**2 Algorithmes**

**1 -Intérêt de l’analyse des séquences**

Actuellement il y a des milliards de séquences génétiques et protéiques.

Seulement peu d'entres elles ont des structures et des fonctions connues. Les autres demeurent encore ambiguës. Mais si jamais on réussi à les aligner, on peut prétendre qu'ellessont similaires par conséquent elles ont les mêmes structures et des fonctions proches.

On peut comparer soit des séquences d'ADN soit des séquences de protéines personnelles (qu'on aura trouver et qu'on voudra déterminer) à des séquences dont on connaît déjà l'origine et les fonctions et qu'on trouve dans des bases de données accessibles à tous dans le but d'établir des homologies structurales et fonctionnelles entres elles.

Attention : On sait qu'il peut toujours y avoir des mutations, des adaptations et des

évolutions ; ce ne sont donc jamais des comparaison exactes.

**2- Les différents types d’algorithmes**

On distingue deux types d'algorithmes se basant sur l'alignement des séquences :

**A-Les algorithmes d'alignement globaux :**  (Needlemann & Wunsch, 1970)

Ils sont qualiés de globaux car ils considèrent l'ensemble des éléments des séquences.

Des insertions devrant être faites dans la séquence la plus courte pour arriver à les

aligner dans le cas où elles auraient des longueurs di\_érentes (exemple de programme

basé sur cette méthode : FASTA).

**B- Les algorithmes d'alignement locaux :** (Smith & Waterman, 1981)

On cherche à trouver les zones les plus similaires entre deux séquences (c'est à dire

deux sous-séquences homologues) sans prédétermination de longueurs. Il ne porte

pas sur la totalité des séquences comme le précédent mais seulement une partie de

chacune d'elles. (ex : BLAST est un programme qui adopte cet algorithme)

Aussi doit on signaler que des fois il peut s'avérer nécessaire de faire des délétions ou des insertions de longueurs variables à des positions différentes au sein des séquences pour en optimiser la comparaison. On parle alors de gap. Une pénalité s'applique suivant que l'alignement s'est fait avec ou sans gap in\_uençant ainsi sur le score qui intervient dans le choix du meilleur alignement. C'est pourquoi il faut introduire un minimum de gap.

Les différentes implémentations de ces algorithmes font appel à la programmation dynamique.

**3-La programmation dynamique :**

1. **L’alignement des séquences nucléiques**

L’alignement, comme nous allons le voir dans les exemples suivants, permet de mesurer la similitude entre les séquences. S’il ya similitude, cela signifie qu’il est possible que les deux séquences présentent la même fonction biologique, ou du moins les deux séquences présente une structure fortement similaire. Ce type d’information est nécessaire dans la mesure où, généralement, nous avons à faire à une séquence inconnue. Sa comparaison avec des séquences de structure et de fonction connues permet de tirer un maximum d’informations quant à la structure et la fonction de la séquence inconnue. Dans certains cas, on peut même confirmer si la séquence inconnue est un gène ou une portion de gène après l’avoir aligné avec des séquences de structure génique connue (régions codantes : codons d’initiation et de terminaison, sites d’épissage, zones de fixation des ribosomes).

* **L'algorithme de Needleman et Wunsch :**

Il permet de réaliser un alignement global entre deux séquences nucléiques. Son expression est de la forme :

 

ࡿ**Exemple :**

Supposons que nous désirons calculer un alignement global des deux séquences suivantes de taille m et n respectivement:

S1 = TAAGTCCG m=8 et S2 = TACGTACG n=8

**Remarque** : Ici, les deux séquences sont de même longueur (8 résidus chacune). On peut calculer un alignement entre deux séquences de tailles inégales.

Pour calculer l’alignement entre les deux séquences S1 et S2, quatre étapes sont nécessaires :

**Etape 1 :** Calcul de la matrice initiale

Il s’agit d’insérer les deux séquences S1 et S2 dans une matrice de sorte que S1 soit à l’horizontal et S2 à la verticale du tableau, puis remplir les cases par 1 (identité des deux nucléotides de S1 et de S2) ou 0 (sinon) :



**Etape 2 :** Calcul de la matrice transformée : Initialisation de la matrice

Construisons une nouvelle matrice à m+2 colonnes et n+2 lignes, dans laquelle la 1ère ligne et la 1ère colonne seront initialisées par un zéro :



L’application de l’algorithme de Needleman et Wunsh permet de remplir les cases de cette matrice. Le résultat est le suivant :



**Etape 3 :** Parcours de la matrice transformée

Parcourir la matrice transformée depuis le plus haut score calculé (ici s=6) jusqu’au score le plus petit (ici s=1)



**Etape 4 :** Alignement des deux séquences et calcul de score



\* Le score global de cet alignement est de 7.( les nucléotides identique )

\* Le pourcentage de l’identité entre les deux séquences S1 et S2 est :

 %id = (7/9)\*100=77,78%

\* Le trou retrouvé entre les nucléotides T et C de la séquence S1 est un GAP ou InDel : il signifie qu’à ce point, la séquence S1 a subi une mutation par **Délétion** au cours de la quelle le nucléotide A est perdu par nécessité évolutive et d’adaptation à l’environnement ; en même temps, il est conservé dans la séquence S2 (à la 6ème position face au gap de S1).

Comme on peut supposer que c’est la séquence S2 qui a subi une mutation par **Insertion** du nucléotide A par nécessité adaptative. Dans un cas ou dans l’autre une des deux séquences a subi une mutation (**In**sertion ou **Dél**étion) ; ce point est appelé **INDEL** 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 l’insertion de C dans la séquence S1.

**\* Une Série de TD.**