# TD1 – Alignements de séquences

Nous allons ici approfondir certains des concepts d'alignement de séquences que nous avons abordés en cours.

Nous commencerons par revoir « le dotplot » qui permet d'établir de façon visuelle des similarités entre deux séquences.

Le « dotplot » n'étant pas une méthode permettant de déterminer le ou les « meilleurs alignements possibles », nous utiliserons ensuite des logiciels implémentant les algorithmes « exacts » de Needleman et Wunsch puis de Smith et Waterman, capables de produire respectivement le meilleur alignement global (les meilleurs alignements locaux). Nous nous interrogerons par ailleurs sur le type de méthode à utiliser quand on souhaite comparer deux séquences, *i.e.*, comparaison globale ou locale ?

## I. Comparaison de deux séquences au moyen d'un « dotplot »

Si l'on souhaite comparer deux séquences entre elles, le « dotplot » est un outil tout à fait approprié et simple d'utilisation. Nous utiliserons ici le logiciel dotpath pour effectuer les « dotplots ».

dotpath (http://www.bioinformatics.nl/cgi-bin/emboss/dotpath) permet de dessiner un dotplot avec une taille de mot fixée (dans le cours, la taille du mot était, « en quelque sorte », fixée à 1, *i.e.*, intersection ligne/colonne).

Lire rapidement le manuel de dotpath (http://www.bioinformatics.nl/cgi-bin/emboss/help/dotpath) afin de saisir les deux idées principales à la base de l'algorithme.

**Résumer ces deux idées.**

Récupérer dans la banque « nucleotide » (https://www.ncbi.nlm.nih.gov/nucleotide/) les séquences nucléiques U09584 et NM_019704 et les sauvegarder au format FASTA.

Utiliser le logiciel dotpath en faisant varier la taille de la fenêtre (Word size), 2 dans un premier temps puis 4, 8, 10 et finalement 20, tout en conservant à chaque fois l'option 'Display the overlapping matches' (no).

Refaire la même chose avec l'option 'Display the overlapping matches' (yes).

**Les résultats obtenus sont-ils en accord avec votre résumé ?**

Observer le résultat pour la taille de fenêtre 10 et l'option 'Display the overlapping matches' (no). Répondre alors aux questions suivantes :

**Les séquences se ressemblent-elles** ? **Si oui, quels sont le ou les segments communs** ? **Dans ce cas, préciser les bornes du ou de ces segments communs.**

## II. Comparaison de deux séquences : alignement global et alignement local

La représentation visuelle des similarités par un « dotplot » ne permet généralement pas de déterminer précisément les bornes d'un alignement. Par ailleurs, un « dotplot » ne garantit aucunement que les similarités trouvées sont les meilleurs possibles. Pour obtenir ce type de garanties, il faut utiliser les algorithmes de Needleman et Wunsch (comparaison globale, *cf*. cours) ou de Smith et Waterman (comparaisons locales, *cf*. cours).

Aligner avec stretcher (alignement global, paramètre par défaut) (http://emboss.bioinformatics.nl/cgi-bin/emboss/stretcher) le " peptide mystère" stocké dans le fichier SEQUENCE_MYSTERE.doc et la séquence protéique stockée dans le fichier SEQUENCE_PL6_HUMAN.doc (*cf.* moodle).

**Quelles sont les parties (positions dans les deux séquences) correctement alignées ? Quelle partie de la séquence mystère n'est pas alignée (*i.e.*, ne trouve pas de " correspondance " avec la séquence pl6) ?**

Lorsque les séquences n'ont pas la même taille (lorsque l'une, par exemple, est beaucoup plus longue que l'autre), mais pas uniquement, il est souvent maladroit de les comparer dans leur globalité. On utilise alors de préférence dans ce cas des algorithmes d'alignements locaux.

Aligner les deux séquences précédentes en utilisan t matcher (http://emboss.bioinformatics.nl/cgi-bin/emboss/matcher) qui implémente un algorithme d'alignement local.

Choisir (paramètre utilisateur) de visualiser les 4 meilleurs alignements locaux (number of alternative matches).

**Comparer l'alignement global avec le résultat précédent. Que constatez-vous ?**

**Dans ce cas, est-ce la taille des séquences qui pose un problème à l'alignement global ?**