IFT3295 - Démo 7

6 novembre 2019

Arbre de suffixes

1. Expliquer comment on pourrait, à partir d'un arbre de suffixe, résoudre les problèmes suivants :
   (a) Vérifier si un mot p est un suffixe de T.
   Illustrez avec $T_1 = gatgaatgg$
   (b) Trouver la plus longue chaîne répétée dans T.
   Illustrez avec T_1
   (c) Trouver le plus petit suffixe (par ordre lexicographique) de T.
   Illustrez avec T_1
   (d) Trouver toutes les séquences $T_i \in \mathcal{T}$ contenant un mot p (et positions des occurrences).
   Illustrez avec $\mathcal{T} = \{T_1, catg, gact\}$
2. Soit la séquence $T = agcgxcga$, expliquez comment trouver la plus longue sous-chaîne de T tel que l'inverse S_r de S est également une sous-chaîne de T

   Élément de réponse 1 *Si S et S_r sont des sous-mots de T alors $S \in T$ et $S_r \in T$, $\equiv$ $S \in T_r$ et $S \in T$ Donc il faut construire l'arbre généralisé des suffixes pour T et T_r et trouver le noeud interne le plus profond avec au moins un suffixe de T et T_r.*

3. On appelle répétition maximale α de T, un sous-mot de T qui apparaît à au moins 2 positions différentes tel que si on rallonge vers la gauche ou vers la droite en ces positions, on n'a plus de répétition. Montrez que tous les noeuds internes ne sont pas des répétitions maximales. Illustrez avec $catgatnatgatp$.

4. Soit S_1, $|S_1| = n_1$ et S_2, $|S_2| = n_2$ deux séquences. On définit la plus longue extension commune $LCE(i, j)$ comme le plus long préfixe commun de $S_1[i...n_1]$ et $S_2[j...n_2]$ Expliquez comment trouver $LCE(i, j)$ pour tout i et j. Illustrez sur $S_1 = axababa$ et $S_2 = ababxbc$

5. Proposez un algorithme pour trouver les répétitions en tandems d'une séquence T, sachant qu'un préfixe répété en tandem est de la forme $\alpha\alpha$ pour un mot α donné. Illustrez sur $T = abaabaabb$

   Élément de réponse 2 *Pour chaque noeud interne v, et pour chaque feuille sous ce noeud correspondant au suffixe à la position i, vérifiez s'il existe une feuille avec un suffixe à la position j sous v tel que $i+depth(v) \geq j$ et $j > i$. Si oui alors répétition en tandem en $S[i, i+2*l]$ de taille $l = j - i$*

6. Proposez un algorithme pour énumérer tous les palindromes présents dans une séquence. Illustrez sur $T = abacdxdca$.

   Élément de réponse 3 *Si on construit un arbre de suffixes généralisé entre S et S_r, on peut retrouver tous les palindromes maximaux de centre i en recherchant LCE(i, $n - i + 1$) pour tout $i|1 < i < n$ (un palindrome ne peut être centré au premier ou dernier caractère).*

7. Il est possible de retrouver un repliement optimal (maximisant le nombre d'appariements G-C et A-U) d'une séquence S en une seule tige-boucle (sans boucles multiples, ni boucles internes ni renflements) à l'aide d'un arbre de suffixes. Expliquez comment utiliser cette méthode pour résoudre le problème. Illustrez sur $S = ccctcagg$.

   Élément de réponse 4 *Ce problème peut se ramener à la recherche du plus long facteur commun entre S et son complémentaire inverse S_c.*