

Séquençage du génome : éléments culturels

Résumé

Nous présentons ici les notions de base nécessaires pour comprendre les objectifs du *Projet génome humain*. Plus précisément, nous donnerons quelques éléments de biologie moléculaire, et décrirons certaines des techniques mises en œuvre pour réaliser le séquençage.

Nous verrons ensuite certains des problèmes algorithmiques posés par l'établissement d'une *carte de restriction*.

Table des matières

1 L'ADN : description sommaire	2
2 La réaction en chaîne polymérase	2
3 Les enzymes de restriction	3
4 Les cartes de restriction	3
5 Quelques pistes d'exploration possibles	4

1 L'ADN : description sommaire

L'ADN (acide désoxyribonucléique) est une molécule présente dans chaque cellule de chaque être vivant (à l'exception notable des virus). Elle est le support de l'information génétique. Cette molécule se présente comme une «double hélice», une sorte d'échelle enroulée dont les barreaux sont des paires de bases A-T ou C-G. Elle comporte donc deux brins, complémentaires l'un de l'autre (voir le document annexé pour des illustrations plus précises) :

```
...ACGGCTGTGAC...
      |||||
...TGCCGACTG...
```

Deux réactions biochimiques essentielles mettent en jeu cette molécule :

- la réplication : les deux brins se détachent l'un de l'autre, et chacun d'eux récupère dans la cellule des bases pour former le brin complémentaire ;
- la transcription : les deux brins s'écartent, autorisant ainsi la recopie d'une petite portion de l'un d'eux, avant de se réassembler.

La réplication intervient lorsqu'une cellule-mère se divise en deux cellules-filles : chacune hérite ainsi d'une copie du code génétique.

La transcription produit un ARN messenger, contenant le code nécessaire pour synthétiser une protéine ; plus précisément, le «texte» contenu dans l'ARN messenger est lu par tranches de trois lettres (les codons), chacun codant la synthèse d'un acide aminé. Les protéines sont de longues chaînes d'acides aminés, qui assurent les fonctions essentielles d'un organisme vivant ; parmi elles, on trouve les hormones, les enzymes, les anti-corps...

La taille du génome d'un être vivant se mesure en paire de bases (bp) ; cette taille est, grossièrement, liée à la complexité de l'être : le génome humain mesure $3 \cdot 10^9$ bp, celui de la bactérie *Escherichia coli* mesure environ $5 \cdot 10^6$ bp. Ceci dit, il est intéressant de noter que le génome du protoptère (poisson vivant dans certains lacs africains) mesure $150 \cdot 10^9$ bp.

2 La réaction en chaîne polymérase

Un élément-clé du séquençage du génome est la PCR (*polymerase chain reaction*) qui permet de réaliser en peu de temps de nombreuses copies d'un fragment d'ADN. C'est donc la «photocopieuse» du biologiste. Le principe est le suivant : on chauffe à 95°C une solution contenant le fragment d'ADN, pour le *dénaturer* : les deux brins se séparent ; la solution contient également, en quantité suffisante, les quatre molécules dATP, dCTP, dGTP et dTTP qui fourniront les bases A, C, T, G ; l'enzyme TAQ-polymerase ; et des amorces (brins d'environ dix à vingt bases, compléments des extrémités des deux brins du fragment).

- fragment à copier :
ACGGCTCGAG.....CCAGTTGTGAC
 ||||| |||||
TGCCGAGCTC.....GGTCAACTG
- bases : A,A,A,A... C,C,C,C... G,G,G,G... T,T,T,T...
- amorces : TGCCGAGC, AGTTGTGAC

On refroidit la solution : chacun des deux brins va s'attacher à une amorce, puis sera complété par action de la TAQ-polymerase. On obtiendra ainsi deux copies identiques du fragment initial. On recommence plusieurs fois ce processus, pour obtenir un grand nombre de copies (ou *clones*) du fragment d'ADN à séquencer. Chaque cycle prend quelques minutes, si bien que l'on peut en une heure obtenir plusieurs millions de copies ; on parle d'*amplification* de l'ADN.

Kary MULLIS a reçu en 1994 un prix NOBEL récompensant son travail de la conception de la PCR.

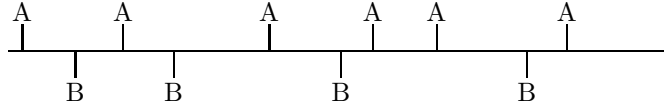


Figure 3: emplacements des sites de restriction

complète par l'enzyme donne l'ensemble de longueurs $B = \{25, 37, 55, 63, 70\}$. Enfin, la double digestion donne l'ensemble de longueurs $A \wedge B = \{5, 12, 15, 18, 19, 20, 24, 27, 34, 36, 40\}$.

On vérifiera que ces données sont compatibles avec les abscisses suivantes des sites de restriction : $a_1 = 5, a_2 = 43, a_3 = 98, a_4 = 137, a_5 = 161, a_6 = 210, b_1 = 25, b_2 = 62, b_3 = 125$ et $b_4 = 195$.

Le *problème de la double digestion* (PDD) consiste à établir une carte de restriction compatible avec ces données ; autrement dit, à trouver un ordre sur l'ensemble A et un ordre sur l'ensemble B , compatibles avec l'ensemble $A \wedge B$. La méthode de la force brutale consiste à essayer chacun des $|A|! \times |B|!$ couples d'ordres possibles : cette méthode est évidemment exclue.

Nous allons montrer que le PDD fait partie de la catégorie des problèmes *NP-complets*, en réduisant au PDD un autre problème classique, celui de l'équipartition d'une famille finie $\mathcal{F} = (E_1, E_2, \dots, E_n)$ d'ensembles ; on cherche une partition de \mathcal{F} en deux sous-ensembles \mathcal{F}_1 et \mathcal{F}_2 vérifiant

$$\sum_{X \in \mathcal{F}_1} |X| = \sum_{Y \in \mathcal{F}_2} |Y|$$

Il suffit de résoudre le PPD pour les ensembles $A = \bigcup_{1 \leq i \leq n} E_i, B = \{L/2, L/2\}$ où $L = \sum_{1 \leq i \leq n} |E_i|$, et $C = A \wedge B = A$. Toute solution du PPD fournit une solution du problème de l'équipartition. Or ce dernier problème est NP-complet.

5 Quelques pistes d'exploration possibles

1. Discutez l'intérêt d'une digestion partielle, précédant une digestion totale avec une enzyme A donnée.
2. Donnez un exemple simple où le PDD admet plusieurs solutions.
3. Résolvez le PDD pour l'exemple donné plus haut (sans regarder la figure) et montrez l'unicité de la solution.
4. La présentation qui a été faite du PDD ne prend pas en compte un certain nombre de réalités physiques : essayez d'en trouver quelques-unes.

FIN