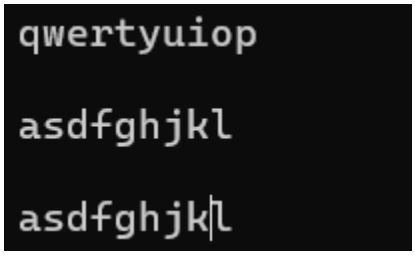
Nabendu Das_24250059
Assignment-3

---

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)

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ vi text-1
```

This will create a file named text-1. Then we entered the file and wrote some text in every alternate line.

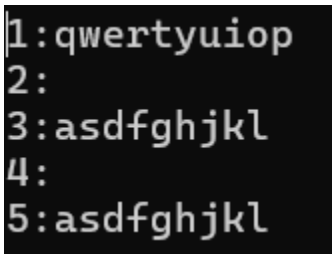



```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ sed '/^$/d' text-1
qwertyuiop
asdfghjkl
asdfghjkl
```

This command will print the lines without empty lines.

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ awk '{print NR ":" $0}' text-1 > text-2.txt
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ vi text-2.txt
```



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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-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.

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-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).

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ sed -n '/CC/p' protein.fasta
MTEYKLVVVGAGCCGKSALTIQLInhfgFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG
MADQLTEEQIAEFKEAFSLFDKDGDGTCCTKELGTVMRSCCQNPTEAELQDMINEVDADGNGQ
```

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ sed '/^>/d
' clock_gene.fasta | awk '{g+=gsub(/[G]/,"")} END {print g}'
355
```

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ awk 'NR>=5 && NR<=28' clock_gene.fasta
GTGGAGGAGGGGAAGGGAAGGGAGGGGGAGGAGGAGCTGGCCACAGGAGCGGCGAATTTTTGGGGGGGTG
GGTGGGGGGCGCCACTCACAGCCCCAGGTGCTGCTGGAGGTGGGAGCCGCGGCGCCTCCTGGACACAGGC
GGGGTAGTGGTTCCGAGTCACCGCAGCGGGAGACCTGGGTGGGGGAGGGAAGAAGCCGGAGCCGCCGCAA
GCCACACGGTGAGGGCGCGGGGAAGGGGAGGGAGCGGGGGGCGGCGTGTGTGGGGCCGGGGGGCGGCGGC
CAAGGGTGGGGAAGGCGGGAGCTGAAGCCCAAGTTTGGCGTGTCGTTCTAGTGTGTCTTTTCCCGGGACT
TCGGGCCGAGGCCCGCCCTGCCTGAGAGGCCCTCTGGGGCAGCTGGGGTTACCTGCGGGGCAGGGGCGGG
AGTGGGGTGCACGGCGGGGCCGGGCGGCTTGAGGGCGCCCGGAGCTGCGGCCGATTCCAGCAGCTGGGAG
GCGGGGAAAGACGGGGACCGGGTGCCGAGAGAGCTTTCGCTGGGGACCCGCTAGGCCTTGTGACCCACTT
```

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ awk '/^>/ {print substr($1,2)}' protein.fasta
seq1|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
```

I use ChatGPT here to know about the substr function.

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ sed -n '/^M.*Q$/p' protein.fasta
MADQLTEEQIAEFKEAFSLFDKDGDGTCCTKELGTVMRSCCQNPTEAELQDMINEVDADGNGQ
MADSQRRLLQNVINKAAGKSSTLLPVDGDKILVVTTGGQVVQSNVLEAMKELLQ
```

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ awk '/^>/{if(seqlen){print header, seql
en}; header=$0; seqlen=0; next} {seqlen+=length($0)} END {print header, 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
```

