IFT2125: Introduction à l'algorithmique

Exercices pour les travaux pratiques

Algorithmes de programmation dynamique

1. Nous avons vu en classe l'algorithme de Floyd qui calcule la longueur des plus courts chemins entre chaque paire de sommets d'un graphe orienté. Comment faut-il modifier l'algorithme pour qu'en plus de retourner cette longueur, nous ayons l'information nécessaire pour connaître tous les sommets faisant partie de ces plus courts chemins?

2. **Problème 8.17, p.279, livre Brassard-Bratley:** Est-ce que l'algorithme de Floyd fonctionne sur des graphes ayant des arêtes de poids négatifs mais aucun cycle négatif? Justifiez votre réponse.

3. Calculer la distance d'édition et tous les alignements optimaux entre les séquences suivantes: $AAGCTAAG$ et $AGGAGGA$.

4. En biologie, il arrive que certains remplacements soient beaucoup plus fréquents que d'autres. Par exemple, le remplacement d'un nucléotide A par un nucléotide G est beaucoup plus plausible que le remplacement de A par C ou T. Ce phénomène s'explique tout simplement par les propriétés chimiques de ces nucléotides. En fait, les nucléotides se divisent en deux classes d'éléments ayant des propriétés chimiques similaires:

   $$\text{les purines} = \{A, G\}$$
   $$\text{et les pyrimidines} = \{C, T\}.$$

   Remplacer une purine par une purine est beaucoup plus probable, et donc beaucoup moins coûteux, que de remplacer une purine par une pyrimidine. La distance d'édition ne tient pas compte de ce fait. Pour obtenir des alignements plus près de la réalité biologique, on peut utiliser une distance d'édition généralisée qui tiendra compte de ce fait. Pour ce faire, nous allons définir une fonction de substitution f, telle que $f(a, b)$ représente le coût de substitution du symbole a par le symbole b. Finalement, pour favoriser l'opération de remplacement d'un symbole a par un symbole b plutôt que la suite d'opérations 'suppression de a' puis 'insertion de b', on pose $f(a, b) < 2f(a, -)$ et $f(a, b) < 2f(-, b)$. (Cette dernière inégalité est essentielle en biologie étant donné que la mutation d'un nucléotide en un autre nucléotide est beaucoup plus plausible que la perte d'un nucléotide en un endroit du génome suivi par l'insertion d'un nouveau nucléotide à ce même endroit.) Tenant compte de la discussion précédente, on peut définir la fonction de substitution f suivante sur cet alphabet:

f	A	C	G	T	-
A	0	1	1/3	1	2
C	1	0	1	1/3	2
G	1/3	1	0	1	2
T	1	1/3	1	0	2
-	2	2	2	2	∞

Si nous utilisons cette distance d'édition généralisée D_f pour calculer un alignement optimal entre 2 séquences on a un coût $f(a,-) = f(-,a) = 2$ pour chaque insertion ou suppression et donc les conditions initiales se modifient comme suit: $D_f(i,0) = 2 * i$ et $D_f(0,j) = 2 * j$. Pour les i et j strictement positifs, on a alors que

$$D_f(i,j) = \min \begin{cases} D_f(i-1,j) & + & 2 \\ D_f(i,j-1) & + & 2 \\ D_f(i-1,j-1) & + & f(x_i, y_j) \end{cases}$$

Calculer la distance d'édition généralisée et tous les alignements optimaux entre les même séquences qu'au numéro 3, i.e. les séquences suivantes: $AAGCTAAG$ et $AGGAGGA$. Sont-ils différents des alignements trouvés en 3? Que remarquez-vous?

5. Soit le texte $T = ACGTAACGAAT$ et le mot $P = AAC$. Calculer par programmation dynamique et en utilisant la distance d'édition, toutes les positions dans le texte où se termine une occurrence du mot P ayant au maximum une erreur. Donnez ces occurrences pour chaque position.