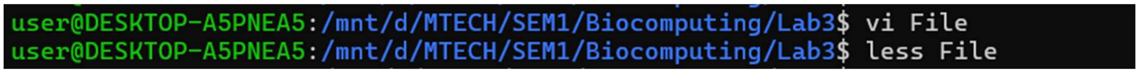
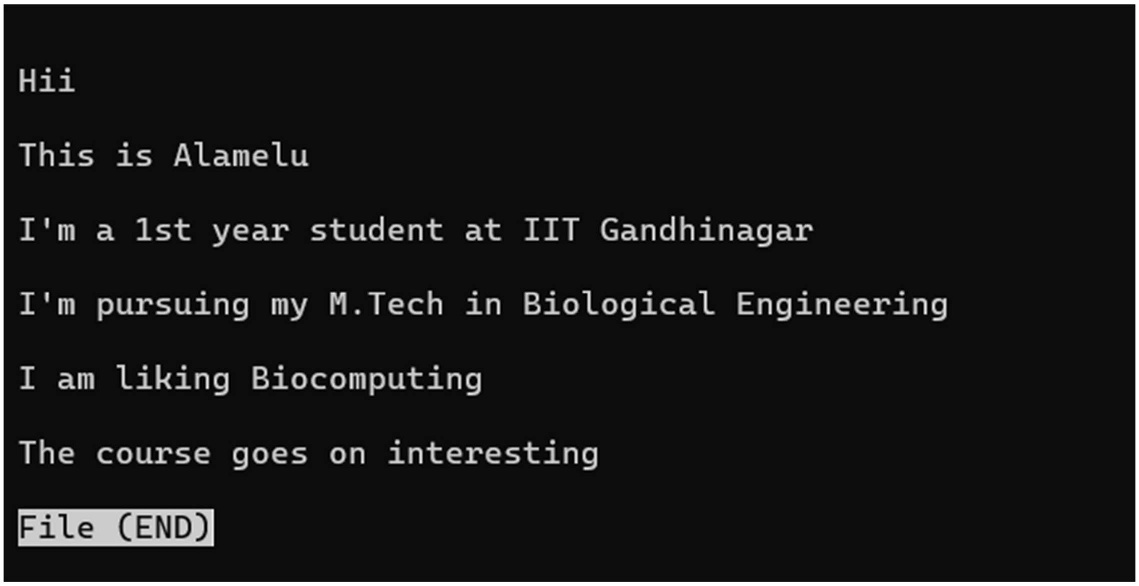
NAME: ALAMELU
ROLL NUMBER: 25210013
LAB-ASSIGNMENT-3
Text processing (sed and awk)

I have used CHATGPT for some difficult questions between 20 and 26 and I have also got hints for some questions from 1-20 from these AI tools only to understand logic which are different from the Lab session 3 codes about sed and awk.

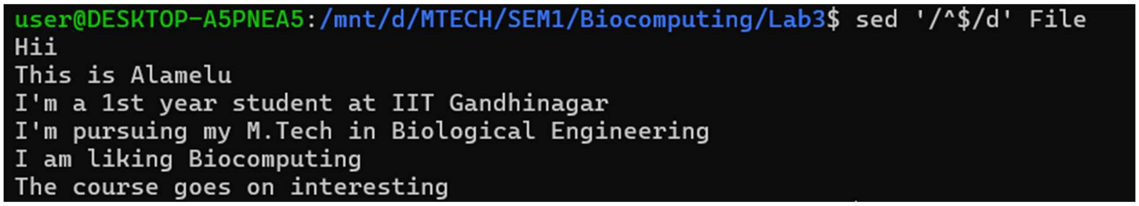
1)



```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ vi File
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ less File
```

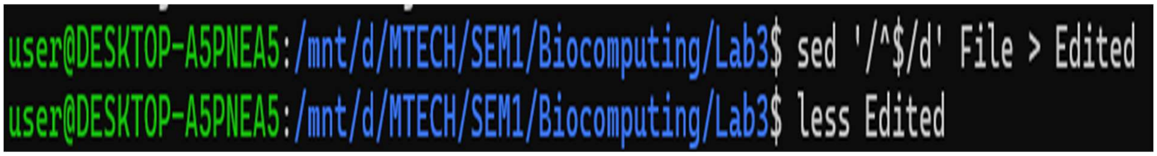


```

Hii

This is Alamelu

I'm a 1st year student at IIT Gandhinagar

I'm pursuing my M.Tech in Biological Engineering

I am liking Biocomputing

The course goes on interesting

File (END)
```



