2021

TD – BLAST – 2ième partie

Rappels : présentation rapide de BLAST au NCBI

Les méthodes d'alignement comme « 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.

2021

I. 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*. moodle 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.

2021

II. « Recherche à définir » (première partie)

Récupérer la séquence stockée dans le fichier “ SEQ_IV.doc ” (*cf*. moodle) et analyser cette au moyen de BLAST, *i.e.*, étant donné cette séquence, quelles questions peut-on se poser et quel Blast doit-on utiliser pour répondre à ces questions.

III. « Recherche à définir » (deuxième partie)

Récupérer la séquence stockée dans le fichier “ SEQ_V.doc ” (*cf*. moodle) et analyser la au moyen de BLAST, *i.e.*, étant donné cette séquence, quelles questions peut-on se poser et quel Blast doit-on utiliser pour répondre à ces questions.