Lab_Assignment_3

1.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ sed '/^$/d' trail.txt > trail_2.txt
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ ls
_trial_temp  clock_gene.fasta  protein.fasta  protein.pdb  trail.txt  trail_2.txt  trial3.txt
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ cat tr
trail.txt    trail_2.txt  trial3.txt
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ cat trail_2.txt
Hello
welcome
see you tomorrow
nice to meet you
see you again
test
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ cat trail.txt
Hello

welcome

see you tomorrow

nice to meet you

see you again

test
```

2.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ awk '{print NR, $0}' trail.txt > numbered_triail
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ cat numbered_triail
1 Hello
2
3 welcome
4
5 see you tomorrow
6
7 nice to meet you
8
9 see you again
10
11 test
12
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ awk '{print NR, $0}' trail_2.txt > numbered_triail
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ ls
_trial_temp  clock_gene.fasta  numbered_triail  protein.fasta  protein.pdb  trail.txt  trail_2.txt  trial3.txt
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ cat numbered_triail
1 Hello
2 welcome
3 see you tomorrow
4 nice to meet you
5 see you again
6 test
```

3.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ sed -n '/^>/p' clock_gene.fasta
>NC_000004.12:c55546909-55427903 Homo sapiens chromosome 4, GRCh38.p14 Primary Assembly
```

4.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ awk '/^>.*CLOCK/ {print}' protein.fasta
>seq1|Homo_sapiens|CLOCK_protein
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$
```

5.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ awk '/CC/ {print}' protein.fasta
MTEYKLVVVGAGCCGKSALTIQLInhfgFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG
MADQLTEEQIAEFKEAFSLFDKDGDGTCCTKELGTVMRSCCQNPTEAELQDMINEVDADGNGQ
```

6.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ awk '{count += gsub(/G/, "G")} END {print count}' clock_gene.fasta
356
```

Note: For this I ask help from chat gpt for syntax writing

Given prompt-how to write syntax in awk for finding number of single character in whole file

Ans- awk '{count += gsub(/A/, "A")} END {print count}' filename

7.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ sed -n '5,28 p' clock_gene.fasta
GTGGAGGAGGGGAAGGGAAGGGAGGGGGAGGAGGAGCTGGCCACAGGAGCGGCGAATTTTTGGGGGGGTG
GGTGGGGGGCGCCACTCACAGCCCCAGGTGCTGCTGGAGGTGGGAGCCGCGGCGCCTCCTGGACACAGGC
GGGGTAGTGGTTCCGAGTCACCGCAGCGGGAGACCTGGGTGGGGGAGGGAAGAAGCCGGAGCCGCCGCAA
GCCACACGGTGAGGGCGCGGGGAAGGGGAGGGAGCGGGGGGCGGCGTGTGTGGGGCCGGGGGGCGGCGGC
CAAGGGTGGGGAAGGCGGGAGCTGAAGCCCAAGTTTGGCGTGTCGTTCTAGTGTGTCTTTTCCCGGGACT
TCGGGCCGAGGCCCGCCCTGCCTGAGAGGCCCTCTGGGGCAGCTGGGGTTACCTGCGGGGCAGGGGCGGG
AGTGGGGTGCACGGCGGGGCCGGGCGGCTTGAGGGCGCCCGGAGCTGCGGCCGATTCCAGCAGCTGGGAG
GCGGGGAAAGACGGGGACCGGGTGCCGAGAGAGCTTTCGCTGGGGACCCGCTAGGCCTTGTGACCCACTT
```

8.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_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
```

Chat gpt promot- awk -F" " '/^>/ {print substr($1,2)}' file name

9.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ awk '!/^>/ && /^M/ && /Q$/ {print}' protein.fasta
MADQLTEEQIAEFKEAFSLFDKDGDGTCCTKELGTVMRSCCQNPTEAELQDMINEVDADGNGQ
MADSQRRLLQNVINKAAGKSSTLLPVDGDKILVVTTGGQVVQSNVLEAMKELLQ
```

10.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ awk '/^>/ {if (seq!="") {print id, length(seq)}; id=substr($1,2); seq=""} !/^>/
{seq=seq $0} END {print id, length(seq)}' 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
```

11.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ awk '$1=="ATOM" && substr($0,22,1)=="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
```

12.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_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
```

I took help from ChatGPT to understand how to write code using substr and field $n in awk.

