

Autour de deux suites d'entiers

Bruno Petazzoni*

23 février 2010

Résumé

Nous nous intéressons à un (double) problème de dénombrement, lié à la fabrication d'un certain type de puces à ADN. À chacun de ces problèmes est associée une suite d'entiers ; aucune des deux n'apparaît «telle quelle» dans la *On-Line Encyclopedia of Integer Sequences* ; en revanche, on trouve dans cette dernière deux suites, très proches de celles que allons découvrir.

Nous commençons par une explication très rapide ce qu'est une puce à ADN, et une description du processus de fabrication de ces puces.

1 Puces à ADN, masques

► Une *puce à ADN* d'ordre n se compose d'un substrat carré (d'environ 1 cm de côté), décomposé en 4^n zones carrées de même taille ; dans chacune de ces zones, on implante un grand nombre de molécules d'ADN monobrin identiques, de longueur n . Chacune de ces molécules est un mot sur l'alphabet $\{A, C, G, T\}$. Deux molécules implantées sur des zones différentes doivent différer par au moins une lettre. La figure 1 présente deux dispositions différentes pour une puce à ADN d'ordre 2.

AA	AC	CA	CC
AG	AT	CG	CT
GA	GC	TA	TC
GG	GT	TG	TT

AA	AC	CC	CA
AG	AT	CT	CG
GG	GT	TT	TG
GA	GC	TC	TA

Figure 1: implantation récursive (à gauche) et avec un code de Gray (à droite)

► La fabrication d'une telle puce requiert une succession de $4n$ étapes, que l'on peut représenter par le mot $m = (ACGT)^n$. Initialement, on dépose sur la puce des molécules dotées d'une base qui va adhérer au substrat, et d'un groupe protecteur photolabile. Lors de la k -ième étape, on pose un *masque* sur la puce et on l'expose aux ultra-violetts : ceci élimine les groupes protecteur des zones exposées ; on enlève le masque et on déverse sur la puce une solution contenant la base m_k , à laquelle est attaché un groupe protecteur. La figure 2 donne une représentation (très approximative).

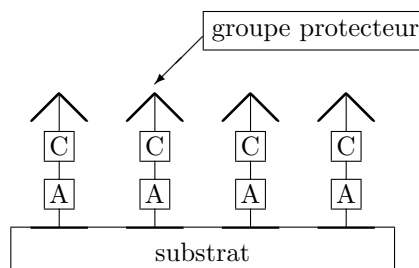


Figure 2: puce à ADN

*Lycée Marcelin-Berthelot — 94100 Saint-Maur-des-Fossés

► La figure 3 donne deux exemples de masques ; ce sont ceux associés à l'étape 5 (dépôt du deuxième A) pour chacune des deux dispositions présentées à la figure 1. Remarque : curieusement, les zones transparentes du masque sont représentées en noir !



Figure 3: masques pour l'étape A-2

► Avec l'implantation récursive, les puces sont définies par les deux dessins ci-dessous ; la notation $X F_n$ désigne une puce d'ordre n , dans laquelle on a ajouté une lettre X en tête de chaque mot.

$$F_1 = \begin{array}{|c|c|} \hline A & C \\ \hline G & T \\ \hline \end{array} \quad F_{n+1} = \begin{array}{|c|c|} \hline A F_n & C F_n \\ \hline G F_n & T F_n \\ \hline \end{array}$$

► À la frontière entre les parties cachées et les parties exposées, se produisent des phénomènes nuisibles (diffraction). On veut donc minimiser la longueur totale des bords des masques. Par exemple, les longueurs de bord des deux masques de la figure 3 sont respectivement 12 et 8 (on ne compte pas ce qui est «au bord» de la puce).

► Pour ce faire, nous disposerons les 4^n mots selon un *code de Gray bidimensionnel* noté G_n et défini par les équations suivantes :

$$G_1 = \begin{array}{|c|c|} \hline A & C \\ \hline G & T \\ \hline \end{array} \quad G_{n+1} = \begin{array}{|c|c|} \hline A G_n & C \overrightarrow{G_n} \\ \hline G G_n \downarrow & T G_n \downarrow \\ \hline \end{array}$$

La notation $\overrightarrow{G_n}$ (respectivement : $G_n \downarrow$) signifie que G_n a subi une symétrie autour de son axe vertical (respectivement : horizontal). La signification de $\overleftarrow{G_n} \downarrow$ est claire !

