

Compression sans perte par anti-dictionnaire

Julien Fayolle¹

¹LRI, Univ. Paris-Sud
julien.fayolle@lri.fr
page ouèbe : www.lri.fr/~fayolle



LIPN, 26 février 2008.

- ① Introduction à DCA ;
- ② Modèles et objectif ;
- ③ Motifs longs et courts ;
- ④ Motifs intermédiaires.

Exemple de compression à l'aide d'anti-dictionnaire

$$T = 011001001$$

Ensemble de tous les mots qui **n'apparaissent pas** dans le texte T .

Sous-ensemble des mots **minimaux** n'apparaissant pas.

Pour la taille 3 :

- 101 car 01**1**001001
- 111 car 01**1**001001 et
- 000 car 011**0**01001.

L'anti-dictionnaire $\mathcal{AD} = \{101, 111, 000, 0011, 010010\}$

$$T = 011001001$$

$$C = 011$$

[Crochemore, Mignosi, Restivo et Salemi'99] **DCA**

[Ota et Morita'04] (Électrocardiogramme)

Mot minimal interdit

Un MMI pour un texte T est un mot w de taille k qui vérifie

- w n'apparaît pas dans le texte T ;
- le **préfixe** de w de taille $k - 1$ apparaît dans T ;
- le **suffixe** de w de taille $k - 1$ apparaît dans T .

Pour un motif $w = w_1 \cdots w_k$, on note $u = w_2 \cdots w_{k-1}$.



- ★ Construction de l'anti-dictionnaire ;
(Build-AD, [Crochemore *et alii*'00] et ST2AD, [Morita et Ota'05])
- ★ Compression du texte ;
- ★ Envoi de (AD, C, L) ;
- ★ Décompression.

```
 $\mathcal{C} := \epsilon;$   
pour  $i := 0$  à  $n$  faire  
  si pour tout suffixe  $v$  de  $T[0 \dots i]$ ,  $v0 \notin \mathcal{AD}$  et  $v1 \notin \mathcal{AD}$   
    alors  $\mathcal{C} := \mathcal{C}.T[i + 1];$   
  fin si  
fin pour
```

$T=011001001$

$\mathcal{C}=011.$

```
 $T := \epsilon; j := 0$   
tant que  $|T| < n$  faire  
  si pour tout suffixe  $v$  de  $T$ ,  $v0 \notin \mathcal{AD}$  et  $v1 \notin \mathcal{AD}$   
    alors  $T := T.T[j++]$ ;  
  sinon  
    si  $v0 \in \mathcal{AD}$   
       $T := T.1$ ;  
    sinon  
       $T := T.0$ ;  
    fin si  
  fin si  
fin tant que
```

$\mathcal{AD} = \{101, 111, 000, 0011, 010010\}$ $L = 9$ $C = 011$

$T = 011001001$

Modèle de génération des textes

Les textes sont engendrés par une source

- binaire $\mathcal{A} = \{0, 1\}$;
- sans mémoire ;
- biaisée (p, q) .

Définition

La *probabilité d'occurrence* p_w du motif w est la probabilité que la source émette un mot qui commence par w .

$$w = 01010010111 \quad \text{et} \quad p_w = p^5 q^6$$

Entropie : $h := -p \log p - q \log q$.

L'entropie renseigne sur la quantité d'incertitude liée à la source.

La **taille** de l'anti-dictionnaire : $\mathcal{S} = \#\{w : w \in \text{MMI}\}$.

$$\begin{aligned} T &= 011001001 \\ \mathcal{AD} &= \{101, 111, 000, 0011, 010010\} \\ \mathcal{S}(T) &= 5. \end{aligned}$$

$$\mathbb{E}_n(\mathcal{S}) = \sum_{w \in \mathcal{A}^*} \mathbb{E}_n(\llbracket w \in \text{MMI} \rrbracket) = \sum_{w \in \mathcal{A}^*} \mathbb{P}_n(w \in \text{MMI}).$$

Taille de l'anti-dictionnaire – sans mémoire [F.'06]

$$\mathbb{E}_n(\mathcal{S}) = K \frac{n}{h} + \frac{n}{h} \epsilon(n) + o(n),$$

avec une constante K explicite.

Trichotomie des motifs selon leur longueur

$$\sum_{w \in \mathcal{A}^*} \mathbb{P}_n(w \in \text{MMI})$$

Motifs courts : $k \leq \frac{5}{6} \frac{1}{\log(1/q)} \log n$

$$\mathbb{P}_n(N_w = 0) \rightsquigarrow o(1)$$

Motifs longs : $k \geq \frac{3}{2} \frac{1}{\log(1/p)} \log n$

$$\mathbb{P}_n(N_u \geq 2) \rightsquigarrow O(n^{0.25})$$

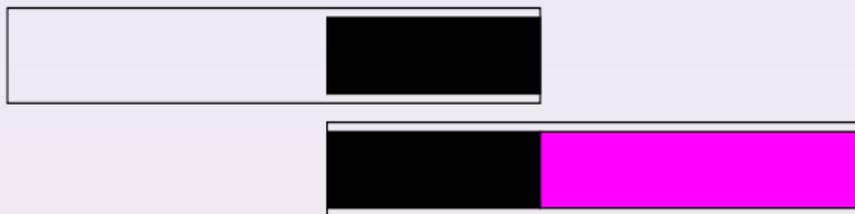
Motifs intermédiaires : (longueurs \mathfrak{J}_n)