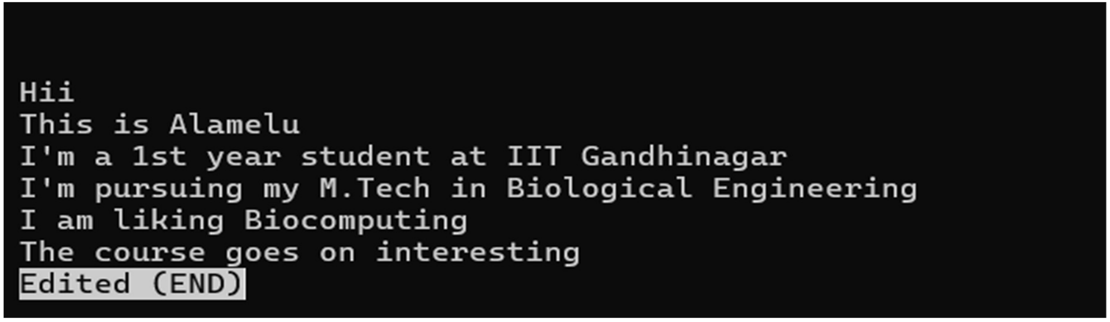
```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ sed '/^$/d' File
Hii
This is Alamelu
I'm a 1st year student at IIT Gandhinagar
I'm pursuing my M.Tech in Biological Engineering
I am liking Biocomputing
The course goes on interesting
```

2)

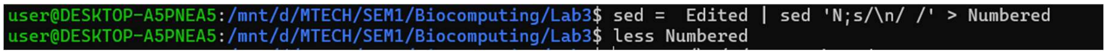


```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ sed '/^$/d' File > Edited
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ less Edited
```

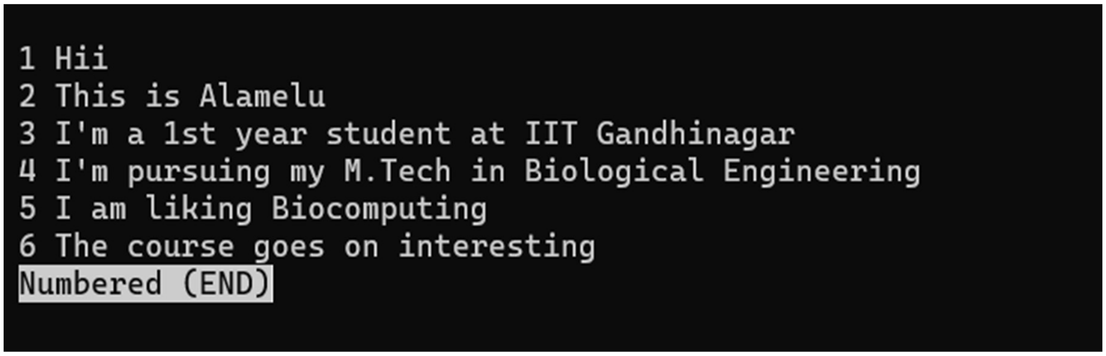
NAME: ALAMELU
ROLL NUMBER: 25210013
LAB-ASSIGNMENT-3
Text processing (sed and awk)



```
Hii
This is Alamelu
I'm a 1st year student at IIT Gandhinagar
I'm pursuing my M.Tech in Biological Engineering
I am liking Biocomputing
The course goes on interesting
Edited (END)
```

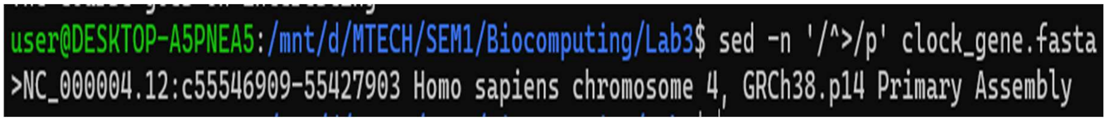


```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ sed =  Edited | sed 'N;s/\n/ /' > Numbered
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ less Numbered
```



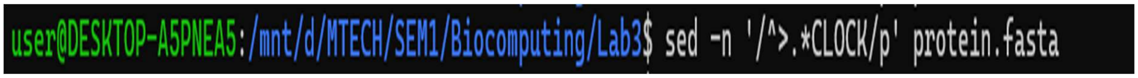
```
1 Hii
2 This is Alamelu
3 I'm a 1st year student at IIT Gandhinagar
4 I'm pursuing my M.Tech in Biological Engineering
5 I am liking Biocomputing
6 The course goes on interesting
Numbered (END)
```