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ 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
```

This will print 642 lines of records.

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ sed -n '/^ATOM/p' protein.pdb | awk '$4=="LYS" || $4=="ARG"'
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
ATOM    155  NZ  LYS A 189     -23.673  -0.401  27.204  1.00 25.98           N
ATOM    228  N   LYS A 200     -30.993   0.420   7.874  1.00 26.73           N
ATOM    229  CA  LYS A 200     -31.745  -0.835   7.833  1.00 24.20           C
ATOM    230  C   LYS A 200     -31.208  -1.820   8.880  1.00 23.56           C
ATOM    231  O   LYS A 200     -30.014  -1.861   9.160  1.00 23.03           O
ATOM    232  CB  LYS A 200     -31.682  -1.479   6.440  1.00 24.17           C
ATOM    233  CG  LYS A 200     -32.216  -0.609   5.294  1.00 23.41           C
ATOM    234  CD  LYS A 200     -32.263  -1.375   3.981  1.00 22.93           C
ATOM    235  CE  LYS A 200     -32.479  -0.443   2.786  1.00 21.93           C
ATOM    236  NZ  LYS A 200     -31.331   0.512   2.647  1.00 19.78           N
ATOM    297  N   LYS A 208     -49.012 -12.189  16.590  1.00 19.70           N
ATOM    298  CA  LYS A 208     -49.580 -11.893  17.916  1.00 20.21           C
ATOM    299  C   LYS A 208     -49.491 -13.063  18.913  1.00 20.08           C
ATOM    300  O   LYS A 208     -49.635 -12.860  20.118  1.00 20.32           O
ATOM    301  CB  LYS A 208     -51.043 -11.459  17.773  1.00 20.47           C
ATOM    302  CG  LYS A 208     -51.935 -12.512  17.115  1.00 20.38           C
ATOM    303  CD  LYS A 208     -53.396 -12.222  17.359  1.00 22.10           C
ATOM    304  CE  LYS A 208     -54.291 -13.221  16.642  1.00 20.94           C
ATOM    305  NZ  LYS A 208     -54.187 -14.607  17.174  1.00 20.34           N
ATOM    357  N   ARG A 215     -43.344 -14.515   6.254  1.00 18.42           N
ATOM    358  CA  ARG A 215     -42.464 -13.537   5.651  1.00 18.42           C
ATOM    359  C   ARG A 215     -41.666 -12.820   6.745  1.00 17.97           C
ATOM    360  O   ARG A 215     -42.240 -12.338   7.726  1.00 19.04           O
ATOM    361  CB  ARG A 215     -43.275 -12.525   4.835  1.00 18.99           C
ATOM    362  CG  ARG A 215     -42.421 -11.489   4.100  1.00 19.30           C
ATOM    363  CD  ARG A 215     -43.301 -10.359   3.594  1.00 20.84           C
ATOM    364  NE  ARG A 215     -43.854  -9.573   4.697  1.00 20.02           N
ATOM    365  CZ  ARG A 215     -44.864  -8.706   4.586  1.00 22.74           C
```

```
ATOM    366  NH1 ARG A 215     -45.467  -8.510   3.418  1.00 23.51           N
ATOM    367  NH2 ARG A 215     -45.282  -8.040   5.656  1.00 23.60           N
ATOM    529  N   LYS A 237     -36.427 -19.755  11.099  1.00 18.90           N
ATOM    530  CA  LYS A 237     -35.253 -20.079  10.303  1.00 20.07           C
ATOM    531  C   LYS A 237     -35.652 -20.086   8.836  1.00 20.62           C
ATOM    532  O   LYS A 237     -36.709 -20.607   8.487  1.00 20.15           O
ATOM    533  CB  LYS A 237     -34.658 -21.438  10.712  1.00 20.03           C
ATOM    534  CG  LYS A 237     -34.152 -21.504  12.151  1.00 19.85           C
ATOM    535  CD  LYS A 237     -33.395 -22.819  12.393  1.00 20.64           C
ATOM    536  CE  LYS A 237     -32.887 -22.927  13.828  1.00 20.64           C
ATOM    537  NZ  LYS A 237     -32.303 -24.281  14.128  1.00 19.43           N
ATOM    538  N   ARG A 238     -34.811 -19.483   7.993  1.00 21.63           N
ATOM    539  CA  ARG A 238     -35.054 -19.421   6.556  1.00 22.74           C
ATOM    540  C   ARG A 238     -35.290 -20.815   5.978  1.00 23.18           C
ATOM    541  O   ARG A 238     -34.580 -21.765   6.321  1.00 23.36           O
ATOM    542  CB  ARG A 238     -33.882 -18.738   5.842  1.00 23.15           C
ATOM    543  CG  ARG A 238     -34.126 -18.455   4.367  1.00 24.78           C
ATOM    544  CD  ARG A 238     -32.909 -17.817   3.729  1.00 29.07           C
ATOM    545  NE  ARG A 238     -33.127 -17.584   2.305  1.00 32.71           N
ATOM    546  CZ  ARG A 238     -32.328 -16.860   1.525  1.00 33.80           C
ATOM    547  NH1 ARG A 238     -31.254 -16.265   2.028  1.00 35.30           N
ATOM    548  NH2 ARG A 238     -32.617 -16.721   0.240  1.00 34.89           N
ATOM    598  N   ARG A 246     -36.004  -7.648  -2.381  1.00 24.57           N
ATOM    599  CA  ARG A 246     -36.526  -6.407  -1.793  1.00 24.09           C
ATOM    600  C   ARG A 246     -37.988  -6.209  -2.186  1.00 23.73           C
ATOM    601  O   ARG A 246     -38.334  -5.370  -3.019  1.00 22.92           O
ATOM    602  CB  ARG A 246     -35.657  -5.200  -2.156  1.00 24.34           C
ATOM    603  CG  ARG A 246     -34.232  -5.365  -1.662  1.00 25.49           C
ATOM    604  CD  ARG A 246     -33.359  -4.136  -1.804  1.00 25.90           C
ATOM    605  NE  ARG A 246     -32.020  -4.466  -1.317  1.00 27.00           N
ATOM    606  CZ  ARG A 246     -31.617  -4.321  -0.057  1.00 28.42           C
ATOM    607  NH1 ARG A 246     -32.447  -3.835   0.870  1.00 27.71           N
ATOM    608  NH2 ARG A 246     -30.378  -4.676   0.281  1.00 29.27           N
```

