Final Project

1. Submit your projects the latest by 30th of March 2021
2. There will be no grade before the project is handed in (except when it is intentionally not done)
3. This project is an individual assignment. You may help each other, but you have to hand in your own code
4. It can be coded in python or perl.
5. Create and work with the following folder structure:

projectname/scriptname
put any scripts in the base folder of your project

projectname/runscript
create a master script in bash, perl or python that runs all individual steps in one go. It should also check if result-files already exist and skip the corresponding step with a message

projectname/data/
holds the raw data for your analysis. Please do not submit them with the project as they are huge

projectname/lib/
place any modules here

projectname/results/
write output to this folder

projectname/doc/
create a logfile where you write down what you did each day you worked towards a final solution. Begin each entry with a date, and then write down enough information to be able to easily retrace your steps. You do not need to submit this logfile, but it's good practice to keep one (analogous to a lab journal)

projectname/tmp/
for files used during testing or execution that are not needed afterwards

Guidelines:

- Develop your program in stages. Once part of it works, save the working version to another file before continuing to improve it.
- Define any constant values at the beginning of your script.
- Use as few hardcoded filenames as possible. Define permanent file paths at the beginning of your script.
- Only use relative file paths.
- Always check before accessing a directory or file, that they are available and that you can perform the operation you intend to. If the check fails, create an error message.
- Try to run your code from a different computer or folder, before submitting.
- If possible, run it on a different operating system or ask someone to do so.
- Use subroutine and variables names that reflect their purpose.
- Use indentation (tab-in of lines) within blocks.

- Comment your code. At the very minimum, precede the code for each task with a comment line indicating the function. Better – explain what it does in plain English.

Data source

Use the downloadable data available from eggNOG (http://eggnogdb.embl.de) as your main data source. EggNOG is a database storing information on which genes in a particular species correspond (by an evolutionary relationship) to which genes in other species. In this analysis, we will consider all genes in an Orthologous Group as equally related/important, and not make any difference between orthologs (genes that diverged by a speciation event) or paralogs (genes that diverged by a duplication event). We will also only work with metazoans (= animals).
The EggNOG version we use is 4.5, as the data structure has vastly changed for 5.0

You will need the following files:
http://eggnogdb.embl.de/download/eggnog_4.5/eggnog4.functional_categories.txt
http://eggnogdb.embl.de/download/eggnog_4.5/eggnog4.species_list.txt
http://eggnogdb.embl.de/download/eggnog_4.5/data/meNOG/meNOG.annotations.tsv.gz
http://eggnogdb.embl.de/download/eggnog_4.5/data/meNOG/meNOG.members.tsv.gz

The two files ending in .gz have to be unpacked first, either with gunzip filename or 7zip on windows.
The relationship and contents of the files are described in eggnog-structure.pdf

Project Tasks

1. Write a program that takes two species names as input and calculates how many genes (proteins) in the first species have at least one corresponding (linked via an ortholog group) homolog in the other species. Note that several proteins (paralogs) in one species can belong to the same ortholog group.

   The species should be chosen at runtime via command line arguments or user input, not hardcoded. Only species listed in eggNOG should be considered.

2. Let's use this code and data to look for genes with a restricted taxonomic distribution. Using human genes as a reference point, examine the following

a) How many of the human (Homo sapiens) genes that do not have a homolog in mouse (Mus musculus), but have at least one homolog in chimp (Pan troglodytes)? For sanity checking: the answer should be around 2500

b) What are their protein ids? Store this in a results file.

c) Is there any corresponding (metazoan level) functional description available from eggNOG for those genes? If so, which functional categories do the orthologous groups (NOGs) that they appear in have, and how many proteins in each category? (store this in a formatted results file)

3. Are there any rodent specific genes, i.e. genes in orthologous groups that only contain mouse (Mus musculus) and rat (Rattus norvegicus) proteins, but no other species? If so, what are the corresponding orthologous groups, protein ids, and what are their (metazoan ortholog group level description) functions? For sanity checking: the answer should be Very Very Few.