TD – Alignements de séquences

Dans ce TD nous allons approfondir les notions d'alignements que nous avons vues en cours. Nous verrons quel type d'alignement il faut utiliser (*i.e.*, global ou local, " exact " ou " heuristique ").

I. Comparaison alignement global et alignement local

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 dans ce cas des algorithmes d'alignements locaux.

Aligner avec stretcher (alignement global) (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.* e-campus).

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) ?

Aligner maintenant les deux séquences précédentes en utilisant **`matcher`** (http://emboss.bioinformatics.nl/cgi-bin/emboss/matcher) qui implémente un algorithme d'alignement local. **Choisissez (paramètre utilisateur) de visualiser les 4 meilleurs alignements locaux (number of alternative matches).**

Comparer les différents alignements (l'alignement global avec les deux meilleurs alignements locaux). Que constatez-vous ? Dans ce cas, est-ce la taille des séquences qui posent un problème à l'alignement global ?

Présentation rapide de BLAST au NCBI

Les méthodes présentées précédemment (stretcher et matcher) sont dites " exactes " dans la mesure où elles donnent les meilleurs alignements possibles. De ce fait, elles sont plutôt lentes et inutilisables dans le cas de comparaison massive (par exemple comparer une séquence à une banque contenant des millions de séquences) si on souhaite obtenir un résultat de l'ordre de la minute maximum. En lieu et place des méthodes " exactes " on préfère utiliser dans ce cadre des méthodes " sous-optimales " ultra-rapides qui donnent néanmoins de bons résultats (*i.e.*, proches de l'optimale). La méthode " heuristique " (sous-optimale) la plus connue et la plus utilisée est BLAST (pour Basic Local Alignment Search Tool).

Plusieurs versions du logiciel sont proposées en fonction de la nature de la séquence requête et de celle de la banque interrogée. Nous ne donnons ci-dessous qu'un très bref aperçu de la suite BLAST au NCBI.

Nucleotide :

BlastN : compare une séquence nucléique à une banque nucléique : utile pour étudier une séquence qui ne code pas une protéine, ou localiser un ARNm sur un génome et *vice versa*.

Translated :

BlastX : compare une séquence nucléique traduite dans les 6 phases de lecture à une banque protéique : utile pour savoir si une séquence nucléique code une protéine et éventuellement localiser les positions de la partie codante.

tBlastN : compare une séquence protéique à une banque nucléique traduite dans les six phases : utile pour identifier le gène et/ou l'ARNm qui code une protéine.

tBlastX : compare une séquence nucléique traduite dans les six phases à une banque nucléique traduite dans les six phases : utile pour comparer une séquence nucléique dont ont ne sait rien à un génome non annoté, ou quand BlastN ne donne pas de résultats. A utiliser avec modération car très long !

Protein :

BlastP : compare une séquence protéique à une banque protéique : recherche les homologues d'une protéine.

II. Recherche dans les banques par similitude de séquence : séquence protéique contre banque protéique

Sur le serveur du NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi), trouvez les outils BLAST. Choisissez le programme Protein Blast (ou blastp), puis dans le formulaire :

1. Sélectionnez la banque SWISSPROT (database)

2. Copiez-collez la séquence protéique mystérieuse suivante (*cf*. e-campus fichier SEQUENCE_SONDE.doc):

```
>sequence sonde mysterieuse
WNTGGGSRYPGQGSPGGNRYPPQGGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQGGGTHS
QWNKPSKPKLMKHMAGAGAGAVVGGLGGYMLGSAMSRPIIHFGSDYEDRYYRENMHRYPINKQVYYRP
MDEYSNQNNFVHDCVNITIKQHTVTTTTKGENFTETDVKMMERVVEQMCITQYERESQAYYQRGSSMV
LFSSPPVILLISFLIFLY
```

3. Cocher l'option permettant d'afficher les résultats dans une autre fenêtre

Blastp organise les résultats en trois parties : Graphic Summary, Descriptions et Alignments.

1. Comment sont distribués les résultats et quels sont les liens entre ces trois composants ?

2. Décrivez les résultats donnés par tous ces liens à savoir :

- Quelle famille de protéines semble similaire à votre séquence mystérieuse ?
- Quelle est l'étendue des E-value?

3. Retenez le meilleur " hit blast " et, en vous aidant de l'alignement correspondant, répondez aux questions suivantes :

- Quel est le pourcentage d'identité ?

- Sur quelle longueur est effectué cet alignement ? Est-ce qu'il recouvre la totalité de votre séquence ? précisez les recouvrements (*i.e.*, les positions) entre ces deux séquences ?

- Que représente, d'après-vous, le pourcentage de "positives" ?

Pour mieux comprendre l'intérêt de la E-value et l'importance de ce qualificatif**, nous vous proposons de faire un test en utilisant BlastP avec une "pseudo" séquence** protéique **aléatoire d'une quarantaine d'acides aminés** (par exemple une phrase en français ou une séquence composée de deux fois la suite des 20 acides aminés (“ ACDEFGHIKLMNPQRSTVWY ”)).

Obtenez-vous un résultat ? Si oui, quelles sont les valeurs associées à vos Hits ?