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ sed -n '/^ATOM/p' protein.pdb | awk '$4=="LYS" || $4=="ARG"' | sed 's/LYS
/ARG/g'
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     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   ARG A 189     -27.943  -1.219  22.313  1.00 19.72           N
ATOM    148  CA  ARG A 189     -26.592  -1.220  22.859  1.00 19.83           C
ATOM    149  C   ARG A 189     -25.535  -0.931  21.783  1.00 19.51           C
ATOM    150  O   ARG A 189     -24.637  -0.121  22.008  1.00 19.20           O
ATOM    151  CB  ARG A 189     -26.300  -2.544  23.584  1.00 19.67           C
ATOM    152  CG  ARG A 189     -24.980  -2.573  24.353  1.00 21.18           C
ATOM    153  CD  ARG A 189     -24.991  -1.568  25.500  1.00 23.97           C
ATOM    154  CE  ARG A 189     -23.703  -1.601  26.298  1.00 25.23           C
ATOM    155  NZ  ARG A 189     -23.673  -0.401  27.204  1.00 25.98           N
ATOM    228  N   ARG A 200     -30.993   0.420   7.874  1.00 26.73           N
ATOM    229  CA  ARG A 200     -31.745  -0.835   7.833  1.00 24.20           C
ATOM    230  C   ARG A 200     -31.208  -1.820   8.880  1.00 23.56           C
ATOM    231  O   ARG A 200     -30.014  -1.861   9.160  1.00 23.03           O
ATOM    232  CB  ARG A 200     -31.682  -1.479   6.440  1.00 24.17           C
ATOM    233  CG  ARG A 200     -32.216  -0.609   5.294  1.00 23.41           C
ATOM    234  CD  ARG A 200     -32.263  -1.375   3.981  1.00 22.93           C
ATOM    235  CE  ARG A 200     -32.479  -0.443   2.786  1.00 21.93           C
ATOM    236  NZ  ARG A 200     -31.331   0.512   2.647  1.00 19.78           N
ATOM    297  N   ARG A 208     -49.012 -12.189  16.590  1.00 19.70           N
ATOM    298  CA  ARG A 208     -49.580 -11.893  17.916  1.00 20.21           C
ATOM    299  C   ARG A 208     -49.491 -13.063  18.913  1.00 20.08           C
ATOM    300  O   ARG A 208     -49.635 -12.860  20.118  1.00 20.32           O
ATOM    301  CB  ARG A 208     -51.043 -11.459  17.773  1.00 20.47           C
ATOM    302  CG  ARG A 208     -51.935 -12.512  17.115  1.00 20.38           C
ATOM    303  CD  ARG A 208     -53.396 -12.222  17.359  1.00 22.10           C
ATOM    304  CE  ARG A 208     -54.291 -13.221  16.642  1.00 20.94           C
ATOM    305  NZ  ARG A 208     -54.187 -14.607  17.174  1.00 20.34           N
ATOM    357  N   ARG A 215     -43.344 -14.515   6.254  1.00 18.42           N
ATOM    358  CA  ARG A 215     -42.464 -13.537   5.651  1.00 18.42           C
ATOM    359  C   ARG A 215     -41.666 -12.820   6.745  1.00 17.97           C
ATOM    360  O   ARG A 215     -42.240 -12.338   7.726  1.00 19.04           O
ATOM    361  CB  ARG A 215     -43.275 -12.525   4.835  1.00 18.99           C
ATOM    362  CG  ARG A 215     -42.421 -11.489   4.100  1.00 19.30           C
ATOM    363  CD  ARG A 215     -43.301 -10.359   3.594  1.00 20.84           C
ATOM    364  NE  ARG A 215     -43.854  -9.573   4.697  1.00 20.02           N
```

