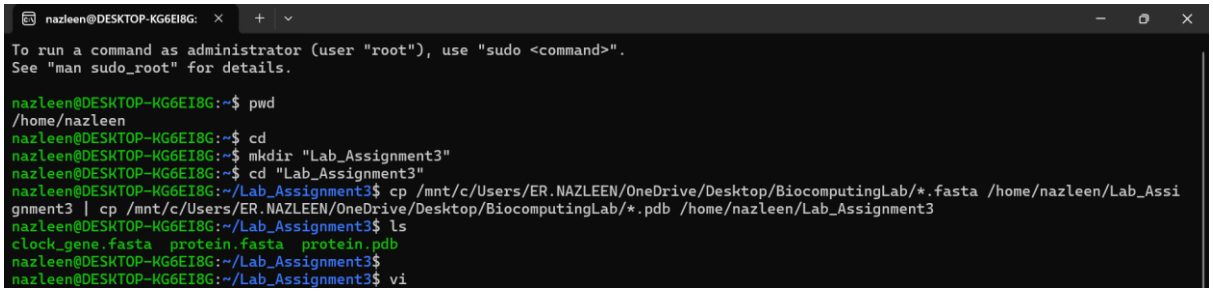
BE623 Biocomputing

Sem1 2025-2026

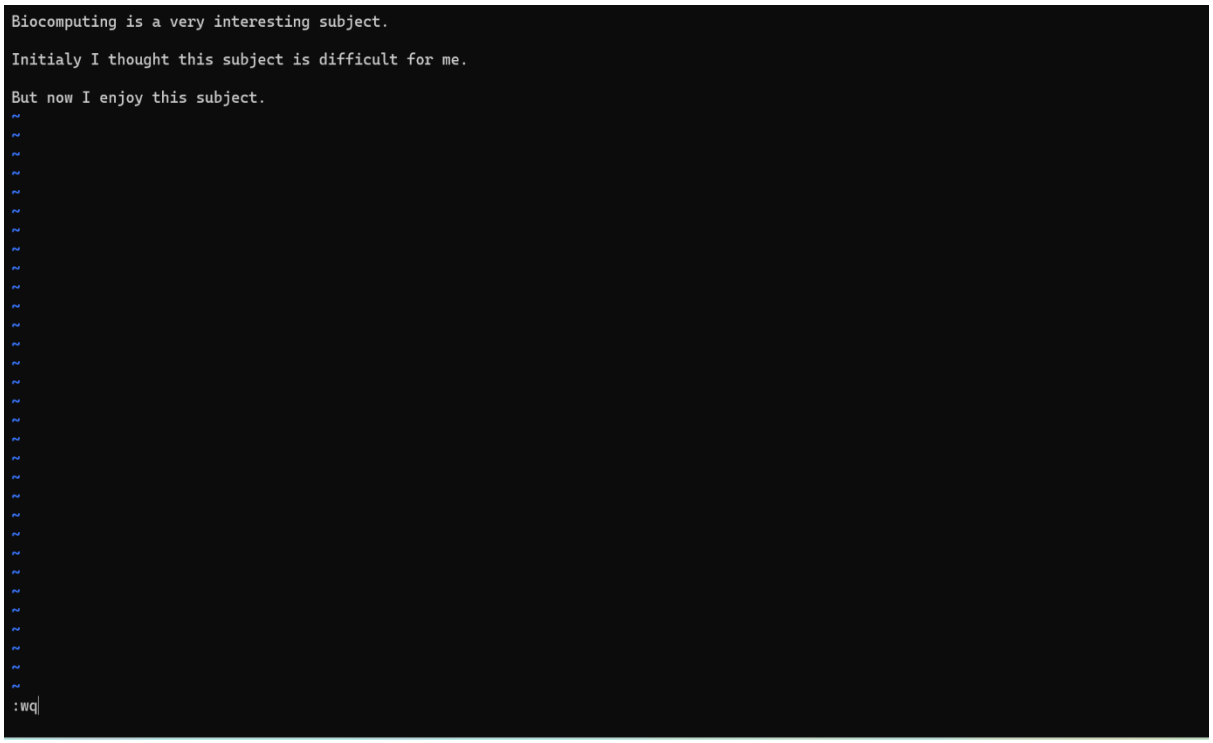
Lab Assignment –3

Text processing (sed and awk)



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)



```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ sed -i '/^$/d' File.txt
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ vi File.txt
```

```
Biocomputing is a very interesting subject.
Initialy I thought this subject is difficult for me.
But now I enjoy this subject.
~
~
~
"File.txt" 3L, 127B                                           1,1          All
```

2. Using the same file created above, add line numbers in front of each line and save in

another file.

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ sed = File.txt | sed '{N;s/\n/ /}' > New_File.txt
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ vi New_File.txt
```

```
1 Biocomputing is a very interesting subject.
2 Initialy I thought this subject is difficult for me.
3 But now I enjoy this subject.
~
~
~
"New_File.txt" 3L, 133B                                       1,1          All
```

3. Print only the header lines from clock_gene.fasta using sed.

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ sed -n '/^>/p' clock_gene.fasta
>NC_000004.12:c55546909-55427903 Homo sapiens chromosome 4, GRCh38.p14 Primary Assembly
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$
```

4. Print all headers from protein.fasta that contain the word CLOCK.

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ sed -n '/^>.*CLOCK/p' protein.fasta
>seq1|Homo_sapiens|CLOCK_protein
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$
```

5. Extract sequences from protein.fasta that contain at least two consecutive C's (CC).

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ awk '/CC/' protein.fasta
MTEYKLVVVGAGCCGKSALTIQLInhfgFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG
MADQLTEEQIAEFKEAFSLFDKDGDGTCCTKELGTVMRSCCQNPTEAELQDMINEVDADGNGQ
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$
```

