42 dans une suite de 42 caractères

Arthur Jacquin

8 juin 2021

1 Énoncé du problème

On considère l'alphabet $\Sigma = \{0, 1\}$. Parmi les mots de longueur 42, quelle est la part de ceux contenant le facteur `101010` ?

2 Principe

Soient $I = [\![0; 2^6[\![$ et $N = [\![6; 42]\!]$. On définit la position d'un facteur dans un mot comme la position de son dernier caractère.

Comptons le nombre C de cas favorables, correspondant au nombre de mots de longueur 42 contenant au moins une fois le facteur `101010`. Pour tout $n \in N$, on pose c_n le nombre de mots de longueur 42 contenant le facteur `101010` pour la première fois en position n. On a bien

$$C = \sum_{k=6}^{42} c_k$$

Le problème est donc ramené au calcul des c_k. Pour les déterminer, on considère la négation de "contenir au moins une fois" qui est "ne pas contenir". On part ainsi de l'ensemble des mots de longueur 6 ne contenant pas le facteur `101010`, puis on construit progressivement les ensembles des mots (qui ne contiennent toujours pas `101010`) de longueur supérieures. Pour cela, il suffit de concatener à chaque mot de l'ensemble précédent une fois avec `1` et une fois avec `0` puis d'éliminer les mots contenant `101010`.

On remarque qu'au cours de la construction, seuls les 6 derniers caractères des mots nous intéressent. En effet, les autres facteurs de longueur 6 sont, par construction, différents du facteur souhaité. Au lieu de considérer l'ensemble des mots, on peut donc se contenter de les rassembler et de les compter selon leurs 6 derniers caractères. La complexité spatiale, précedemment exponentielle, devient linéaire.

3 Résolution

On note :

- bin l'application associant à tout entier de I son écriture binaire à 6 caractères, et dec sa bijection réciproque.
- $\mathcal{A}$ l'ensemble des mots formés par l'alphabet Σ.
- $|.|$ l'application associant à tout mot de $\mathcal{A}$ sa longueur.
- suff_k l'application associant à tout mot de $\mathcal{A}$ son suffixe de longueur k.
- fact_m l'application associant à tout mot m' de $\mathcal{A}$ 1 si m est facteur de m' et 0 sinon.

Pour tout $(n, i) \in N \times I$, on définit

$$T_n^i = \{m \in \mathcal{A} \mid |m| = n,\ \mathrm{suff}_6(m) = \mathrm{bin}(i),\ \mathrm{fact}_{\mathtt{101010}}(m) = 0\}$$

Pour tout $i \in I$, on définit la suite u_i telle que
— $u_i = 0$ si $i = 42$
— $u_i(6) = 1, \quad \forall n \geq 6,\ u_i(n+1) = u_{\lfloor \frac{i}{2} \rfloor}(n) + u_{2^5 + \lfloor \frac{i}{2} \rfloor}(n)$ sinon.

Montrons par récurrence que $H_n :=$ " $\forall i \in I, \quad u_i(n) = \mathrm{Card}\, T_n^i$ " est vraie pour tout $n \in N$. H_6 est trivialement vraie. Soient $n \in N \backslash \{6\}$ tel que H_{n-1} soit vraie et $i \in I$. Si $i = 42$, on a bien le résultat souhaité car $u_i = 0$. Sinon, on note `abcdef` l'écriture binaire de i. Alors tout mot m de T_n^i est tel que $\mathrm{suff}_7(m)$ est de la forme `pabcdef`, autrement dit m est concaténation d'un mot de $T_{n-1}^{\mathrm{dec}(\mathtt{0abcde})}$ ou de $T_{n-1}^{\mathrm{dec}(\mathtt{1abcde})}$ et de `f`. Or $\mathrm{dec}(\mathtt{0abcde}) = \lfloor \frac{i}{2} \rfloor$ et $\mathrm{dec}(\mathtt{1abcde}) = 2^5 + \lfloor \frac{i}{2} \rfloor$. Réciproquement, toute combinaison fonctionne et les mots créés sont deux à deux distincts, d'où le résultat souhaité.

Soit alors la suite v telle que $v(6) = 1$ et $\forall n \geq 6,\ v(n+1) = u_{\lfloor \frac{42}{2} \rfloor}(n) + u_{2^5 + \lfloor \frac{42}{2} \rfloor}(n)$. On peut de même montrer que pour tout $n \in N$, $v(n)$ correspond au nombre de mots de longueur n contenant le motif `101010` une unique fois, en position n.

