UNIVERSITÉ DE ROUEN
Faculté des Sciences
et des Techniques

MASTER PREMIÈRE ANNÉE
G.I.L.
Année 2008-2009

Algorithmique du texte

Contrôle continu — 17 novembre 2008

Documents, calculatrices, mobiles et portables interdits. Durée : 1 h 30. Le barême est donné à titre indicatif.

Exercice I *(8 points.)*

1 Dessinez les automates minimaux qui reconnaissent les langages de la forme A^*p où p est un préfixe du mot $x = \mathbf{aabaabaac}$. Ne faites figurer que les flèches dont la cible n'est pas l'état initial.

2 Dressez la table du bon préfixe du mot x.

3 Dressez la table du meilleur préfixe du mot x.

4 Le *délai* d'un algorithme séquentiel donné de recherche de mots est le nombre comparaisons maximum effectuées sur une lettre de texte. Quel est le délai pour l'algorithme qui utilise la table du bon préfixe dans le cas du mot x ?

5 Même question pour la table du meilleur préfixe.

Exercice II *(6 points.)*

On rappelle que deux mots x et y sont conjugués lorsqu'il existe deux mots u et v tels que $x = uv$ et $y = vu$.

Supposons qu'il existe un mot z non vide tel que $xz = zy$.

1 Montrez par récurrence que $x^k z = z y^k$ pour tout naturel k.

2 Soit n le naturel non nul tel que $(n-1)|x| \leqslant |z| < n|x|$. Montrez alors l'existence de deux mots u et v tels que $x = uv$, $z = x^{n-1}u$ et $vz = y^n$. Illustrez votre propos.

3 Déduisez-en que x et y sont conjugués.

Exercice III *(6 points.)*

Dans l'algorithme de Horspool, la dernière position sur le mot n'est pas nécessairement celle qui donne les plus longs décalages en moyenne.

Voici par exemple les décalages moyens que peuvent proposer les préfixes non vides du mot **AAAACGTA**, en notant p_a la probabilité pour que la lettre a apparaisse dans un texte :

préfixe	décalage moyen
A	1
AA	$p_{\mathtt{A}} \cdot 1 + (1 - p_{\mathtt{A}}) \cdot 2 = 2 - p_{\mathtt{A}}$
AAA	$p_{\mathtt{A}} \cdot 1 + (1 - p_{\mathtt{A}}) \cdot 3 = 3 - 2p_{\mathtt{A}}$
AAAA	$p_{\mathtt{A}} \cdot 1 + (1 - p_{\mathtt{A}}) \cdot 4 = 4 - 3p_{\mathtt{A}}$
AAAAC	$p_{\mathtt{A}} \cdot 1 + (1 - p_{\mathtt{A}}) \cdot 5 = 5 - 4p_{\mathtt{A}}$
AAAACG	$p_{\mathtt{C}} \cdot 1 + p_{\mathtt{A}} \cdot 2 + (1 - p_{\mathtt{C}} - p_{\mathtt{A}}) \cdot 6 = 6 - 4p_{\mathtt{A}} - 5p_{\mathtt{C}}$
AAAACGT	$p_{\mathtt{G}} \cdot 1 + p_{\mathtt{C}} \cdot 2 + p_{\mathtt{A}} \cdot 1 + (1 - p_{\mathtt{G}} - p_{\mathtt{C}} - p_{\mathtt{A}}) \cdot 7 = 7 - 4p_{\mathtt{A}} - 5p_{\mathtt{C}} - 6p_{\mathtt{G}}$
AAAACGTA	$p_{\mathtt{T}} \cdot 1 + p_{\mathtt{G}} \cdot 2 + p_{\mathtt{C}} \cdot 3 + p_{\mathtt{A}} \cdot 4 + (1 - p_{\mathtt{T}} - p_{\mathtt{G}} - p_{\mathtt{C}} - p_{\mathtt{A}}) \cdot 8$ $= 8 - 4p_{\mathtt{A}} - 5p_{\mathtt{C}} - 6p_{\mathtt{G}} - 7p_{\mathtt{T}}$

Si l'alphabet est réduit à ces quatre lettres et si ces lettres sont équibropables, soit $p_{\mathtt{A}} = p_{\mathtt{C}} = p_{\mathtt{G}} = p_{\mathtt{T}} = 1/4$, on a :

préfixe	décalage moyen
A	1
AA	$2 - 1/4 = 1{,}75$
AAA	$3 - 1/2 = 2{,}5$
AAAA	$4 - 3/4 = 3{,}25$
AAAAC	$5 - 1 = 4$
AAAACG	$6 - 1 - 5/4 = 3{,}75$
AAAACGT	$7 - 1 - 5/4 - 3/2 = 3{,}25$
AAAACGTA	$8 - 1 - 5/4 - 3/2 - 7/4 = 2{,}5$

Il s'ensuit que, dans ce cas, la meilleure position est 4 (en commençant la numérotation des positions à 0).

Proposez un algorithme efficace qui permet le calcul d'une meilleure position pour un mot non vide x et un tableau de probabilités (flottant) indicé sur l'alphabet donnés.