6. Count the total number of G's in clock_gene.fasta.

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ sed '/^>/d' clock_gene.fasta | awk '{g+=gsub(/[G]/,""); total+=length($0)} END {print g}'
clock_gene.fasta
356
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$
```

7. Print only lines 5 to 28 from clock_gene.fasta.

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ sed -n '5,28p' clock_gene.fasta
GTGGAGGAGGGGAAGGGAAGGGAGGGGGAGGAGGAGCTGGCCACAGGAGCGGCGAATTTTTGGGGGGGTG
GGTGGGGGGCGCCACTCACAGCCCCAGGTGCTGCTGGAGGTGGGAGCCGCGGCGCCTCCTGGACACAGGC
GGGGTAGTGGTTCCGAGTCACCGCAGCGGGAGACCTGGGTGGGGGAGGGAAGAAGCCGGAGCCGCCGCAA
GCCACACGGTGAGGGCGCGGGGAAGGGGAGGGAGCGGGGGGCGGCGTGTGTGGGGCCGGGGGGCGGCGGC
CAAGGGTGGGGAAGGCGGGAGCTGAAGCCCAAGTTTGGCGTGTCGTTCTAGTGTGTCTTTTCCCGGGACT
TCGGGCCGAGGCCCGCCCTGCCTGAGAGGCCCTCTGGGGCAGCTGGGGTTACCTGCGGGGCAGGGGCGGG
AGTGGGGTGCACGGCGGGGCCGGGCGGCTTGAGGGCGCCCGGAGCTGCGGCCGATTCCAGCAGCTGGGAG
GCGGGGAAAGACGGGGACCGGGTGCCGAGAGAGCTTTCGCTGGGGACCCGCTAGGCCTTGTGACCCACTT
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$
```

8. Print only the sequence ID (without >) from each header in protein.fasta.

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ awk '!/^>/' protein.fasta | sed -n '/ID/p' protein.fasta
MTEYKLVVVGAGCCGKSALTIQLInhfgFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG
MTEYKLVVVGDVGKSTIVKQMQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG
MNVQLKKQLKDLPGVIVLGPPGAGKGTQFVSYVLNQLPQYLKKIDVYRTKGF
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$
```

9. From protein.fasta, extract sequence lines that start with M and end with Q.

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ awk '/^M.*Q$/' protein.fasta
MADQLTEEQIAEFKEAFSLFDKDGDGTCCTKELGTVMRSCCQNPTEAELQDMINEVDADGNGQ
MADSQRRLLQNVINKAAGKSSTLLPVDGDKILVVTTGGQVVQSNVLEAMKELLQ
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$
```

9. Find the length of each sequence in protein.fasta and print it alongside the sequence

ID.

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ awk '/^>/ {if(seqlen) {print seq_id, seqlen}; seq_id=$0; seqlen=0; next}{seqlen+=length($0
)} END{print seq_id,seqlen}' protein.fasta
>seq1|Homo_sapiens|CLOCK_protein 61
>seq2|Mus_musculus|PER_protein 56
>seq3|Drosophila_melanogaster|TIM_protein 63
>seq4|Danio_rerio|BMAL_protein 58
>seq5|Arabidopsis_thaliana|LHY_protein 54
>seq6|Saccharomyces_cerevisiae|CYC_protein 57
>seq7|Caenorhabditis_elegans|CLK_protein 54
>seq8|Gallus_gallus|CRY_protein 54
>seq9|Escherichia_coli|RecA_protein 52
>seq10|Xenopus_laevis|REV-ERB_protein 47
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$
```

10. Print all ATOM lines from protein.pdb that belong to chain A only.

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ awk '/^ATOM/ && $5=="A" {print$0}' protein.pdb
ATOM      1  N   TRP A 172     -39.136 -21.997  24.415  1.00 34.43           N
ATOM      2  CA  TRP A 172     -40.108 -20.907  24.729  1.00 34.28           C
ATOM      3  C   TRP A 172     -41.403 -21.065  23.944  1.00 33.46           C
ATOM      4  O   TRP A 172     -41.385 -21.496  22.789  1.00 33.48           O
ATOM      5  CB  TRP A 172     -39.506 -19.534  24.418  1.00 35.12           C
ATOM      6  CG  TRP A 172     -38.161 -19.292  25.025  1.00 36.34           C
ATOM      7  CD1 TRP A 172     -37.773 -19.568  26.306  1.00 37.69           C
ATOM      8  CD2 TRP A 172     -37.032 -18.693  24.384  1.00 37.47           C
ATOM      9  NE1 TRP A 172     -36.465 -19.190  26.497  1.00 37.97           N
ATOM     10  CE2 TRP A 172     -35.985 -18.650  25.334  1.00 37.83           C
ATOM     11  CE3 TRP A 172     -36.799 -18.192  23.097  1.00 37.57           C
ATOM     12  CZ2 TRP A 172     -34.725 -18.128  25.037  1.00 37.51           C
ATOM     13  CZ3 TRP A 172     -35.545 -17.671  22.802  1.00 37.85           C
ATOM     14  CH2 TRP A 172     -34.523 -17.646  23.769  1.00 37.43           C
ATOM     15  N   LYS A 173     -42.516 -20.697  24.576  1.00 32.18           N
ATOM     16  CA  LYS A 173     -43.842 -20.728  23.949  1.00 31.37           C
ATOM     17  C   LYS A 173     -44.028 -19.604  22.914  1.00 29.85           C
ATOM     18  O   LYS A 173     -44.831 -19.725  21.976  1.00 30.15           O
ATOM     19  CB  LYS A 173     -44.935 -20.645  25.024  1.00 31.31           C
ATOM     20  CG  LYS A 173     -46.343 -20.964  24.519  1.00 32.53           C
ATOM     21  CD  LYS A 173     -47.425 -20.459  25.479  1.00 32.89           C
ATOM     22  CE  LYS A 173     -48.818 -20.684  24.901  1.00 33.96           C
ATOM     23  NZ  LYS A 173     -49.893 -20.189  25.806  1.00 34.66           N
ATOM     24  N   GLU A 174     -43.280 -18.518  23.090  1.00 27.67           N
ATOM     25  CA  GLU A 174     -43.337 -17.366  22.191  1.00 25.77           C
ATOM     26  C   GLU A 174     -41.922 -17.014  21.728  1.00 23.54           C
ATOM     27  O   GLU A 174     -41.381 -15.977  22.138  1.00 23.23           O
ATOM     28  CB  GLU A 174     -43.933 -16.148  22.913  1.00 25.76           C
ATOM     29  CG  GLU A 174     -45.376 -16.258  23.359  1.00 26.89           C
ATOM     30  CD  GLU A 174     -45.777 -15.061  24.206  1.00 27.42           C
ATOM     31  OE1 GLU A 174     -46.102 -14.001  23.639  1.00 29.42           O
ATOM     32  OE2 GLU A 174     -45.756 -15.182  25.445  1.00 30.63           O
ATOM     33  N   PRO A 175     -41.313 -17.867  20.872  1.00 21.55           N
ATOM     34  CA  PRO A 175     -39.891 -17.705  20.564  1.00 20.10           C
ATOM     35  C   PRO A 175     -39.565 -16.385  19.866  1.00 18.58           C
ATOM     36  O   PRO A 175     -38.520 -15.781  20.142  1.00 18.18           O
```