```
ATOM    365  CZ  ARG A 215     -44.864  -8.706   4.586  1.00 22.74           C
ATOM    366  NH1 ARG A 215     -45.467  -8.510   3.418  1.00 23.51           N
ATOM    367  NH2 ARG A 215     -45.282  -8.040   5.656  1.00 23.60           N
ATOM    529  N   ARG A 237     -36.427 -19.755  11.099  1.00 18.90           N
ATOM    530  CA  ARG A 237     -35.253 -20.079  10.303  1.00 20.07           C
ATOM    531  C   ARG A 237     -35.652 -20.086   8.836  1.00 20.62           C
ATOM    532  O   ARG A 237     -36.709 -20.607   8.487  1.00 20.15           O
ATOM    533  CB  ARG A 237     -34.658 -21.438  10.712  1.00 20.03           C
ATOM    534  CG  ARG A 237     -34.152 -21.504  12.151  1.00 19.85           C
ATOM    535  CD  ARG A 237     -33.395 -22.819  12.393  1.00 20.64           C
ATOM    536  CE  ARG A 237     -32.887 -22.927  13.828  1.00 20.64           C
ATOM    537  NZ  ARG A 237     -32.303 -24.281  14.128  1.00 19.43           N
ATOM    538  N   ARG A 238     -34.811 -19.483   7.993  1.00 21.63           N
ATOM    539  CA  ARG A 238     -35.054 -19.421   6.556  1.00 22.74           C
ATOM    540  C   ARG A 238     -35.290 -20.815   5.978  1.00 23.18           C
ATOM    541  O   ARG A 238     -34.580 -21.765   6.321  1.00 23.36           O
ATOM    542  CB  ARG A 238     -33.882 -18.738   5.842  1.00 23.15           C
ATOM    543  CG  ARG A 238     -34.126 -18.455   4.367  1.00 24.78           C
ATOM    544  CD  ARG A 238     -32.909 -17.817   3.729  1.00 29.07           C
ATOM    545  NE  ARG A 238     -33.127 -17.584   2.305  1.00 32.71           N
ATOM    546  CZ  ARG A 238     -32.328 -16.860   1.525  1.00 33.80           C
ATOM    547  NH1 ARG A 238     -31.254 -16.265   2.028  1.00 35.30           N
ATOM    548  NH2 ARG A 238     -32.617 -16.721   0.240  1.00 34.89           N
ATOM    598  N   ARG A 246     -36.004  -7.648  -2.381  1.00 24.57           N
ATOM    599  CA  ARG A 246     -36.526  -6.407  -1.793  1.00 24.09           C
ATOM    600  C   ARG A 246     -37.988  -6.209  -2.186  1.00 23.73           C
ATOM    601  O   ARG A 246     -38.334  -5.370  -3.019  1.00 22.92           O
ATOM    602  CB  ARG A 246     -35.657  -5.200  -2.156  1.00 24.34           C
ATOM    603  CG  ARG A 246     -34.232  -5.365  -1.662  1.00 25.49           C
ATOM    604  CD  ARG A 246     -33.359  -4.136  -1.804  1.00 25.90           C
ATOM    605  NE  ARG A 246     -32.020  -4.466  -1.317  1.00 27.00           N
ATOM    606  CZ  ARG A 246     -31.617  -4.321  -0.057  1.00 28.42           C
ATOM    607  NH1 ARG A 246     -32.447  -3.835   0.870  1.00 27.71           N
ATOM    608  NH2 ARG A 246     -30.378  -4.676   0.281  1.00 29.27           N
```

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-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
```

