Annotation d'un gène eucaryote

Vous allez étudier un fragment du génome humain pour prédire le gène qui est dessus, ainsi que les positions de début et de fin des exons. Il s'agit d'un gène qui subit de l'épissage alternatif. Comme la séquence du génome est longue, je donne les résultats de certains programmes.
La page suivante contient la séquence à étudier.

Recherche de la séquence codante par homologie de séquence

Utilisation de BlastX

Dans un premier temps, nous allons comparer la séquence génomique aux protéines de la banque de données. Cela nous permettra peut-être de trouver des protéines de la famille de celle codée par notre séquence génomique. Pour des séquences eucaryotes, il est plus efficace d'utiliser BlastX, plutôt que passer par ORFfinder puis BlastP.

Voici les résultats trouvés par BlastX, avec les options par défaut.

> Question 1
>
> A quelle famille appartiennent les protéines trouvées par BlastX ?
> Est-ce que ces protéines sont intéressantes ?
> Quelles sont les positions des régions qui ressemblent à ces protéines ?

Recherche des répétitions

Les protéines trouvées font partie de la famille des transcriptases reverses de type LINE. Les LINE (Long INterspersed repeated sequences) sont des répétitions très fréquentes sur le génome humain. Elles couvrent 14% du génome et mesurent 6 à 8 kb de long. Comme il y en a beaucoup dans le génome humain, la banque contient beaucoup de ces séquences. Notre séquence contient peut-être d'autres séquences codantes, mais elles sont masquées par les séquences de type LINE. Il faut donc masquer les séquences répétées connues puis relancer un BlastX.
Le logiciel RepeatMasker compare une séquence à une banque de familles de séquences répétées. Il masque les régions qui ressemblent à des répétitions connues en les remplaçant par des N (lettre qui symbolise n'importe quel nucléotide).

Les résultats de RepeatMasker sont fournis.

> Question 2
>
> Est-ce que notre séquence contient beaucoup de répétitions ?
> Est-ce que les hits de BlastX sont dans les régions masquées par RepeatMasker ? (regarder le fichier d'annotation fourni par RepeatMasker ou bien la séquence avec les répétitions masquées).

Recherche des protéines codées par la séquence

Vous trouverez la séquence avec les répétitions masquées parmi les pages de résultats de RepeatMasker. Il est maintenant possible de copier-coller cette séquence dans BlastX pour obtenir des séquences qui ressemblent aux régions non masquées.

Les résultats de BlastX sont fournis.

> Question 3
>
> Est-ce que d'autres protéines sont trouvées ?
> Est-ce que les protéines trouvées ont toutes (ou presque) la même fonction ?
> Si oui, quelle est cette fonction ? De combien d'exons codants semble être composé le gène (d'après les résultats de BlastX) ? Vérifiez que les différents exons sont sur le même brin. Pourquoi ne sont-ils pas dans la même phase de lecture ?

Prédiction de la structure du gène

Maintenant que nous avons réussi à sélectionner des protéines qui ressemblent à celles codées par notre gène, nous pouvons utiliser WISE pour prédire la position des exons codants présents sur notre séquence.

Les résultats de WISE sont fournis.

> Question 4
>
> Est-ce que les résultats de Wise sont satisfaisants ?

Combien d'exons codants sont prédits par Wise ?
Quelles sont les positions de début et de fin des exons prédits ?

Prédiction de la séquence codante par prédiction statistique

GenScan est le logiciel le plus souvent utilisé pour la prédiction statistique de séquences codantes chez les eucaryotes. Utilsez GenScan sur la séquence nucléique non masquée pour voir quels sont les exons codants prédits.

Voici les résultats trouvés par GenScan, avec les options par défaut.

Question 5

Combien d'exons sont prédits par GenScan ?
Quelles sont leurs positions de début et de fin ?
Est-ce que les exons trouvés par GenScan concordent avec ceux trouvés à l'aide de BlastX couplé à Wise ?

Prédiction de la séquence transcrite

Recherche des ARNm codés par notre gènes

Il est également possible de comparer notre séquence génomique à des séquences d'ARNm humains. Pour cela, il est préférable d'utiliser la page de Blast dédiée aux génomes complets (**"BLAST Genomes"**) **situé sous Web BLAST**, en y précisant `Homo sapiens` celle dédiée à l'humain. Comme on veut comparer une séquence génomique humaine à des ARNm humains, on peut utiliser megablast (prog proposé par défaut). Pour la banque on choisi : **Non-RefSeq RNA**. Il s'agit des ARN non présents dans la banque RefSeq. En fait, RefSeq contient une sélection des ARN compris dans GenBank donc la banque Non-RefSeq est plus complète.

Voici des résultats sur nonrefseq (et sur refseq).

Question 6

Est-ce que des ARNm s'alignent avec notre séquence ?
De combien d'exons notre gène semble être composé ?
Consultez les entrées pour voir si elles contiennent des informations intéressantes.
Est-ce que tous les ARNm trouvés s'alignement avec les mêmes régions du génome ?
Quelle(s) est(sont) la(les) séquence(s) intéressante(s) à étudier de plus près ?

Reconstruction de la structure du gène

Est2genome permet d'aligner l'ARNm trouvé à l'aide de Blast avec notre séquence génomique afin de prédire la structure complète du gène (y compris les parties 5' et 3' UTR). Utilisez les résultats obtenus à l'aide d'Est2genome afin de répondre aux questions suivantes :

Question 7

Combien d'exons sont prédits ?
Quelles sont les positions des exons prédits ?
Est-ce que les prédictions semblent fiables ?
Est-ce que les positions prédites par EST2Genome correspondent aux positions de début et de fin des alignements donnés par Blast ?

Bilan

Compilation des résultats trouvés

Nous avons étudié un fragment de génome humain en le confrontant à des données de protéines et d'ARNm et en faisant tourner un logiciel de prédiction de CDS (GenScan). En rassemblant et comparant les résultats issus des différentes analyses, il est possible d'annoter ce fragment.

Question 8

Est-ce qu'il y a plus d'exons prédits à partir des données de protéines que des données d'ARNm ? Il est possible de trouver des exons en plus avec l'ARNm, ceux-ci correspondent aux régions 3' et 5' UTR (ils doivent être au début et/ou à la fin des exons codants pour une protéine).
Est-ce que vous avez identifié l'exon contenant le codon d'initiation et celui contenant le codon de terminaison de la transcription ?
Est-ce que notre analyse a permis de mettre en évidence de l'epissage alternatif ?
Vous pouvez reconstruire la structure complète du gène.

Vous pouvez même calculer la séquence de la protéine codée par le gène et émettre une hypothèse concernant sa fonction.

Consultation des informations pré-calculées

Il existe des banques qui localisent sur différents génomes toutes les informations connues, issues de différentes sources de données comme des banques de séquences, mais aussi des logiciels de prédiction statistique de CDS, ... Nous allons consulter les informations concernant le gène TCN1, c'est-à-dire le gène que l'on a étudié, sur le site de ensembl.org. Pour cela, recherchez sur *l'humain* le gène *TCN1*, et accédez à la localisation de l'entrée *TCN1 (Human Gene)*: vous obtenez alors toutes les informations liées (alignées) à ce gène sous forme de deux graphiques (le deuxième graphique est disponible avec le lien *Go to Region in Detail for more tracks and navigation options (e.g. zooming)*).

Question 9

Est-ce que certaines annotations montrent la possibilité d'épissage alternatif ?