Exercise 1 – Multiple Alignment using NCBI BLAST online

Objectives:

- to become familiar with the NCBI BLAST online system
- to make use of some very basic multiple alignments options there.
- to learn to download the aligned proteins for future analysis.

Our Test Protein:

This protein sequence comes from a 1000-year old skeleton, found at a monastery in Germany. The protein sequence is the translation of a short open reading frame found on a piece of DNA from that skeleton. More precisely, the DNA was from calcified dental calculus (“tooth-plaque”). We’ll work on this short protein for the next couple of exercises. You can use ‘copy-and-paste’ to copy it from here:

```
GFFGDRVGRKFIIWFSILGTAPFALWLPYADADTTAILVILIGFIISSAFASILVYSQELLPKKIGMISGV
FYGFAFGMGGLASALLGKLIDLTDITFVYKVCSFLPLMGLIAYFLPNLRKVKMKE
```

1) submit the protein for a BLAST search

when confronted with an unknown protein, a good idea is to first search a sequence database online, to find out whether similar proteins have been described already.

- open your browser (Chrome/Firefox), and take it to the NCBI website: https://www.ncbi.nlm.nih.gov
- towards the right side of the page, under “Popular Resources”, click “BLAST”.
- Under the “Web BLAST” section, click on “Protein Blast”.
- now, copy-and-paste our test-protein from above, and paste it into the big search box. Scroll down the page, and click the “BLAST” button – database and algorithm choices are already correct.
- wait until the results show up. Scroll down the page a bit, to get to the detailed list of ‘significant’ similar proteins that are already known.
- let’s see what we can find out about our query protein. Click on the second similarity hit that is reported (‘MFS transporter […]’). This will bring up your first protein alignment … so far, it is not a multiple alignment but just an alignment between two proteins (our “query”, and a similar protein found in the database of known proteins [“Subject”]).
- From which organism is the Subject protein? Is that an organism that we would expect to see on the surface of a tooth? Can you translate its Latin name, or at least parts of it ? … ☺
- how long is the subject protein? Longer or shorter than our test protein? Go back and look at a few other subject proteins. What seems to be a typical length for this protein family? Can you think of any reason why our query

protein might be shorter than the typical family member? Remember under what circumstances we found it …

- using Google, try to found out what “MFS transporter” means. Why would this perhaps be an interesting protein to study?
- now go back to the top of the page. There should be a link called “Multiple Alignment” … follow that one now (you may have to wait some time for the results to show up) and then scroll down a bit to the section “Alignment”. Voilà – this is your first multiple sequence alignment. Our test protein is shown at the very top … try to look around and understand what you see. What do the positions labeled ‘-‘ mean? Sometimes there are square brackets with numbers ‘[20]’, what might those mean?

  Hint: compare amino acid sequence length of different proteins in each block
- Is our protein the only protein that is shorter than the others? It looks like there are some other shorter proteins known as well. Is there a part that in the alignment that every protein indeed covers?

  [indeed there is, towards the end. This is the most ‘conserved’ part, and hence likely the most important and characteristic part of the family. It forms part of a protein ‘domain’, as we will see in the next exercise.]

  You can find more of multiple sequence alignment result explanations in at https://www.ncbi.nlm.nih.gov/tools/msaviewer/

2) download the aligned sequences.

- BLAST has done an alignment for us, but what if we want to use a different algorithm and/or different parameter settings? For that, we need to download the protein sequences and work with them locally on our computer.
- in the BLAST multiple alignment page, towards the very top, there is a link labeled ‘Download’. Follow that link, and then select “Fasta plus gaps”. You will be offered to save the alignment into a file and rename them as “`input_proteins_1.fa`” (the file-ending ‘fa’ stands for ‘fasta’, a frequently used file format for storing biological sequences).
- store the file into a suitable location on your laptop.
- now take a look at the file you just saved: open a command prompt (either gitbash or Terminal). There, go to your folder and look at the first few lines of the file:

```
cd [your directory name]          [brings you to your folder, if not already the case]
head -n 30 input_proteins_1.fa    [prints the first 30 lines of the file]
```