Indian Institute of Technology Gandhinagar

BE623 Biocomputing

Semester-I 2025-2026

Project –1

Total Points: 30

Instructions:

- The deadline to submit the assignment is **19th September 2025, 11:59 PM. No requests for extension of deadline will be considered.**
- The project assignment needs to be completed and a single file with all the screenshots, GitHub link etc. needs to be **submitted at google classroom and turned in** before the deadline.
- Copying of assignment solutions from any source including Gen AI tools, internet sources, books etc. will be considered a violation of honor code and reported for disciplinary action. In any such case, zero marks will be awarded for the entire assignment.
- Use of Gen AI tools for assistance with the assignment is discouraged, however, if used for any specific section, the exact screenshot of the prompt used, answer received, and the subsequent solution utilized for assignment needs to be submitted.

Question 1 (5 Points)

Choose a gene of your interest from human (*Homo sapiens*) and perform the following:

a) Mention the gene name and some description and search for this gene at NCBI. Refine the query using the methods discussed in the lab sessions and download the gene sequence as fasta file (Take screenshots to show the search) (1 Point)
b) Find the orthologous (similar) gene from Mouse (Mus musculus), Rat (Rattus norvegicus), Chimpanzee (Pan troglodytes) and Fruit fly (Drosophila melanogaster) and download the sequences as fasta file. (1 Point)
c) For each of these five sequences, extract the header and save in a separate file (1 Point)
d) calculate the GC content of each sequence (2 Points)

Question 2 (15 Points)

For the same chosen gene:

a. Use the coding sequence (CDS) of this gene to translate to amino acid sequence. You may reduce the gene length to only 30 nucleotides for ease. (Hint: use Genetic code table and search/replace, Coding sequence can be downloaded from NCBI) (3 points)
b. Retrieve the protein sequence for this gene from UniProt in FASTA format and compare the sequence of gene translated by you with that retrieved from Uniprot. (2 points) (Hint: You may use bash commands to compare files or strings)
c. Use the protein sequence from Uniprot to run a BLASTP search at NCBI and download the top 5 homologous sequences as a FASTA file. (1 Point)
d. From this combined FASTA file, extract all the sequence IDs and save them in a separate file. (1 point)
e. Extract sequences that contain a hydrophobic residue followed by a charged residue followed by an aromatic residue and save them in a separate FASTA file. (2 points)
f. Calculate the percentage of cysteine residues in each sequence and report it. (2 points)
g. From the BLAST results, find the longest sequence and report its ID and its length. For this sequence, calculate the amino acid frequency and provide a distribution of each residue type. (4 points)

Question 3 (7 Points)

a. Retrieve the 3D structure of the same protein from the RCSB PDB database in PDB format. If the PDB structure is not present, then download the PDB file of the AlphaFold model from AlphaFold database. (1 Points)
b. Extract the following details of all C-alpha (CA) atoms of Alanine residues in chain A. (3 Points)
    - Atom name, Residue name, Residue ID, Chain ID, x, y, z coordinates and save as a new alanine_info.pdb
c. Using the extracted CA coordinates of Alanine, compute the pairwise distances between all alanine CA atoms in chain A. Report the results in the form of a table with the following format: (3 Points)
    - Residue_i | Residue_j | Distance

Project Submission (3 Points)

a. Create a GitHub repository named **Project-1-Biocomputing**.

   The repository must contain:

   - data/ (all FASTA and PDB files retrieved)
   - scripts/ (all shell scripts used)
   - results/ (output files)
   - README.txt (including the steps you performed, the commands/scripts used to upload the files and folders in GitHub.)

b. A single pdf with all relevant screenshots and link to GitHub repository to be submitted over classroom.