BIN702 - Exercices sur les arbres de suffixe

Manuel Lafond

Exercice 1: Construisez l'arbre de suffixe généralisé pour les trois chaînes $operer, repere, ecopera$.

Exercice 2: On a mentionné sans démonstration que puisqu'un arbre de suffixe A possède n feuilles et que chaque noeud interne a au moins 2 enfants, alors A possède $O(n)$ noeuds au total. En fait, on peut montrer que A a au maximum $2n - 1$ noeuds et au maximum $2n - 2$ arêtes. Pouvez-vous le démontrer?

Exercice 3: Il est possible que deux séquences S et T aient plusieurs sous-chaînes communes de taille maximum. Donnez un algorithme qui retourne le **nombre** de plus longues sous-chaînes communes entre S et T. Quelle est la meilleure complexité que vous pouvez atteindre?

Exercice 4: Supposez que vous avez l'arbuste de suffixes A pour une séquence S. Donnez un algorithme pour transformer A en l'arbre de suffixes de S. Le temps devrait être proportionnel au nombre de noeuds de A.

Exercice 5: Montrez comment calculer la plus longue sous-chaîne commune entre k séquences $S_1, S_2, \ldots, S_k$ en temps $O(kn)$, où n est la somme des longueurs des séquences données.
Si vous aimez les défis, faites-le en temps : $O(n)$. (mais attention ce n'est vraiment pas facile, vous pourriez y passer des heures)

Exercice 6: Un palindrome-complément est une séquence S de taille $2n$ telle que $S[1..n]$ est égale à $S[n+1..2n]$ lue de droite à gauche après avoir fait les échanges de caractères $A \leftrightarrow T$ et $C \leftrightarrow G$. Par exemple, $AACTGCAGTT$.

Montrez comment trouver tous les palindrome-compléments maximaux dans une séquence T.

Exercice 7: La distance Hamming entre deux séquences S et T de même longueur est le nombre de positions auxquelles S et T diffèrent. Par exemple, $S = ACCT$ et $T = AGCT$ ont une distance Hamming de 1.
Soient S un mot à chercher et T une séquence plus longue. Montrez comment on peut trouver toutes les sous-chaînes T' de T de longueur $|S|$ telles que T' et S sont à distance Hamming de 0 ou 1. (on peut le faire en temps $O(|T| + |S| + occ)$)
De façon plus générale, montrez comment trouver toutes les sous-chaînes de T à distance Hamming au plus k de S, et ce en temps $O(k|T| + |S| + occ)$.
(indice: il y a une façon qui ne fait qu'utiliser des appels à PLPC(i, j))