11. Extract all ATOM lines for residues LYS or ARG in protein.pdb.

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ awk '/^ATOM/ && $4=="LYS" || $4=="ARG" {print$0}' protein.pdb
ATOM     15  N   LYS A 173     -42.516 -20.697  24.576  1.00 32.18           N
ATOM     16  CA  LYS A 173     -43.842 -20.728  23.949  1.00 31.37           C
ATOM     17  C   LYS A 173     -44.028 -19.604  22.914  1.00 29.85           C
ATOM     18  O   LYS A 173     -44.831 -19.725  21.976  1.00 30.15           O
ATOM     19  CB  LYS A 173     -44.935 -20.645  25.024  1.00 31.31           C
ATOM     20  CG  LYS A 173     -46.343 -20.964  24.519  1.00 32.53           C
ATOM     21  CD  LYS A 173     -47.425 -20.459  25.479  1.00 32.89           C
ATOM     22  CE  LYS A 173     -48.818 -20.684  24.901  1.00 33.96           C
ATOM     23  NZ  LYS A 173     -49.893 -20.189  25.806  1.00 34.66           N
ATOM     46  N   ARG A 177     -41.200 -13.469  20.062  1.00 17.53           N
ATOM     47  CA  ARG A 177     -41.351 -12.338  20.984  1.00 18.15           C
ATOM     48  C   ARG A 177     -40.135 -12.196  21.880  1.00 18.13           C
ATOM     49  O   ARG A 177     -39.608 -11.088  22.053  1.00 17.51           O
ATOM     50  CB  ARG A 177     -42.634 -12.450  21.807  1.00 18.62           C
ATOM     51  CG  ARG A 177     -42.872 -11.237  22.713  1.00 20.72           C
ATOM     52  CD  ARG A 177     -44.227 -11.292  23.368  1.00 22.66           C
ATOM     53  NE  ARG A 177     -44.366 -10.263  24.391  1.00 24.94           N
ATOM     54  CZ  ARG A 177     -43.848 -10.348  25.616  1.00 25.91           C
ATOM     55  NH1 ARG A 177     -43.147 -11.413  25.983  1.00 25.04           N
ATOM     56  NH2 ARG A 177     -44.030  -9.360  26.477  1.00 26.28           N
ATOM     94  N   ARG A 182     -34.717  -9.406  22.797  1.00 19.68           N
ATOM     95  CA  ARG A 182     -33.268  -9.544  22.849  1.00 20.05           C
ATOM     96  C   ARG A 182     -32.593  -8.739  21.743  1.00 19.42           C
ATOM     97  O   ARG A 182     -31.576  -8.072  21.990  1.00 19.22           O
ATOM     98  CB  ARG A 182     -32.874 -11.019  22.769  1.00 20.66           C
ATOM     99  CG  ARG A 182     -33.592 -11.864  23.806  1.00 23.33           C
ATOM    100  CD  ARG A 182     -32.691 -12.324  24.917  1.00 31.08           C
ATOM    101  NE  ARG A 182     -32.238 -13.693  24.676  1.00 34.53           N
ATOM    102  CZ  ARG A 182     -32.720 -14.777  25.285  1.00 36.34           C
ATOM    103  NH1 ARG A 182     -33.684 -14.685  26.205  1.00 37.09           N
ATOM    104  NH2 ARG A 182     -32.223 -15.966  24.975  1.00 37.59           N
ATOM    147  N   LYS A 189     -27.943  -1.219  22.313  1.00 19.72           N
ATOM    148  CA  LYS A 189     -26.592  -1.220  22.859  1.00 19.83           C
ATOM    149  C   LYS A 189     -25.535  -0.931  21.783  1.00 19.51           C
ATOM    150  O   LYS A 189     -24.637  -0.121  22.008  1.00 19.20           O
ATOM    151  CB  LYS A 189     -26.300  -2.544  23.584  1.00 19.67           C
```

12. Replace every occurrence of LYS with ARG in protein.pdb.

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ sed 's/LYS/ARG/g' protein.pdb
HEADER    PEPTIDE BINDING PROTEIN                 26-MAY-05   1ZT3
TITLE     C-TERMINAL DOMAIN OF INSULIN-LIKE GROWTH FACTOR BINDING PROTEIN-1
TITLE    2 ISOLATED FROM HUMAN AMNIOTIC FLUID
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: INSULIN-LIKE GROWTH FACTOR BINDING PROTEIN 1;
COMPND   3 CHAIN: A;
COMPND   4 FRAGMENT: C-TERMINAL DOMAIN;
COMPND   5 SYNONYM: IGFBP-1, IBP- 1, IGF-BINDING PROTEIN 1, PLACENTAL PROTEIN
COMPND   6 12, PP12
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   3 ORGANISM_COMMON: HUMAN;
SOURCE   4 ORGANISM_TAXID: 9606;
SOURCE   5 OTHER_DETAILS: AMNIOTIC FLUID
KEYWDS    INSULIN-LIKE GROWTH FACTOR BINDING PROTEIN-1, IGFBP-1, AMNIOTIC
KEYWDS   2 FLUID, C-TERMINAL DOMAIN, METAL-BINDING, PEPTIDE BINDING PROTEIN
