

Exercices autour de l'alignement 2 à 2

Exercice 1: Pour commencer

L'algorithme d'alignement global des deux séquences TGTCAAGT et ATTGCAGTAGC donne la table de programmation dynamique suivante :

		T	T	G	T	C	A	A	G	T
	0	-3	-6	-9	-12	-15	-18	-21	-24	-27
A	-3	-1	-4	-7	-10	-13	-13	-16	-19	-22
T	-6	-1	1	-2	-5	-8	-11	-14	-17	-17
T	-9	-4	1	0	0	-3	-6	-9	-12	-15
G	-12	-7	-2	3	0	-1	-4	-7	-7	-10
C	-15	-10	-5	0	2	2	-1	-4	-7	-8
A	-18	-13	-8	-3	-1	1	4	1	-2	-5
G	-21	-16	-11	-6	-4	-2	1	3	3	0
T	-24	-19	-14	-9	-4	-5	-2	0	2	5
A	-27	-22	-17	-12	-7	-5	-3	0	-1	2
G	-30	-25	-20	-15	-10	-8	-6	-3	2	-1
C	-33	-28	-23	-18	-13	-8	-9	-6	-1	1

Quelles sont les valeurs des paramètres pour une identité, une substitution, une insertion et une délétion qui ont été utilisées ? Construire un alignement optimal.

Exercice 2 : Alignements co-optimaux

On considère le problème de l'alignement global entre deux séquences, avec l'algorithme de Needleman et Wunsh. Des alignements sont *co-optimaux* s'ils ont le même score de similarité et que ce score est maximal.

Question 1. Pour le jeu de scores -2 pour une insertion ou une délétion, -1 pour une substitution et +1 pour une identité, construisez un exemple de deux séquences, acceptant au moins deux alignements optimaux distincts.

Question 2. Ecrire un algorithme qui à partir de la matrice de programmation dynamique de l'algorithme de Needleman et Wunsh détermine le nombre d'alignements co-optimaux.

Exercice 3 : Alignement semi-global

L'alignement *semi-global* entre deux séquences U et V est le meilleur alignement global entre un préfixe de U et un suffixe de V , ou entre un suffixe de U et un préfixe de V . Autrement dit, l'alignement semi-global construit un alignement global où les gaps de début et de fin ne sont pas pénalisés.

Donner l'algorithme détaillé (matrice de programmation dynamique et obtention du score de l'alignement) pour l'alignement semi-global.

Exercice 4 : La plus courte super-séquence

Soient U et V , deux séquences. Une *super-séquence* de U et V est un mot S qui contient les lettres de U et V dans le même ordre. Autrement dit, U peut être obtenu à partir de S en appliquant une suite de délétions, et V également. Nous nous intéressons dans cet exercice à la plus courte super-séquence. Par exemple, si $U = \text{ATCGCC}$ et $V = \text{ATTGAC}$, la plus courte super-séquence est $S = \text{ATTGACC}$.

Question 1. Montrer que le meilleur alignement global de U et V sans substitutions fournit la plus courte super-séquence.

On note α le score d'un match (deux lettres identiques), β le coût d'une insertion ou d'une délétion et γ le coût d'une substitution.

Question 2. Quelles valeurs de paramètres choisir pour α , β et γ afin que l'alignement global optimal soit toujours un alignement sans substitutions ?

Question 3. On considère maintenant le problème de la plus longue sous-séquence commune.

Données : deux séquences U et V
Résultat : le plus long mot w qui soit une sous-séquence de U et de V à la fois
(c'est-à-dire que U et V sont des super-séquences de w).

Montrez que l'on peut résoudre ce problème avec un algorithme d'alignement. Lequel ? Avec quel système de score ?

Question 4. Enfin, la dernière question concerne le problème du plus long facteur commun :

Données : deux séquences U et V
Résultat : le plus long mot w qui soit un facteur de U et de V à la fois
(il existe des mots u_1, u_2, v_1, v_2 tels que $U = u_1wu_2$ et $V = v_1wv_2$).

Quel algorithme d'alignement utiliser ? Avec quel système de score ?