BE623 – BIOCOMPUTING

ASSIGNMENT 3

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).

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# vi mytest.txt
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# sed '/^$/d' mytest.txt
Hello everyone. My name is Khushboo Joshi
This is a test file.
Biocomputing assignment third.
Question 1. To remove all the empty lines using the sed command
End of the task.
```

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

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# awk '{ print NR, $0}' mytest.txt > new1.txt
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# head new1.txt
1 Hello everyone. My name is Khushboo Joshi
2
3 This is a test file.
4
5 Biocomputing assignment third.
6
7 Question 1. To remove all the empty lines using the sed command
8
9 End of the task.
```

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

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# sed -n  '/^>/p' clock_gene.fasta
>NC_000004.12:c55546909-55427903 Homo sapiens chromosome 4, GRCh38.p14 Primary Assembly
```

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

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# sed -n '/^>.*CLOCK/p' protein.fasta
>seq1|Homo_sapiens|CLOCK_protein
```

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

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# sed -n '/^[^>].*CC/p' protein.fasta
MTEYKLVVVGAGCCGKSALTIQLInhfgFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG
MADQLTEEQIAEFKEAFSLFDKDGDGTCCTKELGTVMRSCCQNPTEAELQDMINEVDADGNGQ
```

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

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# awk '!/^>/ {g+=gsub(/G/,"")} END {print g}' clock_gene.fasta
23471
```

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

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# sed -n '5,28p' clock_gene.fasta
GTGGAGGAGGGGAAGGGAAGGGAGGGGGAGGAGGAGCTGGCCACAGGAGCGGCGAATTTTTGGGGGGGTG
GGTGGGGGGCGCCACTCACAGCCCCAGGTGCTGCTGGAGGTGGGAGCCGCGGCGCCTCCTGGACACAGGC
GGGGTAGTGGTTCCGAGTCACCGCAGCGGGAGACCTGGGTGGGGGAGGGAAGAAGCCGGAGCCGCCGCAA
GCCACACGGTGAGGGCGCGGGGAAGGGGAGGGAGCGGGGGGCGGCGTGTGTGGGGCCGGGGGGCGGCGGC
CAAGGGTGGGGAAGGCGGGAGCTGAAGCCCAAGTTTGGCGTGTCGTTCTAGTGTGTCTTTTCCCGGGACT
TCGGGCCGAGGCCCGCCCTGCCTGAGAGGCCCTCTGGGGCAGCTGGGGTTACCTGCGGGGCAGGGGCGGG
AGTGGGGTGCACGGCGGGGCCGGGCGGCTTGAGGGCGCCCGGAGCTGCGGCCGATTCCAGCAGCTGGGAG
GCGGGGAAAGACGGGGACCGGGTGCCGAGAGAGCTTTCGCTGGGGACCCGCTAGGCCTTGTGACCCACTT
TATTCCTGTCACCACTCGGGCACGTTTGGAGCAGCGCCCAATGGGGCGCCGGGGCGGCCAGCTCCTCCGG
GGAACCCCCGCCCTCCCGGCGCCCGGCCCGCGTGCCGCAGTCCGCAGTCCGAACGGCCGCCGTTGCCGGC
CGCGGGCTGGTTCCGTTAGTGGTGGTGGTTCCGGGGTTCCGTTCCTAGGCAGCGCGCGGCTATTAGCGTC
TGACTCCAGCGACCGCGCGCGGGTTCGAGGGTTGGCGGCGAGGCGCTCGGTTTCTCTTCTTCCGTCCACC
CGCGCTTCCCGTTCCCCGTCACCGCCGCTGGCGTTTGTAGATTTCTTTCCGCCAGTGAAGCTGGGTTTTC
TGGAGTTGGCTCTGGCGCTCTGGCCCCTGGAGTGTAATTTCCTACACGCAGCGCCGCAGAGTTTATATTC
TTTGAAAGTGTTTGTAGCTTTGTAGAGGTCCTCTTGTTGATGGTAGGTGAGCCTAATTCTGCAAGATAAA
AGCCTAGTCTCTGACCTGGCAGATGAAAGATCAAATCAGATTGTGGTTTCCTGCTATTAGAATGCCGTGC
TATTAGACTTTAAGGCTTTTTAGCCTTCTTTAAAAAATAAAAAAATTTTTACAGTGGAAGAAAAGCACAA
GAAGTAAACTTTTACAGTCGTTGATTTGACTATAACGCTGATCCCCCCAAATCAAAGGTAATTTCACTTT
GAAGATTGCGTTCTGATTTGTAGCTTTAAGCGATTAGAGAAAATTGTGCAATATTCCCCTCTACCTGTTT
GAAAATAAACATTCTTAAAAGGATGTAATTTAGATAATGAATTGCTTTCTCTGAAACTTATCCCTTGGGA
CACCTCAAATCTGATTGGTTTTCAAAGCTTGGGGGAAGGAAATTAATTCCTGTGATAAGTGGTTGGCTGA
ACAGAATGTCTTGAAAGGTTAGCCTGTAGCATTAGGAGAAATACCTAATGTAAACGACGAGTTAATGGGT
GCAGCACACCAGCGTGGCACATTTATACATATGTGACAAACCTGCACGTTGTGCACATGTACCCTAGAAC
TTAAAGTATAATAATAAAAAAAGTAAAAAAAAAAAAAAGTTAGCCTGAAGAAAGCAGACTGAAAATGTTCT
```

8. Print only the sequence ID (without >) from each header in protein.fasta. **(Reference of substr taken from ChatGPT).**

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# awk '/^>/ {print substr($1,2)}' protein.fastaseq1|Homo_sapiens|CLOCK_protein
seq2|Mus_musculus|PER_protein
seq3|Drosophila_melanogaster|TIM_protein
seq4|Danio_rerio|BMAL_protein
seq5|Arabidopsis_thaliana|LHY_protein
seq6|Saccharomyces_cerevisiae|CYC_protein
seq7|Caenorhabditis_elegans|CLK_protein
seq8|Gallus_gallus|CRY_protein
seq9|Escherichia_coli|RecA_protein
seq10|Xenopus_laevis|REV-ERB_protein
```

