Arbres phylogénétiques

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

Durée de l'épreuve: 3 heures 30 minutes

Session 2006

ATTENTION !

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

Important.
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***.

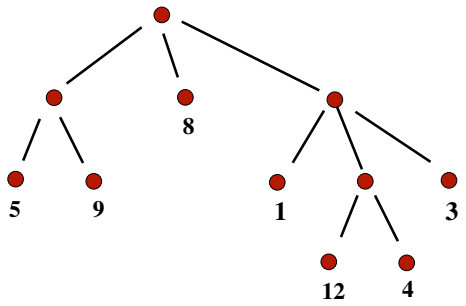

FIG. 1 – Un exemple d'arbre phylogénétique.

Introduction

Ce sujet explore quelques algorithmes pour la manipulation et la comparaison d'arbres phylogénétiques, utilisés en bio-informatique pour la classification des espèces.

Un arbre phylogénétique est un arbre enraciné $\mathcal{A}$ tel que :
- les feuilles (sommets sans fils) ont des étiquettes, et celles-ci sont toutes distinctes. On note $\mathcal{E}$ l'ensemble des étiquettes des feuilles de $\mathcal{A}$. Dans tous le problème, les étiquettes seront des entiers, et on pourra les comparer avec l'ordre usuel sur les entiers.
- les sommets internes (autres que les feuilles) ont au moins deux fils. Les différents fils d'un sommet interne ne sont pas ordonnés.

La figure 1 donne un exemple. La profondeur du sommet racine est 0, et la profondeur d'un autre sommet est celle de son père plus 1. Sur la figure, la profondeur de la feuille d'étiquette 1 est 2.

1 Préambule

Considérons la suite d'entiers (u_n) définie pour $0 \leqslant n \leqslant 10000$ par :

$$u_n = \begin{cases} \textit{votre } \mathtt{u_0} \textit{ (à reporter sur votre fiche)} & \text{si } n = 0 \\ (15\,991 \times u_{n-1}) \mod 65\,539 & \text{si } n \geqslant 1 \end{cases}$$

On pose $v_n = u_n$ et $w_n = u_{n+5000}$ pour $1 \leqslant n \leqslant 5000$

Question 1 *Quelle est la valeur de* ***a)*** u_{100}, ***b)*** v_{5000}, ***c)*** w_{5000} *?* ***d)*** *Les entiers* v_i, $1 \leqslant i \leqslant 5000$, *sont-ils tous distincts ?* ***e)*** *Les entiers* w_i, $1 \leqslant i \leqslant 5000$, *sont-ils tous distincts ?*
Quel est le nombre d'indices i, $100 \leqslant i \leqslant 3000$, *tels que* ***f)*** $v_i \mod 12 \leqslant v_{i+1} \mod 9$ ***g)*** $w_i \leqslant w_{i+3} \leqslant w_{i+7}$ *?*

Question 2 *Pour* $1 \leqslant i < j \leqslant 10000$, *l'intervalle* $[i..j]$ *de longueur* $j - i + 1$ *est dit croissant si* $u_i \leqslant u_{i+1} \leqslant u_{i+2} \leqslant \ldots \leqslant u_j$. ***a)*** *Quelle est la plus grande longueur d'un intervalle croissant ?*

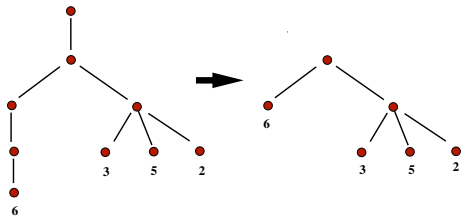

FIG. 2 – Suppression des noeuds internes n'ayant qu'un fils.

2 Construction

On génère l'arbre $\mathcal{A}_m$ comme suit à partir de la suite v_n :
- initialisation : $i \leftarrow 1, n \leftarrow 1$: on génère le sommet racine de l'arbre, et on lui attribue le numéro 1
- à une étape donnée, on traite le sommet i et on a déjà attribué n numéros ; on génère alors $n_i = \min((v_i \text{ div } 21847) + 1, m - n)$ fils pour le sommet i, de numéros $n+1$ à $n+n_i$. Ici v_i div 21847 est le quotient de la division euclidienne de v_i par 21847
- on itère en incrémentant i
- on s'arrête dès qu'on a attribué le numéro m
- on étiquette les feuilles : si le sommet numéro n est une feuille, on lui attribue l'étiquette v_n
- enfin, on supprime tous les sommets internes n'ayant qu'un fils (voir la Figure 2). Noter que les numéros des sommets internes servent uniquement pour décrire la construction, ce ne sont pas des étiquettes.