EXPDTA    X-RAY DIFFRACTION
AUTHOR    A.SALA,S.CAPALDI,M.CAMPAGNOLI,B.FAGGION,S.LABO,M.PERDUCA,A.ROMANO,
AUTHOR   2 M.E.CARRIZO,M.VALLI,L.VISAI,L.MINCHIOTTI,M.GALLIANO,H.L.MONACO
REVDAT   5   16-OCT-24 1ZT3    1       REMARK
REVDAT   4   11-OCT-17 1ZT3    1       REMARK
REVDAT   3   24-FEB-09 1ZT3    1       VERSN
REVDAT   2   30-AUG-05 1ZT3    1       JRNL
REVDAT   1   28-JUN-05 1ZT3    0
JRNL        AUTH   A.SALA,S.CAPALDI,M.CAMPAGNOLI,B.FAGGION,S.LABO,M.PERDUCA,
JRNL        AUTH 2 A.ROMANO,M.E.CARRIZO,M.VALLI,L.VISAI,L.MINCHIOTTI,
JRNL        AUTH 3 M.GALLIANO,H.L.MONACO
JRNL        TITL   STRUCTURE AND PROPERTIES OF THE C-TERMINAL DOMAIN OF
JRNL        TITL 2 INSULIN-LIKE GROWTH FACTOR-BINDING PROTEIN-1 ISOLATED FROM
JRNL        TITL 3 HUMAN AMNIOTIC FLUID
JRNL        REF    J.BIOL.CHEM.                  V. 280 29812 2005
JRNL        REFN                   ISSN 0021-9258
JRNL        PMID   15972819
JRNL        DOI    10.1074/JBC.M504304200
REMARK   2
REMARK   2 RESOLUTION.    1.80 ANGSTROMS.
```

```
ATOM      1  N   TRP A 172     -39.136 -21.997  24.415  1.00 34.43           N
ATOM      2  CA  TRP A 172     -40.108 -20.907  24.729  1.00 34.28           C
ATOM      3  C   TRP A 172     -41.403 -21.065  23.944  1.00 33.46           C
ATOM      4  O   TRP A 172     -41.385 -21.496  22.789  1.00 33.48           O
ATOM      5  CB  TRP A 172     -39.506 -19.534  24.418  1.00 35.12           C
ATOM      6  CG  TRP A 172     -38.161 -19.292  25.025  1.00 36.34           C
ATOM      7  CD1 TRP A 172     -37.773 -19.568  26.306  1.00 37.69           C
ATOM      8  CD2 TRP A 172     -37.032 -18.693  24.384  1.00 37.47           C
ATOM      9  NE1 TRP A 172     -36.465 -19.190  26.497  1.00 37.97           N
ATOM     10  CE2 TRP A 172     -35.985 -18.650  25.334  1.00 37.83           C
ATOM     11  CE3 TRP A 172     -36.799 -18.192  23.097  1.00 37.57           C
ATOM     12  CZ2 TRP A 172     -34.725 -18.128  25.037  1.00 37.51           C
ATOM     13  CZ3 TRP A 172     -35.545 -17.671  22.802  1.00 37.85           C
ATOM     14  CH2 TRP A 172     -34.523 -17.646  23.769  1.00 37.43           C
ATOM     15  N   ARG A 173     -42.516 -20.697  24.576  1.00 32.18           N
ATOM     16  CA  ARG A 173     -43.842 -20.728  23.949  1.00 31.37           C
ATOM     17  C   ARG A 173     -44.028 -19.604  22.914  1.00 29.85           C
ATOM     18  O   ARG A 173     -44.831 -19.725  21.976  1.00 30.15           O
ATOM     19  CB  ARG A 173     -44.935 -20.645  25.024  1.00 31.31           C
ATOM     20  CG  ARG A 173     -46.343 -20.964  24.519  1.00 32.53           C
ATOM     21  CD  ARG A 173     -47.425 -20.459  25.479  1.00 32.89           C
ATOM     22  CE  ARG A 173     -48.818 -20.684  24.901  1.00 33.96           C
ATOM     23  NZ  ARG A 173     -49.893 -20.189  25.806  1.00 34.66           N
ATOM     24  N   GLU A 174     -43.280 -18.518  23.090  1.00 27.67           N
ATOM     25  CA  GLU A 174     -43.337 -17.366  22.191  1.00 25.77           C
ATOM     26  C   GLU A 174     -41.922 -17.014  21.728  1.00 23.54           C
ATOM     27  O   GLU A 174     -41.381 -15.977  22.138  1.00 23.23           O
ATOM     28  CB  GLU A 174     -43.933 -16.148  22.913  1.00 25.76           C
ATOM     29  CG  GLU A 174     -45.376 -16.258  23.359  1.00 26.89           C
ATOM     30  CD  GLU A 174     -45.777 -15.061  24.206  1.00 27.42           C
ATOM     31  OE1 GLU A 174     -46.102 -14.001  23.639  1.00 29.42           O
ATOM     32  OE2 GLU A 174     -45.756 -15.182  25.445  1.00 30.63           O
ATOM     33  N   PRO A 175     -41.313 -17.867  20.872  1.00 21.55           N
ATOM     34  CA  PRO A 175     -39.891 -17.705  20.564  1.00 20.10           C
ATOM     35  C   PRO A 175     -39.565 -16.385  19.866  1.00 18.58           C
ATOM     36  O   PRO A 175     -38.520 -15.781  20.142  1.00 18.18           O
ATOM     37  CB  PRO A 175     -39.594 -18.893  19.632  1.00 20.52           C
```

