# Exercise 2 – Domain Analysis using 'Interpro' online

**Objectives:**

- roughly understand what a 'protein domain' is
- learn to discover and browse domain information using InterPro.
- 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 than one* protein domain, 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", then on "Proteins" and then on to "InterPro" (fifth entry under Data Resources).
- Use copy-and-paste to enter our test protein into the box 'Analyse your protein sequence', and click 'Submit'.

- after a while, a graphical summary will show up, indicating domains and features that have been found for our test protein. The sets of four grey lines towards the bottom indicate that the protein likely has transmembrane sections, so it may be sitting in a membrane. The colored lines are the actual domains. They refer to the very same domain several times, but some of the classifications are more generic ('superfamily'), whereas other ones are more specific ('domain').
- now, click on second line from the top (under "Domains and Repeats", domain 'IPR020846'). 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 sections, 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)
- Ccpy-and-paste the first best hit found in the file "input_proteins_1.fa. It should be the second entry in that file. Submit this protein to InterPro, as before. 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').
- towards the left of the abstract, click on 'Species' … this will tell us in which organisms the domain is found. If you scroll down a bit (section "Taxa"), you will find that it occurs virtually everywhere (Bacteria, Eukaryotes, Archaea, even viruses).
- next, proceed to 'Domain architectures' … this will tell us which other domains can sometimes be seen together with our domain in a protein. Hover over some of them to see their function. Our case is the most boring, most frequent case: only the domain, and nothing else (first row).
- finally, click on the 'Structures' section. This will provide examples of known three-dimensional structures. Select one entry form the 'PDBe' section, 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).

**2) 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.

- open your browser (Chrome/Firefox), and take it to the EBI website: https://pfam.xfam.org
- towards the middle of the page, click on "SEQUENCE SEARCH", and enter the longer version of our test protein, as we did for Interpro before.
- in case the systems complains about "too many search jobs", you can also try your search here: https://www.ebi.ac.uk/Tools/pfa/pfamscan/ (but you would need to additionally change the output format there, to "plain text").
- this should discover the same domain as before ('Major Facilitator Superfamily', here called 'MFS_1').
- click on MFS_1 to bring up the annotation page. After having a look around, click on the Alignment tab located on the left. Now, under the section "Format an alignment", choose the format "FASTA", and set Gaps as "Gaps as '-' (dashes). Under Download/View choose 'Download', and then click on "Generate".
- when offered the file for download, save it into your home directory, and give it the name "`input_proteins_2.fa`".