Indian Institute of Technology Gandhinagar
BE623 Biocomputing
Sem1 2025-2026
Lab Assignment –3

Text processing (sed and awk)

Instructions

- For each question take a screen shot of the command used and the output in the same terminal. Crop the screenshot to only the relevant region of the terminal to reduce the size and past all the screen shots in a single file to uplead the final pdf.
- If you use any help such as copilot, Chat GPT etc. Mention the output of your prompt and how did you proceed from there to solve the question.
- You may be asked how did you solve any of these problems in class as part of the spot quiz. So, even if you have used some help, please ensure that you understand how to solve the problem before proceeding to submit.

1. Create a file with some text written every alternate line using vi. Now delete all empty lines from file using sed (Hint use wildcards for beginning and end of lines)
2. Using the same file created above, add line numbers in front of each line and save in another file.
3. Print only the header lines from clock_gene.fasta using sed.
4. Print all headers from protein.fasta that contain the word CLOCK.
5. Extract sequences from protein.fasta that contain at least two consecutive C's (CC).
6. Count the total number of G's in clock_gene.fasta.
7. Print only lines 5 to 28 from clock_gene.fasta.
8. Print only the sequence ID (without >) from each header in protein.fasta.
9. From protein.fasta, extract sequence lines that start with M and end with Q.
9. Find the length of each sequence in protein.fasta and print it alongside the sequence ID.
10. Print all ATOM lines from protein.pdb that belong to chain A only.
11. Extract all ATOM lines for residues LYS or ARG in protein.pdb.
12. Replace every occurrence of LYS with ARG in protein.pdb.
13. Print only the z-coordinate (third number in coordinates) for each atom from protein.pdb.
14. Count how many lines in protein.pdb contain a GLY residue.
15. Print only the C-alpha (CA) atoms for residues ALA or GLY.

16. Count how many atoms are carbon (element C) in protein.pdb.
17. Print only the HETATM lines from protein.pdb.
18. Extract all residue names that end with “E” (e.g., ILE, PHE).
19. Delete all the lines that contain TER or END from protein.pdb.
20. From protein.pdb, print only the ATOM lines that do not belong to residue ARG.
21. Extract all residues and their frequencies from chain A.
22. From protein.pdb, print only atom name, residue name, and chain ID, separated by commas.
22. Replace all lowercase letters in sequences of protein.fasta with uppercase
23. Find the sequence(s) in protein.fasta with the maximum length.
24. Extract unique residue names from protein.pdb and sort them alphabetically.
25. Find how many distinct chains are present in protein.pdb.
26. From clock_gene.fasta, count nucleotide frequencies (A, T, G, C) separately.