3)



```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ sed -n '/^>/p' clock_gene.fasta
>NC_000004.12:c55546909-55427903 Homo sapiens chromosome 4, GRCh38.p14 Primary Assembly
```

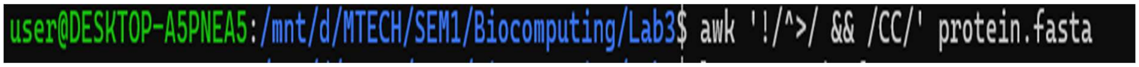
4)



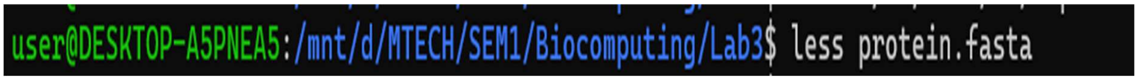
```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ sed -n '/^>.*CLOCK/p' protein.fasta
```

No headers found in protein.fasta that contains the word clock

5)



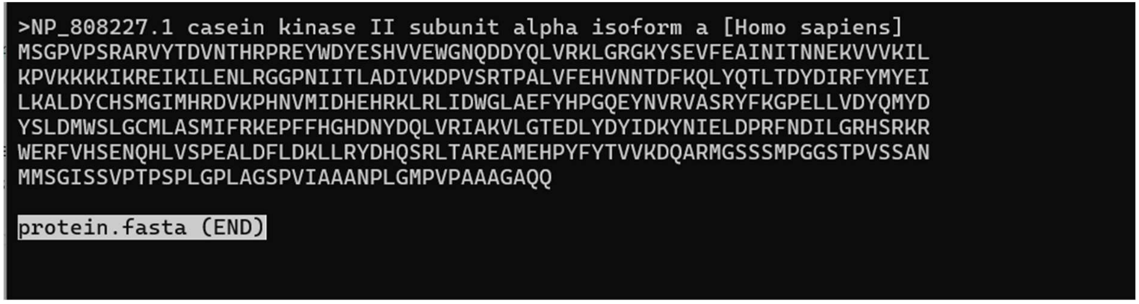
```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ awk '!/^>/ && /CC/' protein.fasta
```



```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ less protein.fasta
```

(less protein.fasta was executed to confirm if no lines with CC were present)

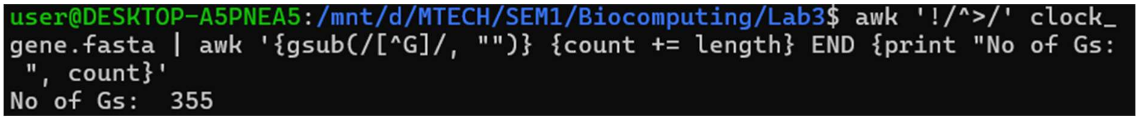
NAME: ALAMELU
ROLL NUMBER: 25210013
LAB-ASSIGNMENT-3
Text processing (sed and awk)



```
>NP_808227.1 casein kinase II subunit alpha isoform a [Homo sapiens]
MSGPVPSRARVYTDVNTHRPREYWDYESHVVEWGNQDDYQLVRKLGRGKYSEVFEAINITNNEKVVVKIL
KPVKKKKIKREIKILENLRGGPNIITLADIVKDPVSRTPALVFEHVNNTDFKQLYQTLTDYDIRFYMYEI
LKALDYCHSMGIMHRDVKPHNVMIDHEHRKLRLIDWGLAEFYHPGQEYNVRVASRYFKGPELLVDYQMYD
YSLDMWSLGCMLASMIFRKEPFFHGHDNYDQLVRIAKVLGTEDLYDYIDKYNIELDPRFNDILGRHSRKR
WERFVHSENQHLVSPEALDFLDKLLRYDHQSRLTAREAMEHPYFYTVVKDQARMGSSSMPGGSTPVSSAN
MMSGISSVPTPSPLGPLAGSPVIAAANPLGMPVPAAAGAQQ

protein.fasta (END)
```

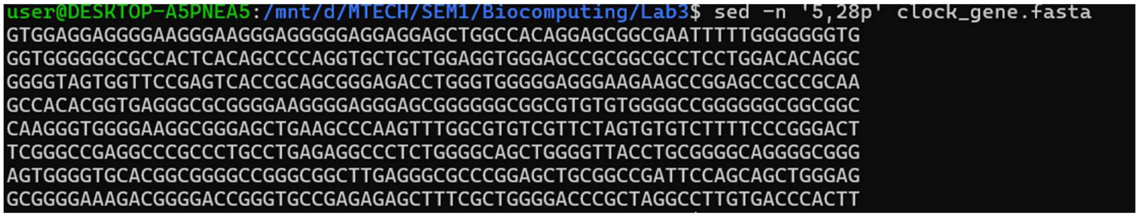
There are no such lines in protein.fasta with two consecutive CC s.

