Luminy, 8 mai 2007

Colloque « Algorithmique et Programmation »

Autour des codes de Gray

Bruno Petazzoni

Lycée Marcelin-Berthelot (94100)

bruno@enix.org

Préambule

quoi de commun entre :

1. les tours de Hanoï
2. un ancien boxeur reconverti en barman
3. un casse-tête chinois
4. des puces à ADN
5. les dérivées successives de $x < \ln(2) \rightarrow \frac{1}{2-e^x}$

sur une même base, cinq textes :

1. un DS d'informatique en 2ème année (cf. site web)
2. un TP d'informatique en 1ère année
3. peut-être un article dans la RMS
4. le prochain DS de maths
5. l'exposé qui va suivre

Une page de publicité

trois sujets sur `bruno.maitresdumonde.com/optinfo/Spe-MP/dmds2006` :

- mots directeurs d'un automate fini
- longueur discriminantes de deux automates finis (Mines-Ponts 2006)
- autour des codes de Gray (cf. plus loin)

Plan

1. Codes de Gray
2. Quelques applications amusantes
3. Puces à ADN
4. Questions subsidiaires

Partie 1
Codes de Gray

Source principale de cette partie : les fascicules 2 à 4 du tome IV du *Art of Computer Programming*

Codes de Gray

code de Gray : énumération $[p(0), \ldots, p(2^n - 1)]$ des entiers de 0 à $2^n - 1$, vérifiant la propriété suivante :

lorsqu'on passe de $p(k)$ à $p(k+1)$, on ne change qu'un bit, dans l'écriture binaire

exemple, avec $n = 3$:

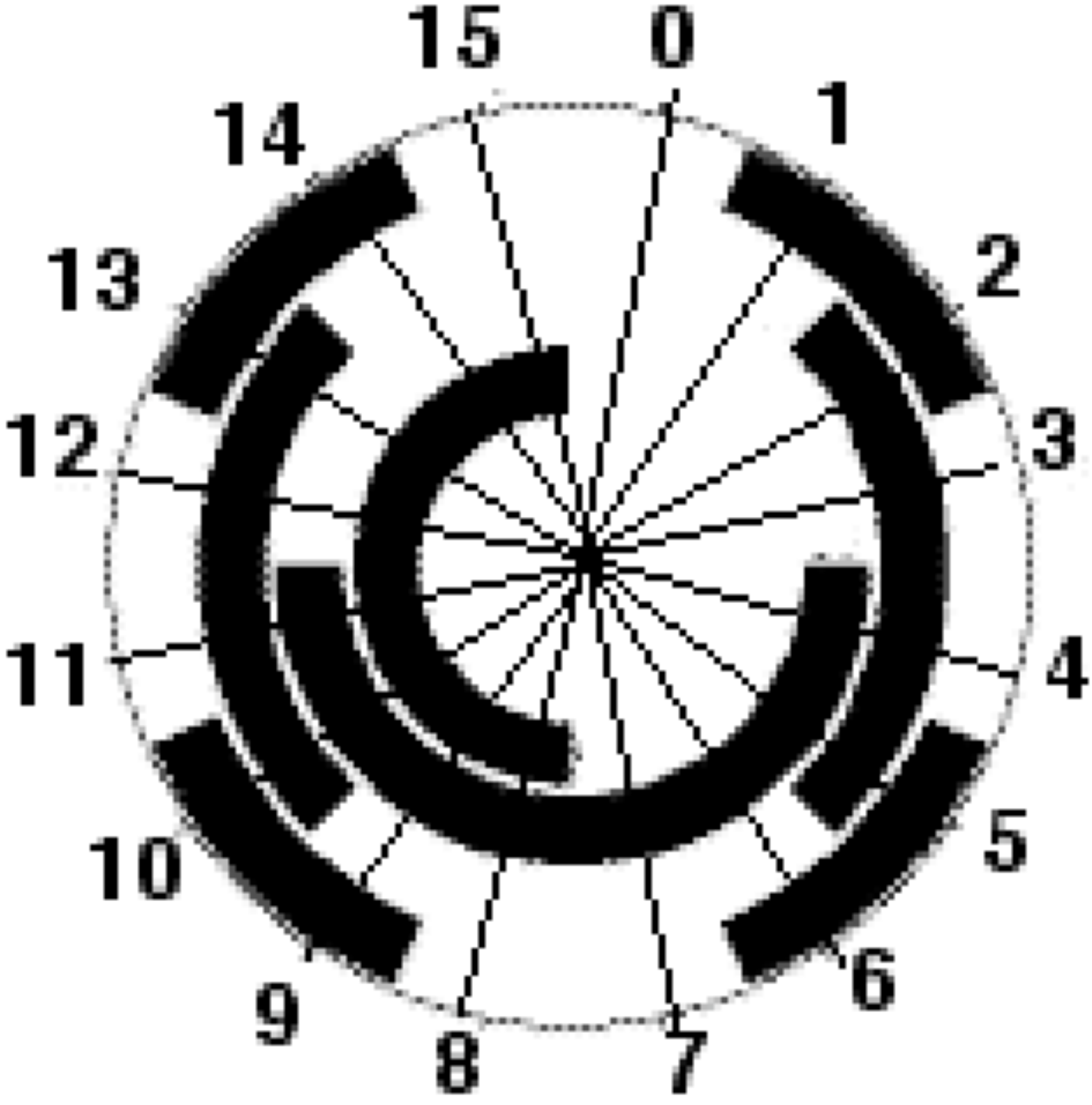
$[000, 001, 011, 010, 110, 111, 101, 100]$

