Comparaison avec BLAST et recherche dans des banques

Liens utiles

- Entrez
- Prosite
- Interpro
- Blast au NCBI

Présentation de BLAST au NCBI

La page d'accueil de BLAST vous guide suivant les expériences que vous souhaitez réaliser. Vous pouvez :

- rechercher votre séquence dans un génome particulier (*BLAST RefSeq Assembled Genomes*) : offre la possibilité d'interroger les données de génomes complets ou en cours de séquençage. Les données interrogées sont non redondantes : chaque région du génome n'apparaît qu'une fois dans la banque. Lorsque les données sont disponibles, il est également possible d'interroger les ARNm, protéines, ... Très utile pour interroger les données d'un seul organisme car il n'y a pas de redondance.
- rechercher votre séquence dans une banque de votre choix avec l'outil de votre choix (*Basic BLAST*)
- faire des recherches spécialisées (*Specialized BLAST*)

Pour le TP nous choisirons sauf indication contraire les liens de la section Basic BLAST.

Plusieurs versions du logiciel sont proposées en fonction de la nature de la séquence requête et de celle de la banque interrogée :

Nucleotide

- Compare une séquence nucléique à une banque nucléique (BlastN) : utile pour étudier une séquence qui ne code pas une protéine, ou localiser un ARNm sur un génome et *vice versa*.
- MegaBlast : programme optimisé pour aligner des séquences qui diffèrent peu et 10 fois plus rapide que BlastN. Ce programme a été conçu pour comparer deux ensembles de séquences entre eux.

Une interface dédiée aux séquences courtes avec peu d'erreurs : les paramètres sont adaptés à ce type de données.

Protein

- Compare une séquence protéique à une banque protéique (BlastP) : recherche les homologues d'une protéine.
- PHI-Blast (Pattern Hit Initiated BLAST) : ce programme prend en entrée une séquence requête protéique et un motif défini par une expression régulière (voir la syntaxe). Il recherche les protéines qui contiennent le motif et qui sont similaires à la protéine requête au voisinage du motif. Ce programme est utile pour rechercher un motif (domaine protéique, site actif,...) dans une banque de séquences protéiques. Son avantage par rapport à une recherche simple de motif réside dans le fait qu'il s'appuie sur la séquence requête (étude du voisinage) pour éliminer les séquences qui contiennent le motif par hasard.
- PSI-Blast (Postion Specific Iterated Blast) : un profil est construit à partir de l'alignement multiple des séquences qui ont obtenu les meilleurs scores avec la séquence requête. Ce profil est comparé à la banque interrogée et est raffiné au fur et à mesure des itérations. Ainsi, la sensibilité du programme est augmentée. PSI-Blast est utile pour détecter des membres éloignés d'une famille protéique et pour étudier la fonction de protéines inconnues.

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 (tBlastX) : 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 !

