \mathrm{Union}(\mathrm{Prod}(a, \mathrm{Prod}(m, w1)), \mathrm{Prod}(b, R))), w2 = \mathrm{Union}(\mathrm{E}, \mathrm{Union}(\mathrm{Prod}(a, w2), \mathrm{Prod}(b, w3))), m = \mathrm{E}, R = \mathrm{Union}(\mathrm{E}, \mathrm{Union}(\mathrm{Prod}(a, w2), \mathrm{Prod}(b, R))), b = \mathit{Atom}, a = \mathit{Atom} \}$$

```
> SysGF:=gfsolve(Auto_Mark,unlabelled,z,[[u,m]]);
```

$$\mathit{SysGF} := \{ \mathrm{w1}(z,u) = -\frac{z^2+1}{-1+2z-z^2+z^3u+z^3}, \mathrm{R}(z,u) = -\frac{z^2+1}{-1+2z-z^2+z^3u+z^3}, \mathrm{w3}(z,u) = -\frac{1-z+z^2+zu}{-1+2z-z^2+z^3u+z^3}, \mathrm{w2}(z,u) = -\frac{1+z^2u}{-1+2z-z^2+z^3u+z^3}, \mathrm{a}(z,u) = z, \mathrm{b}(z,u) = z, \mathrm{m}(z,u) = u \}$$

```
>
```

Démo Maple $R = aba$ (suite)

```
> F(z,u):=normal(subs(z=z/2,subs(SysGF,R(z,u
  ))));
  E(z):=normal(subs(u=1,diff(F(z,u),u)));
```

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

$$E(z) := \frac{1}{2}\,\frac{z^3}{(z-1)\,(-4+4\,z+z^3-z^2)}$$

```
> expand(denom(2*%));
```

$$-8\,z+5\,z^2+z^4-2\,z^3+4$$

```
> Erec:=diffeqtorec(E(z)-y(z),y(z),u(n));Epr
  oc:=rectoproc(Erec,u(n)):
```

$$Erec := \{-2\,u(n)-8\,u(n+2)+n,\ u(3)=\frac{1}{8},\ u(0)=0,$$

$$u(2)=0,\ u(1)=0\}$$

```
> E[4000]:=evalf(Eproc(4000));
  time(evalf(Eproc(4000)));
```

$$E_{4000} := 399.8400000$$

$$.709$$

```
> Mom2F(z):=normal(subs(u=1,diff(u*diff(F(z,
  u),u),u))):
> Mom2rec:=diffeqtorec(Mom2F(z)-y(z),y(z),u(
  n)):Mom2proc:=rectoproc(Mom2rec,u(n)):
> Mom2[4000]:=evalf(Mom2proc(4000));
  Ecart_Type[4000]:=sqrt(Mom2[4000]-E[4000]^
  2);
```

$$Mom2_{4000} := 160103.9616$$

$$Ecart\_Type_{4000} := 15.22944516$$

Démo Maple $R = aba$ (suite)

Asymptotique rapide des coefficients

$L = zT(u)L + 1$

$L^t = (F(z,u), L_2(z,u), \ldots, L_n(u))$

$F(z,u) = \frac{P(z,u)}{Q(z,u)(1-\mu(u)z)}$ et $F(z,1) = \frac{1}{1-z}$

1 est la plus grande valeur propre en $u = 1$

dérivation formelle en u puis en z

$$E(z) = \sum e_n z^n = \frac{\partial F(z,1)}{\partial u} = \frac{H(z)}{D(z)(1-z)^2}$$
$$= \frac{H(1)-(1-z)H'(1)+O(1-z)^2}{(D(1)-(1-z)D'(1)+O(1-z)^2)(1-z)^2}$$
$$= \frac{H(1)}{D(1)(1-z)^2} + \frac{H(1)D'(1)-H'(1)D(1)}{D(1)^2(1-z)} + O(1)$$

$$e_n = \frac{H(1)}{D(1)}(n+1) + \frac{H(1)D'(1)-H'(1)D(1)}{D(1)^2} + o(1)$$

calcul en temps $O(1)$

calcul similaire pour le second moment.

Applications biologiques

Base étudiée: Prosite, environ 1300 Motifs

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

Base de test: ProDom34-2

53500 familles de protéines

On prend 1 séquence moyenne (consensus) par famille.

On concaténe ces séquences

$\rightarrow$ base de consensus,

1 séquence de longueur 6756756

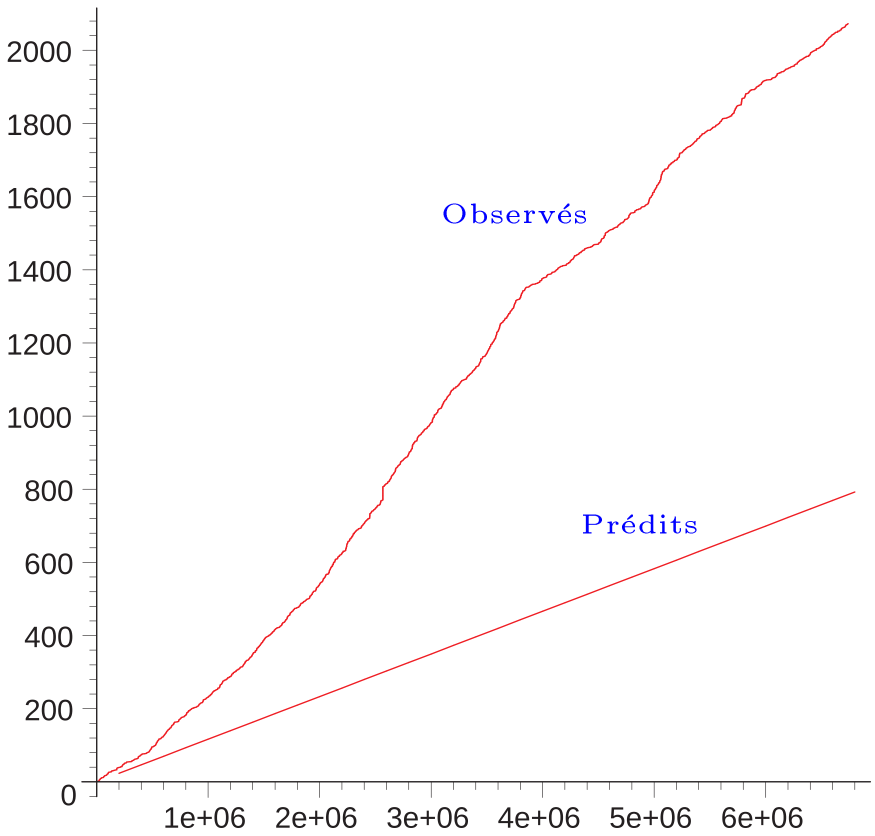
base de consensus $\rightarrow$ distrib. des acides aminés