6)



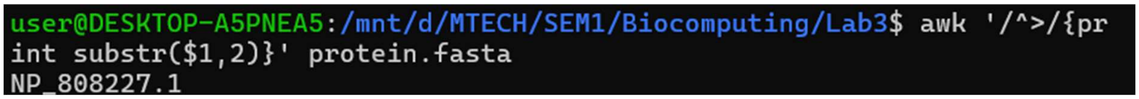
```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ awk '!/^>/' clock_
gene.fasta | awk '{gsub(/[^G]/, "")} {count += length} END {print "No of Gs:
 ", count}'
No of Gs:  355
```

7)



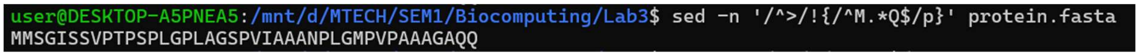
```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ sed -n '5,28p' clock_gene.fasta
GTGGAGGAGGGGAAGGGAAGGGAGGGGGAGGAGGAGCTGGCCACAGGAGCGGCGAATTTTTGGGGGGGTG
GGTGGGGGGCGCCACTCACAGCCCCAGGTGCTGCTGGAGGTGGGAGCCGCGGCGCCTCCTGGACACAGGC
GGGGTAGTGGTTCCGAGTCACCGCAGCGGGAGACCTGGGTGGGGGAGGGAAGAAGCCGGAGCCGCCGCAA
GCCACACGGTGAGGGCGCGGGGAAGGGGAGGGAGCGGGGGGCGGCGTGTGTGGGGCCGGGGGGCGGCGGC
CAAGGGTGGGGAAGGCGGGAGCTGAAGCCCAAGTTTGGCGTGTCGTTCTAGTGTGTCTTTTCCCGGGACT
TCGGGCCGAGGCCCGCCCTGCCTGAGAGGCCCTCTGGGGCAGCTGGGGTTACCTGCGGGGCAGGGGCGGG
AGTGGGGTGCACGGCGGGGCCGGGCGGCTTGAGGGCGCCCGGAGCTGCGGCCGATTCCAGCAGCTGGGAG
GCGGGGAAAGACGGGGACCGGGTGCCGAGAGAGCTTTCGCTGGGGACCCGCTAGGCCTTGTGACCCACTT
```

8)



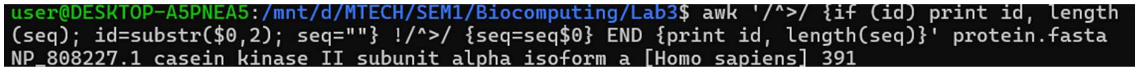
```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ awk '/^>/{pr
int substr($1,2)}' protein.fasta
NP_808227.1
```

9)1)



