Indian Institute of Technology Gandhinagar
BE623 Biocomputing
Sem1 2025-2026

Lab session –3

Advanced shell scripting: Text processing (sed and awk)

1. sed on FASTA files

- sed -n '/^>/p' clock_gene.fasta
  # Print only the header lines (starting with >)
- sed '/^>/! s/T/U/g' clock_gene.fasta
  # Replace all T with U in DNA sequences (convert DNA → RNA, keep headers unchanged)
- sed '/^>/d' protein_sequence.fasta
  # Delete all header lines, keeping only sequences
- sed '/^>/ s/$/ #DNA/' clock_gene.fasta
  # Add the word #DNA at the end of every header line

2. sed on PDB files

- sed -n '/^ATOM/p' protein.pdb # Print only the ATOM lines
- sed -n '/^HETATM/p' protein.pdb # Print only the HETATM lines
- sed '/^HETATM/d' protein.pdb # Delete all HETATM lines
- sed 's/ALA/GLY/g' protein.pdb # Replace all occurrences of ALA (Alanine) with GLY (Glycine)

3. awk on FASTA files

- awk '/^>/' clock_gena.fasta | wc -l
  # Count the number of sequences in a FASTA file (headers only)
- awk '/^>/{if (seqlen){print header, seqlen}; header=$0; seqlen=0; next}{seqlen+=length($0)}END{print header, seqlen}' clock_gene.fasta
  # Print header and sequence length for each record
- awk '!/^>/' protein.fasta # Extract only sequence lines (ignore headers)

4. awk on PDB files

- awk '/^ATOM/ {print $7, $8, $9}' protein.pdb
  # Print all atom coordinates (x, y, z) columns
- awk '/^ATOM/ {print $2, $4}' protein.pdb
  # Print atom serial number and residue name
- awk '/^ATOM/ {res[$4]++} END {for (r in res) print r, res[r]}' protein.pdb
  # Count how many times each residue appears (frequency of amino acids)
- awk '/^ATOM/ && $3=="CA" {print $0}' protein.pdb
  # Extract only C-alpha atoms

5. Combined sed + awk

- sed '/^>/d' clock_gene.fasta | awk '{seqlen+=length($0)} END {print "Total length:", seqlen}'
  # Remove headers from FASTA (sed), then count sequence lengths with awk
- sed 's/HIS/HSE/g' protein.pdb | awk '/^ATOM/ && $3=="CA"'
  # In PDB, replace HIS with HSE (sed), then print only C-alpha lines (awk)
- sed -n '/^ATOM/p' protein.pdb | awk '$3=="N" || $3=="CA" || $3=="C"||$3=="O"'
  # Extract only backbone atoms (N, CA, C, O)
- sed '/^>/d' clock_gene.fasta | awk '{gc+=gsub(/[GC]/,""); total+=length($0)} END {print "GC% =", (gc/total)*100}'
  # % of GC content in sequence