Assignment 1: Practical Computing for Bioinformatics

1 Accessing information

Find the file **ID_list.txt** on Toledo accompanying the assignment

1. From which database are these IDs?
2. What is the corresponding gene and to which organism does it belong ?
3. Where does the protein reside within the cell ? and to which biological process is it associated?
4. Find the homologues gene in mouse. A certain allele of this gene is associated with the development of a cleft palate in a mouse model. What is the name of this allele? How was this allele introduced in the mouse strain? Name the method used, along with the associated literature reference.
5. Retrieve the GO terms assigned to these sequences using BioMart. List the GO-term(s) that are categorized under the GO-domain 'biological process'.

2 Alignment

Find the file **unknown_sequence.fsa**.

1. To which protein and species does this sequence belong?
2. Retrieve the homologuous proteins in fruit fly and human. What are the corresponding genes ?
3. Compose a fasta file of the three proteins and make a multiple sequence alignment at ebi.ac.uk. Observe the result in MView or JalView. What is the degree of homology ? Describe your finding and explain your reasoning.

3 Linux

3.1 Writing a shell script

In the practicals you have been already introduced to the gff-format (see analyzeGFF.sh) and here we will expand on this. Write a shell-script (analyze_GFF_features.sh) with the following specifications:

Input checks:

The script has a mandatory first argument which specifies a chromosome number. For now we will only work with the human genome.

- if no input argument is given, the script should give an error with appropriate useage information

```
$ ./analyze_GFF_features.sh
usage:./analyze_GFF_features.sh <chromosomeID>
```

- if an invalid chromosome number is given, also an error must be given

```
$ ./analyze_GFF_features.sh 99
99 is not a valid chromosomeID (possible values: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 MT X Y)
```

Retrieving the data:

- The gff file for the chromosome must be automatically downloaded from the ensemble ftp site `https://ftp.ensembl.org/pub/current_gff3/homo_sapiens` and subsequently unzipped.
- If this file was already downloaded previously, this step should be skipped.

Feature count

Display all the different types of features present in the file, in alphabetical order, together with the amount of times these appear.

```
Feature count chromosome 3:
--------------------------
8111 biological_region
57575 CDS
1 chromosome
10999 exon
116545 five_prime_UTR
....
(note: for illustration, this data is not correct)
```

Top 10 lists

Display a list of the transcriptIDs containing the highest occurence of exons. For every transcript, display the transcriptID, the number of exons and the gene to which it belongs (geneID + description). Produce the same list for coding sequences, five-prime-UTR and three-prime-UTR. It should look something like this

```
Top 10: chromosome 9:
---------------------

>>>transcriptIDs with the highest number of exon
Transcript ENST00000624552>>> #exon:99  gene:ENSG00000148357 hemicentin 2 [Source:HGNC Symbol%3BAcc:HGNC:21293]
Transcript ENST00000683500>>> #exon:88  gene:ENSG00000148357 hemicentin 2 [Source:HGNC Symbol%3BAcc:HGNC:21293]
Transcript ENST00000360280>>> #exon:72  gene:ENSG00000197969 vacuolar protein sorting 13 homolog A [Source:HGNC Symbol%3BAcc:HGNC:1908]
Transcript ENST00000376636>>> #exon:71  gene:ENSG00000197969 vacuolar protein sorting 13 homolog A [Source:HGNC Symbol%3BAcc:HGNC:1908]
Transcript ENST00000645632>>> #exon:69  gene:ENSG00000197969 vacuolar protein sorting 13 homolog A [Source:HGNC Symbol%3BAcc:HGNC:1908]
Transcript ENST00000643348>>> #exon:69  gene:ENSG00000197969 vacuolar protein sorting 13 homolog A [Source:HGNC Symbol%3BAcc:HGNC:1908]
Transcript ENST00000371820>>> #exon:66  gene:ENSG00000130635 collagen type V alpha 1 chain [Source:HGNC Symbol%3BAcc:HGNC:2209]
Transcript ENST00000371817>>> #exon:66  gene:ENSG00000130635 collagen type V alpha 1 chain [Source:HGNC Symbol%3BAcc:HGNC:2209]
Transcript ENST00000356083>>> #exon:61  gene:ENSG00000196739 collagen type XXVII alpha 1 chain [Source:HGNC Symbol%3BAcc:HGNC:22986]
Transcript ENST00000706939>>> #exon:58  gene:ENSG00000137076 talin 1 [Source:HGNC Symbol%3BAcc:HGNC:11845]

>>>transcriptIDs with the highest number of CDS
Transcript ENST00000624552>>> #CDS:99  gene:ENSG00000148357 hemicentin 2 [Source:HGNC Symbol%3BAcc:HGNC:21293]
Transcript ENST00000683500>>> #CDS:98  gene:ENSG00000148357 hemicentin 2 [Source:HGNC Symbol%3BAcc:HGNC:21293]
Transcript ENST00000360280>>> #CDS:72  gene:ENSG00000197969 vacuolar protein sorting 13 homolog A [Source:HGNC Symbol%3BAcc:HGNC:1908]
....
(note: for illustration, this data is not correct)
```

Remarks:

- The ReadMe file specifies all you need to know about the format and content of the file.
- The program should only take some seconds to run (excluding the downloading of the data). If you find it takes minutes or longer, maybe consider a different approach.

3.2 Running on the VSC

Now write a job script (.slurm) to launch your script on the cluster.

- Analyse chromosome 1 for this VSC run.
- The slurm script should specify a 5 minute walltime, 1GB of memory, and a single processor.
- Also specify the error and standard output files.
- don't forget to make your script executable:
  chmod u+x analyze_ GFF_features.sh

Provide the shell script (.SH), the job script (.slurm) and the standard + error output file from the VSC run

Good luck :)