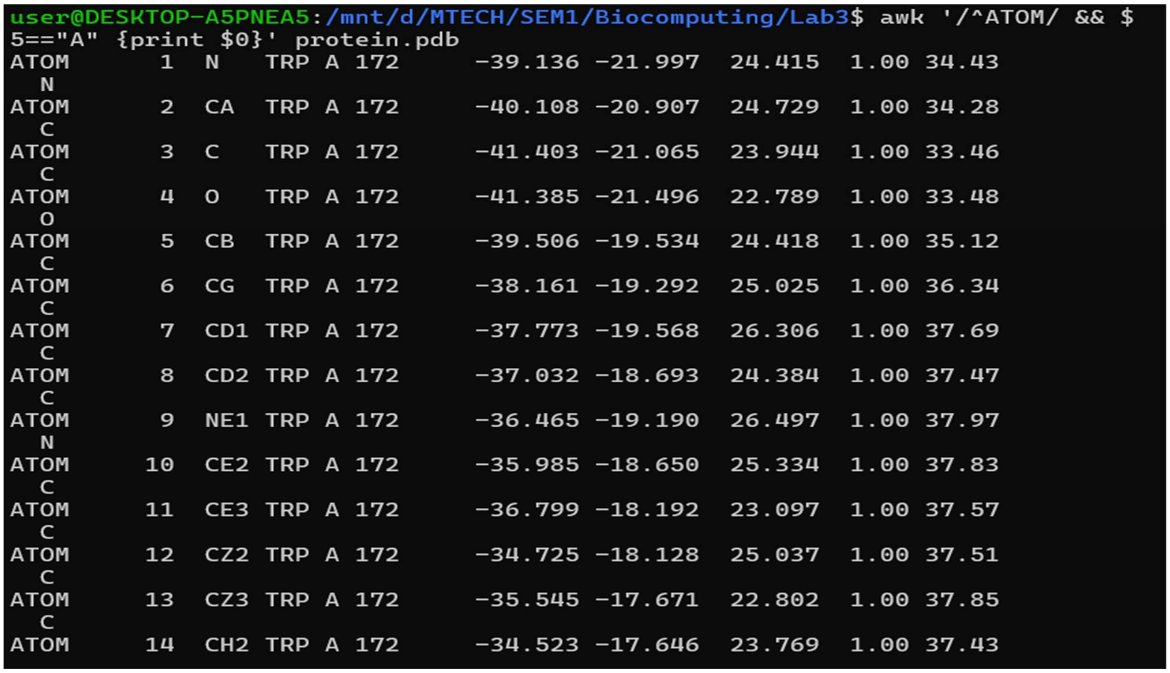
```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ sed -n '/^>/!{/^M.*Q$/p}' protein.fasta
MMSGISSVPTPSPLGPLAGSPVIAAANPLGMPVPAAAGAQQ
```

9)2)



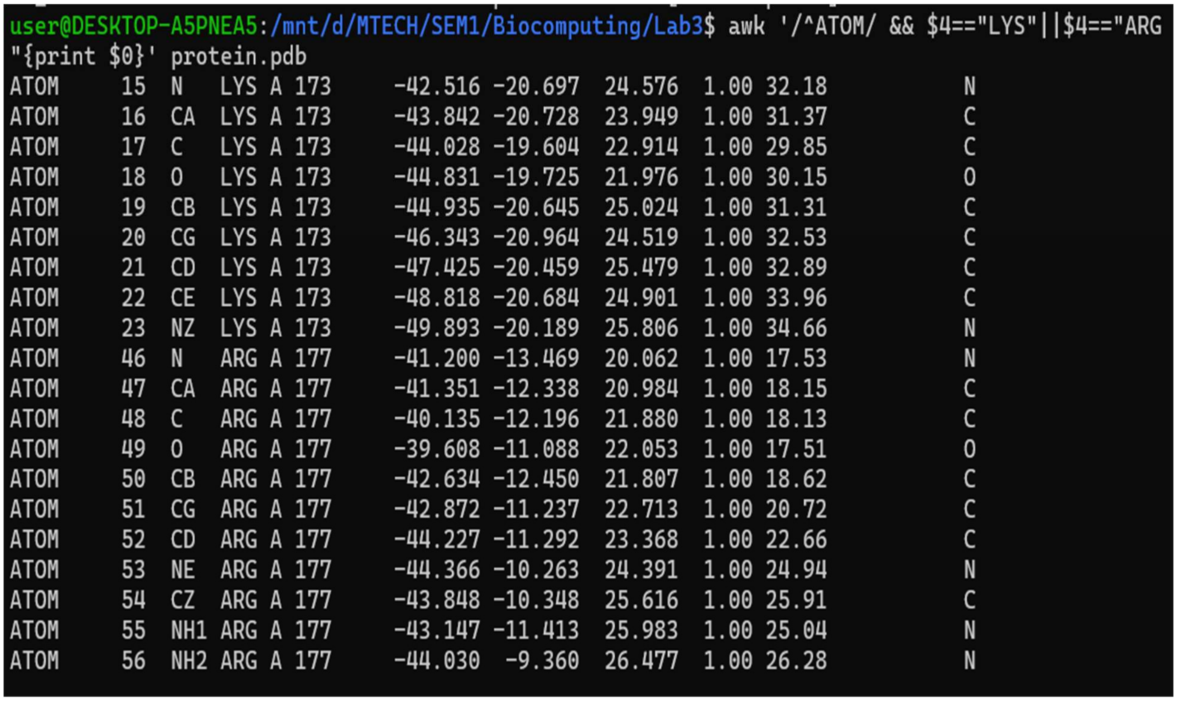
```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ awk '/^>/ {if (id) print id, length
(seq); id=substr($0,2); seq=""} !/^>/ {seq=seq$0} END {print id, length(seq)}' protein.fasta
NP_808227.1 casein kinase II subunit alpha isoform a [Homo sapiens] 391
```