On génère de manière similaire l'arbre $\mathcal{A}'_m$ à partir de la suite w_n.

Dans toute la suite du problème, il vous sera demandé de répondre aux questions pour les arbres $\mathcal{A}_m$ ou $\mathcal{A}'_m$ pour $m \in \{10, 100, 1500\}$.

Question 3 *Quel est le nombre de sommets internes de l'arbre $\mathcal{A}_m$ pour* **a)** $m = 10$, **b)** $m = 100$, *et* **c)** $m = 1500$ *? Quel est le nombre de feuilles de l'arbre $\mathcal{A}'_m$ pour* **d)** $m = 10$, **e)** $m = 100$, *et* **f)** $m = 1500$ *? Quelle est la somme des profondeurs des feuilles de l'arbre $\mathcal{A}_m$ pour* **g)** $m = 10$, **h)** $m = 100$, *et* **i)** $m = 1500$ *?*

Question 4 *Combien de paires d'étiquettes $(e, e') \in \mathcal{E} \times \mathcal{E}'$ (e est une étiquette de $\mathcal{A}$ et e' une étiquette de $\mathcal{A}'$) vérifient-elles que $e + e'$ est divisible par 100 pour* **a)** $m = 10$, **b)** $m = 100$, *et* **c)** $m = 1500$ *?*

3 Écriture

En biologie, on décrit un arbre phylogénétique au format NH (de New Hampshire), à l'aide d'une chaîne de symboles. Ici un symbole est soit un entier, soit l'un des trois caractères '(', ',' ou ')'. La chaîne est obtenue par un parcours en profondeur de l'arbre $\mathcal{A}$, à partir de la racine. On part de la chaîne vide et à chaque visite on ajoute un symbole :
- quand on rencontre une feuille, on ajoute son étiquette

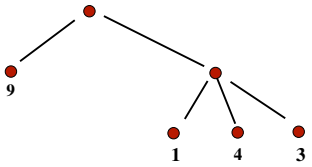

FIG. 3 – Un exemple de restriction.

- quand on rencontre un sommet interne :
(i) pour la première fois, on ajoute une parenthèse ouvrante '('
(ii) pour la dernière fois, on ajoute une parenthèse fermante ')'
(iii) pour toutes les (éventuelles) autres fois, on ajoute une virgule ','
On indice les symboles ajoutés de 1 à $L(\mathcal{A})$. Par exemple, une chaîne obtenue pour l'arbre de la figure 1 est $((5,9),8,(1,(12,4),3))$, et $L(\mathcal{A}) = 21$. On peut obtenir plusieurs chaînes pour décrire $\mathcal{A}$ en jouant sur l'ordre des fils, mais celles-ci ont toutes même longueur $L(\mathcal{A})$.

Question 5 *Que vaut $L(\mathcal{A})$ dans une écriture au format NH de $\mathcal{A}_m$ pour* ***a)*** $m = 10$, ***b)*** $m = 100$, *et* ***c)*** $m = 1500$ *?*

On définit la min-étiquette d'un sommet interne s de l'arbre comme la plus petite étiquette des feuilles qui sont dans le sous-arbre dont s est racine. On obtient une description au format NH unique, dite minimale, si à chaque visite d'un sommet interne, on visite ses fils dans l'ordre croissant de leur min-étiquettes. Pour l'arbre de la figure 1, on obtient ainsi $((1,3,(4,12)),(5,9),8)$.

Question 6 *Quels sont les cinq symboles d'indices* 10 *à* 14 *dans l'écriture au format NH minimale de $\mathcal{A}_m$ pour* ***a)*** $m = 10$, ***b)*** $m = 100$, *et* ***c)*** $m = 1500$ *? Quels sont les cinq symboles d'indices $L(\mathcal{A}'_m) - 20$ à $L(\mathcal{A}'_m) - 16$ dans l'écriture au format NH minimale de $\mathcal{A}'_m$ pour* ***d)*** $m = 10$, ***e)*** $m = 100$, *et* ***f)*** $m = 1500$ *?*

★ *Vous présenterez à l'oral l'algorithme que vous avez utilisé, ainsi que sa complexité.*

