TP8 – Fin du Dot-plot

Nous allons terminer le “ Dot-plot ” en introduisant de la “ souplesse ” à la version réalisée au TP7. Cette souplesse est nécessaire car si l’introduction d’une fenêtre permet bien de nous “ débarrasser ” des « points » qui viennent parasiter/bruiter la lecture du “ Dot-plot ” (et donc l’émergence des diagonales qui marquent le partage d’une similarité entre les deux séquences considérées), l’égalité stricte est trop stringente ce qui a tendance à “ casser ” (ou rompre) les diagonales.

Pour introduire cette souplesse, en lieu et place d'une comparaison stricte, nous calculerons le pourcentage d'identité entre les deux « mots » des séquences « x » et « y » que nous comparons à chaque itération (la taille de ces deux mots correspond à la longueur de la fenêtre rentrée par l'utilisateur). Si ce pourcentage dépasse un seuil fixé par l'utilisateur, un point sera affiché.

Avant d’introduire cette nouvelle fonctionnalité dans le code du « Dot-plot », l’exercice 1 va nous permettre de préciser ce qui vient d’être dit.

Exercice 1 :

Ecrire un programme permettant de calculer le % d’identité entre deux séquences de même longueur.

Exemple : la séquence "ATGCATGCAT" a 80% d’identité avec la séquence "TTGCATTCAT" :

ATGCATGCAT
| | | | | | | |
TTGCATTCAT

Nous voyons bien 8 nucléotides identiques (même position) dans les 2 séquences sur les 10 nucléotides, soit 80% d’identité.

L’utilisateur devra saisir les 2 séquences à comparer. En sortie, on affichera le % d’identité entre les 2 séquences.

Exercice 2 : suite du Dot-plot

Introduire, dans cette version du « Dot-plot », la souplesse telle que nous l'avons définie à l'exercice précédent. Un “ seuil d’identité ” (pourcentage) sera spécifié par l’utilisateur ; si le pourcentage d’identité entre les deux « mots » des séquences « x » et « y » comparés est supérieur au seuil, un “ point ” sera affiché (rien dans le cas contraire).

Exercice 3 : fin du Dot-plot

Avant d'utiliser votre programme sur de vraies séquences, le modifier afin qu'il soit en mesure de lire des fichiers au format fasta.

Exercice 4 : Test “ grandeur nature ”

Récupérer les séquences (format “ fasta ”) des deux protéines suivantes :

http://www.uniprot.org/uniprot/P0AA25.fasta

http://www.uniprot.org/uniprot/P14949.fasta

Faites tourner votre “ Dot-plot ” en faisant varier la taille de la fenêtre et le seuil d’identité afin de faire apparaître les similarités pertinentes entre les deux séquences (si elles existent).

Exercice 5 : Test “ grandeur nature ” (suite)

Récupérer les trois séquences sur moodle. Faites tourner votre “ Dot-plot ” en faisant varier la taille de la fenêtre et le seuil d’identité afin de faire apparaître les similarités pertinentes entre ces différentes séquences (si elles existent). Les résultats obtenus sont-ils en accord avec l'annotation des séquences (*cf*. partie « info » du format fasta) ?