10)



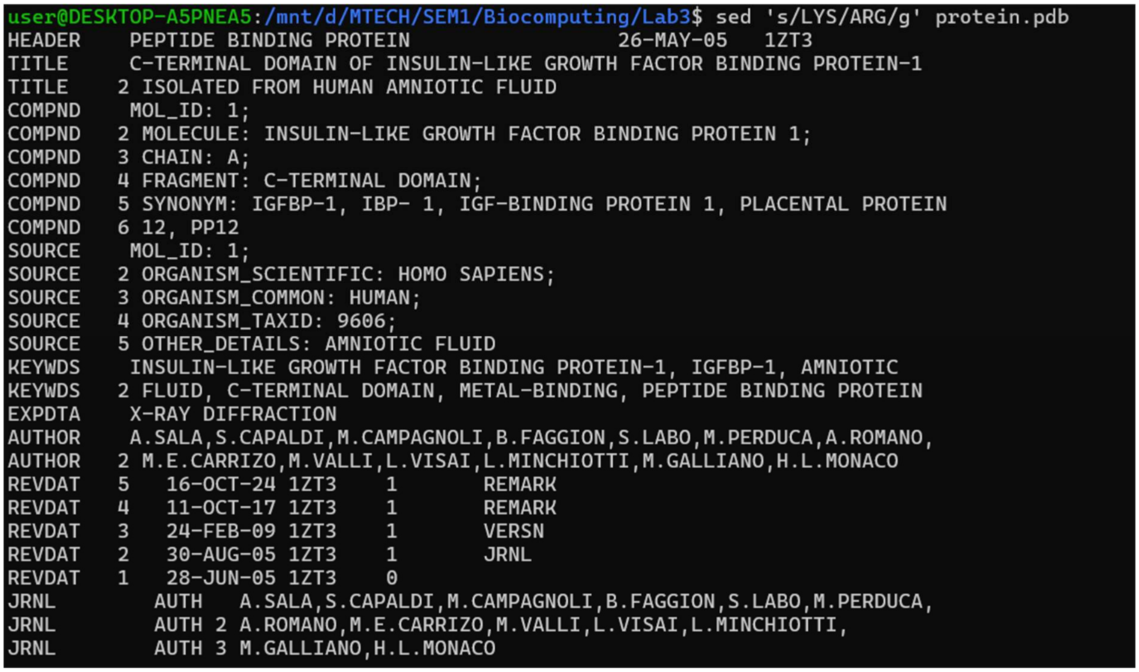
```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ awk '/^ATOM/ && $
5=="A" {print $0}' protein.pdb
ATOM      1  N   TRP A 172     -39.136 -21.997  24.415  1.00 34.43
  N
ATOM      2  CA  TRP A 172     -40.108 -20.907  24.729  1.00 34.28
  C
ATOM      3  C   TRP A 172     -41.403 -21.065  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
```

11)