remarque : on passe du dernier (100) au premier (000) en ne changeant qu'un bit. . .

Application

roue codeuse : quand on passe de k à $k+1$, un seul bit change (cf. vue)

montrer le fichier `roue-codeuse-gray.pdf`


15
0
1
14
2
13
3
12
4
11
5
10
6
9
8
7

Le baguenaudier

casse-tête ancien ; solution liée au code de Gray (cf. vue)

apparemment, Louis Agathon Gros serait le véritable inventeur du code de Gray (fin XiXème)

Wallis (fin XVIIème) note que la combinaison la plus difficile à atteindre est 10^{n-1}

montrer le fichier `baguenaudier.pdf`

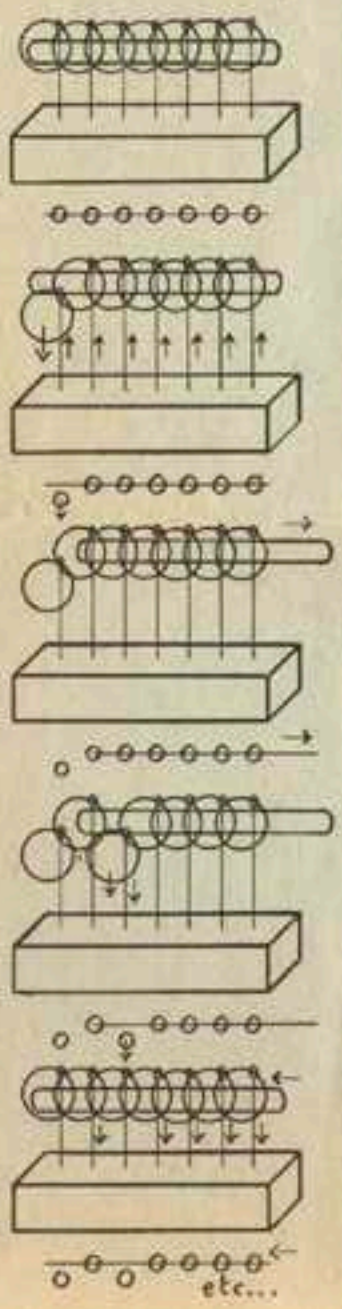
montrer le fichier `baguenaudier-howto.pdf`

Casse-tête

Le Baguenaudier

GRANDJOUAN

Modèle déposé



Codes de Gray réfléchis

famille $(G_n)_{n\geqslant 1}$ de codes définis par les formules $\Gamma_1 = (0, 1)$ et $\Gamma_{n+1} = (0 \cdot \Gamma_n)@(1 \cdot \widetilde{\Gamma_n})$ où :

– @ est la concaténation de listes
– $x \cdot \ell$ est l'ajout de x en tête de chaque élément de ℓ
– $\widetilde{\ell}$ est le miroir de la liste ℓ

Exemples

$\Gamma_2 = [00, 01, 11, 10]$

Γ_3 a été décrit plus haut

$\Gamma_4 = [0000, 0001, 0011, 0010, 0110, 0111, 0101, 0100,$
$1100, 1101, 1111, 1110, 1010, 1011, 1001, 1000]$

Règle pratique

on passe de $p(k)$ à $p(k+1)$ en :