Si non, augmentez le seuil “ Expect threshold ” dans les paramètres “ Search parameters ” des E-value jusqu'à 10 000.

Décrire ces nouveaux résultats (notamment E-value, alignements, …) ?

III. Recherche dans les banques par similitude de séquence : séquence d'ADN contre banque séquences protéiques

Cet exercice porte sur l'analyse de séquences d'enzymes de conversion de l'angiotensine I en angiotensine II, aussi appelées ACE. Ci-dessous, la séquence nucléotidique de l'ARNm de l'ACE de sangsue (*cf*. e-campus SEQUENCE_SANGSUE.doc) :

```
>Sangsue, ACE
aatttaaaaatgaatttaataaatttttcatacttaaatttgctttttggtgccggtttatttagcgttttagaa
agcgctacaatattaaataccgaatcggatgctaaaaaatggctgacaacgtataacgatgaagccggaaaatat
atttacgatgcaactgaagcagaatggaattacaacaccaacctgactgatcacaatttaggaatttctattaaa
aaatcaaatgatttggctacttttacggaacaaaaggcaatcgaggccaataaaaaatttgtatggaaaaatttt
actgatccacttttgaaaagagaattttcaaaaataactgacattggtactgctagcctttcagatgaagacttt
caaaagatgtcaggtttgaactctgatctaacaaaaatttacagcactgcaaaagtttgtaacaagcctaacgac
ccatctggaaaatgctatcctttagatcctgatttgtccgacataatctccaagtcaaacgatctcgaggaattg
acctgggcatggaaaggttggagggatgcgtctggcaaacatatgcccgataaatatgatgaatttgttcaactg
ctcaacaaagctgctaagattcatggatatgaagacaacggggattattggaggtcctggtacgagtcccccacg
ttcagaaaggattgtgaagatttgtggcaggagatcaaaccattctacgaacaactgcatgcatacgtcagaagg
aagctgcagaagaagtatccccaaattgcattccccaaggaggggcccatccctgctcatctgctcggcaacatg
tgggcccaatcgtgggagaacatagagtacttgttatgggcccaatcgtgggagaacatagagtacttgttaagg
cccgctcctgaccttcctagcatggacatcactgaggaactcgtcaaacagaactacacggcattgaaactcttc
caactgtcggacacatttttcaaatccttgggtctcatccagatgcctcagccgttttgggaaaagtcgatgatc
gagaaaccagctgatcgggatgtgttcagaatcaaacaatgcgtttgccatgcgtcagcctgggacttctacaat
cgcaaggatacggttgtggacatgcactggttcatgacgactcaccatgagatgggacacatcgaatactacctc
cactacaaggaccaacccatcagtttcagatctggcgctaatccaggatttcatgaggccattgccgatattgca
tcactgtcagtggccacacctgaatatatgcaatccgtcagcctgttgcctaatttcactgacgatccaaatggc
gatttaaacttcttaatgaaccaagccttaacgaaggtggccttcctaccattcggttacctgatcgaccagtgg
agatgggacgtgttctcgggagataccctcgaccaaaatacaactccaagtggtggcacaacaggtgtaagtac
cagggcatatatcctccagtgaaaaggtcagagcaagattttgatgccggttccaagttccatgtacccaacaac
actccatacatcaggtactttgttgctcacgtcatccaattccaattccatgaagccctgtgcaaggctgccaac
aacagcagacctctacatagatgtaacatcgccaattccaaggaagctggagagaaactggctgaattgatgaaa
tctggatcttcaattccgtggcctaaagttctagaaaatcttactggatcggaaaaaatgtcagcgaaatctctc
atggcctattacaaaccgttgatcgattggcctgaaaaaagaaaaccaagggcagaaaattggatgggaggaaaa
atgtcctcctggatcatttgaaccatgaaattatttatttgattttatgtcatttcataatttttctaccacttt
tttaataaacttaggtgcctattgaatatgttcttgcaatttgaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
aaaaa
```

Lancez un BlastX avec la séquence de sangsue et les options par défaut. Dans ce cas, la séquence requête est traduite à l'aveugle dans les six phases. Les six peptides obtenus sont alignés avec les protéines de la banque, y compris les codons stop qui sont remplacés par une étoile.

Combien de séquences de la banque ressemblent à la nôtre ?

Quelle est la E-value (E) des 2 premières séquences de la liste ? Celle des 2 dernières ?

Pour le premier "hit Blast" :

- A quelle phase correspond ce premier " hit Blast " ? Que remarquez-vous au niveau des bornes de l'alignement ('Query' et 'Sbjct') ?

- Quelle est la longueur de la protéine de la banque correspondant à ce premier "hit Blast"? Est-ce que notre séquence correspond à la totalité de la protéine?

- Où se trouve le codon d'initiation et le codon stop ? Indiquer leurs positions dans la séquence d'ARNm.

IV. Recherche à définir (première partie)

Récupérer et analyser au moyen de BLAST, la séquence stockée dans le fichier " SEQ_IV.doc ".

IV. Recherche à définir (deuxième partie)

Récupérer et analyser au moyen de BLAST, la séquence stockée dans le fichier " SEQ_V.doc ".