13. Print only the z-coordinate (third number in coordinates) for each atom from protein.pdb.

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ awk '/^ATOM/ {print$9}' protein.pdb
24.415
24.729
23.944
22.789
24.418
25.025
26.306
24.384
26.497
25.334
23.097
25.037
22.802
23.769
24.576
23.949
22.914
21.976
25.024
24.519
25.479
24.901
25.806
23.090
22.191
21.728
22.138
22.913
23.359
24.206
23.639
25.445
20.872
20.564
19.866
20.142
```

14. Count how many lines in protein.pdb contain a GLY residue.

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ awk '/^ATOM/ && $4=="GLY" {print$0}' protein.pdb | wc -l
28
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$
```

15. Print only the C-alpha (CA) atoms for residues ALA or GLY.

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ awk '/^ATOM/ && $4=="GLY" || $4=="ALA" {print$0}' protein.pdb | awk '$3=="CA"'
ATOM    143  CA  ALA A 188     -29.906  -0.273  21.249  1.00 19.62           C
ATOM    157  CA  ALA A 190     -24.689  -1.402  19.528  1.00 20.13           C
ATOM    193  CA  GLY A 195     -19.179   3.890  13.965  1.00 34.45           C
ATOM    315  CA  GLY A 210     -45.353 -14.753  19.536  1.00 18.56           C
ATOM    422  CA  GLY A 223     -36.815   5.170   1.658  1.00 21.58           C
ATOM    435  CA  ALA A 225     -37.186  -1.492   0.463  1.00 20.30           C
ATOM    440  CA  GLY A 226     -35.705  -3.955   2.980  1.00 18.85           C
ATOM    526  CA  GLY A 236     -37.957 -18.276  12.295  1.00 18.22           C
ATOM    565  CA  GLY A 241     -34.199 -22.463  -1.334  1.00 28.67           C
ATOM    610  CA  GLY A 247     -40.259  -7.039  -1.851  1.00 24.01           C
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$
```

16. Count how many atoms are carbon (element C) in protein.pdb.

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ awk '/^ATOM/ && $12=="C" {print$0}' protein.pdb | wc -l
401
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$
```