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

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# awk '/^M.*Q$/' protein.fasta
MADQLTEEQIAEFKEAFSLFDKDGDGTCCTKELGTVMRSCCQNPTEAELQDMINEVDADGNGQ
MADSQRRLLQNVINKAAGKSSTLLPVDGDKILVVTTGGQVVQSNVLEAMKELLQ
```

10. Find the length of each sequence in protein.fasta and print it alongside the sequence ID. **(Reference of substr taken from ChatGPT)**

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# awk '/^>/ {id=substr($1,2); next} {print id, length($0)}' protein.fasta
seq1|Homo_sapiens|CLOCK_protein 61
seq1|Homo_sapiens|CLOCK_protein 0
seq2|Mus_musculus|PER_protein 56
seq2|Mus_musculus|PER_protein 0
seq3|Drosophila_melanogaster|TIM_protein 63
seq3|Drosophila_melanogaster|TIM_protein 0
seq4|Danio_rerio|BMAL_protein 58
seq4|Danio_rerio|BMAL_protein 0
seq5|Arabidopsis_thaliana|LHY_protein 54
seq5|Arabidopsis_thaliana|LHY_protein 0
seq6|Saccharomyces_cerevisiae|CYC_protein 57
seq6|Saccharomyces_cerevisiae|CYC_protein 0
seq7|Caenorhabditis_elegans|CLK_protein 54
seq7|Caenorhabditis_elegans|CLK_protein 0
seq8|Gallus_gallus|CRY_protein 54
seq8|Gallus_gallus|CRY_protein 0
seq9|Escherichia_coli|RecA_protein 52
seq9|Escherichia_coli|RecA_protein 0
seq10|Xenopus_laevis|REV-ERB_protein 47
```

