TD 2 : Alignement de séquences

***Alignement de séquences* : une affaire de distance entre chaîne de caractères**

Etude préliminaire. *Valeurs discrètes.*

Soient les deux individus suivants correspondant à des séquences ADN :
X = AGGGTGGC et Y = AGGCGTAA

1. Dans quel espace « vivent » les points X et Y ? A quelle dimension ? Si on code A=0, G=1, C=2 et T=3, quelle est la distance euclidienne d(X,Y) ? Cela a-t-il un sens en terme de similitude entre les séquences ADN X et Y ? Expliquez notamment en comparant d(A,G) et d(A,T) ?

2. En bioinformatique, la comparaison de séquences ADN deux à deux doit permettre de trouver des homologies c'est-à-dire comment les séquences ont muté à travers les espèces durant l'évolution. Pour cela, on a regroupé les séquences ADN par famille (clustering). A l'intérieur de ces familles, on a réalisé des mesures statistiques. On s'est aperçu que les mutations trans-nucléotides sont déséquilibrées à l'intérieur de famille de séquences appariées. Des matrices de substitution sont utilisées en guise d'heuristiques à la recherche de séquences homologues. Soit par exemple, la matrice de pondération suivante (s'inspirant des matrices de substitution type BLOSUM62 utilisées en bioinformatique) :

$$S=\begin{pmatrix} & A & C & G & T \\ A & 0 & 1 & 0.01 & 1 \\ C & 1 & 0 & 0.01 & 0.01 \\ G & 0.01 & 0.01 & 0 & 1 \\ T & 1 & 0.01 & 1 & 0 \end{pmatrix}$$

   Le coefficient 0.01 à la croisée de la ligne G et de la colonne A traduit la très grande fréquence observée de ce type de substitution dans les séquences déjà appariées. A l'inverse, un coefficient 1 indique une très grande rareté observée.
   Proposez une nouvelle mesure de proximité entre deux séquences ADN et l'appliquer aux deux séquences proposées.

3. En quoi tout cela aide-t-il à la découverte de séquences semblables ?

4. Que se passe-t-il si les séquences sont de longueurs différentes ?

Google, Bing, Duchduckgo : indexation, analyse de documents et « text mining »

Essayez de faire une recherche sur un moteur de recherche avec les mots suivants : *loovres* puis *kouvres*, puis *souvres*. Quelle est la connaissance capturée par le coefficient de substitution k ↔ l ou s ↔ d ? Essayez : *tartable* puis *tirtable* puis *tistable*. Concluez. Pensez au web sémantique.