Assignment 1

Practical Computing for Bioinformatics

Retrieving information, sequence alignment and Linux

1. Find the file IDs.txt on Toledo accompanying the assignment.
   (a) From which database are these IDs?
   (b) Describe what these IDs refer to, the type of sequences, and scientifically describe to which species these sequences belong.
   (c) Retrieve the GO terms assigned to these sequences. Which annotation do they all share?
   (d) Do these sequences have homologs in humans? And in *Saccharomyces cerevisiae*?
2. Find the file with a sequence in mystery.fsa.
   (a) To which gene and species does this sequence belong?
   (b) Retrieve the protein sequence for this gene. It is claimed that the gene is homologous to RAD51 in eukaryotes. Retrieve also the protein sequences for RAD51 of *Saccharomyces cerevisiae* and *Homo sapiens sapiens*. Make a fasta file, and make a multiple sequence alignment at https://www.ebi.ac.uk/Tools/msa/. Observe the result in MView or JalView. Do you see homology? Describe your finding and explain your reasoning.
3. Download the dataset mmp_mut_strains.txt from Toledo. This file contains part of the data provided in the Million Mutation Project for C. elegans (http://genome.sfu.ca/mmp/). This fille contains one mutation per line. The columns describe the details of the mutation: the chromosome (chr) and allele where it occurs, the gene where the mutation occurs, the eect of the mutation, etc. Answer the following questions using linux shell commands, report your commands as well as your answers.
   (a) How many chromosomes are there in the dataset?
   (b) How many mutations are situated in an intron and how many in an exon?
   (c) What type of non-synonymous mutations do you find in exons?
4. Download the file "media.zip" from Toledo. Unpack the zip file in your linux environment. You will see a folder with several files inside. Each file describes the composition of a growth medium for yeast.
   (a) Write a shell script that, given a substance provided as input argument, lists all the media (by number and by name) that do *not* contain this substance.
   (b) Using your script, list the media that do not contain 'Yeast extract'. Report your script as well as your answers.