– inversant le bit de droite, si le nombre de 1 dans $p(k)$ est pair
– inversant le bit à gauche du 1 le plus à droite, sinon

preuve par récurrence sur n

Une formule efficace

notons $g(k)$ le k-ième terme dans l'énumération d'un code de Gray réfléchi d'ordre n (avec $2^n > k$)

alors $g(2^n + r) = 2^n + g(2^n - 1 - r)$, pour $0 \leqslant r < 2^n$

preuve : facile

Une autre écriture

soient k et n tels que $0 \leqslant k < 2^n$; l'écriture de $g(k)$ dans Γ_{n+1} se déduit de l'écriture de $g(k)$ dans Γ_n par ajout d'un zéro en tête

ceci suggère une écriture « infinie », selon les puissances croissantes

exemple : $k = 22$; son écriture binaire est 10110 ; dans le code de Gray Γ_5, le mot numéro 22 est 11101 ; on leur associe $\psi(k) = 011010^\omega$ et $\overline{\psi}(k) = 101110^\omega$

Formules

notons $\psi(k) = a_0 a_1 \dots a_n \dots$ et $\overline{\psi}(k) = b_0 b_1 \dots b_n \dots$; alors $a_j = b_j \oplus b_{j+1}$ et $b_j = a_j \oplus a_{j+1} \oplus \dots$

le signe $\oplus$ désigne l'addition modulo 2

Une autre formule

$\overline{\psi}(k) = \psi(k) \oplus \psi(\lfloor k/2 \rfloor)$

le $\oplus$ est à comprendre « composante par composante »

implémentation facile d'un convertisseur binaire usuel $\rightarrow$ Gray, avec des portes XOR

Application banale

le code de Gray d'ordre n fournit un circuit hamiltonien sur l'hypercube à 2^n sommets

La méthode de la porte à tambour

Nous voulons énumérer les s-parties d'un ensemble de taille $n = s + t$, en respectant la contrainte suivante : pour passer de la k-ième à la $(k+1)$-ième partie, on enlève un élément puis on en ajoute un.

L'idée est semblable à celle des codes de Gray : utiliser des modifications « minimales » .

Partons de la relation $\Gamma_{n+1} = (0\Gamma_n)@(1\widetilde{\Gamma_n})$; chaque terme décrit une partie de $[\![1, n]\!]$. Ne gardons que les mots qui contiennent s bits 1. Par exemple, avec $s = 3$ et $t = 2$, nous avons $\Gamma_5 =$

00000 00001 00011 00010 00110 00111 00101 00100
01100 01101 01111 01110 01010 01011 01001 01000
11000 11001 11011 11010 11110 11111 11101 11100
10100 10101 10111 10110 10010 10011 10001 10000

Par projection, nous gardons $\Gamma_{2,3} =$

00011 00110 00101 01100 01010
01001 11000 10100 10010 10001

Observons les numéros des bits à 1 (chaque mot étant lu de gauche à droite) :

10 21 20 31 30 42 41 40

Le premier numéro va croissant ; le deuxième, à valeur constante du premier, va décroissant. Ceci se généralise au cas de s parmi $s+t$: notant $c_1, \ldots, c_s$ les numéros des bits à 1, la suite des c_1 est croissante ; à c_1 fixé, la suite des c_2 est décroissante ; à c_2 fixé, la suite des c_3 est croissante et ainsi de suite.

il existe des codes de Gray pour les permutations, les partitions, les arbres binaires, les forêts, etc.

cf. TAOCP volume 4

Partie 2

Quelques applications curieuses

Nous présentons quelques applications curieuses ; la dernière vous permettra de briller en société !

La serrure du professeur Macheprot

Le professeur Macheprot a muni la porte de son laboratoire d'une curieuse serrure de son invention. La chose se compose d'un clavier à cinq touches $\boxed{\mathbf{A}}$, $\boxed{\mathbf{B}}$, $\boxed{\mathbf{C}}$, $\boxed{\mathbf{D}}$ et $\boxed{\mathbf{X}}$; et de quatre tiges a, b, c et d, qui peuvent chacune être dans deux positions (haute ou basse). Appuyer sur une touche autre que $\boxed{\mathbf{X}}$ a pour effet d'inverser la position de la tige associée. Appuyer sur la touche $\boxed{\mathbf{X}}$ a pour effet d'ouvrir la porte si les quatre tiges sont dans la même position. Enfin, un dispositif de sécurité bloque la serrure pour une durée d'une heure, après huit essais infructueux.