- Motivation du modèle simplifié ;
- Modèle simplifié ;
- Calcul sous ce modèle ;
- Retour au modèle “réel” .

w	=	01011010 ✓	$i = 0$
		01011010	1
		01011010	2
		01011010	3
		01011010	4
		01011010 ✓	5
		01011010	6
		01011010 ✓	7

Polynôme d'auto-corrélation :

$$\mathfrak{C}_w(z) = 1 + z^5 + z^7.$$



Polynôme d'auto-corrélation probabilisé

Pour un motif w de taille k ,

$$c_w(z) := \sum_{i=0}^{k-1} \mathbb{I}[w_1^{k-i} = w_{i+1}^k] p_{w_{k-i+1}} z^i. \quad (1)$$

Sous un modèle sans mémoire (p, q) :

$$\sum_{w \in \mathcal{A}^k} c_w(1) = 2^k + k - 1.$$

Modèle avec deux propriétés :

P_1 – occurrences de u distantes

$$\mathbb{P}_n(w \in \text{MMI}) \simeq \mathbb{P}_n(\{w \in \text{MMI}\} \cap \mathfrak{D}_u)$$

\mathfrak{D}_u = ensemble des textes avec au moins deux occurrences de u et pour lesquels ces occurrences sont à distance au moins 2.

P_2 – approximation de Poisson

$$\forall j \in \mathbb{N}, \forall u \in \mathcal{A}^*, \quad \mathbb{P}_n(\tilde{N}_u = j) \simeq \frac{(np_u)^j}{j!} \exp(-np_u).$$

\tilde{N}_u vaut le nombre d'occurrences de u dans les textes où les occurrences de u sont à distance au moins 2 et 0 sinon.

Calcul sous le modèle simplifié

$$\sum_{k \in \mathcal{I}_n} \sum_{w \in \mathcal{A}^k} \mathbb{P}_n(w \in \text{MMI}) \stackrel{P_1}{=} \sum_{k \in \mathcal{I}_n} \sum_{w \in \mathcal{A}^k} \sum_{j \geq 2} \mathbb{P}_n(\{w \in \text{MMI}\} | \{\tilde{N}_u = j\})$$

$$\stackrel{P_2}{=} \sum_{k \in \mathcal{I}_n} \sum_{\alpha, \beta \in \mathcal{A}} \sum_{w \in \mathcal{A}^k} \sum_{j \geq 2} \mathbb{P}_n(\alpha u \beta \in \text{MMI} | \{\tilde{N}_u = j\}) \frac{(np_u)^j}{j!} \exp(-np_u).$$



$$\sum_{j \geq 2} \frac{C_j^{\alpha, \beta}}{j!} z^j = 1. (\exp(zp_\alpha p_\beta) - 1) (\exp(zp_{\bar{\alpha}} p_\beta) - 1) \exp(zp_{\bar{\alpha}} p_\beta)$$

$$\sum_{k \geq 0} \sum_{i=0}^k \binom{k}{i} \left[(\exp(zp q p^i q^{k-i}) - 1)^2 \exp(zq^2 p^i q^{k-i}) \exp(-z p^i q^{k-i}) \right]$$

Définition

$$f^*(s) := \mathcal{M}(f; s) := \int_0^{\infty} f(x)x^{s-1} dx.$$

développement asymptotique de f	information complexe sur f^*
terme en x^{-c}	localisation d'un pôle (en c)
facteur en $\log^{k-1} x$	ordre k du pôle
coefficient du terme en x^{-c}	résidu en c

[de Bruijn, Knuth et Rice'72]

[Flajolet, Gourdon et Dumas'95]

Kicécuilà →

Hjalmar Mellin, suomalainen
matemaatikko (1854–1933)



$$\sum_{k \geq 0} \sum_{i=0}^k \binom{k}{i} (\exp(npqp^i q^{k-i}) - 1)^2 \exp(nq^2 p^i q^{k-i}) (\exp(-np^i q^{k-i}))$$

$$f(z) := \exp(z(-1+2pq+q^2)) - 2 \exp(z(-1+pq+q^2)) + \exp(z(-1+q^2)).$$

Mellin : (0, 1), (1, 0), (0, 0) et (1, 1).

