Arbres, préfixes et suffixes.

Épreuve pratique d'algorithmique et de programmation
Concours commun des écoles normales supérieures

Durée de l'épreuve: 3 heures 30 minutes

Juin/Juillet 2011

ATTENTION !

N'oubliez en aucun cas de recopier votre u_0
à l'emplacement prévu sur votre fiche réponse

Important.

Sur votre table est indiqué un numéro u_0 qui servira d'entrée à vos programmes. Les réponses attendues sont généralement courtes et doivent être données sur la fiche réponse fournie à la fin du sujet. À la fin du sujet, vous trouverez en fait deux fiches réponses. La première est un exemple des réponses attendues pour un $\widetilde{u_0}$ particulier (précisé sur cette même fiche et que nous notons avec un tilde pour éviter toute confusion!). Cette fiche est destinée à vous aider à vérifier le résultat de vos programmes en les testant avec $\widetilde{u_0}$ au lieu de u_0. Vous indiquerez vos réponses (correspondant à votre u_0) sur la seconde et vous la remettrez à l'examinateur à la fin de l'épreuve.

En ce qui concerne la partie orale de l'examen, lorsque la description d'un algorithme est demandée, vous devez présenter son fonctionnement de façon schématique, courte et précise. Vous ne devez en aucun cas recopier le code de vos procédures!
Quand on demande la complexité en temps ou en mémoire d'un algorithme en fonction d'un paramètre n, on demande l'ordre de grandeur en fonction du paramètre, par exemple: $O(n^2)$, $O(n \log n)$,...
Il est recommandé de commencer par lancer vos programmes sur de petites valeurs des paramètres et de ***tester vos programmes sur des petits exemples que vous aurez résolus préalablement à la main ou bien à l'aide de la fiche réponse type fournie en annexe***. Enfin, il est recommandé de lire l'intégralité du sujet avant de commencer afin d'effectuer les bons choix de structures de données dès le début.

1 Introduction

On s'intéresse dans ce sujet aux problèmes de recherche de motifs dans des séquences de caractères. Ces problèmes sont très présents dans le domaine de la génomique : l'ADN est représenté comme une séquence de caractères A, T, G et C, et on souhaite rechercher un motif donné dans une séquence d'ADN, les occurrences de plusieurs motifs dans une ou plusieurs séquences, ou encore les motifs communs qui apparaîssent dans plusieurs séquences. On se propose d'étudier un outil permettant la conception d'algorithmes efficaces pour ces problèmes.
Dans ce sujet, on travaillera toujours sur des séquences de caractères qui ne peuvent prendre que les quatre valeurs A, T, G et C.

2 Génération aléatoire de séquences

On considére la suite d'entiers (u_k) définie pour $k \geqslant 0$ par :

$$u_k = \begin{cases} \textit{votre } \mathtt{u_0} \textit{ (à reporter sur votre fiche)} & \text{si } k = 0 \\ 15\,091 \times u_{k-1} \mod 64\,007 & \text{si } k \geqslant 1 \end{cases}$$

Question 1 *Que valent :* **a)** u_{10} **b)** u_{100} **c)** u_{1000}

Pour une séquence de caractères S de longueur l et $1 \leqslant i \leqslant l$, on note $S[i]$ le $i^{\text{ème}}$ caractère de S. On définit des séquences de caractères aléatoires de la manière suivante :

Définition 1 *Pour $k, l \in \mathbb{N}^2$ on note $S_{k,l}$ la séquence de l caractères sur l'alphabet $\{A, T, G, C\}$, telle que pour $1 \leqslant i \leqslant l$, le $i^{\text{ème}}$ caractère de cette séquence est donné par :*

$$S_{k,l}[i] = \begin{cases} A & \text{si } u_{i+k-1} \bmod 4 = 0 \\ T & \text{si } u_{i+k-1} \bmod 4 = 1 \\ G & \text{si } u_{i+k-1} \bmod 4 = 2 \\ C & \text{si } u_{i+k-1} \bmod 4 = 3 \end{cases}$$

Question 2 *Donnez les séquences suivantes :* **a)** $S_{0,5}$ **b)** $S_{10,5}$ **c)** $S_{20,5}$

On définit la liste de n séquences $L_{k,n,l} = (S_{k,l}, S_{k+l,l}, S_{k+2l,l}, \ldots, S_{k+(n-1)\times l,l})$.

