Aina Montalbán
Practical 3

Exercise 1: Analyze the following sequences using InterProScan.

a- What is the difference between the two sequences?

To compare the two sequences, we can use a bash command. Ww will use:

```
diff patient.fasta patient_cardio.fasta
```

The output we obtain is:

```
1c1
< >patient
---
> >patient_cardio
12c12
< EAVREFAKEIDVSYVKIEEVIGAGEFGEVCRGRLKAPGKKESCVAIKTLKGGYTERQRRE
---
> EAVREFAKEIDVSYVKIEEVIGAGEFGEVCRGRLKAPGKKESCVAISTLKGGYTERQRRE
18a19
>
```

The main difference is found on line 12. In patient fasta file, we can find a Lysine (K), whereas in the patient_cardio.fasta there is a serine.

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?

To obtain the output in html format, we use the next command:

```
./interproscan.sh -i one_protein.fasta -f HTML
```

We open the patient.html and the patient_cardio.html in the browser. There, we can observe one big difference: the patient_cardio doesn't have a protein kinase, an ATP binding site with IPR017441 Interpro accession number. So, a possible reason for the patient disease is that they do not have the protein just mentioned.

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

The Interpro accession is IPR016257.

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

The GO terms are found on the html results.

GO term prediction

Biological Process

GO:0006468 protein phosphorylation
GO:0007169 transmembrane receptor protein tyrosine kinase signaling pathway

Molecular Function

GO:0004672 protein kinase activity
GO:0004713 protein tyrosine kinase activity
GO:0005003 ephrin receptor activity
GO:0005515 protein binding
GO:0005524 ATP binding

Cellular Component

GO:0005887 integral component of plasma membrane
GO:0016021 integral component of membrane

Exercise 2
Go to the directory named exercise2. There you will see two fasta files called YEAST.fasta and PICAN.fasta. Each of the files contains between 200 and 300 proteins.

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

If we analyze all the sequences together, we get an error.

Invalid FASTA format - or sequence contains illegal characters.

We must introduce only 1 sequence.

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

```
grep '>' YEAST.fasta -c
```

We obtain 219 sequences.

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).

```
#!/bin/bash
echo '$1 = ' $1
echo '$2 = ' $2

grep '>' -c $1
grep '>' -c $2
```

Output:

```
$1 =  YEAST.fasta
$2 =  PICAN.fasta
219
209
```

d- Now, you want to scan your sequences (in YEAST.fasta and PICAN.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).

```
./interproscan.sh -i Practical3/YEAST.fasta -f tsv -goterms -o out.YEAST.tsv

./interproscan.sh -i Practical3/PICAN.fasta -f tsv -goterms  -o out.PICAN.tsv
```

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/)

```
cut PICAN_GOTERMS.tsv -f 14 > PICAN_only_GOTERMS.txt
sort PICAN_only_GOTERMS.txt | uniq -c | sort -n
```

The GO term that appears more time in PICAN.tsv is GO:0005515. It appears 86 times.

```
cut YEAST_GOTERMS.tsv -f 14 > YEAST_only_GOTERMS.txt
sort PICAN_only_GOTERMS.txt | uniq -c | sort -n
```

In yeast, we find that is again the same GO term: GO:0005515. It appears 79 times.

The GO term is related with molecular function. In this case, **protein binding**. The definition of the GO term is:

"*Interacting selectively and non-covalently with any protein or protein complex (a complex of two or more proteins that may include other nonprotein molecules).*"

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

Yes, we can filter the results with this command: `awk'{if($9>0.001){print}}' PICAN_GOTERMS.tsv.` However, the result will not be reliable because all the e-values come from different databases, such as Pfam, SMART, PANTHER, among others.

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?

```
grep '>' uniprot_sprot.fasta > Human.fasta
```

There are 553941 headers.

To know the number of headers:

```
grep '>'  Human.fasta -c
```

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)

```
#!/bin/bash

filename1="$1"
while read -r line; do
    x="$line"
    y=`grep -c $x $2`
    echo "The keyword $x was found $y in the fasta file"
done < "$filename1"
```

or

```
#!/bin/bash

for x in $(cat $1);
do y=$(grep -c $x $2);
echo "The keyword $x was found $y in the fasta file.";
done
```

```
The keyword Phosphatidylserine was found 492 in the fasta file.
The keyword helicase was found 3793 in the fasta file.
The keyword acetylcholine was found 115 in the fasta file.
The keyword Cancer was found 45 in the fasta file.
The keyword Saur was found 104 in the fasta file.
The keyword Caspase was found 81 in the fasta file.
The keyword Isochorismatase was found 25 in the fasta file.
The keyword cysteine was found 1819 in the fasta file.
The keyword Endomucin was found 4 in the fasta file.
```

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).

```
sh search_key_words.sh key_words.txt Human.fasta | sort -k 6 -r -n
```

The -k specifies a key, in this case, the column number six, which is the number of occurrences per keyword. Moreover, -n specifies a numeric sort and -r specifies a sort in reverse order.