17. Print only the HETATM lines from protein.pdb.

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ sed -n '/^HETATM/p' protein.pdb
HETATM  644  C1  DIO A 400     -29.064  -6.946  17.132  1.00 36.16           C
HETATM  645  C2  DIO A 400     -28.073  -9.061  16.720  1.00 36.92           C
HETATM  646  C1' DIO A 400     -27.687  -6.281  17.202  1.00 35.99           C
HETATM  647  C2' DIO A 400     -26.684  -8.437  16.825  1.00 36.68           C
HETATM  648  O1  DIO A 400     -28.996  -8.072  16.254  1.00 36.78           O
HETATM  649  O1' DIO A 400     -26.726  -7.251  17.629  1.00 36.28           O
HETATM  650  O   HOH A   1     -37.255  -6.228  10.647  1.00 14.97           O
HETATM  651  O   HOH A   2     -22.012  -0.788  22.336  1.00 20.64           O
HETATM  652  O   HOH A   3     -38.877  -3.391   4.471  1.00 20.33           O
HETATM  653  O   HOH A   4     -34.212 -23.871   7.998  1.00 18.39           O
HETATM  654  O   HOH A   5     -20.730  -0.315  24.894  1.00 20.65           O
HETATM  655  O   HOH A   6     -44.936 -13.438   1.965  1.00 28.30           O
HETATM  656  O   HOH A   7     -48.895 -18.702  15.563  1.00 27.48           O
HETATM  657  O   HOH A   8     -21.393  -0.854  17.811  1.00 24.13           O
HETATM  658  O   HOH A   9     -32.124   5.776   0.506  1.00 29.82           O
HETATM  659  O   HOH A  10     -46.186 -13.792   6.539  1.00 23.52           O
HETATM  660  O   HOH A  11     -29.575  -1.996  25.245  1.00 28.23           O
HETATM  661  O   HOH A  12     -45.642 -11.444  19.694  1.00 25.61           O
HETATM  662  O   HOH A  13     -49.384 -20.064  17.570  1.00 29.28           O
HETATM  663  O   HOH A  14     -30.137  -4.552   3.329  1.00 27.31           O
HETATM  664  O   HOH A  15     -42.693  -7.945  15.244  1.00 19.76           O
HETATM  665  O   HOH A  16     -35.906 -28.174   5.866  1.00 31.98           O
HETATM  666  O   HOH A  17     -44.171  -7.687  17.621  1.00 22.18           O
HETATM  667  O   HOH A  18     -47.265 -12.454  21.564  1.00 29.40           O
HETATM  668  O   HOH A  19     -36.430   3.094  -3.026  1.00 25.02           O
HETATM  669  O   HOH A  20     -29.553  -5.969  12.150  1.00 34.06           O
HETATM  670  O   HOH A  21     -42.686  -4.398  27.240  1.00 25.96           O
HETATM  671  O   HOH A  22     -43.889  -9.382  19.695  1.00 29.00           O
HETATM  672  O   HOH A  23     -43.476  -6.477  -2.563  1.00 30.73           O
HETATM  673  O   HOH A  24     -28.999   3.283  21.951  1.00 26.71           O
HETATM  674  O   HOH A  25     -50.516 -11.430  14.190  1.00 25.35           O
HETATM  675  O   HOH A  26     -27.306   5.304  20.576  1.00 30.44           O
HETATM  676  O   HOH A  27     -48.424 -14.440  -0.286  1.00 61.67           O
HETATM  677  O   HOH A  28     -43.808 -10.099   7.884  1.00 28.89           O
HETATM  678  O   HOH A  29     -35.566  -5.200  24.698  1.00 29.22           O
HETATM  679  O   HOH A  30     -34.679  -7.575  -4.768  1.00 25.20           O
```

18. Extract all residue names that end with “E” (e.g., ILE, PHE). **{taken idea from ChatGPT}**

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ awk '/^ATOM/ && /[A-Z][A-Z]E/ {print$4}' protein.pdb
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
PHE
PHE
PHE
PHE
PHE
PHE
PHE
PHE
PHE
PHE
PHE
PHE
PHE
PHE
PHE
PHE
PHE
PHE
PHE
PHE
PHE
PHE
```

19. Delete all the lines that contain TER or END from protein.pdb. **{taken help from ChatGPT for how to separate both residues}**

Beginning:

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ sed '/TER/d; /END/d' protein.pdb
HEADER    PEPTIDE BINDING PROTEIN                 26-MAY-05   1ZT3
TITLE    2 ISOLATED FROM HUMAN AMNIOTIC FLUID
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: INSULIN-LIKE GROWTH FACTOR BINDING PROTEIN 1;
COMPND   3 CHAIN: A;
COMPND   5 SYNONYM: IGFBP-1, IBP- 1, IGF-BINDING PROTEIN 1, PLACENTAL PROTEIN
COMPND   6 12, PP12
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   3 ORGANISM_COMMON: HUMAN;
SOURCE   4 ORGANISM_TAXID: 9606;
SOURCE   5 OTHER_DETAILS: AMNIOTIC FLUID
KEYWDS    INSULIN-LIKE GROWTH FACTOR BINDING PROTEIN-1, IGFBP-1, AMNIOTIC
EXPDTA    X-RAY DIFFRACTION
AUTHOR    A.SALA,S.CAPALDI,M.CAMPAGNOLI,B.FAGGION,S.LABO,M.PERDUCA,A.ROMANO,
AUTHOR   2 M.E.CARRIZO,M.VALLI,L.VISAI,L.MINCHIOTTI,M.GALLIANO,H.L.MONACO
REVDAT   5   16-OCT-24 1ZT3    1       REMARK
REVDAT   4   11-OCT-17 1ZT3    1       REMARK
REVDAT   3   24-FEB-09 1ZT3    1       VERSN
REVDAT   2   30-AUG-05 1ZT3    1       JRNL
REVDAT   1   28-JUN-05 1ZT3    0
JRNL        AUTH   A.SALA,S.CAPALDI,M.CAMPAGNOLI,B.FAGGION,S.LABO,M.PERDUCA,
JRNL        AUTH 2 A.ROMANO,M.E.CARRIZO,M.VALLI,L.VISAI,L.MINCHIOTTI,
JRNL        AUTH 3 M.GALLIANO,H.L.MONACO
JRNL        TITL 2 INSULIN-LIKE GROWTH FACTOR-BINDING PROTEIN-1 ISOLATED FROM
JRNL        TITL 3 HUMAN AMNIOTIC FLUID
JRNL        REF    J.BIOL.CHEM.                 V. 280 29812 2005
JRNL        REFN                   ISSN 0021-9258
JRNL        PMID   15972819
JRNL        DOI    10.1074/JBC.M504304200
REMARK   2
REMARK   2 RESOLUTION.    1.80 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3   PROGRAM     : REFMAC 5.2.0005
REMARK   3   AUTHORS     : MURSHUDOV,SKUBAK,LEBEDEV,PANNU,STEINER,
```

Ending:

