BIN702 - Série d'exercices #1 : l'alignement de séquences

Manuel Lafond

Exercice 1: Soient les séquences

$$S = ACAATCG$$
$$T = CTCATGC$$

Considérez les scores $+2$ pour un "match" (deux caractères identiques alignés), et -1 pour toute autre paire de caractères alignés (gaps et mutations).

a. Remplissez la table de programmation dynamique pour calculer l'alignement **global** entre S et T. Donnez ensuite un alignement global.

b. Remplissez la table de programmation dynamique pour calculer l'alignement **local** entre S et T. Donnez ensuite un alignement local.

Exercice 2: Une *sous-chaîne* d'une séquence S est un segment contigu de caractères de S. Par exemple si $S = ACCT$, alors les sous-chaînes possibles sont la chaîne vide, $A, C, C, T, AC, CC, CT, ACC, CCT, ACCT$.
Étant donné deux séquences S et T, le problème de la *plus longue sous-chaîne commune* demande de trouver une sous-chaîne C de S telle que C est aussi une sous-chaîne de T, et telle que C est de longueur maximum parmi les choix possibles.

a. Considérez l'algorithme naïf suivant: pour chaque sous-chaîne S' de S et chaque sous-chaîne T' de T, retenir S' et T' si $S' = T'$. Retourner la paire (S', T') retenue qui était de longueur maximum.

   Quelle la complexité en temps de cet algorithme?

b. Donnez un algorithme qui prend un temps $O(nm)$, où $|S| = n$ et $|T| = m$. (plus tard dans le cours, nous verrons que ceci peut se faire en temps $O(n + m)$)

Exercice 3: Soit S une séquence. Une *sous-séquence* de S est une séquence que l'on peut obtenir en supprimant des caractères de S. En d'autres termes, une sous-séquence est une séquence de caractères qui apparaissent dans le même ordre que dans S, mais pas nécessairement de façon contigüe dans S. Par exemple, si $S = ACGGTCA$, alors $AGTC$ ou encore $ACTCA$ sont des sous-séquences de S.
Étant donné deux séquences S et T, le problème de la *plus longue sous-séquence commune* demande de trouver une sous-séquence R de S telle que R est aussi une sous-séquence de T, et telle que R est de longueur maximum parmi les choix possibles.

a. Considérez l'algorithme naïf suivant: pour chaque sous-séquence S' de S et chaque sous-séquence T' de T, retenir S' et T' si $S' = T'$. Retourner la paire (S', T') retenue qui était de longueur maximum.

   Quelle la complexité en temps de cet algorithme? (ceci est une façon détournée de vous faire compter le nombre de sous-séquences d'une séquence)

b. Donnez un algorithme qui prend un temps $O(nm)$, où $|S| = n$ et $|T| = m$. (n'essayez pas de faire mieux que $O(nm)$, car ce n'est probablement pas possible)

Exercice 4: Donnez un algorithme qui retourne le **nombre** d'alignements globaux optimaux entre deux séquences S et T.

Exercice 5: L'alignement *semi-global* est comme l'alignement global, mais on ne pénalise pas une série de gaps en début et en fin de l'alignement de l'une ou l'autre des séquences. Par exemple, considérez des scores de $+2$ pour un "match" et -1 pour toute autre paire. Supposons qu'on ne pénalise pas les séries de gaps en début et fin de T seulement. Soient

$$S = GCGCGATTATAACGGGGG$$
$$T = ATTCTATA$$

Un bon alignement semi-global serait

```
GCGCGATTATA-ACGGGGG
-----ATTCTATA------
```

Puisqu'on ne pénalise pas les séries de gap en début/fin de T, le score serait de 10 (2×6 matches, $2 \times (-1)$ mutations/gaps).
Donnez un algorithme en temps $O(nm)$ pour effectuer l'alignement semi-global.
Si vous voulez vous amuser, considérez tous les cas où on ne veut pas pénaliser les gaps en début et/ou fin de S, et/ou les gaps en début et/ou fin de T.

Exercice 6: Considérez l'alignement avec une *pénalité affine* pour des suites de gap. C'est-à-dire, on vous donne des constantes h et s telles que le coût d'un gap de taille k est $h + sk$. Ici, h est le coût d'ouverture du gap, et s le coût d'extension du gap.
Étant donné deux chaînes S et T, h, s et une matrice de mutations, donnez un algorithme en temps $O(nm)$ qui calcule l'alignement global avec une pénalité affine.
(note : cette question n'est pas facile, mais encore une fois je recommande la programmation dynamique)