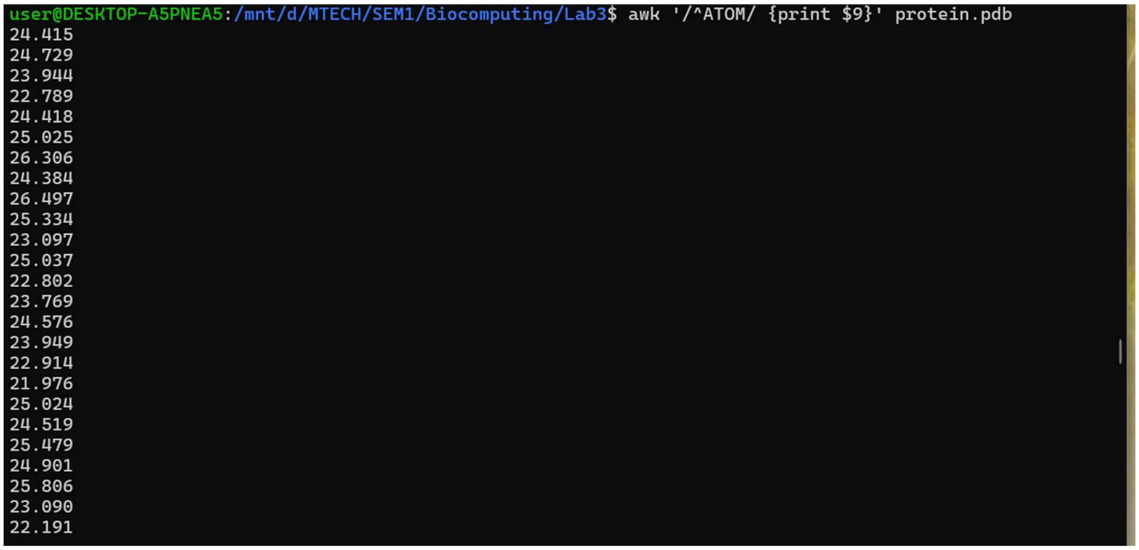
```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ awk '/^ATOM/ && $4=="LYS"||$4=="ARG
"{print $0}' protein.pdb
ATOM     15  N   LYS A 173     -42.516 -20.697  24.576  1.00 32.18           N
ATOM     16  CA  LYS A 173     -43.842 -20.728  23.949  1.00 31.37           C
ATOM     17  C   LYS 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
```

12)



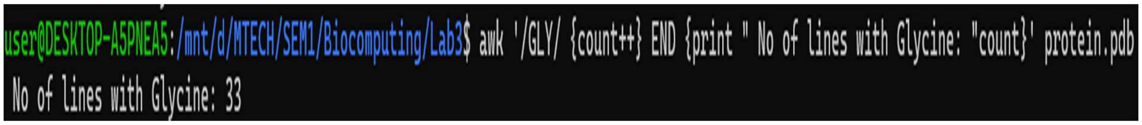
```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ sed 's/LYS/ARG/g' protein.pdb
HEADER    PEPTIDE BINDING PROTEIN                 26-MAY-05   1ZT3
TITLE     C-TERMINAL DOMAIN OF INSULIN-LIKE GROWTH FACTOR BINDING PROTEIN-1
TITLE    2 ISOLATED FROM HUMAN AMNIOTIC FLUID
COMPND    MOL_ID: 1;
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
```

13)



```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ 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
```

14)



```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ awk '/GLY/ {count++} END {print " No of lines with Glycine: "count}' protein.pdb
 No of lines with Glycine: 33
```

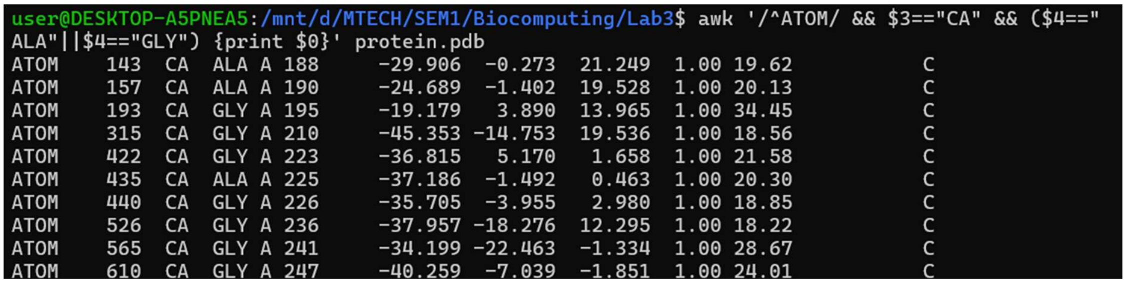
NAME: ALAMELU

ROLL NUMBER: 25210013

LAB-ASSIGNMENT-3

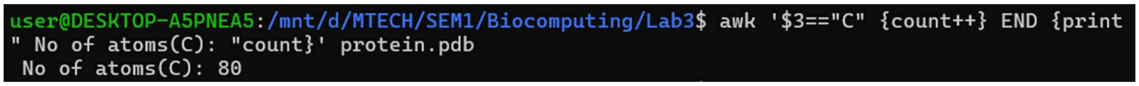
Text processing (sed and awk)

15)