Pour tout $n \in N$, considérer c_n revient à considérer le nombre de mots permettant d'obtenir `101010` une unique fois en position n ($v(n)$) ainsi que l'ensemble des mots de longueur $42 - n$ à concatener (2^{42-n}), mots sur lesquels il n'existe pas de contraintes. On a alors :

$$c_n = 2^{42-n} v(n)$$

En sommant les termes :

$$C = \sum_{k=6}^{42} c_k = 2^{42} \sum_{k=6}^{42} \frac{v(k)}{2^k}$$

En notant P la probabilité d'obtention d'un mot contenant au moins une fois le facteur `101010` lors du tirage d'un mot aléatoire de longueur 42, on obtiens par équiprobabilité :

$$P = \sum_{k=6}^{42} \frac{v(k)}{2^k}$$

4 Mise en œuvre algorithmique

Pour un n fixé, on peut stocker les valeurs des $u_i(n)$ dans une liste indexée par I.

```python
# Initialisation
total = 0
act, new = [1]*64, [0]*64

# Traitement
for k in range(6, 42 + 1):
    total += act[42] * 2**(42-k)
    act[42] = 0
    for i in range(32):
        new[2*i] = new[2*i+1] = act[i] + act[32 + i]
    act, new = new, [0]*64

# Resultat
print(total)
```

Listing 1 – Dénombrement des cas favorables

5 Résultats

L'exécution de cet algorithme affiche `1660901974812`, d'où $P \simeq 0.37765$.

6 Version matricielle

Soient $I = [\![0, 2^6 [\![$, $N = [\![6, 42]\!]$, les matrices $A \in \mathcal{M}_{2^6}$, $P \in \mathcal{M}_{1,2^6}$ et la suite U de matrices dans $\mathcal{M}_{2^6,1}$ tels que pour tout $(i,j) \in I^2$ et $n \geq 6$:

$$\begin{aligned} a_{i+1,j+1} &= \begin{cases} 1 \text{ si } i \neq 42 \text{ et } j \equiv \lfloor \frac{i}{2} \rfloor \ [2^5] \\ 0 \text{ sinon} \end{cases} \\ p_{1,j+1} &= \begin{cases} 1 \text{ si } j \equiv \lfloor \frac{42}{2} \rfloor \ [2^5] \\ 0 \text{ sinon} \end{cases} \\ [U_6]_{i+1,1} &= 1 \\ U_{n+1} &= A \cdot U_n \end{aligned}$$

On montre par récurrence que pour tout $(n,i) \in N \times I$, $[U_n]_{i+1,1} = \operatorname{Card} T_n^i$. On a ainsi :

$$P = \sum_{k=6}^{42} \frac{[PU_k]_{1,1}}{2^k}$$

7 Version formule du crible (non terminée)

Soit $N = [\![6, 42]\!]$. On définit les événements suivants :

$$\begin{aligned} R &:= \text{"Le mot contient le facteur } \texttt{101010}\text{"} \\ \forall i \in N, \quad A_i &:= \text{"Le facteur } \texttt{101010} \text{ apparaît en position } i\text{"} \end{aligned}$$

On a alors :

$$P(R) = P\left(\bigcup_{i \in N} A_i\right) = \sum_{S \in \mathcal{P}(N)} (-1)^{\operatorname{Card} S + 1} \cdot P\left(\bigcap_{i \in S} A_i\right)$$

Il reste donc à déterminer le nombre de caractères fixés par une intersection de A_i. On note pour tout $t \in [\![1, 42]\!]$:

— l'événement $C_t :=$ "Le caractère en position t est fixé par $\bigcap_{i \in S} A_i$"

— la quantité d_t correspondant au nombre de positions pour lesquelles un facteur de longueur 6 affecte le caractère en position t.

On constate que $d_t = \min(t, 6, 43 - t)$:

t	1	2	3	4	5	6	7	8	$\cdots$	36	37	38	39	40	41	42
d_t	1	2	3	4	5	6	6	6	$\cdots$	6	6	5	4	3	2	1

En passant par le complémentaire et en comptant le nombre de cas favorables :

$$P(C_t) = 1 - P(\overline{C_t}) = 1 - \frac{\binom{42 - d_t}{\operatorname{Card} S}}{\binom{42}{\operatorname{Card} S}} = 1 - \frac{\binom{42 - \min(t,6,43-t)}{\operatorname{Card} S}}{\binom{42}{\operatorname{Card} S}}$$

8 Généralisation

A venir... Idées de paramètres à modifier :
- Motif
- Taille du motif
- Taille de l'alphabet