2 Calcul de la longueur de bord avec l'implantation récursive

Notons $L_f(n)$ la longueur de bord totale des $4n$ masques requis pour la barication d'une puce à ADN régulière d'ordre n , avec la méthode d'implantation récursive. Nous avons $L_f(n+1) = 4L_f(n) + 8 \cdot 2^n + 8n \cdot 2^n$. Le premier terme est la contribution des $4n$ premiers masques, avant de les raccorder ; le deuxième terme est la contribution des quatre derniers masques ; enfin, le troisième terme provient des raccords entre les $4n$ premiers masques : on montre avec une récurrence immédiate que, par exemple, le bord inférieur de la zone nord-ouest et le bord supérieur de la zone sud-ouest sont deux mots de longueur 2^n dont la distance de Hamming est égale à 2^n .

La résolution de cette relation de récurrence, avec la condition initiale $L_f(1) = 8$, nous donne $L_f(n) = 2 \cdot 4^{n+1} - (n+2)2^{n+2}$; notons que cette résolution peut être effectuée avec un logiciel de calcul formel (fonction `rsolve` de Maple, par exemple), mais aussi «à la main».

Les huit premiers termes de cette suite sont 8, 64, 352, 1664, 7296, 30720, 126464, 514048. Ces termes étant tous divisibles par 8, nous cherchons dans l'OEIS la suite 1, 8, 44, 208, 912, 3840, 15808, 64256 et nous tombons sur la suite A100575 ; plus précisément, $L_f(n) = 8a_{n+2}$, où a_n désigne le terme d'indice n de cette suite.

3 Calcul de la longueur de bord avec l'implantation «Gray»

Notons $L_g(n)$ la longueur de bord totale des $4n$ masques requis pour la barication d'une puce à ADN régulière d'ordre n , avec la méthode d'implantation «Gray». Nous avons $L_g(1) = 8$; et $L_g(n+1) = 4L_g(n) + 8 \cdot 2^n$; en effet, les masques des $4n$ premières étapes exploitent la propriété des codes de Gray, si bien que les «raccords» entre eux ne coûtent rien. Le terme $8 \cdot 2^n$ provient des 4 derniers masques.

La résolution de cette relation de récurrence est immédiate ; il vient $L_g(n) = 4^{n+1} - 2^{n+2}$. Les huit premiers termes de cette suite sont 8, 48, 224, 960, 3968, 16128, 65024, 261120 ; comme dans le cas précédent, nous divisons tous les termes par 8 ; la recherche dans l'OEIS nous donne la suite A006516. Plus précisément, nous avons $L_g(n) = 8c_{n+1}$ pour $n \geq 1$, où c_n désigne le terme d'indice n de la suite A006516.

Curieusement, c_n est le nombre de mots de longueur n sur un alphabet à quatre lettres (disons Σ , par exemple) dans lesquels la lettre A apparaît un nombre impair de fois.

c_n est aussi le nombre de droites passant par deux sommets d'un hypercube à 2^n sommets, ce qui est assez banal. Sauf qu'avec un code de Gray, on définit un cycle hamiltonien sur les sommets dudit hypercube ...

Nous constatons que $\frac{L_g(n)}{L_f(n)} \xrightarrow{n \rightarrow \infty} \frac{1}{2}$: asymptotiquement, la longueur de bord totale est divisée par 2 lorsque l'on applique la méthode «Gray».

4 Retour sur la suite de terme général a_n

Notons $f : x \mapsto \frac{1}{2 - e^x}$. Nous montrons que $f^{(n)}(x) = \frac{P_n(e^x)}{(2 - e^x)^{n+1}}$, où P_n est un polynôme de degré n ; puis, nous montrons que a_n est le coefficient de X^2 dans l'expression de $P_n(X)$.

Curieusement, la définition donnée par l'OEIS était doublement incorrecte : a_n était défini comme *la valeur absolue du coefficient de e^x dans l'expression de $P_n(e^x)$* . De plus, aucune forme close n'était donnée pour a_n .

L'assertion est vérifiée pour $n = 0$, avec $P_0 = 1$; elle l'est aussi pour $n = 1$, avec $P_1 = X$ puisque $f'(x) = \frac{e^x}{(2 - e^x)^2}$.

