HLIN608 Algorithmique du texte

TD Assemblage

1 Méthode gloutonne

Exercice 1 *Exécuter l'algorithme glouton sur la famille de mots :*
$F_1 =$ `ACCTGAG`
$F_2 =$ `TGCATTGC`
$F_3 =$ `GCAGACC`
$F_4 =$ `AGCAAT`
$F_5 =$ `CAATG`

Exercice 2 *Formalisez l'algorithme en précisant vos structures de données. Quelle est la complexité de cette méthode ?*

2 Méthode utilisant le graphe de chevauchement

Exercice 3 *Construire le graphe de chevauchement de l'exemple précédent. Donnez un algorithme de construction d'un graphe de chevauchement à partir d'une famille quelconque $F = \{F_1, \ldots, F_n\}$. Quelle est la complexité de votre algorithme ?*

Le problème SSP avec cette formalisation revient à rechercher dans le graphe de chevauchement un circuit hamiltonien de poids maximum. On considère le graphe comme étant complet, en ajoutant au besoin des arcs de poids 0 là où il n'y a pas d'arc dans le graphe de chevauchement. On appelle ce problème MAX-TSP (TSP pour Travelling Salesman Problem, le problème du voyageur de commerce). Ce problème est NP-complet…

Exercice 4 *Ecrire l'algorithme glouton qui donne un cycle hamiltonien. Appliquez-le au graphe de chevauchement de l'exemple. Quelle est la complexité de cet algorithme ? Cet algorithme ne vous rappelle-t-il rien ?*

Exercice 5 *(*) Ecrire l'algorithme heuristique appelé 2-interchange et appliquez-le à l'exemple. Quelle est la complexité de cet algorithme ?*

3 Graphe de De Bruijn vs. graphe des k-mers

Un des défauts du graphe de chevauchement est qu'il peut être très gourmand en espace mémoire. Pour pallier ce défaut, on passe à la structure du graphe des k-mers, qui permet des simplifications soit dans le graphe lui-même, soit dans la formulation du problème. On appelle k-mers d'un mot $w = x_1 \ldots x_n$ tout les facteurs de taille k de ce mot.

Exercice 6 *Combien y a-t-il de k-mers possible sur un alphabet de taille l ?*

Exercice 7 *Lorsque deux mots w_1 et w_2 se chevauchent avec un chevauchement de taille m, combien de k-mers partagent-ils dans ce chevauchement ?*

On part maintenant des hypothèses suivantes, destinées à se simplifier le problème (évidemment, on s'éloigne des cas réels) :

- on suppose le génome de départ circulaire,
- on suppose les reads sans erreurs,
- on suppose que l'on a exactement tous les k-facteurs apparaissant dans le génome parmi les k-mers extraits des reads.

Étant donné un ensemble R de mots de longeur k satisfaisant les conditions précédentes, on considère le graphe H_R suivant :

— Les sommets sont les mots de R

— On connecte le mot R_1 au mot R_2 si le $k-1$-suffixe de R_1 correspond au $k-1$-préfixe de R_2.

Exercice 8 *Construire le graphe des 3-mers construit à partir des mots de l'exemple de l'exercice 1. Matérialisez les chemins dans ce graphe qui correspondent aux mots de l'ensemble F. Retrouve-t-on tous les chevauchements observés dans le graphe de chevauchement ?*

Exercice 9 *Donnez un algorithme permettant de construire le graphe de k-mers dans le cas général. Quelle est sa complexité ? Que se passe-t-il quand k grandit ?*

Le graphe des k-mers est une forme particulière de graphe de De Bruijn. Le graphe de De Bruijn d'ordre k est construit à partir de tous les mots de taille $k-1$ sur un alphabet donné. Par exemple, pour un alphabet de taille 2, et $k = 3$, le graphe de De Bruijn B(2,4) est donné en Figure 1.

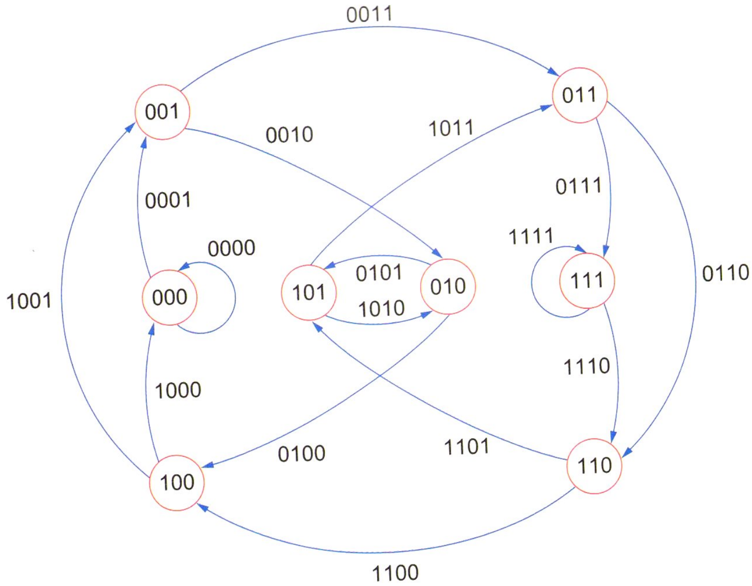

FIGURE 1 – Le graphe de De Bruijn B(2,4).

Exercice 10 *Pour un alphabet de taille n, combien y a-t-il de sommets dans le graphe de De Bruijn d'ordre k ? Quel est le degré entrant et le degré sortant de chaque sommet dans ce graphe ? Comparer B(4,4) (sur l'alphabet $\Sigma = \{A, C, G, T\}$ et avec le graphe de chevauchement obtenu à l'exercice 8, en terme de nombre de sommets (on ne demande pas de construire B(4,4)), sauf si vous vous ennuyez fortement).*