3 Arbres préfixes

On rappelle qu'un arbre enraciné est un ensemble de nœuds et d'arêtes tels que chaque nœud est :
– soit une feuille f ;
– soit un nœud interne u, qui possède un ou plusieurs fils qui sont des nœuds de l'arbre ; u est relié à chacun de ses fils v par une arête (u, v) et u est appelé père de v.
On nomme racine d'un arbre le seul nœud qui ne possède pas de père. Les nœuds comme les arêtes d'un arbre peuvent être munis d'étiquettes.
On définit les arbres préfixes de la façon suivante :

Définition 2 *L'arbre préfixe d'un ensemble de séquences E est un arbre enraciné tel que :*
- *chaque arête est étiquetée par une lettre de l'alphabet ;*
- *pour un nœud u et deux de ses fils v et w, l'arête (u, v) ne porte pas la même étiquette que l'arête (u, w) ;*
- *à chaque séquence S de l'ensemble E correspond une feuille f de l'arbre telle que la concaténation des caractères étiquetant les arêtes sur le chemin de la racine à f est égale à S ; de même chaque feuille correspond à une séquence de E.*

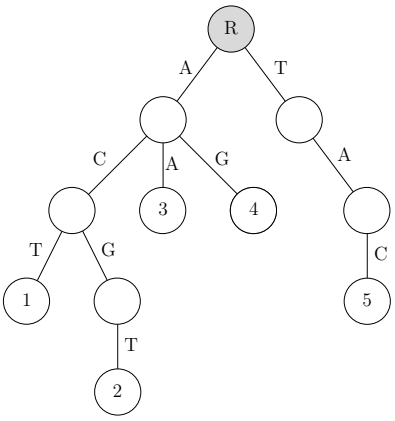


FIGURE 1 – Exemple d'arbre préfixe. La racine est marquée par la lettre R.

La figure 1 représente l'arbre préfixe des séquences ACT, $ACGT$, AA, AG, et TAC. La feuille notée 1 correspond par exemple à la séquence ACT, obtenue en lisant les caractères depuis la racine jusqu'à cette feuille.

Question à développer pendant l'oral : Montrez que dans un arbre préfixe, un nœud a au plus quatre fils.

Question 3 *Donnez le nombre de nœuds ainsi que le nombre d'arêtes étiquetées par la lettre "A" dans les arbres préfixes des listes de séquences suivantes :*
a) $L_{0,5,4}$ **b)** $L_{0,50,10}$ **c)** $L_{0,50,40}$

Question à développer pendant l'oral : Quelle est la complexité de votre algorithme ?

Dans l'étude de complexité des algorithmes suivants, on supposera désormais que les arbres préfixes des listes de séquences ci-dessus ont été construites, et on ne prendra plus en compte la complexité de leur création.
On munit les lettres de l'alphabet d'un ordre total arbitraire : $A < T < G < C$ et on étend cet ordre aux séquences à l'aide de l'ordre lexicographique : pour deux séquences S_1 et S_2, si i est l'index du premier caractère différent dans S_1 et S_2 ($\forall k < i, S_1[k] = S_2[k]$) et que $S_1[i] < S_2[i]$, alors on a $S_1 < S_2$. Par exemple $ATAGC < ATACA$ car les trois premiers caractères sont identiques et $G < C$ pour le troisième.

Question 4 *Pour les ensembles de séquences suivants, donnez la séquence qui apparaît en $4^{\text{ème}}$ position dans l'ordre lexicographique :*
a) $L_{0,5,4}$ **b)** $L_{0,50,10}$ **c)** $L_{0,50,40}$ *(on ne donnera que les 10 premiers caractères pour cette dernière séquence)*

Question 5 *Donnez le plus long préfixe commun à au moins k séquences distinctes des listes suivantes (s'il existe plusieurs tels préfixes, on donnera le plus petit au sens de l'ordre lexicographique) :*
a) $L_{0,5,4}, k = 2$ **b)** $L_{0,50,10}, k = 4,$ **c)** $L_{0,50,40}, k = 10$

Question 6 *Donnez la plus courte séquence qui n'est préfixe d'aucune séquence des listes suivantes (en cas d'égalité, on donnera la plus petite au sens de l'ordre lexicographique) :*
a) $L_{0,5,4}$ **b)** $L_{0,50,10}$ **c)** $L_{0,50,40}$

Question à développer pendant l'oral : Pour les questions 4, 5 et 6, donnez la complexité de vos algorithmes.

4 Arbre des suffixes et recherche de motifs dans un texte

On considère une séquence S de n caractères. On appelle suffixe de S commençant au caractère i, et on note $S[i \dots n]$ la séquence de caractères $S[i], \dots S[n]$ extraite de S. À partir d'une séquence S de n caractères sur l'alphabet $\{A, T, G, C\}$, on considère la séquence $\overline{S}$ obtenue en rajoutant un caractère de terminaison X à la fin de la séquence : $\overline{S} = S[1], \dots, S[n], X$. On définit l'arbre des suffixes de $\overline{S}$ de la façon suivante :

Définition 3 *Un arbre des suffixes pour la séquence $\overline{S}$ est un arbre enraciné comprenant exactement $n + 1$ feuilles, qui sont numérotées de 1 à $n + 1$. Les arêtes sont étiquetées à l'aide des caractères A, T, G, C et X, de telle sorte que pour un nœud interne u et deux de ces fils v et w, (u, v) et (u, w) ne sont pas étiquetées par le même caractère et que la concaténation des caractères étiquetant les arêtes sur le chemin de la racine à la feuille i pour $i \leqslant n$ correspond à la séquence $S[i \dots n]X$, et la feuille $n + 1$ correspond à la séquence X.*

La figure 2 montre un exemple d'arbre des suffixes.