This will print all z coordinates for each atom.

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ awk '/GLY/ {count++} END {print count}' protein.pdb
33
```

Or,

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ grep -c "GLY" protein.pdb
33
```

I know how to solve this question using grep, but solved it by using awk, I use ChatGPT.

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ awk '/^ATOM/ && $3=="CA" {print $0}' protein.pdb | awk '$4=="ALA" || $4=="GLY"
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
```

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ awk '/^ATOM/ && $NF=="C" {count++} END{print coun
t}' protein.pdb
401
```

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-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
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
```

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ awk '($1=="ATOM" || $1=="HETATM") && $4 ~ /E$/ {print $4}' protein.pdb
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
PHE
PHE
PHE
PHE
PHE
PHE
PHE
PHE
```

I use ChatGPT for this question to know how we find E at the end of residue names.

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ sed '/^TER/d; /^END/d' protein.pdb
```

This code will be deleted the line containing TER or END.

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ awk '/^ATOM/ && $4!="ARG"' protein.pdb
```

This code prints the ATOM lines that do not belong to residue ARG.

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ awk '/^ATOM/ && $5=="A" {res[$4]++} END
{for(r in res) print r, res[r]}' protein.pdb
GLN 18
SER 36
HIS 10
ILE 32
TYR 48
VAL 21
GLU 81
ALA 15
GLY 28
LYS 45
CYS 37
PRO 42
THR 14
LEU 32
MET 8
ARG 55
PHE 22
ASN 40
ASP 16
TRP 42
```

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ 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
```

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ sed '/^>/! s/[a-z]/\U&/g' protein.fasta > protein_u.fasta
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ vi protein_u.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
```

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ awk '/^>/ {if (len > max) {max=len; id=
hdr} hdr=$0; len=0} /^[^>]/ {len+=length($0)} END {if (len > max) {max=len; id=hdr} print id, max}' protein.fa
sta
>seq3|Drosophila_melanogaster|TIM_protein 63
```

I use ChatGPT here to know how to define the max length and then write the code accordingly.

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ awk '/^ATOM/ || /^HETATM/ {print $4}' p
rotein.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
```

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ awk '/^ATOM/ || /^HETATM/ {print $5}' p
rotein.pdb | sort -u
A
```

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

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ awk '!/^>/ {count+=gsub(/A/,"")} END{pr
int "A:",count}' clock_gene.fasta
A: 114
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ awk '!/^>/ {count+=gsub(/T/,"")} END{pr
int "T:",count}' clock_gene.fasta
T: 100
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ awk '!/^>/ {count+=gsub(/G/,"")} END{pr
int "G:",count}' clock_gene.fasta
G: 355
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Assignment-3$ awk '!/^>/ {count+=gsub(/C/,"")} END{pr
int "C:",count}' clock_gene.fasta
C: 201
```