Tours de Hanoï

Dans le casse-tête des tours de Hanoï, n disques sont rangés par tailles décroissantes sur l'axe 1. Il s'agit d'amener cette pile sur l'axe 3, en ne déplaçant qu'un disque à la fois, et en ne formant que des piles décroissantes. Il est connu que la solution optimale requiert $2^n - 1$ déplacements.

Numérotons les disques par taille croissante, de 1 à n. Dans un code de Gray réfléchi d'ordre n, le numéro du bit modifié à la k-ième transition est le numéro du disque à déplacer à la k-ième étape.

Le barman aveugle avec des gants de boxe

Padraig O'Shugrue est un ancien boxeur qui a perdu la vue au cours de son dernier combat. Aujourd'hui, il est le barman d'un lieu branché de Manhattan. Parfois, il propose à un client un jeu étrange qu'il a mis au point : sur un plateau tournant, il dispose quatre verres à pied, formant les sommets d'un carré. Le client peut alors en retourner certains (ou aucun).

Padraig enfile les gants du dernier combat (qui sont accrochés au mur du bar), ce qui l'empêche de savoir, au toucher, si un verre est « pied en bas » ou « pied en haut ». Il peut alors retourner un, ou deux, ou trois verres ; si les quatre verres se trouvent tous dans la même position, le barman a gagné ; sinon, le client peut faire tourner le plateau d'un ou deux quarts de tour, dans un sens arbitraire ; puis c'est au tour de Padraig de jouer.

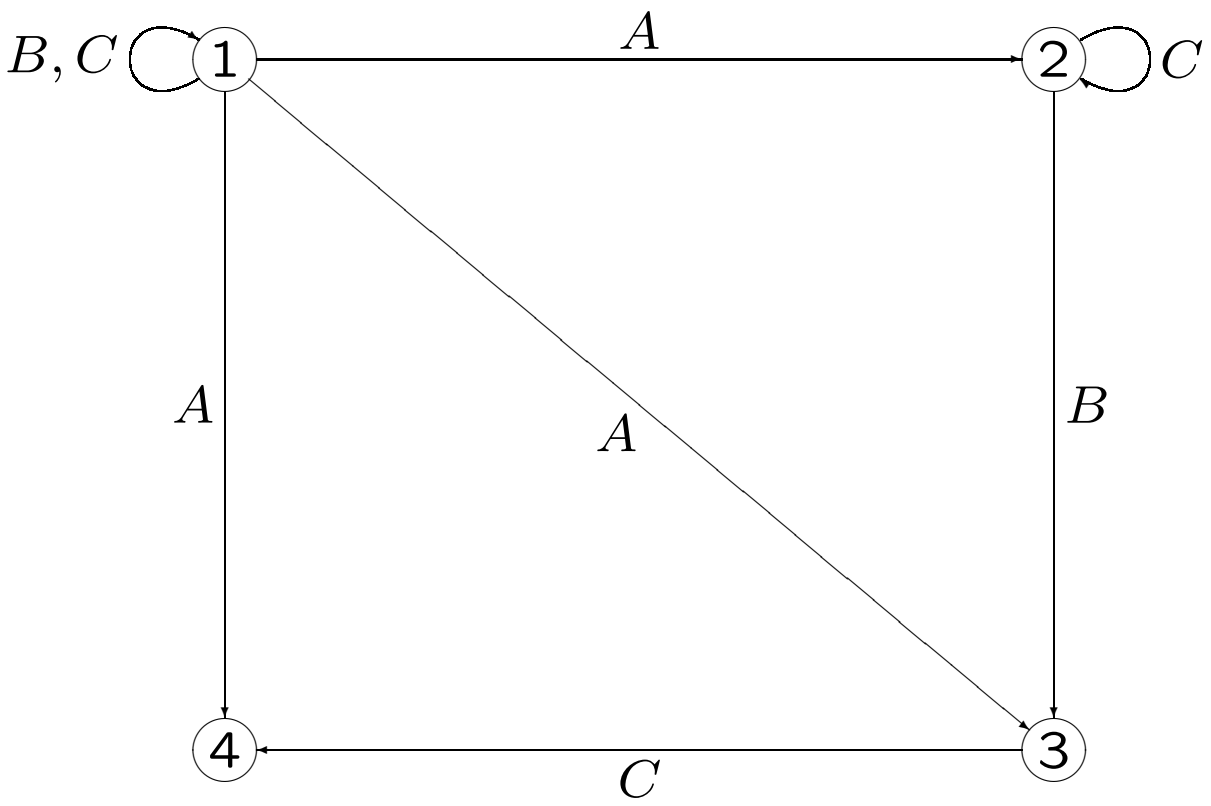
Si ce dernier n'a pas trouvé au bout de huit essais, le client a gagné. Curieusement, aucun client n'a jamais gagné.

