

TP2 Part A : Alignement de séquences

Vous trouverez le nécessaire ici : <http://www.math-info.univ-paris5.fr/~lomn/Cours/BI/>

Ce tableau résume les grands familles d'alignement de séquences textuelles utilisées en bio-informatique :

Méthodes algorithmiques d'alignement de séquences	Optimale	Heuristique (rapide)
Globale	Needleman-Wunsch	FASTA / FASTP
Locale	Smith-Waterman	BLAST

Exercice 1. Étude d'une séquence ADN

Copiez-collez ci-dessous ou récupérez sur le site du cours la séquence suivante *unknown.fasta* (au format FASTA) afin d'essayer d'identifier cette séquence dans les bases de connaissance disponible en ligne.

```
ATGGCGAACCTTGGCTGCTGGATGCTGGTTCTCTTTGTGGCCACATGGAGTGACCTGGGCCTCTGCAAGAAGCG
CCCGAAGCCTGGAGGATGGAACACTGGGGGCAGCCGATACCCGGGGCAGGGCAGCCCTGGAGGCAACCGCTACC
CACCTCAGGGCGGTGGTGGCTGGGGGCAGCCTCATGGTGGTGGCTGGGGGCAGCCTCATGGTGGTGGCTGGGGG
CAGCCCCATGGTGGTGGCTGGGGGCAGCCTCATGGTGGTGGCTGGGGGCAGCCTCATGGTGGTGGCTGGGGGCA
GCCTCATGGTGGTGGCTGGGGGCAGCCCCATGGTGGTGGCTGGGGACAGCCTCATGGTGGTGGCTGGGGTCAAG
GAGGTGGCACCCACAGTCAGTGAACAAGCCGAGTAAGCCAAAAACCAACATGAAGCACATGGCTGGTGTGCTGCA
GCAGCTGGGGCAGTGGTGGGGGGCCTTGGCGGCTACATGCTGGGAAGTGCCATGAGCAGGCCCATCATAACATTT
CGGCAGTGACTATGAGGACCGTTACTATCGTGAACACATGCACCGTTACCCCAACCAAGTGTACTACAGGCCCA
TGGATGAGTACAGCAACCAGAACAACTTTGTGCACGACTGCGTCAATATCACAATCAAGCAGCACACGGTCACC
ACAACCACCAAGGGGGAGAAC TTCACCGAGACCGACGTTAAGATGATGGAGCGCGTGGTTGAGCAGATGTGTAT
CACCCAGTACGAGAGGGAATCTCAGGCCTATTACCAGAGAGGATCGAGCATGGTCTCTTCTCCTCTCCACCTG
TGATCCTCCTGATCTCTTTCCTCATCTTCTCCTGATAGTGGGATGA
```

- Pour faire une recherche avec l'algorithme FASTA:
(<https://www.ebi.ac.uk/Tools/sss/fasta/nucleotide.html>) proposé par EBI-EMBL
- Essayez enfin avec l'algorithme BLAST :
<https://www.ebi.ac.uk/Tools/sss/ncbiblast/nucleotide.html>
- Et comparer avec l'outil proposé par le NCBI en prenant soin de sélectionner l'espèce humaine: <http://blast.ncbi.nlm.nih.gov/>
- Qu'est ce que la E-value ? (voir annexe de ce TP si nécessaire)

Interprétation/Commentaire de résultats : par exemple, de quel gène s'agit-il ? Le temps de réponse ?

Exercice 2 : Alignement de deux séquences ADN

- Récupérez le fichier *mutation.fasta* proposé sur le site du TP. Que contient-il a priori ?
- Procédez à un alignement des deux séquences ADN normale et mutée en visualisant les mutations avec l'outil EMBOSS proposé par EBI-EMBL : par exemple https://www.ebi.ac.uk/Tools/psa/emboss_water/nucleotide.html
- On peut aussi utiliser d'autres outils comme <https://www.ebi.ac.uk/Tools/psa>.
- Faites de même avec l'outil blast proposé par le NCBI (voir exercice précédent) en cochant « Align two or more sequences »

Interprétation/Commentaire de résultats :

Exercice 3 : Les protéines

- Connectez-vous à Uniprot : <http://www.uniprot.org/>
- Combien de protéines contient cette base ? (combien annotées ? combien identifiées automatiquement ?)
- Qu'est-ce que le projet TREMBL ?

