

Examen en Production Automatique de Graphiques, Statistiques et Documents

Question 1 : k -mers

On appelle k -mer toute chaîne de k caractères. On nomme k -mers réalisés dans une chaîne de caractères toutes les sous-chaînes de k caractères consécutifs présentes dans cette chaîne. Ainsi pour la chaîne définie par $S="TACA"$, les deux seuls 3-mers réalisés sont "TAC" et "ACA".

Question 1.1

Donner l'ensemble de tous les 3-mers réalisés des deux chaînes $S1="TACAGT"$ et $S2="AGTCAG"$. On les donnera par ordre alphabétique avec leur nombre d'occurrences.

Question 1.2

Combien y a-t-il de k -mers possibles et distincts en tout que l'on peut former avec n lettres différentes ?

Et combien y a-t-il de k -mers possibles au maximum (éventuellement égaux) dans une chaîne de p caractères ?

Question 1.3

On suppose que le fichier "chaines.txt" contient des chaînes de caractères, à raison d'une chaîne par ligne. Expliquer comment on peut lire ce fichier pour construire un tableau qui contient tous les k -mers réalisés dans l'ensemble de ces chaînes pour k fixé puis donner le code d'un programme *PHP* qui lit ce fichier et qui produit le tableau $\$tkmer$ de tous les k -mers réalisés dans ces chaînes pour k fixé.

Question 2 : Graphes de *de Bruin**j*

On appelle *graphe de de Bruin**j* d'ordre k pour un ensemble de chaînes le graphe dont les sommets n_i sont les $k - mers$ réalisés par les chaînes. Il y a un arc du sommet n_i vers le sommet n_j si et seulement si les $(k - 1)$ derniers caractères de n_i sont égaux aux $(k - 1)$ premiers caractères de n_j .

Par exemple pour le *graphe de de Bruin**j* d'ordre $k = 3$ associé à la chaîne $E = \text{"TACA"}$, les deux seuls sommets sont $n_1 = \text{"TAC"}$ et $n_2 = \text{"ACA"}$ et le seul arc va de n_1 à n_2 .

Question 2.1

Donner le *graphe de de Bruin**j* d'ordre $k = 3$ associé aux deux chaînes $E = \text{"TACAGT"}$ et $E = \text{"TCAGA"}$. Indication : il y a sans doute 6 sommets et 6 arcs.

Question 2.2

On suppose que le fichier "chaines.txt" contient des chaînes de caractères, à raison d'une chaîne par ligne et qu'on a déjà construit le tableau $\$tkmer$ de tous les $k - mers$ réalisés dans ces chaînes pour k fixé.

Expliquer comment on peut construire par programme le tableau $\$tsom$ de tous les sommets du *graphe de de Bruin**j* associé au tableau $\$tkmer$.

Donner ensuite le code *PHP* qui construit le tableau $\$tsom$ de tous les sommets du *graphe de de Bruin**j* associé au tableau $\$tkmer$.

3. Production d'un *graphe de de Bruin**j*

On suppose qu'on dispose d'un *graphe de de Bruin**j* défini en *PHP* par :

- un tableau $\$tsom$ de chaînes de caractères correspondants aux sommets du graphe ;
- un tableau $\$tarc$ correspondant aux arcs du graphe.

Chaque élément de $\$tarc$ est lui-même un tableau de deux éléments numériques. Le premier correspond au numéro du sommet de départ de l'arc et le second correspond au numéro du sommet d'arrivée de l'arc. Par exemple, si l'arc d'indice 12 relie le sommet d'indice 35 au sommet d'indice 61, alors $\$tarc[12]$ est un tableau $\$t$ tel que $\$t[0]=35$ et $\$t[1]=61$.

Question 3.1

Expliquer comment on peut produire par programme un fichier au format ".dot" pour le *graphe de de Bruin* associé aux tableaux $\$tsom$ et $\$tarcs$.

Donner le code *PHP* qui produit le fichier "gdb.dot" du *graphe de de Bruin* associé aux tableaux $\$tsom$ et $\$tarcs$. Il s'agit bien sûr d'un graphe orienté. On ne demande aucun affichage.

Question 3.2

Donner le code *PHP* qui produit le fichier "gdb.png" du *graphe de de Bruin* associé au fichier "gdb.dot". On supposera que le programme *PHP* doit s'exécuter dans une page *Web*, que le fichier doit être produit dans le répertoire "/tmp/" du serveur Web sans qu'il n'y ait de problème d'accès concurrents. Là encore, on ne demande aucun affichage.

