Indian Institute of Technology Gandhinagar

BE623 Biocomputing

Sem1 2025-2026

Lab Assignment –2

Linux & Shell Scripting with Biological Data Files

Important:

- Do not copy commands from the internet — your answer will be checked for correctness by running it in the lab environment.
- Some questions require looking at the actual files before deciding the correct command.
- You may combine commands using pipes (|) or use intermediate variables where needed.

Part 1 - vi Basics & File Editing

1. Open a new file called notes.txt in vi.
   - Insert exactly one line of text:
     Have a nice day
   (Make sure there is no trailing space at the end.)
   - Save and exit.
   - Verify that the file contains exactly one line and 15 characters.

Part 2 - Pattern Matching in FASTA Files

2. Display the last four lines of sequence.fasta without opening the file in an editor.

3. In sequence5.fasta, print all header lines (lines starting with >).

4. Find all matches in sequence5.fasta where A is followed by any single character and then G.

5. Find all matches in sequence5.fasta where P is followed by any character except A, then L.

6. Print all lines in sequence5.fasta that have exactly 2 consecutive Vs anywhere in the line.

7. Print all lines in sequence5.fasta that contain either AA or DD.

8. Print only the sequence lines (ignore headers) from sequence5.fasta that contain the letter P.

Part 3 - Using Variables

9. Store the filename sequence5.fasta in a variable called seq and print the number of sequences in it (headers count as sequences).

10. Store the pattern G\{2,\} in a variable and search protein.fasta for sequence lines (ignore headers) with 2 or more consecutive Gs.

11. Store "Biocomputing" in a variable, export it, and verify that it is available inside a new shell started using:

```
bash -c 'echo $VARIABLE_NAME'
```

Part 4 - File Existence & Loops

12. Write a shell script that checks if sequence3.fasta exists in the current folder. If yes, print the number of lines. If no, print "Missing file".

13. Using a for loop, go through all .fasta files in the current directory and print: filename, number of sequences, and file size in characters.

14. Modify the above loop so that it only prints files with more than 3 sequences.

Part 5 - Applied Data Extraction

15. From sequence5.fasta, extract only the sequence lines (no headers) that contain 3 or more cysteines (C). Save the output to a file named cys_rich.txt. Ensure the output file contains no empty lines.

Extra Challenge (Optional)

Write a single shell command that finds the file in the current directory with the largest number of sequences (by header count) and prints:

```
<filename> has <count> sequences
```

Hint: You will likely need wc, grep, sort, and head.