Interprétation/Commentaire de résultats :

- Trouvez la séquence de la protéine du prion humain (Homo Sapiens) PrP : code P04156 dans cette base. Sauvegardez la séquence au format *fasta*.
- Consultez l'ensemble des documents fournis pour comprendre la fonction de la protéine et les maladies auxquelles elle est associée.
- Étudiez le réseau d'interaction de la protéine avec l'outil BioGrid <http://thebiogrid.org/> (cherchez pour le gène PRNP et sélectionnez l'onglet Network)
- Comparez avec le réseau d'interaction du gène PRNP trouvé avec l'outil STRING

Interprétation/Commentaire de résultats :

Exercice 4 : Alignement de deux séquences de protéines

- Récupérer les deux séquences d'acides aminés correspondant aux deux protéines du prion normal et muté que vous avez sauvegardé au format *fasta* (voir annexe de ce TP)
- Utilisez l'utilitaire ncbi/blast pour aligner les séquences.
- Utilisez l'utilitaire *pair-linux* en modifiant le fichier *targlist* comme il se doit.
<http://www.ks.uiuc.edu/Training/Tutorials/science/bioinformatics-tutorial/bioinformatics/pairData/>

```
$vi targlist #(ou un éditeur graphique comme gedit pour le modifier)
$pair-linux targlist
```

Interprétation/Commentaire de résultats :

TP2 Part B : Alignement de famille de séquences

Exercice 5 : Alignement de famille de séquences et Arbres phylogénétiques

1. Dans les résultats du BLAST du prion contre SWISSPROT, sélectionnez 10 protéines prions, allant des plus *proches* aux plus *éloignées*
 - Obtenez chaque séquence au format FASTA
 - Ajoutez au fur et à mesure chaque séquence dans un fichier texte
 - Sauvez votre liste de séquence au format FASTA
2. Sur le serveur du EBI, trouvez l'outil CLUSTAL OMEGA. Copiez-collez les 10 séquences au format FASTA du fichier texte
 - Alignez les 10 séquences
 - Repérez les régions conservées/divergentes
 - Sauvegardez cet alignement sur votre disque.
 - Observez l'arbre de l'évolution de type PHYLIP. Qu'est-ce que l'outil PHYLIP ?
3. Refaite l'exercice avec l'outil T-Coffee (<http://www.tcoffee.org/>) accessible aussi d'ici <https://www.ebi.ac.uk/Tools/msa/tcoffee/>
4. Utilisez l'utilitaire *multiple-linux* toujours sur le site <http://www.ks.uiuc.edu/Training/Tutorials/science/bioinformatics-tutorial/bioinformatics/>

Interprétation des résultats :

Annexe

Séquence de la protéine PRNP traduite d'un cDNA expérimental issu d'un patient atteint de CJD

>protein prion CDJ

MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQPHG
GGWGQPHGGGGWGQPHGGGGWGQPHGGGGWGQPHGGGGWGQPHGGGGWGQPHGGGGWGQPHGG
GGWGQGGGTHSQWNKPSKPKTNMKHMAGAAAAGAVVGLGGYMLGSAMSRPIIHFGSDYEDRY
RENMHRYPNQVYYRPMDEYSNQNNFVHDCVNITIKQHTVTTTTKGENFTETDVKMMERVVEQMC
I
TQYERESQAYYQRGSSMVLFSPPVILLISFLIFLIVG

Séquence de la protéine **prion normale** (SWISSPROT:P04156)

>SWISSPROT:P04156 HUMAN MAJOR PRION PROTEIN PRECURSOR (PRP)

MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQPHG
GGWGQPHGGGGWGQPHGGGGWGQPHGGGGWGQGGGTHSQWNKPSKPKTNMKHMAGAAAAGAV
VGLGGYMLGSAMSRPIIHFGSDYEDRYRENMHRYPNQVYYRPMDEYSNQNNFVHDCVNITIK
QHTVTTTTKGENFTETDVKMMERVVEQMCITQYERESQAYYQRGSSMVLFSPPVILLISFLIFLIVG

E-value : Dans les alignements de séquence, une **E-value** de 4 signifie que si l'on avait soumis une séquence aléatoire, on s'attendrait à trouver 4 alignements avec un score aussi élevé, par le simple jeu du hasard. Un tel alignement ne peut donc en aucun cas être considéré comme significatif. En revanche des valeurs très faibles sont une bonne indication d'un alignement significatif.