Ce problème nous avait été présenté par Antoine Petit, il y a quelques années !

Le jeu peut être modélisé par un automate fini (non déterministe) à quatre états (vue suivante).

- A : retourner un seul verre ;
- B : retourner deux verres adjacents ;
- C : retourner deux verres opposés.

- 1 : un verre isolé ;
- 2 : deux verres en quadrature ;
- 3 : deux verres opposés.
- 4 : les 4 verres dans le même sens


B, C
1
A
2
C
A
A
B
4
C
3

Partie 3
Puces à ADN

Source principale de cette partie : le livre de Pevzner *Computational molecular biology*

Nous présentons une application des codes de Gray à un problème d'optimisation lié à la fabrication de puces à ADN.

Puces à ADN : qu'es aco ?

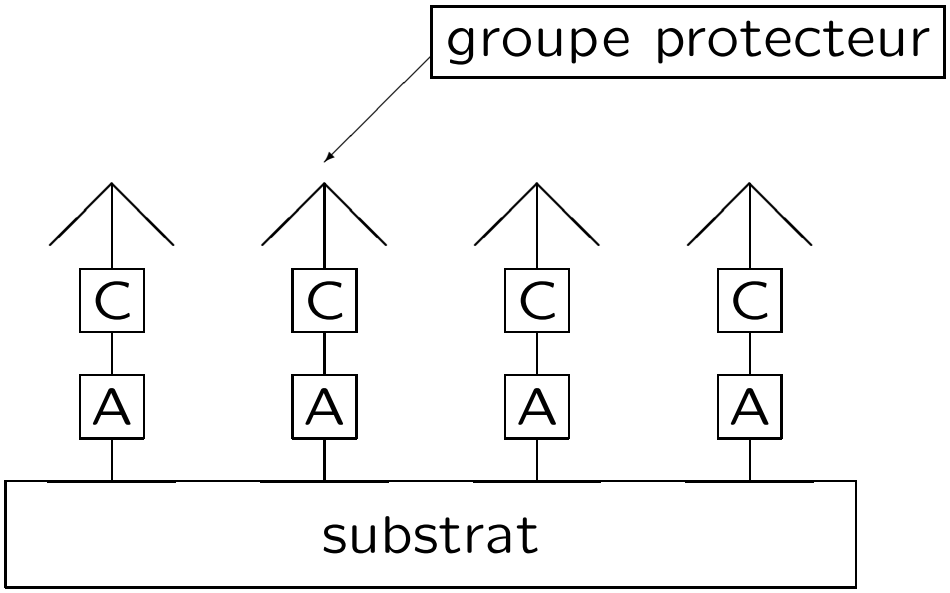
Une *puce à ADN* d'ordre n se compose d'un substrat carré (d'environ 1 cm de côté), décomposé en 4^n zones carrées de même taille ; danas chacune de ces zones, on implante un grand nombre de molécules d'ADN monobrin identiques, de longueur n. Chacune de ces molécules est un mot sur l'alphabet $\{\mathsf{A},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

A A	A C	C A	C C
A G	A T	C G	C T
G A	G C	T A	T C
G G	G T	T G	T T

A A	A C	C C	C A
A G	A T	C T	C G
G G	G T	T T	T G
G A	G C	T C	T A

à gauche : implantation « fractale » ; à droite : implantation « code de Gray bidimensionnel »

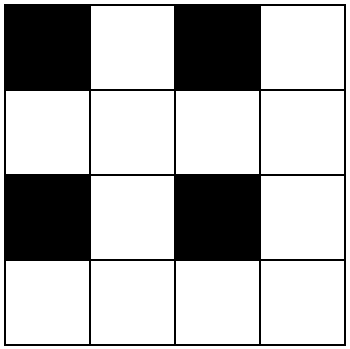
La fabrication d'une telle puce requiert une succession de $4n$ étapes, que l'on peut représenter par le mot $m = (\mathsf{A}\,\mathsf{C}\,\mathsf{G}\,\mathsf{T})^n$. Initialement, on dépose sur la puce des molécules dotées d'une base qui va adhérer au substrat, et d'un groupe protecteur photolabile. Lors de la k-ième étape, on pose un *masque* sur la puce et on l'expose aux ultra-violets : 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.

