Assignment 1

Practical Computing for Bioinformatics

1. Find the file *ids.txt* on Toledo accompanying the assignment.  
   1. The file contains several IDs. From which database are these IDs?  
      NCBI refseq / reference sequences
   2. Look up some of these IDs. Describe what they have in common, including type of sequence and the scientific name of the organism.  
      Peptide chains of Mus Musculus actin   
      <https://www.ncbi.nlm.nih.gov/protein/NP_031418.1>
   3. Use BioMart to retrieve all GO terms assigned to these sequences. Download the table as a TSV and use a linux pipe command to rank the GO terms on unique appearance.  
      Report the most common GO terms and your pipe command.  
      Biomart search on Mouse genes  
      Filter refseq peptide id list

Add gene name and GO terms -> 410 lines tsv file

cat mart\_export.txt | sort | uniq | cut -f3,4 | sort | uniq -c | sort -r  
Text

Description automatically generated

alternative

Text

Description automatically generated

* 1. Still in BioMart, search for orthologs in humans. Do all sequences have an ortholog?  
     Gene names only differ in capitalization, epsilon transcripts have no human ortholog.  
     Biomart alternative attributes -> Orthologues with human, return gene name of both  
     Table

     Description automatically generated
  2. Look up the original gene names on the database specific for this organism. Which gene is expressed in the least tissues?

Epsilon actin is only highly expressed in half of the tissues, the others mostly everywhere.

1. Find the file *sequence.fasta* on Toledo.  
   1. Identify this sequence and the species it belongs to.  
      Human APAF1 mRNA, apoptotic peptidase activating factor

<https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Get&RID=NFWV0PW2013>

* 1. Describe the single point mutation that happened.

A->T in the middle of the sequence

* 1. On which chromosome of the corresponding organism can the gene be found?  
     Chromosome 12  
     <https://www.ncbi.nlm.nih.gov/nucleotide/NM_013229.3>
  2. PDB structure 3J2T is a complex that contains the peptide of this gene. Describe the components and symmetry of the complex.

Hetero-14-mer, A7B7, 7-fold symmetry, 7 x APAF1, 7 x Cytochrome c

* 1. Download the PDB file. Use grep or awk to get the number of disulfide bonds. Report both the result and the command(s) used.

35

awk '($1=="SSBOND")' 3j2t.pdb | wc -l

1. Download *in.zip* from Toledo, send it to the cluster using SCP, and unzip it in your $VSC\_DATA folder. Solve the following questions while working from the cluster as well.  
   1. Write a shell script to manage the files, performing the following tasks:
      * Read the files one by one
      * Sort them in new folders, according to their organism (create folders when needed)
      * Rename them to a better description, like the header of the files
      * Summarize in a new TSV file: original name, new name, gene ID, length of file.   
        (include a header, one line per input file, tab separated)
   2. Text

      Description automatically generatedProvide the TSV file, the shell script (.SH), and the job script (.PBS) that you wrote.

Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence