Cartographie physique du génome, matrices ayant la propriété des 1 consécutifs, arbres PQ

Résumé

Nous nous intéressons au problème de l'établissement de la carte physique d'un segment d'ADN. Nous présentons rapidement les notions de biologie moléculaire nécessaires, puis formalisons le problème. Enfin, nous décrivons les arbres PQ et donnons une idée de l'algorithme de Booth et Lueker.

Table des matières

1 Cartographie physique

L'une des premières étapes du projet génome humain a été la réalisation d'une carte physique, donnant les emplacements relatifs de marqueurs (STS, ou *sequence-tagged sites*) Un tel marqueur est, en gros, un fragment d'ADN d'environ 200 bp (paires de bases), dont on sait qu'il n'apparaît qu'une fois dans le génome. Plusieurs centaines de milliers de STS ont été répertoriés, et un fragment d'ADN de longueur 100 kbp contient au moins un STS.

L'objectif est donc de placer ces marqueurs les uns par rapport aux autres. Ensuite, quand on a réussi à séquencer un fragment assez long, il sera facile de le positionner, pourvu qu'il contienne au moins un STS.

Une *bibliothèque de clones* est un ensemble de fragments (les *clones*) provenant d'un segment d'ADN que l'on souhaite cartographier. On ne connaît pas l'emplacement des clones dans le segment, ni même leur ordre respectif. Deux clones distincts peuvent se recouvrir partiellement, voire totalement ; bien entendu, il est souhaitable que la bibliothèque couvre la totalité du segment. Pour repérer la présence d'un marqueur dans un clone, on dénature celui-ci (séparation des deux brins), on l'immobilise dans un gel et on le met en présence d'une *sonde* (complément du STS, modifié pour être fluorescent ou radioactif). Si le marqueur est présent dans le clone, il y aura *hybridation* entre celui-ci et la sonde (les bases complémentaires s'apparient) ; après rinçage du gel pour éliminer les sondes qui n'ont pas servi, on constatera éventuellement l'hybridation.

Quelques remarques de bon sens : si le marqueur j est présent dans les clones i et i', ceux-ci se recouvrent en partie. D'autre part, si deux marqueurs j et j' apparaissent simultanément dans plusieurs clones, ils sont vraisemblablement proches l'un de l'autre.

2 Formalisation du problème

Chaque fragment est amplifié par PCR (voir la section 5) ; on présente une fraction de l'échantillon obtenu à chaque sonde. On sait ainsi quels STS apparaissent dans ce fragment, mais on ne connaît pas leur ordre relatif. La figure 1 donne un exemple, avec quatre clones et sept sondes.

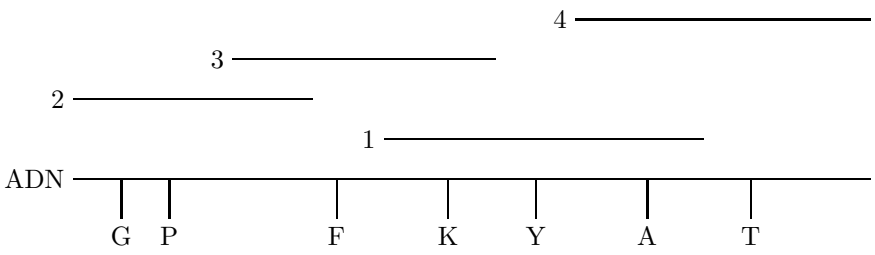

Figure 1: quatre clones, sept sondes.

On résume les résultats obtenus par une matrice M, comportant autant de lignes que de clones et autant de colonnes que de sondes ; le coefficient $M_{i,j}$ vaut 1 ou 0 selon que la sonde j a ou non réagi avec le clone i. La figure 2 montre la matrice qui correspond au fragment de la figure 1, les sondes étant placées dans l'ordre alphabétique.

	A	F	G	K	P	T	Y
1	1	0	0	1	0	0	1
2	0	0	1	0	1	0	0
3	0	1	0	1	0	0	0
4	1	0	0	0	0	1	0

Figure 2: la matrice obtenue.

Ainsi, le clone 4 nous dit que les marqueur A et T sont consécutifs : il n'y a aucun autre marqueur entre eux. Le clone 1 nous dit alors que les marqueurs K et Y sont situés d'un côté de A, et T de l'autre côté. Ceci donne quatre possibilités : (K, Y, A, T) ; (Y, K, A, T) ; (T, A, K, Y) et (T, A, Y, K). En exploitant le clone 3, on peut affirmer que c'est Y qui est à l'extérieur, ce qui ne laisse que deux possibilités : (F, K, Y, A, T) et (T, A, Y, K, F). Le clone 2 indique que G et P sont consécutifs, ce qui au final fera huit possibilités : deux façons de lire la séquence de cinq marqueurs reliés, deux façons de lire la séquence de deux marqueurs, et deux façons de placer ces deux séquences l'une par rapport à l'autre.

Suggestion : expliquez en détail le raisonnement.

