2021

TD – BLAST – 1ère partie

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 protéique contre banque protéique

Sur le serveur du NCBI, accédez aux outils BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi).
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-moodle 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 e-value associées à vos Hits ?

2021

II. « Matcher » et « Blast 2 sequences »

Récupérer sous moodle les séquences stockées dans les fichiers “ SEQ_1 ” et “ SEQ_2 ”.

Analyser ces séquences au moyen de « Matcher » (*cf.* TD1) puis de « BlastP 2 sequences » qui est un Blast permettant de comparer deux séquences entre elles (plutôt qu'une séquence contre une banque). On prendra garde à utiliser le Blast adapté au type des séquences à comparer. On analysera les résultats en utilisant « Graphic Summary », « Dot Matrix View », « Descriptions » et « Alignements ».

Note : URL du « BlastP 2 sequences » :

https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&BLAST_SPEC=blast2seq&LINK_LOC=align2seq