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

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# awk '$1=="ATOM" && $5=="A"' 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
ATOM     37  CB  PRO A 175     -39.594 -18.893  19.632  1.00 20.52           C
ATOM     38  CG  PRO A 175     -40.909 -19.247  19.043  1.00 19.77           C
ATOM     39  CD  PRO A 175     -41.896 -19.015  20.148  1.00 21.28           C
ATOM     40  N   CYS A 176     -40.455 -15.942  18.986  1.00 16.73           N
ATOM     41  CA  CYS A 176     -40.212 -14.710  18.226  1.00 16.80           C
ATOM     42  C   CYS A 176     -40.222 -13.501  19.159  1.00 16.78           C
ATOM     43  O   CYS A 176     -39.363 -12.626  19.053  1.00 16.20           O
ATOM     44  CB  CYS A 176     -41.244 -14.528  17.116  1.00 16.50           C
ATOM     45  SG  CYS A 176     -40.885 -13.084  16.044  1.00 15.20           S
ATOM     46  N   ARG A 177     -41.200 -13.469  20.062  1.00 17.53           N
```

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

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# awk '$1=="ATOM" && ($4=="LYS" || $4=="ARG")' 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
ATOM    152  CG  LYS A 189     -24.980  -2.573  24.353  1.00 21.18           C
ATOM    153  CD  LYS A 189     -24.991  -1.568  25.500  1.00 23.97           C
ATOM    154  CE  LYS A 189     -23.703  -1.601  26.298  1.00 25.23           C
```

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

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# sed 's/LYS/ARG/g' protein.pdb > ARG.pdb
```

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

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# 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
```

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

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# awk '/GLY/ {count++} END {print count}' protein.pdb
33
```

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

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# awk '$1=="ATOM" && $3=="CA" && ($4=="ALA" || $4=="GLY")' protein.pdb
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
```

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

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3#
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# awk '$1=="ATOM" && $NF=="C" {count++} END {print count}' protein.pdb
401
```

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

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# 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
```

19. Extract all residue names that end with “E” (e.g., ILE, PHE).

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3#
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# awk '$4 ~ /E$/ {print $4}' protein.pdb
INSULIN-LIKE
INSULIN-LIKE
RANGE
RANGE
VALUE
VALUE
FREE
FREE
ANGLE
ANGLE
ANGLE
ANGLE
RANGE
RANGE
FREE
PROBE
PROBE
TYPE
SOURCE
TYPE
SOFTWARE
RANGE
RANGE
MERGE
THE
MERGE
SURFACE
THE
SOFTWARE
DIOXIDE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
```

20. Delete all the lines that contain TER or END from protein.pdb.

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# 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
```

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

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# awk '$1=="ATOM" && $4!="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
```

22. Extract all residues and their frequencies from chain A. **(Refrence of sort and uniq taken from ChatGPT.)**

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3#
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# awk '$1=="ATOM" && $5=="A" {print $4}' protein.pdb | sort | uniq -c
     15 ALA
     55 ARG
     40 ASN
     16 ASP
     37 CYS
     18 GLN
     81 GLU
     28 GLY
     10 HIS
     32 ILE
     32 LEU
     45 LYS
      8 MET
     22 PHE
     42 PRO
     36 SER
     14 THR
     42 TRP
     48 TYR
     21 VAL
```

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

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS# cd /mnt/c/Users/ASUS/BE623_labsession_3
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# awk '$1=="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
CB,PRO,A
CG,PRO,A
CD,PRO,A
N,CYS,A
CA,CYS,A
C,CYS,A
O,CYS,A
CB,CYS,A
SG,CYS,A
N,ARG,A
```

24. Replace all lowercase letters in sequences of protein.fasta with uppercase.

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3#
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# 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
```

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

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# awk '/^>/ { if(seqlen){print seqlen}; seqlen=0; next} {seqlen+=length($0)} END {print seqlen}' protein.fasta | sort -nr | head -n 1
63
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# awk '/^>/ { if(seqlen){print seqlen}; seqlen=0; next} {seqlen+=length($0)} END {print seqlen}' protein.fasta | sort -nr
63
61
58
57
56
54
54
54
52
47
```

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

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3#
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# awk '$1=="ATOM" || $1=="HETATM" {print $4}' protein.pdb | sort -u
ALA
ARG
ASN
ASP
CYS
DIO
GLN
GLU
GLY
HIS
HOH
ILE
LEU
LYS
MET
PHE
PRO
SER
THR
TRP
TYR
VAL
```

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

```
> ^C
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# awk '$1=="ATOM" || $1=="HETATM" {print $5}' protein.pdb | sort -u | wc -l
1
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3#
```

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

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_3# awk '/^[^>]/ {seq = seq $0} END {
print "A:", gsub(/A/,"",seq)
print "T:", gsub(/T/,"",seq)
print "G:", gsub(/G/,"",seq)
print "C:", gsub(/C/,"",seq)}' clock_gene.fasta
A: 35332
T: 39197
G: 23471
C: 21007
```