GEBZE TECHNICAL UNIVERSITY, FACULTY OF ENGINEERING,
DEPARTMENT OF BIOENGINEERING
2024-2025 FALL SEMESTER
BENG451: Introduction to Bioinformatics
BSB511: Bioinformatics Fundamentals

HOMEWORK II

Please submit a single Jupyter notebook including your answers/solutions.

Multiple sequence alignment and motif identification

Note: For this assignment, you should perform steps 1, 2, 3, 9, 11, and 12 using command-line tools or Python progamming.

1. Align the sequences provided in the fasta file (homework_II_2024.fasta).
2. Identify the motif that match the descriptions provided below:
    - The length of the first motif is 8. It starts with two subsequent C (Cysteine) residues, and ends with another one. The motif also contains a H (Histidine= somewhere near the middle of the motif. The third position of the motif should NOT contain a P (Proline) residue. The 6th position of the motif should NOT contain L, G, or Y.
3. Construct these two motifs as per the style described during the class. A quick refresher:

- [AC]-x-V-x(4)-{ED}.
    - This pattern/motif is translated as: [Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}

4. Go to scanProSite tool at https://prosite.expasy.org/scanprosite/, and use the motifs you constructed, and the **unaligned** fasta file (i.e. the file provided to you) as queries to scan the Prosite database. For this, you need to use the "Option 3 – Submit PROTEIN sequences and MOTIFS to scan them against each other.", and paste the constructed motif string into the Step 2 input query textbox, and the content of the fasta file to the Step 1 input query box.
    - Note: This tool will search for your motif in the provided input sequences. Thus, you can check whether the motif you constructed makes sense for the provided input sequences (i.e. whether it is actually included in the provided sequences).
5. How many hits were found in the input sequences?
    - Note, if no hits have been found, go back to Step 2-3 and re-construct your motif. Then, repeat step 4.
6. Go to the details on the first hit (this is somewhere down the page within a section titled ("hits by patterns"). Click on "View all PROSITE motifs hits on sequence".
7. Under the "hits by patterns" of the popup window that opens, find a link starting with "PS", and click on it. This should take you to the PROSITE description of the respective ProRule (in our case this should be a pattern/motif).

8. The resulting page gives a description of the protein (family) that features this motif, along with a “consensus pattern”. If this pattern/motif same/similar to the motif you constructed? What is the function of this pattern/motif? In which protein family is it found?
9. Go to NCBI Blast service and perform a blastp (Protein BLAST) query using the first protein sequence found in the fasta file provided to you. Choose UniProtKB/SwissProt as the database to search within. Leave all other settings at their default values.
10. Describe the results briefly, including the protein(s) returned, the range of max. identities with the input query, and E values.
11. Retrieve the sequences of these hits.
12. Repeat steps 1-3, this time using the downloaded fasta file as input.
13. Were you able to identify the same motif as before? Were you able to identify the PROSITE pattern mentioned above? Why, or why not? Discuss.