```
HETATM  668  O   HOH A  19     -36.430   3.094  -3.026  1.00 25.02           O
HETATM  669  O   HOH A  20     -29.553  -5.969  12.150  1.00 34.06           O
HETATM  670  O   HOH A  21     -42.686  -4.398  27.240  1.00 25.96           O
HETATM  671  O   HOH A  22     -43.889  -9.382  19.695  1.00 29.00           O
HETATM  672  O   HOH A  23     -43.476  -6.477  -2.563  1.00 30.73           O
HETATM  673  O   HOH A  24     -28.999   3.283  21.951  1.00 26.71           O
HETATM  674  O   HOH A  25     -50.516 -11.430  14.190  1.00 25.35           O
HETATM  675  O   HOH A  26     -27.306   5.304  20.576  1.00 30.44           O
HETATM  676  O   HOH A  27     -48.424 -14.440  -0.286  1.00 61.67           O
HETATM  677  O   HOH A  28     -43.808 -10.099   7.884  1.00 28.89           O
HETATM  678  O   HOH A  29     -35.566  -5.200  24.698  1.00 29.22           O
HETATM  679  O   HOH A  30     -34.679  -7.575  -4.768  1.00 25.20           O
HETATM  680  O   HOH A  31     -41.964 -17.506  25.641  1.00 37.16           O
HETATM  681  O   HOH A  32     -34.312  -2.922  25.191  1.00 31.83           O
HETATM  682  O   HOH A  33     -51.606 -11.651  21.823  1.00 29.90           O
HETATM  683  O   HOH A  34     -32.561 -16.311  28.119  1.00 50.80           O
HETATM  684  O   HOH A  35     -34.469 -16.004   9.163  1.00 24.01           O
HETATM  685  O   HOH A  36     -31.585 -23.210   8.833  1.00 26.89           O
HETATM  686  O   HOH A  37     -49.015 -19.802  20.176  1.00 31.69           O
HETATM  687  O   HOH A  38     -30.973 -14.980   5.105  1.00 43.06           O
HETATM  688  O   HOH A  39     -47.022 -17.146  11.346  1.00 28.11           O
HETATM  689  O   HOH A  40     -30.833  -7.743  14.123  1.00 34.35           O
HETATM  690  O   HOH A  41     -25.168   6.080  14.148  1.00 49.89           O
HETATM  691  O   HOH A  42     -51.167 -14.258  13.359  1.00 47.34           O
CONECT   45  288
CONECT  288   45
CONECT  382  456
CONECT  456  382
CONECT  476  641
CONECT  641  476
CONECT  644  646  648
CONECT  645  647  648
CONECT  646  644  649
CONECT  647  645  649
CONECT  648  644  645
CONECT  649  646  647
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$
```

20. From protein.pdb, print only the ATOM lines that do not belong to residue ARG.

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ awk '/^ATOM/ && !/ARG/' protein.pdb
ATOM      1  N   TRP A 172     -39.136 -21.997  24.415  1.00 34.43           N
ATOM      2  CA  TRP A 172     -40.108 -20.907  24.729  1.00 34.28           C
ATOM      3  C   TRP A 172     -41.403 -21.065  23.944  1.00 33.46           C
ATOM      4  O   TRP A 172     -41.385 -21.496  22.789  1.00 33.48           O
ATOM      5  CB  TRP A 172     -39.506 -19.534  24.418  1.00 35.12           C
ATOM      6  CG  TRP A 172     -38.161 -19.292  25.025  1.00 36.34           C
ATOM      7  CD1 TRP A 172     -37.773 -19.568  26.306  1.00 37.69           C
ATOM      8  CD2 TRP A 172     -37.032 -18.693  24.384  1.00 37.47           C
ATOM      9  NE1 TRP A 172     -36.465 -19.190  26.497  1.00 37.97           N
ATOM     10  CE2 TRP A 172     -35.985 -18.650  25.334  1.00 37.83           C
ATOM     11  CE3 TRP A 172     -36.799 -18.192  23.097  1.00 37.57           C
ATOM     12  CZ2 TRP A 172     -34.725 -18.128  25.037  1.00 37.51           C
ATOM     13  CZ3 TRP A 172     -35.545 -17.671  22.802  1.00 37.85           C
ATOM     14  CH2 TRP A 172     -34.523 -17.646  23.769  1.00 37.43           C
ATOM     15  N   LYS A 173     -42.516 -20.697  24.576  1.00 32.18           N
ATOM     16  CA  LYS A 173     -43.842 -20.728  23.949  1.00 31.37           C
ATOM     17  C   LYS A 173     -44.028 -19.604  22.914  1.00 29.85           C
ATOM     18  O   LYS A 173     -44.831 -19.725  21.976  1.00 30.15           O
ATOM     19  CB  LYS A 173     -44.935 -20.645  25.024  1.00 31.31           C
ATOM     20  CG  LYS A 173     -46.343 -20.964  24.519  1.00 32.53           C
ATOM     21  CD  LYS A 173     -47.425 -20.459  25.479  1.00 32.89           C
ATOM     22  CE  LYS A 173     -48.818 -20.684  24.901  1.00 33.96           C
ATOM     23  NZ  LYS A 173     -49.893 -20.189  25.806  1.00 34.66           N
ATOM     24  N   GLU A 174     -43.280 -18.518  23.090  1.00 27.67           N
ATOM     25  CA  GLU A 174     -43.337 -17.366  22.191  1.00 25.77           C
ATOM     26  C   GLU A 174     -41.922 -17.014  21.728  1.00 23.54           C
ATOM     27  O   GLU A 174     -41.381 -15.977  22.138  1.00 23.23           O
ATOM     28  CB  GLU A 174     -43.933 -16.148  22.913  1.00 25.76           C
ATOM     29  CG  GLU A 174     -45.376 -16.258  23.359  1.00 26.89           C
ATOM     30  CD  GLU A 174     -45.777 -15.061  24.206  1.00 27.42           C
ATOM     31  OE1 GLU A 174     -46.102 -14.001  23.639  1.00 29.42           O
ATOM     32  OE2 GLU A 174     -45.756 -15.182  25.445  1.00 30.63           O
ATOM     33  N   PRO A 175     -41.313 -17.867  20.872  1.00 21.55           N
ATOM     34  CA  PRO A 175     -39.891 -17.705  20.564  1.00 20.10           C
ATOM     35  C   PRO A 175     -39.565 -16.385  19.866  1.00 18.58           C
ATOM     36  O   PRO A 175     -38.520 -15.781  20.142  1.00 18.18           O
```