Premiers pas

Cet exercice porte sur l'analyse de séquences d'enzymes de conversion de l'angiotensine I en angiotensin II, aussi appelées ACE. Ci-dessous, la séquence nucléotidique de l'ARNm de l'ACE de sangsue :

```
>Sangsue, ACE
aatttaaaaatgaatttaataaatttttcatacttaaatttgctttttggtgccggtttatttagcgttttagaaagcgc
tacaatattaaataccgaatcggatgctaaaaaatggctgacaacgtataacgatgaagccggaaaatatatttacgatg
caactgaagcagaatggaattacaacaccaacctgactgatcacaatttaggaatttctattaaaaaatcaaatgatttg
gctacttttacggaacaaaaggcaatcgaggccaataaaaaatttgtatggaaaaattttactgatccacttttgaaaag
agaattttcaaaaataactgacattggtactgctagcctttcagatgaagactttcaaaagatgtcaggtttgaactctg
atctaacaaaaatttacagcactgcaaaagtttgtaacaagcctaacgacccatctggaaaatgctatcctttagatcct
gatttgtccgacataatctccaagtcaaacgatctcgaggaattgacctgggcatggaaaggttggagggatgcgtctgg
caaacatatgcccgataaatatgatgaatttgttcaactgctcaacaaagctgctaagattcatggatatgaagacaacg
gggattattggaggtcctggtacgagtcccccacgttcagaaaggattgtgaagatttgtggcaggagatcaaaccattc
tacgaacaactgcatgcatacgtcagaaggaagctgcagaagaagtatccccaaattgcattccccaaggaggggcccat
ccctgctcatctgctcggcaacatgtgggcccaatcgtgggagaacatagagtacttgttatgggcccaatcgtgggaga
acatagagtacttgttaaggcccgctcctgaccttcctagcatggacatcactgaggaactcgtcaaacagaactacacg
gcattgaaactcttccaactgtcggacacatttttcaaatccttgggtctcatccagatgcctcagccgttttgggaaaa
gtcgatgatcgagaaaccagctgatcgggatgtgttcagaatcaaacaatgcgtttgccatgcgtcagcctgggacttct
acaatcgcaaggatacggttgtggacatgcactggttcatgacgactcaccatgagatgggacacatcgaatactacctc
cactacaaggaccaacccatcagtttcagatctggcgctaatccaggatttcatgaggccattgccgatattgcatcact
gtcagtggccacacctgaatatatgcaatccgtcagcctgttgcctaatttcactgacgatccaaatggcgatttaaact
tcttaatgaaccaagccttaacgaaggtggccttcctaccattcggttacctgatcgaccagtggagatgggacgtgttc
tcgggagataccctcgaccaaaatacaactccaagtggtggcacaacaggtgtaagtaccagggcatatatcctccagt
gaaaaggtcagagcaagattttgatgccggttccaagttccatgtacccaacaacactccatacatcaggtactttgttg
ctcacgtcatccaattccaattccatgaagccctgtgcaaggctgccaacaacagcagacctctacatagatgtaacatc
gccaattccaaggaagctggagagaaactggctgaattgatgaaatctggatcttcaattccgtggcctaaagttctaga
aaatcttactggatcggaaaaaatgtcagcgaaatctctcatggcctattacaaaccgttgatcgattggcctgaaaaaa
gaaaaccaagggcagaaaattggatgggaggaaaaatgtcctcctggatcatttgaaccatgaaattatttatttgattt
tatgtcatttcataatttttctaccacttttttaataaacttaggtgcctattgaatatgttcttgcaatttgaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaaaaaa
```

Séquence requête ADN contre banque nucléique

Découverte

Suivez le lien *nucleotide blast* de la section *Basic BLAST*. Copiez-collez la séquence ci-dessus dans la boîte correspondante, choisissez la banque de données (database) *Other : Nucleotide collection (nr/nt)*, sélectionnez le programme *Somewhat similar sequences (blastn)*, limitez la recherche aux entrées issues de l'humain. Pour cela dans le champ *Organism*, entrez `Homo sapiens`, et lancez la requête.