Supposons l'assertion acquise au rang n . Alors :

$$\begin{aligned} f^{(n+1)}(x) &= \frac{d}{dx}(f^{(n)}(x)) = \frac{d}{dx}\left(\frac{P_n(e^x)}{(2 - e^x)^{n+1}}\right) = \frac{(2 - e^x)^{n+1}e^x P_n'(e^x) + (n+1)e^x(2 - e^x)^n P_n(e^x)}{(2 - e^x)^{2n+2}} \\ &= \frac{(2 - e^x)P_n'(e^x) + (n+1)e^x P_n(e^x)}{(2 - e^x)^{n+2}} \end{aligned}$$

Notons $P_{n+1} = (2 - X)XP_n' + (n+1)XP_n$: nous avons $f^{(n+1)}(x) = \frac{P_{n+1}(e^x)}{(2 - e^x)^{n+2}}$, où P_{n+1} est un polynôme.

Nous avons déjà $\deg(P_0) = 0$ et $\deg(P_1) = 1$. Soit $n \geq 1$; supposons que P_n est de degré n , alors $(2 - X)XP_n'$ et $(n+1)XP_n$ sont de degré $n+1$; donc P_{n+1} est de degré $n+1$ au plus. Notons λ le coefficient dominant de P_n ; alors les coefficients dominants de $(2 - X)XP_n'$ et $(n+1)XP_n$ sont respectivement $-\lambda$ et $(n+1)\lambda$; donc le coefficient de X^{n+1} dans le polynôme P_{n+1} est λ , réputé non nul. Conclusion : P_{n+1} est de degré $n+1$. Par récurrence, $\deg(P_n) = n$ pour tout $n \in \mathbb{N}$.

Notons au passage que tous les P_n sont unitaires.

Un calcul (manuel ou avec un logiciel de calcul formel) donne $P_2 = X^2 + 2X$ et $P_3 = X^3 + 8X^2 + 4X$.

La formule exprimant P_{n+1} en fonction de P_n et P_n' montre que P_{n+1} est divisible par X ; ceci revient à dire que son terme constant est nul.

Conjecture : pour $n \geq 1$, définissons Q_n par $P_n = XQ_n$; alors Q_n est scindé sur \mathbb{R} , à racines simples, strictement négatives, séparant celles de Q_{n+1}/X .

Notons b_n le coefficient de X dans le polynôme P_n ; ainsi, $P_n(x) = b_n X + a_n X^2 + X^3 Q_n$, où Q_n est un polynôme de degré $n-3$ si $n \geq 3$, de degré nul sinon. Alors :

$$\begin{aligned} P_{n+1} &= (2 - X)XP_n' + (n+1)XP_n \\ &= (2 - X)X(b_n + 2a_n X + 3X^2 Q_n + X^3 Q_n') + (n+1)X(b_n X + a_n X^2 + X^3 Q_n) \end{aligned}$$

Nous en déduisons $b_{n+1} = 2b_n$, et $a_{n+1} = 4a_n + nb_n$. Comme $b_1 = 1$, la première relation nous donne

$\boxed{b_n = 2^{n-1} \text{ pour } n \geq 1}$. Nous en déduisons $\frac{a_{k+1}}{4^{k+1}} = \frac{a_k}{4^k} + \frac{k}{2^{k+3}}$. Un télescopage nous donne :

$$\frac{a_n}{4^n} = \frac{a_1}{4} + \sum_{1 \leq k < n} \frac{k}{2^{k+3}} = \frac{1}{4} - \frac{n+1}{2^{n+2}}$$

Nous en déduisons $\boxed{a_n = 4^{n-1} - (n+1)2^{n-2}}$.

5 Réflexions finales

Que la même suite (c_n) apparaisse dans un calcul lié à l'ADN et dans un autre calcul faisant intervenir un alphabet à quatre lettres est vraisemblablement fortuit.

Le lien avec la suite (b_n) est curieux, de prime abord; toutefois, la relation définissant cette suite est suffisamment «simple» pour que, parmi toutes les suites construites sur le même principe, il en existe une qui convient ...

6 Références bibliographiques

La référence sur l'algorithmique du génôme est le livre de Pevzner *Computational Molecular Biology: An Algorithmic Approach*.

L'article *Combinatorial algorithms for design of DNA arrays* présente d'autres problèmes liés à la conception des puces à ADN, et en particulier celui le problème de la décomposition d'un masque en un nombre minimal de rectangles.