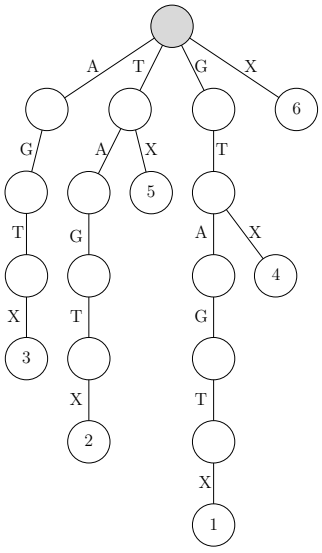


FIGURE 2 – Arbre des suffixes pour la séquence $GTAGT$.

Question à développer pendant l'oral : Montrez que si la séquence S ne se termine pas par un caractère spécial X, il n'existe pas nécessairement d'arbre des suffixes pour S.

Question 7 *Construisez l'arbre des suffixes pour les textes suivants, donnez leur nombre de nœuds, ainsi que le nombre d'arêtes étiquetées par la lettre A :*
a) $\overline{S_{0,10}}$ **b)** $\overline{S_{10,100}}$ **c)** $\overline{S_{110,1000}}$

Question à développer pendant l'oral : On souhaite rechercher un grand nombre de motifs dans un même texte. Proposez une méthode efficace pour ce problème, et donnez sa complexité. Dans quels cas cette méthode est elle intéressante ?

Question 8 *On appelle sous-séquence consécutive d'une séquence S toute séquence de caractères L de taille l telle qu'il existe $k \geqslant 0$ avec $L[i] = S[i + k]$ pour tout $1 \leqslant i \leqslant l$. Donnez la séquence de caractères la plus courte qui n'est pas une sous-séquence des séquence suivantes (en cas d'égalité, on donnera la séquence la plus petite au sens de l'ordre lexicographique) :* **a)** $\overline{S_{0,10}}$ **b)** $\overline{S_{10,100}}$ **c)** $\overline{S_{110,1000}}$

5 Comparaison de deux séquences

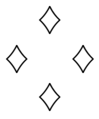
On cherche à déterminer la ressemblance de deux séquences de caractères. On appelle sous-séquence non consécutive commune à deux séquences de caractères S_1 et S_2 une séquence L telle que $L[i] = S_1[f(i)] = S_2[g(i)]$, où f et g sont deux fonctions strictement croissantes. Par exemple $TATAGC$ est une sous-séquence non consécutive commune à $\underline{TA}G\underline{TA}C\underline{G}A\underline{C}$ et à $AG\underline{T}G\underline{A}CG\underline{TAG}A\underline{C}$.

Question à développer pendant l'oral : On note $l(i, j)$ la plus longue sous-séquence non consécutive commune à $S_1[1..i]$ et $S_2[1..j]$. Exprimez $l(i, j)$ en fonction de $l(i', j')$ pour $i' \leqslant i$ et $j' \leqslant j$ et des caractères de S_1 et S_2.

Question 9 *Donnez la <u>longueur</u> de la plus longue sous-séquence non consécutive des séquences suivantes :*
a) $S_{0,10}$ *et* $S_{10,10}$ **b)** $S_{0,100}$ *et* $S_{100,100}$ **c)** $S_{0,1000}$ *et* $S_{1000,1000}$

On s'intéresse maintenant aux sous-séquences <u>consécutives</u> communes à deux séquences, comme définies dans la partie précédente.

Question 10 *Donnez la plus longue sous-séquence consécutive commune aux paires de textes suivants (en cas d'égalité, on donnera la plus petite dans l'ordre lexicographique) :*
a) $S_{0,10}$ *et* $S_{10,10}$ **b)** $S_{0,100}$ *et* $S_{100,100}$ **c)** $S_{0,1000}$ *et* $S_{1000,1000}$



Fiche réponse type: Arbres, préfixes et suffixes.

$\widetilde{u_0}$: 115

Question 1

a) 52892

b) 37008

c) 14225

Question 2

a) CACAC

b) AGATA

c) GAATC

Question 3

a) 17,7

b) 404,101

c) 1897,499

Question 4

a) CACT

b) ATAGTTAATT

c) ATGATGATCT...

Question 5

a) CAC

b) GAA

c) T

Question 6

a) G

b) TT

c) CT

Question 7

a) 50, 15

b) 4891, 1533

c) 497366, 129684

Question 8

a) G

b) AAA

c) AGGT

Question 9

a) 5

b) 62

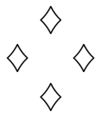
c) 649

Question 10

a) TA

b) AAGAGC

c) ACATGGCCGAA



Fiche réponse: Arbres, préfixes et suffixes.

Nom, prénom, u_0: .....

Question 1

a)

b)

c)

Question 2

a)

b)

c)

Question 3

a)

b)

c)

Question 4

a)

b)

c)

Question 5

a)

b)

c)

Question 6

a)

b)

c)

Question 7

a)

b)

c)

Question 8

a)

b)

c)

Question 9

a)

b)

c)

Question 10

a)

b)

c)