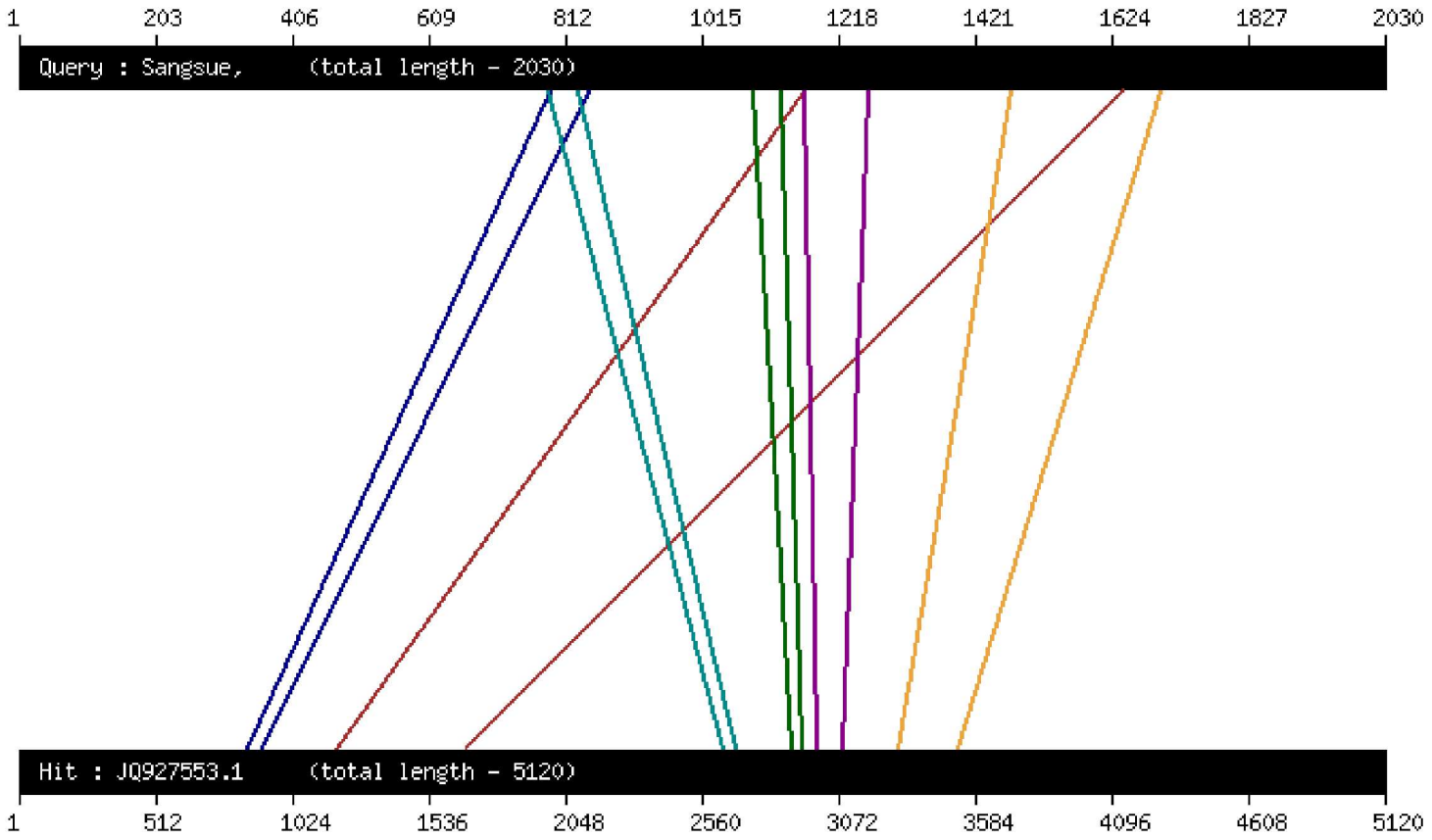
Question 1

Combien de séquences humaines de la banque ressemblent à la nôtre (voir le nombre de 'hits') ?
Est-ce que les alignements obtenus semblent pertinents d'un point de vue biologique ?
Est-ce que les séquences trouvées font partie de la famille des ACE ?

La représentation graphique des résultats indique les régions de la séquence requête (échelle) qui s'alignent avec les séquences de la banque (rectangles colorés). Seuls les rectangles reliés par un trait noir font partie de la même entrée de la banque.

Question 2

Combien de régions communes entre la séquence de numéro d'accession `NM_000789` et celle de la sangsue sont représentées sur le graphique ? (Faites glisser le curseur sur les traits colorés pour que le numéro d'accession s'affiche juste au dessus de *Color key for alignment scores*)
Combien de régions sont réellement communes entre les deux séquences ? (Voir les alignements en cliquant, soit sur un des traits colorés associés dans le graphique, soit sur le score de l'alignement dans le tableau des résultats situé sous le graphique)
Pourquoi y-a-t-il des différence ?



Question 3

Que signifie les termes "Total Score" et "Query coverage" dans le tableau des résultats ?

La E-value

Question 4

Quel est le score du meilleur alignement obtenu pour la séquence humaine `NM_000789` ?
Quelle est la e-value correspondante?
Comment varie la e-value en fonction du score (comparez pour différents alignements de cette séquence) ?

Sauvergarder cette page résultat au format *page web complète* (ou conservez la dans un onglet) : nous en aurons besoin par la suite dans l'ensemble du TP.

Faites maintenant tourner BLAST en limitant les données à la banque `refseq_rna` (en conservant la limite aux entrées issues de `Homo sapiens`). Pour cela choisissez la banque de données qui convient dans le formulaire d'interrogation.

> Question 5
>
> Est-ce que l'on retrouve la séquence d'ACE `NM_000789` ?
> Est-ce que le score du meilleur alignement a changé ?
> Comment a varié la e-value ?
> Pourquoi y t-il eu changement de E-value sans changement de score ?

Le choix du logiciel

> Question 6
>
> Comment vérifier (le plus efficacement possible) si la séquence dont nous disposons est présente dans la banque nr entière ?
> Est-elle effectivement présente ?