13.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ 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;
```

14.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ awk '$1=="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
```

15.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ awk '$1=="ATOM" && $4=="GLY"' protein.pdb |wc -l
28
```

16.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ awk '$1=="ATOM" && $3=="CA" && ($4=="ALA" || $4=="GLY") {print}' prote
.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.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ awk '$1=="ATOM" && $12=="C" ' protein.pdb| wc -l
401
```

18.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ awk '/^HETATM/' 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
```

19.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ awk '$1=="ATOM" && substr($4,3,1)=="E" {print $4}' protein.pdb
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
```

20.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ sed  '/TER|END/d' 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
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ sed -n '/TER|END/d' protein.pdb
```

21.

```
aval@DESKTOP-50ECNBT:~/Lab_session_3$ sed -n '/TER|END/d' protein.pdb
aval@DESKTOP-50ECNBT:~/Lab_session_3$ awk '/^ATOM/ && $4 != "ARG" {print}' protein.pdb
OM      1  N   TRP A 172     -39.136 -21.997  24.415  1.00 34.43           N
OM      2  CA  TRP A 172     -40.108 -20.907  24.729  1.00 34.28           C
OM      3  C   TRP A 172     -41.403 -21.065  23.944  1.00 33.46           C
OM      4  O   TRP A 172     -41.385 -21.496  22.789  1.00 33.48           O
OM      5  CB  TRP A 172     -39.506 -19.534  24.418  1.00 35.12           C
OM      6  CG  TRP A 172     -38.161 -19.292  25.025  1.00 36.34           C
OM      7  CD1 TRP A 172     -37.773 -19.568  26.306  1.00 37.69           C
OM      8  CD2 TRP A 172     -37.032 -18.693  24.384  1.00 37.47           C
OM      9  NE1 TRP A 172     -36.465 -19.190  26.497  1.00 37.97           N
OM     10  CE2 TRP A 172     -35.985 -18.650  25.334  1.00 37.83           C
```

22.

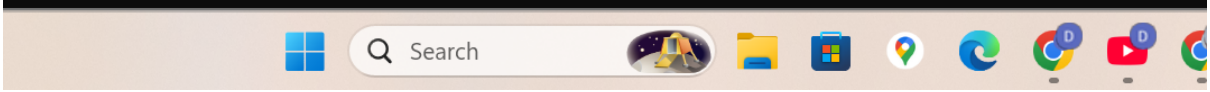
```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ awk '$1 == "ATOM" && $5 == "A" {residues[$4]++} END {for (res in residues) print res, residues[res]
' protein.pdb
GLY 28
CYS 37
LEU 32
THR 14
GLN 18
PRO 42
ILE 32
MET 8
ASN 40
TYR 48
LYS 45
ASP 16
SER 36
PHE 22
HIS 10
GLU 81
ARG 55
TRP 42
ALA 15
VAL 21
```

I took help from Perplexity AI to understand how to find the frequency of each residue (molecule) in a PDB file.

awk '$1 == "ATOM" && $5 == "A" {residues[$4]++} END {for (res in residues) print res, residues[res]}' filename

23.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_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
CE2,TRP,A
CE3,TRP,A
CZ2,TRP,A
CZ3,TRP,A
CH2,TRP,A
```



24.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_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

>seq10|Xenopus_laevis|REV-ERB_protein
MADEEKLPPGWEKRMSRSSGRVYYFNHITNASQWERPSGNSSSGSLS
```

25.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ awk '/^>/ {if(seqlen>maxlen){maxlen=seqlen; maxheader=header} header=$0; seqlen=0; next} {seqlen+=le
ngth($0)} END {print maxheader, maxlen}' protein.fasta
>seq3|Drosophila_melanogaster|TIM_protein 63
```

26.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ 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
```

27.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ awk '/^ATOM/ {print $5}' protein.pdb | sort | uniq
A
```

28.

```
dhaval@DESKTOP-50ECNBT:~/Lab_session_3$ awk '/^>/ {next} {for(i=1;i<=length($0);i++){c=substr($0,i,1); count[c]++}} END{print "A:"count["A"]
+0, "T:"count["T"]+0, "C:"count["C"]+0, "G:"count["G"]+0}' clock_gene.fasta
A:114 T:100 C:201 G:355
```

Note: I used Perplexity AI to assist with understanding questions 26 and 28, and for guidance on interpreting PDB data. In question 28, I did not fully understand the solution.