Notons c le nombre de clones, et s le nombre de sondes ; on peut supposer les uns et les autres numérotés à partir de 1. Le problème consiste à trouver une permutation π de l'intervalle discret $[\![1, s]\!]$ telle que la matrice Q à c lignes et s colonnes définie par $Q_{i,j} = M_{\pi(i),j}$ possède la *propriété des 1 consécutifs* : dans chaque ligne, les 1 apparaissent consécutivement. π décrit donc une permutation des colonnes permettant de transformer M en une matrice Q ayant cette propriété.

La méthode de la force brutale consisterait à essayer les $s!$ permutations possibles. Ceci n'est pas envisageable : pour le projet génome humain, les STS sont séparés par environ 100 kbp, or la longueur d'un chromosome varie entre 50 Mbp et 250 Mbp.

Suggestion : construisez une matrice qui ne peut pas être transformée en une matrice ayant la propriété des 1 consécutifs au moyen d'une permutation des colonnes.

3 Les arbres PQ

En 1976, Booth et Lueker ont proposé, pour tester la planarité d'un graphe, l'emploi d'une structure de données qu'ils ont appelée *arbre PQ*. L'emploi d'un arbre PQ permet de résoudre le problème posé à la section précédente, pour un coût total linéaire par rapport à la somme $c + s$ du nombre de clones et du nombre de sondes.

Considérons des arbres dont les nœuds sont de deux types, P ou Q, et dont les feuilles sont étiquetées par les entiers de 1 à n. Définissons sur l'ensemble de ces arbres la relation suivante : t est semblable à t' si l'on peut passer de l'un à l'autre en effectuant une suite d'opérations de l'un des deux types suivants :

- appliquer une permutation arbitraire à la liste des fils d'un nœud de type P ;
- ≪retourner≫ la liste des fils d'un nœud de type Q.

On définit ainsi une relation d'équivalence ; un arbre PQ est une classe modulo cette relation. Il est d'usage de représenter un nœud P par un cercle, et un nœud Q par un rectangle. Remarquons qu'un nœud P possède au moins deux fils, et un nœud Q au moins trois fils.

Suggestion : expliquez cette dernière affirmation.

La figure 3 présente un arbre possédant deux nœuds P et nœud Q ; il est clair que la classe de cet arbre possède 24 éléments : on peut échanger les deux sous-arbres qui pendent de la racine, appliquer une permutation parmi six au nœud de type P, et retourner la liste des fils du nœud de type Q.

Suggestion : définissez un type Caml pour décrire les arbres PQ, puis rédigez une fonction qui calcule le cardinal de la classe d'équivalence d'un tel arbre.

4 L'algorithme de Booth et Lueker

Au départ, on connaît le nombre s de sondes ; on construit un arbre ayant une racine de type P, et s fils numérotés de 1 à s. Chaque examen d'un clone dans lequel apparaissent au moins deux marqueurs nous apporte une information du type $\{m_1, m_2, \ldots, m_k\}$; cette information nous indique que les marqueurs s_{m_1} à s_{m_k} apparaissent ≪groupés≫, dans un ordre qui reste à déterminer. Il s'agit de réorganiser l'arbre PQ pour prendre en compte cette contrainte.

Booth et Lueker ont montré que onze cas de figure pouvaient se produire. Leur méthode consiste à faire remonter l'information des feuilles vers la racine ; un nœud est *pertinent* si l'une

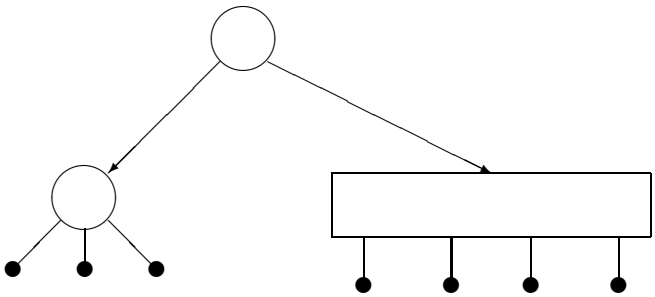

Figure 3: un arbre PQ.

au moins des feuilles qui lui sont rattachées (directement ou indirectement) apparaît dans le clone. Il est clair que seul les nœuds pertinents seront (éventuellement) touchés par des modifications.

Observons par exemple ce qui se passe lors de l'examen du premier clone : un nœud de type P est créé et les feuilles concernées passent sous ce nœud.

Suggestion : dessinez les arbres PQ successifs, qui correspondent à l'exemple de la figure 1.

Suggestion : construisez un jeu de clones (avec sept marqueurs) qui mènerait à l'arbre PQ de la figure 3.

Un blocage peut se produire au cours du déroulement de l'algorithme : c'est que les données fournies sont incohérentes.

Suggestion : suivez le déroulement de l'algorithme, appliqué au jeu de données incohérent que vous avez construit dans la section 2.

5 Annexe : la PCR

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 ≪photo-copieuse≫ 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..........GGTCAACACTG
```

- 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 heures 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.

FIN