Le filtre de faible complexité

Les régions de faible complexité sont des parties de séquences composées de peu de lettres différentes. Par défaut, dans Blast, l'option "Low complexity" est active (cochée). Les régions de faible compléxité présentes dans la séquence requête sont remplacées par des `N` dans le cas de l'ADN et par des `X` dans le cas des protéines. Ces régions ne sont donc pas alignées avec les séquences de la banque. Nous allons étudier l'intérêt de cette option.

Vous pouvez observer la région de faible complexité présente dans la séquence de sangsue en faisant un dotplot de la séquence contre elle-même à l'aide d'un dotplot.

Utilisez de nouveau BLAST avec la séquence de sangsue contre toutes les séquences de la banque (en continuant à se limiter à l'humain et à `refseq_rna`), mais en décochant l'option "Low complexity regions" dans "Algorithm parameters".

Les résultats obtenus sont alors différents des précédents. Pourtant, la séquence requête et la banque sont les mêmes.

> Question 7

Combien de séquences de la banque ressemblent désormais à la nôtre ?
Quelle partie de notre séquence est détectée plus qu'auparavant ? Sur quel type de séquences "matche" t'-elle ? A votre avis pourquoi ?
Quelles entrées, parmi celles du précédent graphique de résultats, n'apparaissent alors plus dans les 100 premiers hits lorsque l'on desactive les filtres de faible complexité. (Utilisez CTRL + F pour faire une recherche) ? A votre avis, où sont-ils désormais ?

Séquence requête ADN contre banque protéique.

Dans ce cas, la séquence requête est d'abord traduite à l'aveugle dans les six phases de lecture par BlastX. Les six peptides obtenus sont ensuite recherchés et alignés avec les protéines de la banque (y compris les codons stop des peptites requêtes qui sont remplacés par une étoile).

BlastX

Lancez un BlastX avec la séquence de sangsue et en restreignant la rechercherche aux mammifères (*Mammalia*).

Question 8

Combien de séquences de la banque ressemblent à la notre ?
Quelle est la E-value des 2 premières séquences de la liste ?
Celle des 2 dernières ?
Comparez les valeurs trouvées à celles obtenues avec BlastN.

Question 9

Est-ce que plus de séquences de la famille des ACE sont trouvées (ne comptez pas!) ?

Question 10

De quel organisme provient la première séquence trouvée ?

Trouvez dans les résultats, la protéine ACE correspondant à la séquence humaine.

Question 11

Combien de hits y a-t-il sur cette séquence protéique.
Comparez cet alignement avec celui obtenu précédement à l'aide de BlastN.
Qu'observez-vous (Evaleur, couverture du/des alignements, qualité du résultat)?.

Sensibilité aux paramètres

Modification de la taille de mots

Faites une requête à l'aide de **BlastN** ("nucleotide blast") contre la banque *nr* avec le gène de MAKORIN1 chez *Seriola quinqueradiata* (conservez la fenêtre pour la suite).
Changer maintenant, dans "Algorithm parameters", la taille des mots exacts recherchés afin d'être plus spécifique.

[résultat pour w=7,résultat pour w=11, résultat pour w=15]

> Question 12
>
> Quelle taille choisissez-vous ?
> Comparez les résultats à ceux obtenus avec la taille de mots par défaut. Qu'observez-vous ?

Modification de la valeur de "Expect threshold"

> Question 13
>
> Quelle est la valeur par défaut de ce paramètre. Que cela signifie-t-il ?

Changer maintenant la valeur de *Expect threshold* pour être plus spécifique.

[résultat]

> Question 14
>
> Quelle valeur choisissez-vous ?
> Comparez les résultats à ceux obtenus avec la valeur par défaut. Qu'observez-vous ?

Modification des pénalités et de la matrice de score dans BlastN

Chercher comment modifier ces paramètres.

> Question 15
>
> Quels sont les paramètres par défaut (pénalité d'ouverture et d'extension de gap, score de match et de mismatch) ?

Faites une requête BlastN contre la banque *nr* avec le gène MAKORIN1 chez *Seriola quinqueradiata* avec les pénalités de gap les moins pénalisantes (conservez la fenêtre pour la suite).

[résultat]

> Question 16

Quels différences observez-vous par rapport à la requête avec les paramètres par défaut ?
Cherchez les hits sur le gène MAKORIN en limitant la recherche au cochon (*Sus scrofa*). Expliquez les résultats obtenus avec les deux jeux de paramètres.

