IFT6299 A05 — Devoir 1

Miklós Csűrös

26 septembre 2005

À remettre au cours du 13 octobre.

1 Une baignoire d'ADN (10 points)

Soit $\mathcal{G}$ un graphe régulier orienté avec n sommets où de chaque sommet sortent k arêtes. Supposons on veut utiliser la procédure d'Adleman pour vérifier l'existence d'un chemin Hamiltonien de sommet u à sommet v. Quel est le nombre de parcours différents avec $(n-1)$ arêtes à partir de sommet u ? Leonard Adleman utilisait une solution de concentration $\rho = 50\,\mu$mol/l de chaque oligonucléotide. Si le graphe contient exactement un chemin Hamiltonien, et les parcours sont représentés uniformement (en vérité, la concentration des parcours n'est pas uniforme en général), quelle est la quantité de solution nécessaire pour assurer la formation d'au moins une molécule qui correspond au chemin Hamiltonien ? Soit $k = 10$. Pour quelle taille n on a besoin d'une baignoire d'ADN ?

2 Un verre d'automates finis (30 points)

La méthode de Kobi Benenson permet l'implantation d'un automate fini à deux états sur un alphabet binaire à l'aide de l'ADN et des enzymes comme *FokI*. Expliquez comment la méthode peut être généralisée (a) à plus de deux états, et (b) à plus de deux caractères dans l'alphabet. Montrez un exemple détaillé pour un automate à trois états et un alphabet binaire. Si vous avez besoin d'un autre enzyme de restriction pour l'implantation, vous pouvez en chercher un de type IIS (c'est la classe d'enzymes où il y a des cas avec un site de clivage loin du site reconnu) : `http://rebase.neb.com/cgi-bin/asymmlist`. Pour **5 points de boni** et du fun, calculez le coût du *shopping list* qui permet l'implantation de votre automate : votre enzyme de restriction (*FokI* et beaucoup d'autres peuvent être achetés chez New England Biolabs `http://www.neb.com/`), les

oligonucléotides synthétiques, etc. (Cherchez «Custom Oligonucleotide Synthesis» dans la catalogue d'une des compagnies comme Integrated DNA Technologies. Pour comparaison de prix, vous pouvez consulter `http://www.biocompare.com/matrix/8797/Oligo-Synthesis-Services.html`.)

3 Phred m'a dit que c'était OK (30 points)

Le logiciel le plus souvent utilisé pour produire des séquences à partir d'un electropherogram s'appelle Phred. Phred calcule des valeurs de qualité pour tous les nucléotides d'une séquence. Les valeurs de qualité sont des entiers non-négatifs : une valeur de q signifie que le nucléotide y correspondant est incorrect avec une probabilité de $10^{-q/10}$. Par exemple $q = 10$ signifie une probabilité d'erreur de 10%, $q = 40$ signifie une probabilité d'erreur de 0.01%, etc. Typiquement, les deux extrémités de la séquence comprennent des nucléotides de basse qualité, c'est-à-dire des caractères avec $q \leq 20$. Ces extrémités ne sont pas utiles pour calculer les chevauchements, donc souvent on les enlève avant d'aligner les fragments. Le but de cet exercice est de concevoir un algorithme qui fait exactement ça.

a. (8 points) Soit $S = s_1 \cdots s_m$ une séquence de longueur m avec valeurs de qualité q_i pour chaque nucléotide s_i. Quelle est l'espérance du nombre d'erreurs dans S ? (Indice : écrivez le nombre d'erreurs comme une somme de variables indicateurs d'erreurs en positions $i..j$.)

b. (22 points) Soit E_{ij} l'espérance du nombre d'erreurs dans chaque sous-mot $S_{ij} = s_i s_{i+1} \cdots s_j$ avec $i \leq j$. Donnez un algorithme qui trouve le sous-mot le plus long avec une erreur moyenne de 1%, c'est-à-dire un sous-mot S_{ij} avec $E_{ij} \leq \frac{|S_{ij}|}{100}$ et $|S_{ij}| = j - i + 1$ maximale. Il est important de démontrer que votre solution est correcte, ainsi que d'analyser son temps de calcul. L'algorithme naïf qui essaie tous les sous-mots prend un temps de $O(m^2)$. Votre solution doit prendre un temps strictement sous-quadratique. Une solution avec temps de calcul $O(m)$ vaut **5 points de boni**.

4 Et la digestive arrive. . . (30 points)

Une librairie de clones peut être construit par la digestion d'une molécule d'ADN. Soit x le site reconnu par un enzyme de restriction qui clive le site entre $x[1..d]$ et $x[d+1..|x|]$. Soit S est la séquence de la molécule ciblée. La digestion produit des fragments $S[1..i_1], S[i_1+1..i_2], \ldots, S[i_n+1..|S|]$ où les i_j sont les sites de clivage : $S[i_j - d + 1..i_j + |x| - d] = x$. Pas tous les fragments peuvent être inserés dans une vecteur de clonage : seulement ceux de longueur entre A et B où $A..B$ est la rangée de tailles d'inserts acceptés par la vecteur. Par exemple, pour

une librairie de BACs, on a $A = 50000$ et $B = 300000$.

(a) Démontrez que quand $|x| \ll |S|$, et S est une séquence aléatoire uniforme, alors la fraction de positions clonables peut être approximée par

$$f(p, A, B) = (1 + pA)e^{-pA} - (1 + pB)e^{-pB}$$

où p est la probabilité du site reconnu. Par exemple pour EcoRI (site `GAATTC`) $p = 4^{-6} = 1/4096$, pour HindII (site `GTYRAC`, $p = 1/(4 \cdot 4 \cdot 2 \cdot 2 \cdot 4 \cdot 4) = 1/1024$.

(b) Calculez le p qui maximise f (en fonction de A, B). Donnez des valeurs numériques pour ce p est le maximum f dans le cas de BACs (50k..300k) et de cosmids (35k..45k).

Indice [d'autres solutions sont possibles aussi] : (1) Ignorez les extremités $S[1..i_1]$ et $S[i_n + 1..|S|]$. (2) Introduisez les variables indicateurs $X_{i,l}$ pour l'événement qu'un fragment de longueur l commence en position i. Calculez la probabilité de cet événement. (3) Démontrez que l'espérance de nombre de positions clonables est $\mathbb{E} \sum_i \sum_l l X_{i,l}$ et évaluez cette somme en utilisant la linéarité de l'espérance. (4) Il peut être utile de se servir de l'égalité $\sum_{k=1}^{n} kq^{k-1} = \frac{1-(n+1)q^n+nq^{n+1}}{(1-q)^2}$ $(q \neq 1, n \geq 1)$ [pour **5 points de boni** : démontrez cette égalité], et de l'approximation $(1 - p)^n \approx e^{-np}$ $(0 < p \ll 1, n \gg 1)$.