Exercise 4 – cut out a domain of interest before aligning

Objectives:

- to learn how to identify domain positions in a protein with '`hmmsearch`'.
- to cut domains from the input proteins before aligning, using Python.

Introduction:

This exercise is somewhat more difficult than the first three, and it involves some serious Python coding. It should be stressed that there are much simpler ways to cut out domains from proteins, but we'll use Python here while the memory is still fresh from the Python introduction. Also, the generic task of parsing some large data files and manipulating them is frequently encountered in Bioinformatics, so this is a good 'real-world' training case for Python.

- on the course-website, there should be the two files below. Please download them and store them into your home directory

`mfs_domain_proteins.fa`	[this is a collection of 15 bacterial proteins, which have at least three domains each: one of our MFS-1 domains, and two enzymatic domains. From each, we want to cut out and align the MFS domain only, and discard the rest]
`MFS_1.hmm`	[this is a 'hidden markov model' describing how to identify "mfs" domains]

1) identify the MFS domains with 'hmmsearch':

- next, will use the 'hmm' file to find the positions of each MFS domain in the set of proteins. We'll run the '`hmmsearch`' utility to scan the hidden markov model along the sequences and to report any hit it finds. You should still have `hmmsearch` installed from the previous exercise.

```
./hmmsearch --domtblout domains_found.tsv MFS_1.hmm mfs_domain_proteins.fa
```

`ls -lart`	[list the contents of the directory, most recent last. there should be a new file]
`head domains_found.tsv`	[check the first few lines of that new file; the domain positions are in 'env coord']

2) use Python to cut out the domains:

- now, we have all the information we need: one file with the protein sequences, and a second file containing the domain coordinates on these sequences.
- here is the challenge: can you write a Python script that uses this information to cut out the domain sequences and to print them into a third file?

hints: first, read the protein sequences, and store these into a dictionary (one 'string' of amino-acids per protein name). Then, read the second file and parse the coordinates. Whenever you've parsed one valid line from the second file, retrieve the corresponding protein sequence from the hash you've made earlier, and cut out the part that contains the domain. Use the 'substr' function in Python to cut out the relevant part of the string. Then, print the substring and proceed to the next line.

- the above task is about as difficult as it ever gets in terms of file-processing for Bioinformatics. So, do play around a bit with the task. Try to write a Python script that does at least part of it … and don't give up too easily … ;)
- when you do get stuck, on the course website you will find the solution (which you can download and run at any time):

```
python cut_domains_from_proteins.py > input_proteins_mfs_domain_only.fa
```

- finally, using `muscle` and `clustalx` as in the exercises before, align and visualize the proteins. You should get a very nice and very compact alignment: