Lab Session 7

23.09.2024

BT3051 - DSA Biology Lab

Problem 1:

Write a Python program to compute the edit distance between 2 sequences.

- Given a set of 10 gene sequences, compute the edit distances between all pairs of sequences and find the most closely related pair of sequences.
- Given a query protein sequence and a set of other protein sequences, find the protein to which the query is closely related.

Instructions:

- Download gene sequences (FASTA) of insulin (INS) from humans and other organisms for the comparison from https://www.ncbi.nlm.nih.gov/gene/
- Download protein sequences (FASTA) of insulin from https://www.uniprot.org/

Problem 2:

Read about Needleman-Wunch algorithm for sequence alignment and comment on its similarity/difference with Levenshstein distance

Problem 3:

SMILES is a string representation format for chemical compounds. Download SMILES strings for some compounds and compute the distance between them. List out the scenarios where this is useful. What other metrics can be used to compute distance between SMILES. You can download SMILES from https://pubchem.ncbi.nlm.nih.gov/