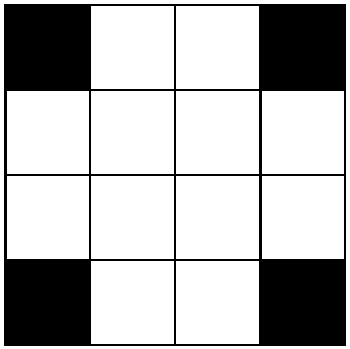

groupe protecteur
C
C
C
C
A
A
A
A
substrat

Avec l'implantation fractale, les puces sont définies par les deux dessins ci-dessous ; la notation $\mathsf{X}\,\mathsf{F}_n$ désigne une puce d'ordre n, dans laquelle on a ajouté une lettre X en tête de chaque mot.

$$\mathsf{F}_1 = \begin{array}{|c|c|}\hline \mathsf{A} & \mathsf{C} \\ \hline \mathsf{G} & \mathsf{T} \\ \hline \end{array} \qquad \mathsf{F}_{n+1} = \begin{array}{|c|c|}\hline \mathsf{A}\,\mathsf{F}_n & \mathsf{C}\,\mathsf{F}_n \\ \hline \mathsf{G}\,\mathsf{F}_n & \mathsf{T}\,\mathsf{F}_n \\ \hline \end{array}$$

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. Les longueurs de bord des deux masques ci-dessous 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 :

$$\mathsf{G}_1 = \begin{array}{|c|c|}\hline \mathsf{A} & \mathsf{C} \\ \hline \mathsf{G} & \mathsf{T} \\ \hline \end{array} \qquad \mathsf{G}_{n+1} = \begin{array}{|c|c|}\hline \mathsf{A}\,\mathsf{G}_n & \mathsf{C}\,\overrightarrow{\mathsf{G}_n} \\ \hline \mathsf{G}\,\mathsf{G}_n\downarrow & \mathsf{T}\,\overrightarrow{\mathsf{G}_n\downarrow} \\ \hline \end{array}$$

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

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 « fractale ». 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 $L_f(n) = 2 \cdot 4^{n+1} - (n+2)2^{n+2}$.

La suite $8, 64, 352, 1664 \ldots$ n'apparaît pas dans le site *On-Line Encyclopedia of Integer Sequences*, mais elle est reliée à la suite numéro A100575 par la relation $L_f(n) = 8a_n$, où a_{n+2} est le terme d'indice n de la suite numéro A100575.

L'OEIS définit a_n comme coefficient de e^{2x} dans le numérateur de $\frac{d^n}{dx^n}(\frac{1}{2-e^x})$.

La formule $a_n = 4^{n-1} - (n+1)2^{n-2}$ est assez facile à conjecturer et à prouver.

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 $L_g(n) = 4^{n+1} - 2^{n+2}$.

La suite de terme général $L_g(n)$ n'est pas connue de l'EOIS, mais elle est très proche de la suite numéro A006516 ; plus précisément, si nous notons b_n le terme d'indice n de celle-ci, nous avons $L_g(n) = 8b_{n+1}$ pour $n \geqslant 1$.

Curieusement, b_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.

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

Partie 4
Questions subsidiaires

Question 1 (ouverte)

Question : montrer que la fonction $x < \ln(2) \rightarrow \frac{1}{2-e^x}$ est absolument convexe (toutes les dérivées sont strictement positives)

on prouve facilement que le numérateur de la dérivée n-ième est de la forme $P_n(e^x)$ avec $P_{n+1} = (2-X)XP_n' + (n+1)XP_n$

Question : montrer que tous les coefficients de P_n sont positifs (sauf le terme constant)

Question 2

On a évoqué plus haut une question de combinatoire (nombre de mots de longueur n sur un alphabet à quatre lettres A, B, C, D la lettre A apparaissant un nombre impair de fois)

Question analogue : sur le même alphabet, combien existe-t-il de mots de longueur n comportant un nombre pair de A et un nombre pair de B ? On fait apparaître une matrice carrée d'ordre 4, qui est diagonalisable ; ceci se généralise-t-il ? (question floue)