Exercise 4 : Align proteins to the genome using Exonerate, and run an ab-initio gene prediction program (Genscan) to predict gene models

(i) Align proteins to the genome using Exonerate

- First, we're going to align a protein sequence (ENSMUSP00000102876) to a part of a chromosome. We will use the exonerate program for this.

- Exonerate was written by Guy Slater (http://www.ebi.ac.uk/~guy/exonerate/). We have installed an older version of Exonerate here: /usr/local/ensembl/bin/exonerate-0.9.0.

- You'll find some example input files here :
    - Protein-file (*query*) : /home/training/ensembl-genebuild-data/tests/protein.fa
    - DNA sequence (*target*) : /home/training/ensembl-genebuild-data/tests/chromosome:NCBIM37:3:94665450:94866948:1.fa
- You can download more protein sequence from the ensembl website using Biomart.

- Exonerate has a built-in model to align proteins to the genome. To use this model use the --model protein2genome option if you run exonerate. Here's an example commandline :

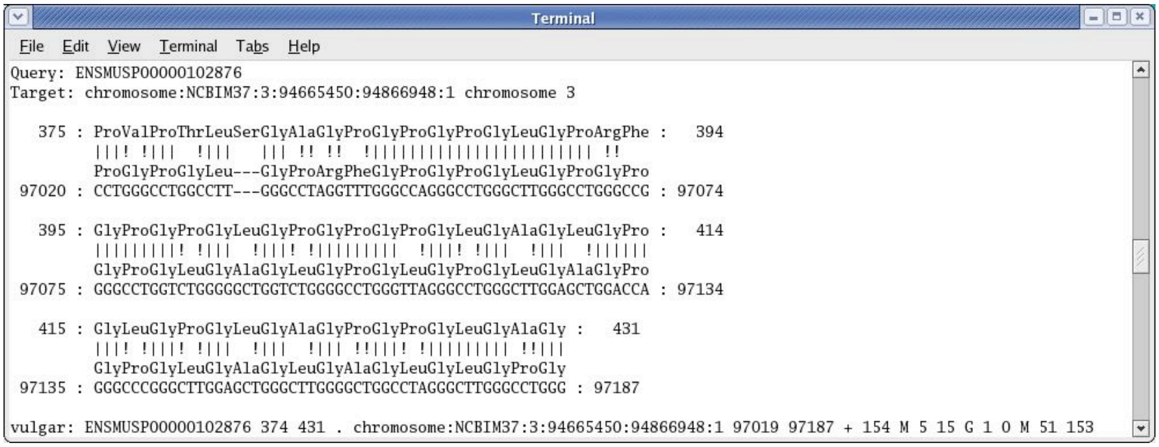
```
/usr/local/ensembl/bin/exonerate-0.9.0 --model protein2genome \
   --query /path/to/your/protein/sequence.fa \
   --target  /path/to/your/genome_sequence_file
```

(You'll get more information about the different options with the --help or -h flag.)

Instructions

- Modify the commandline above and run Exonerate on the example sequence manually - it will run very quickly. Exonerate's results are a sequence alignment.

- We'll see that the protein sequence aligns a few times on chromosome 3 - the alignments differ in their quality.



```
Query: ENSMUSP00000102876
Target: chromosome:NCBIM37:3:94665450:94866948:1 chromosome 3

   375 : ProValProThrLeuSerGlyAlaGlyProGlyProGlyProGlyLeuGlyProArgPhe :    394
         |||! !|||  !|||   ||| !! !!  !|||||||||||||||||||||||| !!
         ProGlyProGlyLeu---GlyProArgPheGlyProGlyProGlyLeuGlyProGlyPro
 97020 : CCTGGGCCTGGCCTT---GGGCCTAGGTTTGGGCCAGGGCCTGGGCTTGGGCCTGGGCCG : 97074

   395 : GlyProGlyProGlyLeuGlyProGlyProGlyProGlyLeuGlyAlaGlyLeuGlyPro :    414
         |||||||||! !|||  !|||! !|||||||||  !|||! !|||  !|||  !||||||
         GlyProGlyLeuGlyAlaGlyLeuGlyProGlyLeuGlyProGlyLeuGlyAlaGlyPro
 97075 : GGGCCTGGTCTGGGGGCTGGTCTGGGGCCTGGGTTAGGGCCTGGGCTTGGAGCTGGACCA : 97134

   415 : GlyLeuGlyProGlyLeuGlyAlaGlyProGlyProGlyLeuGlyAlaGly :   431
         |||! !|||! !|||  !|||  !||| !!|||! !||||||||| !!|||
         GlyProGlyLeuGlyAlaGlyLeuGlyAlaGlyLeuGlyLeuGlyProGly
 97135 : GGGCCCGGGCTTGGAGCTGGGCTTGGGGCTGGCCTAGGGCTTGGGCCTGGG : 97187

vulgar: ENSMUSP00000102876 374 431 . chromosome:NCBIM37:3:94665450:94866948:1 97019 97187 + 154 M 5 15 G 1 0 M 51 153
```

(ii) Run an ab-initio gene prediction program (Genscan) to predict gene models

- Genscan is an ab-initio algorithm that predicts gene models based on DNA sequence only.

Instructions

- Run Genscan on the commandline:

```
  /usr/local/ensembl/bin/genscan\
  /usr/local/ensembl/bin/genscan_matrices/HumanIso.smat \
/home/training/ensembl-genebuild-data/tests/chromosome:NCBIM37:3:94665450:94866948:1.fa
```

(Genscan will predict different gene structures on the chromosome 3 region.)

- If you have time, run the tandem repeat finder TRF on the commandline :

```
/usr/local/ensembl/bin/trf \
/home/training/ensembl-genebuild-data/tests/chromosome:NCBIM37:3:94665450:94866948:1.fa  \
2 5 7 80 10 40 500 -d -h
```

- TRF writes html-files as output in the directory where you're running the commandline - have a look at some of the html files in your browser.

* * * End of Exercise * * *