Exercise 2 – Domain Analysis using 'Interpro' online

Objectives:

- to roughly understand what a 'protein domain' is
- to learn to discover and browse domain information using InterPro.
- to learn how to download the 'seed' alignment sequences for a given domain.

Our test protein:

same protein as in exercise 1, from the 1000-year old skeleton:

```
GFFGDRVGRKFIIWFSILGTAPFALWLPYADADTTAILVILIGFIISSAFASILVYSQELLPKKIGMISGV
FYGFAFGMGGLASALLGKLIDLTDITFVYKVCSFLPLMGLIAYFLPNLRKVKMKE
```

1) What is a 'protein domain':

in short, a *protein domain* is a short stretch of multiple sequence alignment that somehow seems 'important' enough to be given its own name and annotation. Most proteins typically consist of one or more protein domains, but also of non-domain sections that align less well and may perhaps be less important or at least less diagnostic of function.
In practice, protein domains are discovered by experts staring at alignments all day. As part of describing and naming a new domain, these experts will usually provide a summary on where the domain can be found, and what function it might have. Within the actual proteins, domains often represent autonomously folding structural subunits. Many proteins are consisting of multiple different domains, which can be rearranged and exchanged over evolutionary timescales ('domain shuffling'). As of today, most domains in existence have probably been discovered and described already.

2) submit the test protein for a domain search

after exploring an unknown protein against sequence databases (exercise 1), the next typical strategy is to search it for previously described domains. This may provide a broader idea of its function, and often also some three-dimensional structure information.

- open your browser (Chrome/Firefox), and take it to the EBI website: https://www.ebi.ac.uk/
- Click on "Services" (on the very top of the page), then type "InterPro" into the search box on the newly opened page, click "Search" and from the results navigate to the 'InterPro' page.

- Use copy-and-paste to enter our test protein into the box '`Search by sequence`', and click 'Search'. You may need to open the 'Advanced options' section once to make the 'Search' button active and available. No settings need to be changed.
- after a while, your results should become available, and you can first click on the jobId ("iprscan…") and then on "Sequence1" to get a graphical summary, indicating domains and features that have been found for our test protein. First choose "Options" and then "Color by" (drop-down box just above graph) "Domain Relationship". Then, under "Feature Display Mode", choose 'Full'.
- towards the bottom of the display, you can see grey horizontal lines with the predicted protein features. Hover your mouse over some of those grey lines in this section, and you will find at least four of them are likely transmembrane regions, so the protein may be sitting in a membrane. The colored lines above are the actual domains. Here, they refer to the very same domain several times, but some of the classifications are more generic ('family'), whereas other ones are more specific ('domain').
- now, click on domain accession 'IPR020846' on the right of first line under "Domains". This brings up the so-called 'Interpro-abstract', which provides an excellent summary of the general function of this protein family. As you will see, it is a transporter, which the cell uses to transport a variety of 'small molecules' across the membrane. Now we know a lot more already, but of course the annotation is too generic to let us know which is the preferred 'small molecule' that our protein wants to transport.
- now, notice that in the annotation, it said that these proteins typically have 12 trans-membrane regions, but we only found four … this is another indication that our test protein from the skeleton is indeed incomplete. Hence, let's repeat the analysis with the closest relative it has in the sequence databases (this is the 'best hit' we found in exercise 1)
- Copy-and-paste the first best hit found in the file "`input_proteins_1.fa`" (the file generated from exercise 1). It should be the second entry in that file. Submit this protein to InterPro, as before. You may have to clean the sequence before submitting, if it contains illegal characters such as "-". How many transmembrane sections do you find now? (take care to remove the gap characters ["-"] before submitting)
- Next, let's find out a bit more about this domain. Proceed to the Interpro abstract again, like before ('IPR020846').
- On the left of the abstract, click on 'Taxonomy' … this will tell us in which organisms the domain is found. You will find that it occurs virtually everywhere (Bacteria, Eukaryotes, Archaea).
- finally, click on the 'Structures' section. This will provide examples of known three-dimensional structures. Select one of the entries and look at the protein structure. Now we know how our protein generally looks like … but again no useful information about the substrate (in this case, the proteins with the solved structure transport mainly sugars).

3) download a 'seed' alignment from Pfam

The Pfam database (= "Protein families") is similar to InterPro, but is actually one of the original databases for protein domains (in contrast, InterPro is a 'meta-resource' bundling several original databases such as Pfam).
For many of its domains, Pfam maintains a 'seed' alignment, which is the original alignment that was made by the discoverer of the domain (or an updated version thereof). It is made up of non-redundant, representative sequences, and often has been quality-checked manually. For further work on alignments and trees, we'll get one such seed alignment from Pfam now.

- go back to the search results.
- now, click on Pfam accession 'MFS_1 - PF07690' on the right of second line under "Family".
- Next, on the left, find "Alignment", then from the "Available alignments" box choose "seed (192)". Click the Download button to download the seed alignment file.
- Uncompress the downloaded file on the command-line, with command "`gunzip PF07690.alignment.seed.gz`" (on PC: "`gzip -d PF07690.alignment.seed.gz`").
- one thing to notice here: the downloaded file is a "STOCKHOLM"-format file instead of the FASTA format file that we will need. We thus use python to transform the downloaded file to FASTA format. Download the python script "`stoTransform.py`" from OLAT or from github to your laptops folder.
- the script requires the "biopython" module, so let's install that. Make sure your python environment is active, then use this command:

  ```
  pip install biopython
  ```

- after that, we can run the stoTransform.py script:

  ```
  python stoTransform.py -i PF07690.alignment.seed -o input_proteins_2.fa
  ```

- In case the InterPro website happens to be down or too slow, please feel free to proceed directly to Exercise #3 and use the file "`input_proteins_2.fa`" that we have provided through OLAT.