

Faculté des Sciences

Plus courts chemins dans un graphe pondéré

Projet réalisé dans le cadre
de la 1er Master en Sciences informatiques

réalisé par
Simon Olbregts



Faculté
des Sciences

Sous la direction de:
Véronique Bruyère

novembre 2014



Table des matières

Note préliminaire	2
1 Introduction	2
2 Implémentation	2
2.1 Structures de données	2
2.2 Construction du graph de chevauchement	2
2.3 Algorithme de l'alignement semi-global	3
2.4 Algorithme Greedy	4
2.5 Construction de la séquence	4

Note préliminaire

La rédaction de ce document est basé sur le livre *Introduction to algorithm* [?]

1 Introduction

2 Implémentation

2.1 Structures de données

Lors de la lecture des fichiers, les fragments sont insérés dans des objets *Fragment*.

Ces objets *Fragment* sont ensuite insérés dans des objets *Node* correspondant aux noeuds du graph de chevauchement. Ces objets en plus de contenir un fragment contiennent aussi un booléen *in* et un booléen *out* nécessaire pour le *Greedy*, une référence vers le noeud de son complémentaire inversé et une liste des noeuds dont les fragments sont inclus au fragment de ce noeud.

Pour appliquer l'algorithme *Greedy* sur le graph, il faut tout d'abord que ce graph possède des arcs. Pour ce faire, un objet *Edge* a été créé et contient le noeud de début et le noeud de fin de cet arc ainsi qu'un objet *Alignment* contenant les informations de l'alignement entre le fragment du noeud source et le fragment du noeud destination.

L'objet *Alignment* comprend donc : Le type d'alignement (préfixe, suffixe, inclu), le coût de l'alignement, l'indice du début de l'alignement, l'alignement du fragment du noeud source et l'alignement du fragment du noeud destination.

La dernière structure de données concerne le graph de chevauchement. Il contient une liste des noeuds ainsi qu'une liste des arcs.

2.2 Construction du graph de chevauchement

Cette partie de l'application est la plus gourmande en terme de ressources et consiste à créer les arcs du graph de chevauchement.

En effet, lorsque nous avons n fragments et que nous devons calculer les complémentaires inversés, $n * (n - 2)$ arcs sont à créer. Si les fragments complémentaires inversés ne sont pas à calculer, $n * (n - 1)$ arcs sont à créer.

Pour réduire le temps de calcul, nous appliquons donc le multithreading.

Deux contraintes sont à respecter lors de la construction des arcs. La première est qu'il ne faut pas créer d'arc entre un noeud et son complémentaire inversé. La seconde étant qu'il ne faut pas créer un arc ayant comme source et destination le même noeud. D'où le $n * (n - 2)$

La création des arcs introduit aussi l'algorithme de l'alignement semi-global.

2.3 Algorithme de l'alignement semi-global

Lors de la création des arcs, l'alignement entre les deux fragments concernés doit être calculé pour attribuer un coût à l'arc et pour déterminer le type d'alignement entre ces deux fragments.

Lors de la construction d'un arc de n_1 à n_2 , la matrice calculée et utilisée pour déterminer l'alignement entre les deux fragments sera également utilisée pour déterminer l'alignement de l'arc allant de n_2 à n_1 .

Pour construire l'alignement de n_1 à n_2 , on se positionne sur la valeur maximale de la dernière ligne de la matrice tandis que pour construire l'alignement de n_2 à n_1 , on se positionne sur la valeur maximale de la dernière colonne.

Une fois l'algorithme exécuté, nous obtenons deux chaînes de caractères représentant l'alignement entre les fragments :

ACTGCGTA***** : Fragment du noeud source (f1)

****CGTAAAA*** : Fragment du noeud destination (f2)

Grâce à ces deux chaînes de caractères, nous pouvons déterminer le type d'alignement qui sera nécessaire lors de l'exécution du *Greedy*

Si le premier et dernier caractère de la première chaîne est '*', alors f1 est inclu à f2.

Sinon, si le premier caractère de f1 est '*' et son dernier n'est pas '*', alors f1 est préfixe de f2.

Sinon, si le premier caractère de f1 n'est pas '*' et son dernier caractère l'est, alors f1 est suffixe de f2.

Sinon, si le premier caractère de f2 est '*' ou son dernier caractère est '*', alors f2 est inclu à f1.

Sinon les deux fragments sont égaux.

Il est également possible de déterminer la position du début de l'alignement qui sera utile lors de la construction de la séquence.

Dans l'exemple, le début de l'alignement se trouve à la 4^{ème} position dans la seconde chaîne.

Une fois les arcs construits et les caractéristiques des alignements connues, l'algorithme *Greedy* peut être appliqué sur les arcs du graph.

2.4 Algorithme Greedy

Avant d'appliquer le *Greedy*, les arcs doivent tout d'abord être triés par ordre décroissant en fonction de leur coût.

Pour ce faire, la méthode *Collection.sort()* est utilisé sur la liste des arcs.

Le problème lors de ce tri est que les arcs ayant le même coût ont des positions à chaque fois différentes lors de l'exécution de *Collection.sort()* ce qui fait que les arcs sélectionnés par *Greedy* varient à chaque exécution de l'application.

L'algorithme a dû être modifié pour gérer les fragments inclus.

Pour ce faire, lors du traitement d'un arc $n_1 \rightarrow n_2$, si n_1 est inclu à n_2 , le noeud n_1 est inséré dans la liste des fragments inclus à n_2 . De plus, il faut empêcher de prendre par la suite le noeud n_1 et son complémentaire inversé en mettant ses booléen *in* et *out* à *TRUE* et en insérant n_1 et son complémentaire inversé dans l'ensemble de n_2 . Cela est uniquement faisable si le noeud n_1 n'est pas déjà un noeud destination d'un arc choisit par *Greedy* aussi non il serait impossible de reformer la séquence.