MegaBLAST

Lancez une requête MegaBLAST avec le gène de MAKORIN1 chez *Seriola quinqueradiata*.

[résultat]

Question 17

Quelles sont les différences avec BlastN ?
Observez les résultats obtenus sur les séquence de *Zebrafish*, pourquoi ne voit-on pas les mêmes hits avec MegaBLAST ?

PSI-Blast

Voici une protéine de *E. coli* :

```
>trpc
MMQTVLAKIVADKAIWVEARKQQQPLASFQNEVQPSTRHFYDALQGARTAFILECKKASP
SKGVIRDDFDPARIAAIYKHYASAISVLTDEKYFQGSFNFLPIVSQIAPQPILCKDFIID
PYQIYLARYYQADACLLMLSVLDDDQYRQLAAVAHSLEMGVLTEVSNEEEQERAIALGAK
VVGINNRDLRDLSIDLNRTRELAPKLGHNVTVISESGINTYAQVRELSHFANGFLIGSAL
MAHDDLHAAVRRVLLGENKVCGLTRGQDAKAAYDAGAIYGGLIFVATSPRCVNVEQAQEV
MAAAPLQYVGVFRNHDIADVVDKAKVLSLAAVQLHGNEEQLYIDTLREALPAHVAIWKAL
SVGETLPAREFQHVDKYVLDNGQGGSGQRFDWSLLNGQSLGNVLLAGGLGADNCVEAAQT
GCAGLDFNSAVESQPGIKDARLLASVFQTLRAY
```

Faites tournez PSI-blast dessus sur la banque NR.

Question 18

Lors de la première itération, quelles sont les fonctions protéiques trouvées ? Est-ce qu'il y en a plusieurs ? Enregistrez la page web complète dans votre répertoire de travail.

Lancez la deuxième itération (bouton disponible sur la page de résultat obtenue lors de la première itération, sous le diagramme des hits).

Question 19

Est-ce que les résultats sont différents ? D'après-vous, pourquoi ? Quelles sont les séquences qui sont apparues ? Celles qui ont disparues ?

Analyse d'un cDNA

La séquence que nous allons étudier provient d'un poisson (dont vous chercherez la description sur fishbase). C'est la copie ADN d'un ARNm.

```
>Gasterosteus aculeatus cDNA clone CLJ188-G12 5', mRNA sequence
AATTGGACATGACAGTTCGGTCCGGAATCCCGGGATGGAGATGCCATCCGTTGGATCCGGATCTTCAGAA
GATCATGGCGGAGTCCAGGGATTATGACGAACTGCTGTTTGCGTGGAAAGGATGGAGAGATTCTGCCGGC
AAAGTGCTTCGCCAGGATTACAAGAGATATGTTGAACTGGCCAACATGGCCGCCAAACTCAACGGTCACT
CCGACAACGGGGCTTCCTGGCGCTCCCTGTATGAAACCCCCAGCTTCGAGGAGGACCTGGAGGCTCTGTG
GAAGGAGCTGGAGCCGCTCTATCAGAATGTGCACGCCTATGTGCGCAGGGCCCTGTACAAAAAGTATGGC
TCCCAGCACATCAACCTGAAGGGAGCCATCCCGGCTCATTTGCTGGGCAACATGTGGGCCCAGACGTGGT
CGGGCATAATGGATTTGGTCATGCCCTACCCGCATGCCACGCAGGTGGACGCCACGCCCGCCATGGTTTC
ACAGGGCTGGAACGCCACCAGAATGTTCCAGGAATCCGACAATTTTTTCACCTCTCTGGGTCTTTTGCCA
ATGCCCCAAGAGTTCTGGGACAAATCCATGCTAGAGAAGCCGTCTGGTGGACGCCAGGTGGTGTGCCACG
CCTCCGCATGGGACTTCTATAACCGAAAAGACTTCAGGATCAAACAGTGCACCGTGGTGACTATGGACGA
TTCCGCGACGGAGCCAACCCCGGCTTCCACGAGGCCATTGGCGACGTGTTGGCCCTGTCAGTGTCTACGC
CCTGATCACGGCGCACCATGAGATGGGCCACATTCAGTACTTCCTGCAGTACAAAGACCAGCCCGTGTCC
CCAAACACCTGCAGAGCATCGGCCTGCTGGACAAAGTGGAGAGCAACCATGAGAGCGATATCAACTTCCT
GATGAGCATGGCGCTCGACAAGATCGCCTTCTTACCCTTTCGCTACCTGATGGATCAGTGNAGATGGAAG
GGTGTCGATGGNCGTATCCCATCGACTGAGTANCATAAAGAATGGNTGGAACTCAGAATGAAGTACCAGG
GCCTCTGTCCCACTGTAACCCGCACAGAGGAAGACTTCGACCANGTGCAAAAGTCACATCCCTGCTACGT
GCCATACGTGAAGAACTTTGTCACCTCATCATCAGGTCCAGGTTCCAAAGCTCTCTGGGATGCCCCAAAA
CGAAGGGGCCTGGAAACTGGAAATTTTAAATTCGGAAAACCCGGACCTCTTGCGGACGATGAAACCCGTT
TCTAAAACCTGGCCCGGGGAAA
```

