Université de Rouen	Année 2011–2012
UFR Sc. & Tech.	Master Info. GIL et ITA 1re année
Algorithmique du texte	Examen

Durée : 2 heures
Documents, calculatrices, téléphones mobiles, ordinateurs portables interdits (liste non exhaustive)

Exercice 1. *(1 point)* Soit $x = \texttt{bababbab}$. Donner un automate fini déterministe $\mathcal{A}$ tel que $\mathcal{L}(\mathcal{A}) = \mathit{Fact}(x)$ (utiliser un *trie*).

Exercice 2. *(6 points)* Effectuer la recherche de $x = \texttt{babbaa}$ dans $y = \texttt{abbaababbababbabbaa}$ avec :

a. un automate de recherche (donner l'automate reconnaissant A^*x et les états atteints lors de l'examen de chaque lettre de y) ;

b. l'algorithme de Knuth, Morris et Pratt (donner la table *meil-préf* et les comparaisons effectuées lors de chaque tentative) ;

c. l'algorithme de Boyer-Moore (donner les tables *bon-suff* et *dern-occ* et les comparaisons effectuées lors de chaque tentative).

Exercice 3. *(3 points)* Effectuer la recherche exacte de $x = \texttt{agac}$ dans $y = \texttt{agagatgac}$ avec une inégalité avec l'algorithme Shift-Or.

Donner les vecteurs S sur 4 bits pour les lettres `a`, `c`, `g` et `t`. Donner les vecteurs initiaux R^0 et R^1 sur 4 bits et pour chaque position sur y.

Exercice 4. *(3 points)* Construire la machine de Aho-Corasick pour l'ensemble fini de mots $X = \{\texttt{ababa}, \texttt{abba}, \texttt{bba}, \texttt{babb}\}$.

Utiliser cette machine pour rechercher les occurences des mots de X dans $y = \texttt{abbaababbababbabbaa}$ (donner les états atteints lors de l'examen de chaque lettre de y).

Exercice 5. *(3 points)* Compression

Rappel : les codes ASCII de `a`, `c`, `g` et `t` sont respectivement 97, 99, 103 et 116.
Coder `agagatgac` avec la méthode de Huffman.
Rappels : on code à la fois l'arbre et le texte à coder.

Exercice 6. *(4 points)* Soit $A = \{a_0, a_1, a_2, \ldots\}$ un alphabet ordonné. Soit u_n la suite définie comme suit :

- $u_0 = \varepsilon$;
- $u_i = u_{i-1}a_{i-1}u_{i-1}$ pour $i \geqslant 1$.

a. Donner u_1, u_2, u_3 et u_4.

b. Donner les longueurs de u_0, u_1, u_2, u_3, u_4. Généraliser pour u_i.

c. Donner le nombre de bords de u_0, u_1, u_2, u_3, u_4. Généraliser pour u_i.

d. Soit $D(\{x\}) = (A, Q, q_0, F, T)$ l'automate fini déterministe reconnaissant le langage $A^*\{x\}$ pour un mot $x \in A^*$.

On note $C(p)$ le nombre d'états distincts cibles de transition de source l'état p.
Formellement :

$$C(p) = \mathrm{card}\{q \in Q \mid (p, a, q) \in F\}\,.$$

Donner $C(u_i)$ pour $D(\{u_i\})$.