On compare

1) les occurrences des Motifs observés dans la base de consensus ProDom

2) la statistique (espérance, écart-type) du nombre d'occurrences prédits dans une séquence aléatoire de même distribution de lettres

Nb. occurrences

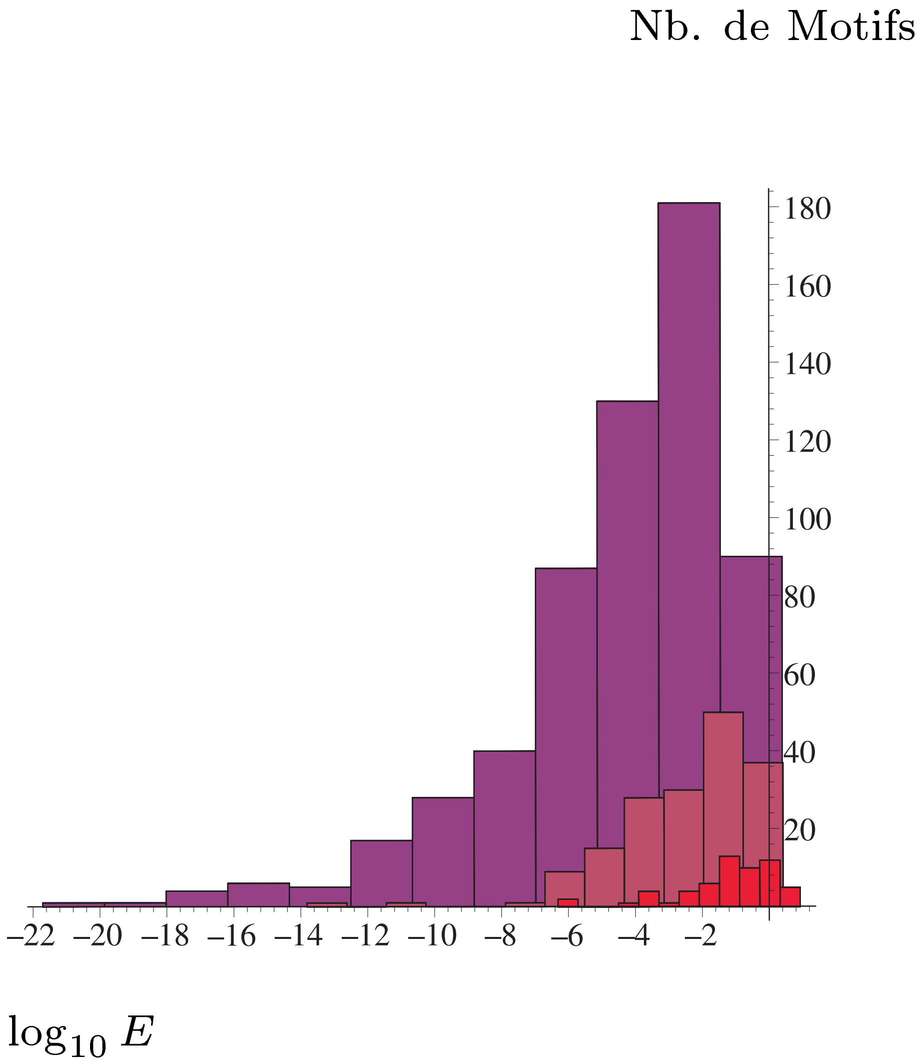


pos. dans ProDom

Motif PS00013

{DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C.

membrane lipoprotein lipid attachment site



Espérance des Motifs observés 1 fois, 2 fois (violet), 3 fois

E	O	Z
S-G-x-G		
2149	3302	25
[RK](2)-x-[ST]		
11209	13575	22
DERK(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C		
788	2073	46
[KR]-x(1,3)-[RKSAQ]-N-x(2)-[SAQ](2)-x-[RKTAENQ]-x-R-x-[RK]		
2.75	37	20
C-CPWHF-CPWR-C-H-CFYW		
25	173	29
[ST]-x-[RK]		
99171	90192	-30

Motifs avec des Z-scores exceptionnels

$$(Z = \frac{O-E}{\sigma})$$

Extensions envisagées

1) Motifs

- Traitement du cas Markovien pour l'ADN

2) Expressions régulières, automates

- Mise en forme star-normal d'une expression régulière
- Minimisation d'automate en $n \log n$

3) Extension vers des grammaires non-contextuelles?

Référence

Motifs Statistics, RR n^o 3606 - INRIA

`http://algo.inria.fr/papers`