21. Extract all residues and their frequencies from chain A.

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ awk '/^ATOM/ && $5=="A" {res[$4]++} END {for (r in res) print r,res[r]}' protein.pdb
GLY 28
CYS 37
LEU 32
THR 14
GLN 18
PRO 42
ILE 32
MET 8
ASN 40
TYR 48
LYS 45
ASP 16
SER 36
PHE 22
HIS 10
GLU 81
ARG 55
TRP 42
ALA 15
VAL 21
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$
```

22. From protein.pdb, print only atom name, residue name, and chain ID, separated by

commas.

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ awk '/^ATOM/ {print $3 "," $4 "," $5}' protein.pdb
N,TRP,A
CA,TRP,A
C,TRP,A
O,TRP,A
CB,TRP,A
CG,TRP,A
CD1,TRP,A
CD2,TRP,A
NE1,TRP,A
CE2,TRP,A
CE3,TRP,A
CZ2,TRP,A
CZ3,TRP,A
CH2,TRP,A
N,LYS,A
CA,LYS,A
C,LYS,A
O,LYS,A
CB,LYS,A
CG,LYS,A
CD,LYS,A
CE,LYS,A
NZ,LYS,A
N,GLU,A
CA,GLU,A
C,GLU,A
O,GLU,A
CB,GLU,A
CG,GLU,A
CD,GLU,A
OE1,GLU,A
OE2,GLU,A
N,PRO,A
CA,PRO,A
C,PRO,A
O,PRO,A
```

22. Replace all lowercase letters in sequences of protein.fasta with uppercase **{use ChatGPT for uppercase symbol}**

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ sed '/^>/! s/[a-z]/\U&/g' protein.fasta
>seq1|Homo_sapiens|CLOCK_protein
MTEYKLVVVGAGCCGKSALTIQLINHFGFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG

>seq2|Mus_musculus|PER_protein
MSDDEEVQPSLLTKDGRVLQVLQSLFFGKNSDQLQSLENQLQDLLTAAQNNYSSST

>seq3|Drosophila_melanogaster|TIM_protein
MADQLTEEQIAEFKEAFSLFDKDGDGTCCTKELGTVMRSCCQNPTEAELQDMINEVDADGNGQ

>seq4|Danio_rerio|BMAL_protein
MLSRAVCGTSGTGKSTLSRIIAQYFKKTDVVLVGPSGAGKTTISKLLEQLDYLNQKNV

>seq5|Arabidopsis_thaliana|LHY_protein
MSEQNGVVVDDGSIKVLVTGNKCDPQQRVTSQPVLQAGLDRIFGVIRDLGGSSS

>seq6|Saccharomyces_cerevisiae|CYC_protein
MTEYKLVVVGDVGKSTIVKQMQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG

>seq7|Caenorhabditis_elegans|CLK_protein
MADSQRRLLQNVINKAAGKSSTLLPVDGDKILVVTTGGQVVQSNVLEAMKELLQ

>seq8|Gallus_gallus|CRY_protein
MPGSGYVVRAGTVAGQLRIMNNKVVVVGDLGAGKTTLLQSVIEMLKLLGEKGTA

>seq9|Escherichia_coli|RecA_protein
MNVQLKKQLKDLPGVIVLGPPGAGKGTQFVSYVLNQLPQYLKKIDVYRTKGF

>seq10|Xenopus_laevis|REV-ERB_protein
MADEEKLPPGWEKRMSRSSGRVYYFNHITNASQWERPSGNSSSGSLS
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$
```

23. Find the sequence(s) in protein.fasta with the maximum length.

(not understand)

24. Extract unique residue names from protein.pdb and sort them alphabetically.

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ awk '/^ATOM/ {print $4}' protein.pdb | sort | uniq
ALA
ARG
ASN
ASP
CYS
GLN
GLU
GLY
HIS
ILE
LEU
LYS
MET
PHE
PRO
SER
THR
TRP
TYR
VAL
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$
```

25. Find how many distinct chains are present in protein.pdb.

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ awk '/^ATOM/ || /^HETATM/ {print $5}' protein.pdb | sort -u | wc -l
1
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$
```

26. From clock_gene.fasta, count nucleotide frequencies (A, T, G, C) separately.

```
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$ sed '/^>/d' clock_gene.fasta | awk '{A+=gsub(/A/, ""); T+=gsub(/T/, ""); G+=gsub(/G/, "");
 C+=gsub(/C/, ""); total+=length($0)} END {print "A:", A, "T:", T, "G:", G, "C:", C}' clock_gene.fasta
A: 115 T: 100 G: 356 C: 203
nazleen@DESKTOP-KG6EI8G:~/Lab_Assignment3$
```