Eukaryotic gene annotation

You will study a fragment from the human genome to predict the gene it contains, as well as the start and end positions of the exons. This is a gene that is subject to alternative splicing.
The following webpage contains the sequence to study.

Searching for the coding sequence by sequence homology

Using BlastX

Firstly, we will compare the genomic sequence to proteins from the NR databank. This may allow us to find proteins from the same family as the one encoded by our genomic sequence. For eukaryotic sequences, it is more efficient to use BlastX, instead of using ORFfinder then BlastP, because the exons can be in different reading frames.
Launch BlastX with default options.

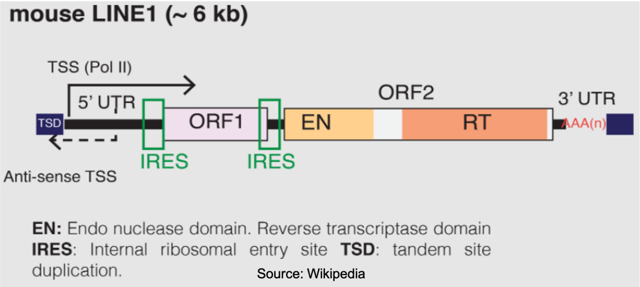
[BlastX results]

Question 1

To which family do the proteins found by BlastX belong ?
Are these proteins usually encoded by eukaryotic genomes ?
What are the positions of the regions that look like those proteins ?

Searching for repetitions

The identified proteins are reverse transcriptases encoded by LINE (Long INterspersed repeated sequences) - type retrotransposons. LINEs, which can be several kilobases long, are represented by a very high number of copies in the human genome. They indeed make up about 20% of our genome, and the LINE-1 superfamily is represented by no less than 100,000 copies (truncated and inactivated for most of them!)
Here is the structure of a mouse LINE1 element:


mouse LINE1 (~ 6 kb)
TSS (Pol II)
5' UTR
TSD
ORF1
ORF2
EN
RT
3' UTR
AA(n)
IRES
IRES
Anti-sense TSS
EN: Endo nuclease domain. Reverse transcriptase domain
IRES: Internal ribosomal entry site TSD: tandem site duplication.

Source: Wikipedia

Since they are present in high numbers in the human genome, the NR bank contains many of these sequences. Our sequence may contain other coding sequences, but they are masked by LINE type sequences. It is therefore necessary to mask known repeated sequences, and then to re-launch BlastX.
The RepeatMasker software compares a sequence to a repeated sequences families bank. It masks the regions which look like known repetitions by replacing them with N

(letter that symbolizes any nucleotide).
Launch RepeatMasker with default parameters on the genomic sequence (the "Homo sapiens" repetition base is selected by default).

[RepeatMasker results]

> **Question 2**
>
> Does our sequence contain many repetitions ?
> Are the BlastX hits within the regions masked by RepeatMasker (check the annotation file provided by RepeatMasker, or the sequence with the masked repetitions)?

Searching for proteins encoded by the sequence

You will find the sequence with the masked repetitions among the results pages of RepeatMasker (".masked" file).
Relaunch BlastX from the masked sequence.

[masked sequence BlastX results]

> **Question 3**
>
> Are other proteins found ?
> Do the found proteins all have the same (or almost the same) function ?
> If yes, what is this function ?
> Of how many coding exons does the gene seem to be composed (according to the BlastX results) ?
> Check that the different exons are on the same strand. Why are they not in the same reading frame ?

Gene structure prediction

Now that we achieved, thanks to Blast, identifying proteins which look like the ones encoded by our gene, we can use Wise to predict the position of coding exons within our sequence.
Launch Wise using the protein sequence which showed the most significant alignment with the genomic sequence and make the comparison against the unmasked genomic sequence.

[Wise results]

> **Question 4**
>
> Are the Wise results satisfying ?
> How many coding exons are predicted by Wise ?
> What are the positions of the predicted exons' start and end ?

Coding sequence prediction by statistical prediction

GenScan is a software often used for the statistical prediction of coding sequences in eukaryotes. Use GenScan on the unmasked nucleic sequence as to see which are the predicted coding exons(**Be careful** to provide GenScan with the sequence in "raw sequence" format, without the header).

[GenScan results]

> **Question 5**
>
> How many exons are predicted by GenScan ?
> What are the start and end positions ?
> Do the exons found by GenScan concur with the ones found by BlastX coupled with Wise ?

Transcribed sequence prediction

Searching for mRNAs encoded by our genes

It is also possible to compare our genomic sequence to human mRNA sequences. To do so, in Blast, use the masked genomic sequence, choose the `refseq_RNA` bank and limit the request to "Homo sapiens". You can also use the page dedicated to humans and choose the `refseq_RNA` base. Since we want to compare a human genomic sequence to human mRNAs, we can also use megablast (software suggested by default).

[BlastN results with the refseq_RNA base]

> **Question 6**
>
> Do mRNAs align with our sequence ?
> Of how many exons does our gene seem to be made up of ?
> Check the entries to see whether they contain interesting information.

Gene structure reconstruction

Est2genome allows us to align the mRNA we found thanks to Blast to our genomic sequence in order to predict the full gene structure (including the 5' and 3' UTR regions). Launch EST2genome by using the unmasked genomic sequence.

[EST2genome results]

> **Question 7**
>
> How many exons are predicted ?
> What are the positions of the predicted exons ?
> Do the predictions seem reliable ?
> Do the positions predicted by EST2genome correpsond to the start and end positions of the alignments given by Blast ?

Conclusion

Compiling the results

We studied a human genome fragment by confronting it to protein and mRNA sequence data, and by statistically predicting CDS (GenScan). By gathering and comparing results from different analyses, it is possible to annotate this fragment.

> **Question 8**
>
> Are there more exons predicted by comparing to protein sequences, or by comparing to mRNAs ?
> It is possible to find supplementary exons by comparing to mRNAs, those corresponding to 3' and 5' UTR regions (they must therefore be at the start and/or at the end of protein coding exons).
> Have we identified the exon containing the initiation codon, and the one containing the transcription termination codon ?
> Has our analysis allowed us to showcase alternative splicing ?
> You can reconstruct the complete gene sequence.
> You can even calculate the sequence of the protein encoded by the gene, and formulate an hypothesis regarding its function.

Consulting pre-calculated informations

There are banks that localize, on different genomes, all the known informations taken from different data sources such as sequence banks, but also statistical CDS prediction softwares, ... We will consult the informations concerning the TCN1 gene, in other words the gene we studied, on the ensembl.org website. To do so, search in *humans* the *TCN1* gene, and get access to the *TCN1 (Human Gene)* entry location: you then obtain all the informations linked to (aligned to) this gene in the form of 2 graphs (the second graph is available with the link *Go to Region in Detail for more tracks and navigation options (e.g. zooming)*).

> **Question 9**
>
> Do some annotations show the possibility of alternative splicing ?

Web page created by the teams bonsai and EEP, updated in November 2022

Université de Lille