Alignement des séquences nucléiques

Algorithme de Needleman et Wunsch

L'algorithme de Needleman-Wunsch est un algorithme qui effectue un alignement global maximal entre deux séquences nucléiques et/ou protéiques. Son expression est de la forme :

$$S(i,j) = \text{Max} \begin{cases} S(i-1, j-1) + s(i,j) \\ S(i-1, j) \\ S(i, j-1) \end{cases}$$

j : numéro des cases dans l'axe des abscisses

i : numéro des cases dans l'axe des ordonnées

S(i,j) = score de **Needleman-Wunsch** dans la case i,j

s (i,j)= score élémentaire dans la case i,j

NB : dans le cas des acides nucléiques, le score élémentaire (s) varie en fonction de la matrice à utilisé (matrice d'identité, matrice de transition transevrion…etc. par exemple entre deux nucléotides identiques s=1 si on utilise la matrice d'identité et 3 si on utilise la matrice de transition transversion).

L'alignement entre deux séquences nucléiques s'effectue en quatre étapes complémentaires :

1ère étape : Construction de la matrice initiale :

Dans un premier temps, les deux séquences S1 et S2 sont insérées dans une matrice dite **initiale** de sorte que S1 soit à l'horizontal (axe des abscisses) et S2 à la verticale du tableau (axe des ordonnées). Puis, les cases de cette matrice doivent être remplies par des scores élémentaires selon la matrice choisie (matrice d'identité ou de transition transversion).

Exemple :

Soit les deux séquences :

S1 : TAAGTCCG

S2 : TACGTACG

Remplir la matrice initiale en utilisant la **matrice d'identité** (1 dans le cas d'identité des deux nucléotides des séquences S1 et S2 et 0 si non).

	T	A	A	G	T	C	C	G
T	**1**	0	0	0	**1**	0	0	0
A	0	**1**	**1**	0	0	0	0	0
C	0	0	0	0	0	**1**	**1**	0
G	0	0	0	**1**	0	0	0	**1**
T	**1**	0	0	0	**1**	0	0	0
A	0	**1**	**1**	0	0	0	0	0
C	0	0	0	0	0	**1**	**1**	0
G	0	0	0	**1**	0	0	0	**1**

2ème étape : construction de la matrice transformée

Dans un deuxième temps, il faut créer une deuxième matrice à i+2 colonnes et j+2 lignes, dans laquelle la 1ère ligne et la 1ère colonne seront initialisées à zéro comme suit :

		j	1	2	3	4	5	6	7	8
			T	A	A	G	T	C	C	G
i		0	0	0	0	0	0	0	0	0
1	T	0								
2	A	0								
3	C	0								
4	G	0								
5	T	0								
6	A	0								
7	C	0								
8	G	0								

C'est à ce niveau qu'on applique l'algorithme de **Needleman-Wunsch** afin d'aligner les deux séquences S1 et S2 :

$$S(i,j) = \text{Max} \begin{cases} S(i-1, j-1) + s(i,j) \\ S(i-1, j) \\ S(i, j-1) \end{cases}$$

Si on commence par la case (1,1), l'algorithme est appliqué comme suit :

$$S(1,1)= \text{Max} \begin{cases} S(0,0) + s(1,1) = 0 + 1 = \mathbf{1} \\ S(0,1) = 0 \\ S(1,0) = 0 \end{cases}$$

Le maximum entre 1, 0 et 0 c'est bien 1. Donc le score à mettre dans la case i= 1 et j=1c'est **1.**

		T	A	A	G	T	C	C	G
	0	0	0	0	0	0	0	0	0
T	0	**1**							
A	0								
C	0								
G	0								
T	0								
A	0								
C	0								
G	0								

L'application de l'algorithme de **Needleman-Wunsch** permet de remplir la matrice transformée comme suit :

		T	A	A	G	T	C	C	G
	0	0	0	0	0	0	0	0	0
T	0	1	1	1	1	1	1	1	1
A	0	1	2	2	2	2	2	2	2
C	0	1	2	2	2	2	3	3	3
G	0	1	2	2	3	3	3	3	4
T	0	1	2	2	3	4	4	4	4
A	0	1	2	3	3	4	4	4	4
C	0	1	2	3	3	4	5	**5**	5
G	0	1	2	3	4	4	5	5	**6**

3ème étape : traceback (traçage en arrière)

le parcours de la matrice transformée commence par le plus haut score, vers le plus haut score parmi les trois cases (i-1, j-1) (i-1,j) et (i, j-1) et ainsi de suite jusqu'à la case (1,1). Dans cet

exemple, on commence par la case (8,8) ayant le plus haut score = 6. Dans ce cas, les scores des 3 cases (7,7), (7,8) et (8,7) sont les mêmes (5). Dans ce cas, le parcours en diagonal est recommandé (vers la case (7,7)). Le parcours qui en résulte est le suivant :

		T	A	A	G	T	C	C	G
	0	0	0	0	0	0	0	0	0
T	0	**1**	1	1	1	1	1	1	1
A	0	1	**2**	2	2	2	2	2	2
C	0	1	2	**2**	2	2	3	3	3
G	0	1	2	2	**3**	3	3	3	4
T	0	1	2	2	3	**4**	4	4	4
A	0	1	2	3	3	**4**	4	4	4
C	0	1	2	3	3	4	**5**	**5**	5
G	0	1	2	3	4	4	5	5	**6**

4ème étape : génération de l'alignement et calcul des scores

En suivant le parcours de la matrice transformée tracé précédemment, les nucléotides en diagonal représentent soit appariement (identité **|**) ou une substitution (**:**).

S1	**T**	**A**	**A**	**G**	**T**	**—**	**C**	**C**	**G**
	|	**|**	**:**	**|**	**|**	*****	**|**	*****	**|**
S2	**T**	**A**	**C**	**G**	**T**	**A**	**C**	**—**	**G**

Le trou (**—**) retrouvé entre les nucléotides T et C de la séquence S1 est un GAP ou InDel (représenté en * lors de l'alignement): il signifie qu'à ce point, la séquence S1 a subi une mutation par **DEL**étion au cours de la quelle un nucléotide est perdu par nécessité évolutive et d'adaptation à l'environnement ; en même temps, ce nucléotide A est conservé dans la séquence S2 (à la 6ème position). Comme on peut supposer que c'est la séquence S2 qui a subi une mutation par **IN**sertion du nucléotide A par nécessité adaptative. Dans un cas ou dans l'autre une des deux séquences a subi une mutation (**IN**sertion ou **DEL**étion) ; ce point est appelé **INDEL** (en anglais **gap**) pour dire INSERTION dans la séquence S2 ou DELETION dans la séquence S1. La même interprétation concerne le deuxième gap retrouvé 8ème

position : il s'agit d'une délétion du nucléotide C dans la séquence S2 ou de l'insertion de C dans la séquence S1.

Il est également important de signaler qu'une substitution a eu lieu au niveau de la troisième position où le nucléotide A de la séquence S1 a été substitué par un C dans la séquence S2 et ce par nécessité évolutive et d'adaptation à l'environnement.

Calcul des scores :

Le pourcentage d'identité (%id) = (nombre d'identités / taille de la séquence après alignement)

Dans cet exemple %id= (6/9)* 100 = 66.66%

Le pourcentage des gaps : = (nombre gaps / taille de la séquence après alignement)*100

= (2/9)*100= 22.22%

Le pourcentage des substitutions : (nombre de substitutions / taille de la séquence après alignement)*100

= (1/9)*100= 11.11%.