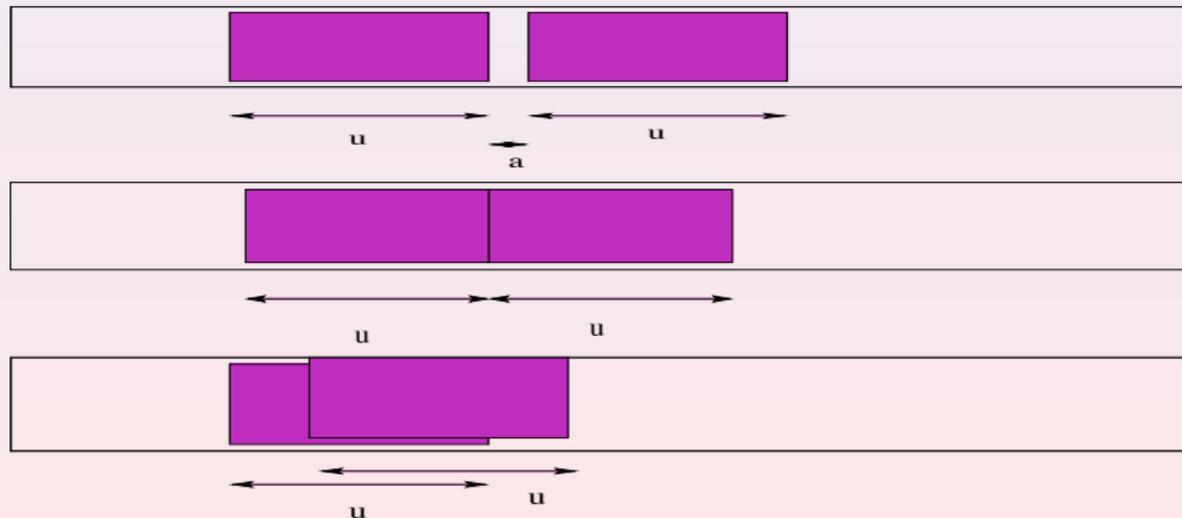
$$\frac{K}{h} n + o(n)$$

où $K := 2h + (1-p^2) \log(1-p^2) + (1-q^2) \log(1-q^2) + 2(1-pq) \log(1-pq)$

Validation du modèle : P_1

Le résultat sous modèle approché et le “vrai” résultat diffèrent d'un terme sous-linéaire.

$$\Delta_1(n) := \sum_{k \in \mathcal{J}_n} \sum_{w \in \mathcal{A}^k} \mathbb{P}_n(w \in \text{MMI}) - \mathbb{P}_n(w \in \text{MMI} \cap \mathfrak{M}_u).$$



$$\Delta_1(n) = O\left(n^{1 - \frac{5 \log p}{12 \log q}}\right).$$

On évalue la distance entre \tilde{N}_u et X_u ($X_u \rightsquigarrow \mathcal{P}(np_u)$).

$$\Delta_2(n) = \sum_{k \in \mathcal{I}_n} \sum_{w \in \mathcal{A}^k} \sum_{j \geq 2} \mathbb{P}_n(w \in \text{MMI} \mid \tilde{N}_u = j) \left[\mathbb{P}_n(\tilde{N}_u = j) - \mathbb{P}_n(X_u = j) \right]$$

- ① [Barbour, Holst et Janson'92] *via* Stein-Chen :

$$\sum_{j \geq 0} \left| \mathbb{P}_n(N_u = j) - \frac{(np_u)^j}{j!} \exp(-np_u) \right| \leq 4(c_u(1)-1) + 2(2|u|-1)p_u$$

N_u est le nombre d'occurrences du motif u

②

$$\mathbb{P}_n(N_u = j) - \mathbb{P}(\tilde{N}_u = j).$$

Textes avec exactement j occurrences de u et où au moins deux occurrences sont à au plus une lettre d'écart (cf. P_1).

$$\Delta_2(n) = O\left(n^{1 - \frac{5}{12} \frac{\log p}{\log q}}\right)$$

Théorème [Fayolle, Flajolet, Morita et Ota'06]

Le comportement asymptotique de la taille moyenne de l'anti-dictionnaire pour un texte de taille n produit par une source **sans mémoire** (p, q) s'écrit

$$\mathbb{E}_n(\mathcal{S}) = K \frac{n}{h} + \frac{n}{h} \epsilon(n) + o(n),$$

où la constante K vaut

$$2h + (1 - p^2) \log(1 - p^2) + (1 - q^2) \log(1 - q^2) \\ + 2(1 - pq) \log(1 - pq),$$

l'entropie $h = -p \log p - q \log q$ et $\epsilon(n)$ est une fonction oscillant de très faible amplitude autour de zéro.

Autre résultat : [Morita et Ota'04]

- ★ Étendre les résultats du modèle sans mémoire à un modèle de source dynamique [Vallée] ;
- ★ Obtenir l'asymptotique de la variance de la taille et des moments d'ordre supérieurs;
- ★ Obtenir l'asymptotique des moments de la longueur de cheminement;
- ★ Montrer que la compression par anti-dictionnaire atteint le taux entropique de compression [Crochemore *et alii*].