**BT302: Bioinformatics Lab Assignment No. 1 10.08.2022**

**Note 1:** Submit the assignment online through Moodle either in .doc or .pdf format. Your final report file should be named as "**YourName_BT302_Lab1_10082022**". Make sure that your name and roll numbers are written at the first page of your final report. Note that you can upload only one file; thus, put together all the answers in a single file.

**Note 2:** There are two parts of this assignment. In this first part, students are expected to answer the questions in sections A-C. In the second part, students are expected to write a script to achieve the intended results. Each student should choose only one of the parts.

**--------------------------------------------------PART 1-----------------------------------------------**

**1A. The goal of this exercise is to get the statistics (number and types) of nucleotide sequences in the NCBI database.**

**Visit the NCBI homepage; select the taxonomy database from the drop-down menu; click on search; click on the statistics link and answer the following questions:**

1. What is the total number of bacterial, archaeal and viral species present in the database till date?
2. How many new bacterial, archaeal and viral species were added to the database in the year 2022?
3. What is the total number of eukaryotic species present in the database for the year you joined the institute for your current degree?
4. How many more species have been added till date?

**1B. The goal of this exercise is to compare the contents of different databases**

**Go to the NCBI homepage; select the nucleotide database from the drop-down menu; search for the protein "Cas9 from *Streptococcus thermophilus*"; Select the strain 24853 and open in a new tab.**

1. Note down the GenBank id, locus no., number of base pairs, type of molecule, date of deposition, date of revision and version no.
2. Extract and highlight the coding DNA sequence(s) (CDS) part of the DNA.
3. How many amino acids are encoded by this gene(s)? Note down the protein_id(s).
4. What are differences between the contents of the nucleotide (GenBank) and protein (GenPept) database of this gene?

**Go to the UniProtKB database; search for "Cas9"; select the *Streptococcus thermophilus* and reviewed entries; and open in a new tab. [Hint: UniProt id-Q03LF7]**

5. Note down its UniProt id.
6. What are the Molecular function and Biological processes of this gene?

7. What are the pathological roles of this gene?
8. Is there any Post-Translational Modification of this protein?
9. What are the different proteins which interact with this protein?
10. Whether the tertiary structure of this protein is available? If yes, what are the PDB id(s)?
11. Can you find the links for GenBank and RefSeq ids from this database? If yes, which subsection provides that?

Write a note of your experience about the utilities of these databases considering the information available for this gene.

**1C. The goal of this exercise is to understand the file format and content of a protein structure in PDB database**

**Go to the RCSB (PDB) database; search for "Cas9"; select the entry from *Streptococcus thermophilus* and open in a new tab.**

1. In RCSB, how many structures are available for this proptein? Is the PDB id(s) match with that of the UniProtKB database?
2. Open a crystal structure in a new tab. Note down the PDB id, protein length, resolution and year of deposition, release and modification.
3. Is there any ligand bound to the structure? If yes, name them.
4. Can you find the UniProtKB id of this protein in this database? If yes, whether it matches with that of the previous question.
5. What is the expression system used?
6. Which crystal condition was used to crystallize this protein?
7. What was the pH at which the protein was crystallized? What is the pI of this protein (use the program ProtParam)?

**----------------------------------------------------END of PART 1----------------------------------------**

**----------------------------------------------------PART 2-------------------------------------------------**

**2A.** Write a script to download a user-input genome using its "genome id" or "genome name".

**2B.** Write a script to extract the user-intended genes using the keywords such as gene name, gene id and so on.

**Note:** Copy paste or attach your written code in the report and create a README file on how to use the code.

**----------------------------------------------------END of PART 2----------------------------------------**