Exercise 1: Analyze the following sequences using InterProScan.

Go to the directory named exercise2 and look at this patient sequence:

patient.fasta

a- Now, look at this second sequence that is from a patient with a cardio-vascular disease. What is the difference in the sequence?

patient_cardio.fasta

b- Which is the command for interprot that you might use to have the output in html format? (You don't need to launch the command, results are **patient.html** and **patient_cardio.html**. Also, Remember that you can override the default output formats using the -f option). Can you infer a possible reason for the patients disease?

c- What is the Interpro accession of the family this patient belongs to?

d- Which are the GO terms associated to the patient protein?

Exercise 2

Go to the directory named exercise2. There you will see two fasta files called **PENCH.fasta** and **ASPCL.fasta**. Each of the files contains between 200 and 300 proteins.

a- Try to use the web server of Interproscan to analyze the PENCH.fasta sequences altogether (https://www.ebi.ac.uk/interpro/search/sequence-search). Did you have any error? Why?

b- Check how many proteins are in a PENCH.fasta?

c- Write a shell Script that given the filename of a file in FASTA format you want to know how many proteins are in it. (Remember that arguments are accessed inside a script using the variables $1, $2, $3, etc., where $1 refers to the first argument, $2 to the second argument, and so on).

d- Now, you want to scan your sequences (in PENCH.fasta and ASPCL.fasta) for matches against the Interpro database. Show the Interpro command that you might use to have the output in tsv and to include the GO terms (tab-separated values).

e- Which is the GO term that appears more times in each file? How many times? Are both the same GO term? Could you provide information about these GO terms? (You might use: https://www.ebi.ac.uk/QuickGO/)

f- Could you filter the results in ASPCL_GOTERMS.tsv by a particular e-value? Why?

Exercise 3

a- Unzip uniprot_sprot.fasta.gz from the directory: exercise3. From this fasta file extract only the fasta header and save it to a file called Human.fasta (write the command to do that). How many headers are?

b- Write a bash script, named **search_key_words.sh**. The script must go through the keywords file (**key_words.txt**) and for each keyword it should search in **Human.fasta** and it has to print the following: “The keyword **X** was found **Y** times in the fasta file” (where X=keyword Y:=currence of the keyword in Human.fasta)

c- Using the script from exercise 3b and without modifying it, you have to show the keywords and how many times it appears in the file but ordered alphabetically (by occurrence).

Ex:

acetylcholine 400

Cancer 250

helicase 70

….