Avec Discontiguous MegaBLAST

Comparez ce cDNA (de *l'épinoche*) à la banque **nucléique** "nr/nt" en utilisant la version *More dissimilar sequences (discontiguous megablast)*.

[résultat]

> Question 20
>
> Est-ce que les séquences trouvées ressemblent vraiment à la séquence requête ? Est-ce que l'on a réussi à localiser l'ARNm sur le génome de *Gasterosteus aculeatus* ? Peut-on avoir une idée de la fonction de la protéine codée par l'ARNm étudié ?

Le génome de *Gasterosteus aculeatus* est en cours de séquençage: les données ne sont pas dans la banque "nr/nt", mais dans la banque "wgs" ("whole genome shotgun").

Interroger (toujours avec *More dissimilar sequences, discontiguous megablast*) les séquences de la banque "wgs" pour l'organisme *Gasterosteus aculeatus*.

[résultat pour tout WGS, résultat pour WGS limité à Gasterosteus]

Question 21

Est-ce qu'on parvient à localiser l'ARNm ? Peut on le positionner sur le génome ? Pourquoi cela n'est t-il pas possible ? Quel est le numéro d'accession et la taille du contig trouvé ? Pourquoi l'alignement entre l'ARNm et le contig est-il morcelé ? Notez sur le graphique les barres verticales noires entre les hits rouges indiquant que ces hits proviennent de la même séquence de la banque.

Avec Ensembl

Nous allons faire une recherche sur le site Ensembl du cDNA contre le génome de *Gasterosteus aculeatus* afin d'obtenir plus d'information sur les positions prédites par *Ensembl* pour le gène complet.

Cliquez sur le lien *BLAST/BLAT* de Ensembl, donnez votre séquence, et choisissez l'espèce *Gasterosteus aculeatus* sur laquelle cette séquence sera localisée. Lancez la comparaison, puis cliquez sur "[View results]".

Question 22

Sur quel chromosome le gène est-il présent ? Dans quel sens est représenté le gène sur le contig ?

Dans *Genomic location*, il est possible de visualiser les hits trouvés par BLAT *visuellement* sur la séquence du chromosome; Pour cela, cliquez pour l'une des entrées, sur le lien situé juste avant " [sequence]" : il vous menera à l'interface "Region in detail" pour cette entrée.

Vous devez avoir une image assez complexe décrite comme "Region in detail", disposant **du** hit (**des** hits si vous dezoomez) de BLAT/BLAST dans le sens direct, ainsi que dans le sens complementaire inverse.

Comparez les positions des exons trouvés sur notre séquence ("BLAT/BLAST hits" en rouge/marron) avec celles des exons *prédits* par Ensembl ("Genes [Ensembl]").

Question 23

Les hits principaux de notre requête sont-ils plutôt au début ou la fin du gène ?

Sur la banque protéique nr

En revenant sur l'interface du NCBI, comparez désormais la séquence **cDNA** à la banque **protéique** "nr", en choisissant le **bon** programme.

Question 24

Est-ce que les résultats sont plus satisfaisants (meilleure similarité) que ceux obtenus avec le cDNA contre la banque nucléique ? D'après-vous, pourquoi ?
Que peut-on apprendre sur la fonction de la protéine ? Qu'est-ce que vous constatez en consultant les alignements obtenus (consultez par exemple le premier hit sur une entrée "Swiss-Prot" sp)? Est-ce que la séquence étudiée contient une séquence codante complète ?

Page réalisée par les membres des équipes Bonsai et EEP, mise à jour en octobre 2016