Mainenant, si n_2 est inclu à n_1 , le noeud n_2 est inséré dans la liste des fragments inclus à n_1 . Il faut également empêcher de prendre par la suite le noeud n_2 et son complémentaire inversé en mettant ses booléen *in* et *out* à *TRUE* et en insérant n_2 et son complémentaire inversé dans l'ensemble de n_1 . Ici, cela est uniquement faisable si le noeud n_2 n'est pas déjà un noeud source d'un arc choisit par *Greedy*.

Une fois l'algorithme du *Greedy* terminé, la liste des meilleurs arcs est disponible pour construire la séquence.

2.5 Construction de la séquence

Le but est de créer un tableau contenant le nombre de fois que les nucléotides A, C, T, G et le nombre de gaps sont présents à la $i^{ème}$ position dans la séquence et ce pour chaque position. Si la séquence fait 1000 caractères, nous aurons donc 1000 tableaux. A noter que les tableaux contiendrons une

2 IMPLÉMENTATION

cellule permettant de savoir si à la position i , nous sommes en présence ou non d'un gap. Cela permettra de propager les gaps vers le bas.

Une fois tous les arcs traités, pour reconstruire la séquence, il suffira de regarder dans chaque tableau quel est le symbole le plus représenté et ce symbole sera celui à insérer dans la séquence à la position i . A noter que pour empêcher d'avoir le même nombre d'occurrences de plusieurs symboles à la position i et ainsi effectuer un choix aléatoire, on utilise le score de l'alignement (coût de l'arc) comme valeur à insérer dans le tableau.

La première étape consiste à trouver l'arc débutant la séquence c'est-à-dire, l'arc dont le noeud source est lié qu'au noeud destination et dont le noeud destination n'est lié qu'au noeud source de cet arc et à un noeud source d'un autre arc. Autrement dit, les booléens *in* et *out* du noeud source de cet arc doivent valoir respectivement *FALSE* et *TRUE* et ceux du noeud destination doivent valoir *TRUE* et *TRUE*.

Une fois cet arc trouvé, on le traite et une fois cet arc traité, on cherche l'arc dont le noeud source équivaut au noeud destination de l'arc venant d'être traité pour ainsi former une chaîne.

Lors de la construction de la séquence, deux éléments importants doivent être pris en compte : la propagation des gaps vers les fragments précédents et la propagation des gaps vers les fragments futurs.

En effet, lors de la gestion d'un arc $n_1 \rightarrow n_2$, si la chaîne de caractères du fragment de n_1 possède un gap à la i^{eme} position cela signifie que le caractère $i + 1$ de cette chaîne correspond au caractère à la position i dans la séquence. Il s'agit donc de décaler vers la droite le tableau correspondant au i^{eme} caractère de la séquence ainsi que tous les tableaux correspondant aux caractères se trouvant après le i^{eme} caractère et d'insérer un nouveau tableau à la position i . Un exemple de ce type de propagation de gaps est visible à la figure 1.

Un problème équivalent survient lorsque dans l'arc $n_1 \rightarrow n_2$, la chaîne de caractère du fragment de n_2 possède un gap à la i^{eme} position. Les futurs fragments à insérer dans la séquence devront faire correspondre leur i^{eme} caractère au $i + 1^{eme}$ caractère de la séquence. Un exemple de ce type de propagation de gaps est visible à la figure 2.

2 IMPLÉMENTATION

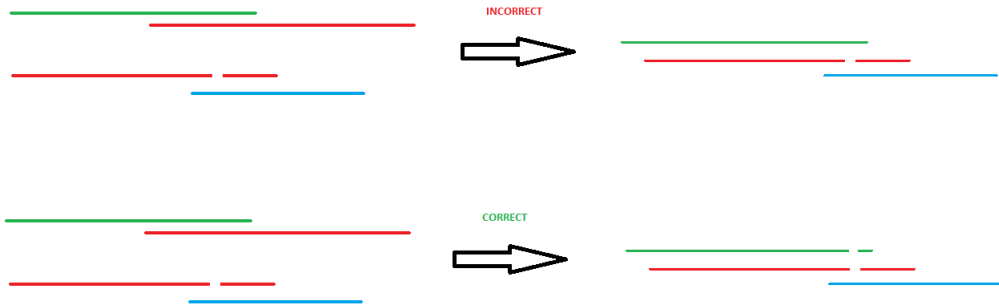


FIGURE 1 – Exemple de propagation d'un gap vers les fragments précédents

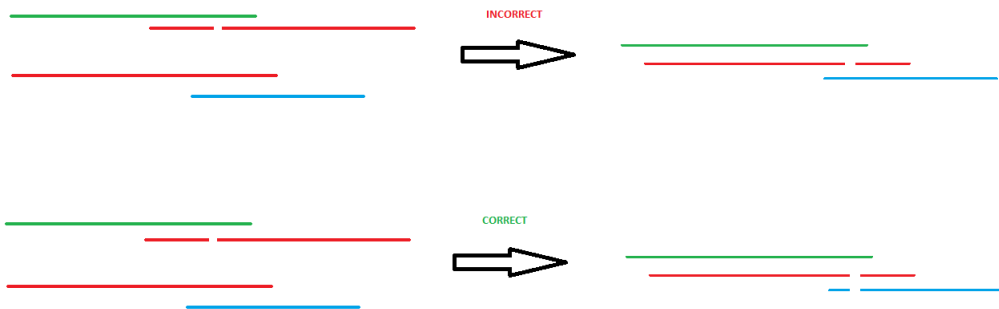


FIGURE 2 – Exemple de propagation d'un gap vers les fragments futurs