Question 3.3

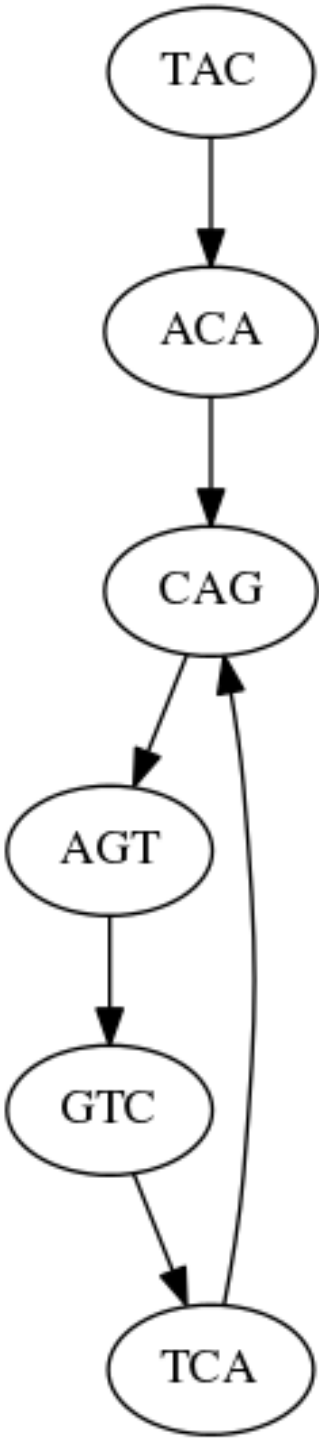
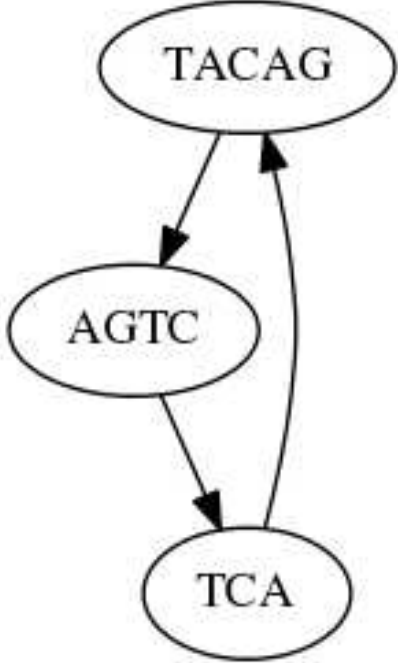
Afin de réduire le nombre d'arcs dans un *graphe de de Bruin*, on décide de regrouper les "suites linéaires" d'arcs consécutifs. Par exemple la suite des trois sommets et deux arcs "TAC" \rightarrow "ACA" et "ACA" \rightarrow "CAT" sont réduits au seul sommet "TACAT". Si un arc aboutissait à "TAC", il aboutit désormais à "TACAT". Si un arc partait de "CAT", il part désormais de "TACAT".

Expliquer comment, à partir des tableaux $\$tsom$ et $\$tarcs$ classiques d'un *graphe de de Bruin*, on peut produire par programme les tableaux $\$tsomR$ et $\$tarcsR$ correspondant aux sommets et aux arcs réduits. Ici, on ne demande pas le code *PHP*.

Question 4 : Discussion sur le format HTML

HTML, comme LaTeX, est un format de description de page. Selon vous, que manque-t-il vraiment à HTML pour être plus complet ?

Votre réponse devra mettre en évidence votre culture sur la qualité des documents produits par programme ainsi que votre recul et votre esprit de synthèse en matière de traitement de l'information et de programmation. Votre réponse devra faire 10 lignes au minimum, sans limite de maximum. On utilisera au moins 3 mots de 4 syllabes ou plus.

Exemple de <i>graphe de de Bruin</i>	Graphe réduit correspondant
 <pre> graph TD TAC((TAC)) --> ACA((ACA)) ACA --> CAG((CAG)) CAG --> AGT((AGT)) AGT --> GTC((GTC)) GTC --> TCA((TCA)) CAG --> TCA </pre>	 <pre> graph TD TACAG((TACAG)) --> AGTC((AGTC)) AGTC --> TCA((TCA)) TCA --> TACAG </pre>