4 Restriction

Soit $\mathcal{F}$ un sous-ensemble de $\mathcal{E}$, l'ensemble des étiquettes de l'arbre $\mathcal{A}$. La restriction $\mathcal{A}(\mathcal{F})$ de l'arbre $\mathcal{A}$ aux feuilles dont les étiquettes sont dans $\mathcal{F}$ est l'arbre obtenu à partir de $\mathcal{A}$ en éliminant les feuilles de $\mathcal{E} \setminus \mathcal{F}$, et en détruisant les sommets internes n'ayant plus qu'un fils. La figure 3 donne la réduction de l'arbre de la figure 1 pour $\mathcal{F} = \{9,1,4,3\}$.
On définit $\mathcal{R}(p)$ comme la restriction de l'arbre $\mathcal{A}$ aux p feuilles dont les étiquettes sont les p plus petits éléments de $\mathcal{E}$ (de même pour $\mathcal{R}'(p)$ avec $\mathcal{A}'$).

Question 7 *Quel est le nombre de sommets internes de la restriction $\mathcal{R}(p)$ de l'arbre $\mathcal{A}_m$ pour* ***a)*** $m = 10$ *et* $p = 3$, ***b)*** $m = 100$ *et* $p = 10$, *et* ***c)*** $m = 1500$ *et* $p = 20$ *?*
Quelle est la longueur $L(\mathcal{R}'(p))$ dans l'écriture au format NH de de la restriction $\mathcal{R}'(p)$ de l'arbre $\mathcal{A}'_m$ pour ***d)*** $m = 10$ *et* $p = 3$, ***e)*** $m = 100$ *et* $p = 10$, *et* ***f)*** $m = 1500$ *et* $p = 20$ *?*

★ *Vous présenterez à l'oral l'algorithme que vous avez utilisé, ainsi que sa complexité.*

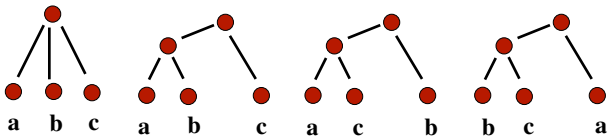

FIG. 4 – Les quatre réductions possibles pour le triplet de feuilles $\{f(a), f(b), f(c)\}$.

5 Comparaison

On veut étudier ici la comparaison d'arbres phylogénétiques qui ont le même ensemble d'étiquettes. Pour ce faire, on commence par modifier les étiquettes dans $\mathcal{A}$ et dans $\mathcal{A}'$. Soit $\mathcal{E} = \{e_1, e_2, \ldots, e_k\}$ l'ensemble trié des étiquettes de $\mathcal{A}$, avec $e_i \leqslant e_{i+1}$ pour $1 \leqslant i < k$. De même, soit $\mathcal{E}' = \{e'_1, e'_2, \ldots, e'_{k'}\}$ l'ensemble trié des étiquettes de $\mathcal{A}'$. Pour $1 \leqslant i \leqslant \min(k, k')$, on remplace les entiers e_i et e'_i par l'entier i. Enfin, si $k \leqslant k'$ on remplace $\mathcal{A}'$ par sa restriction $R'(k)$, et sinon on remplace $\mathcal{A}$ par sa restriction $\mathcal{R}(k')$.
On étudie deux mesures pour comparer $\mathcal{A}$ et $\mathcal{A}'$. La première mesure est rudimentaire, et concerne les distances des paires de feuilles. On note $f(e)$ la feuille d'étiquette e. La distance dans $\mathcal{A}$ d'une paire de feuilles $\{f(a), f(b)\}$ (avec $a \neq b$), est le nombre de sommets internes sur le chemin qui relie $f(a)$ et $f(b)$ dans l'arbre.

Question 8 *Combien y-a t-il de paires de feuilles dont les distances dans $\mathcal{A}_m$ et $\mathcal{A}'_m$ ne sont pas les mêmes, pour* ***a)*** $m = 10$, ***b)*** $m = 100$, *et* ***c)*** $m = 1500$ *?*

★ *Vous présenterez à l'oral l'algorithme que vous avez utilisé, ainsi que sa complexité.*

La deuxième mesure est plus fine et concerne les triplets de feuilles. La réduction de $\mathcal{A}$ au triplet de feuilles distinctes $\{f(a), f(b), f(c)\}$, avec $a < b < c$, peut être de quatre types, comme le montre la figure 4.

Question 9 *Combien y-a t-il de triplets de feuilles dont les réductions dans $\mathcal{A}_m$ et $\mathcal{A}'_m$ ne sont pas de même type, pour* ***a)*** $m = 10$, ***b)*** $m = 100$, *et* ***c)*** $m = 1500$ *?*

★ *Vous présenterez à l'oral l'algorithme que vous avez utilisé, ainsi que sa complexité.*