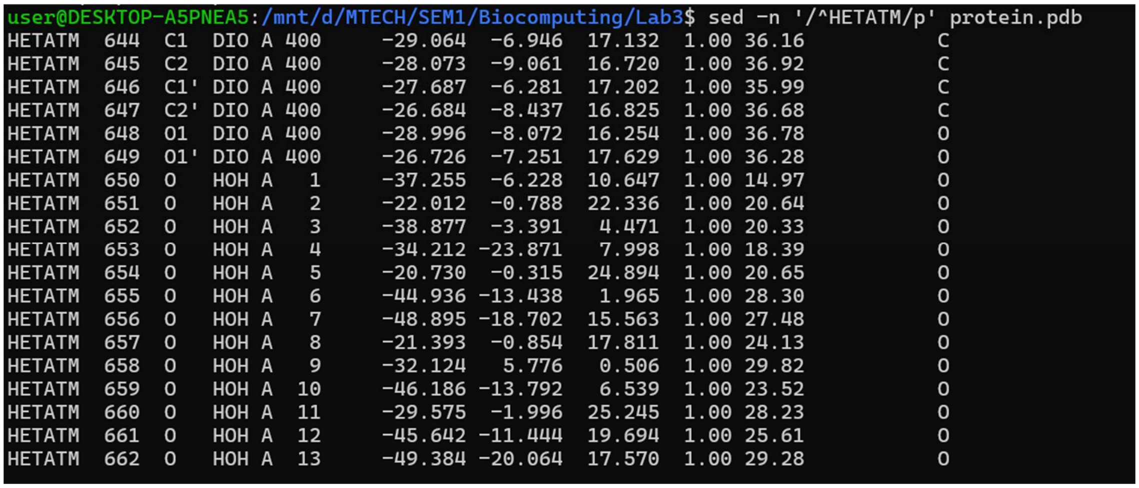
```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ awk '/^ATOM/ && $3=="CA" && ($4=="
ALA"||$4=="GLY") {print $0}' protein.pdb
ATOM    143  CA  ALA A 188     -29.906  -0.273  21.249  1.00 19.62           C
ATOM    157  CA  ALA A 190     -24.689  -1.402  19.528  1.00 20.13           C
ATOM    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
```

16)



```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ awk '$3=="C" {count++} END {print
" No of atoms(C): "count}' protein.pdb
 No of atoms(C): 80
```

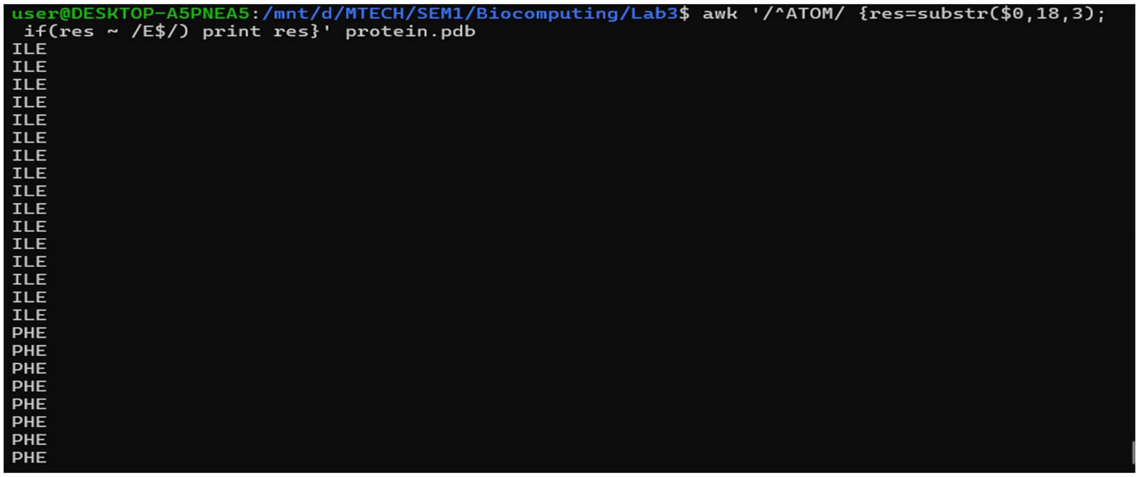
17)



```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ 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
```

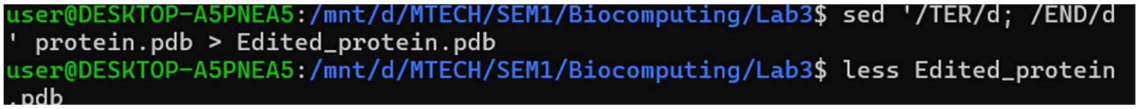
18)

awk '$1=="ATOM" {res=substr($0,18,3); if(res ~ /E$/) print res}' protein.pdb | sort | uniq (CODE from ChatGPT)



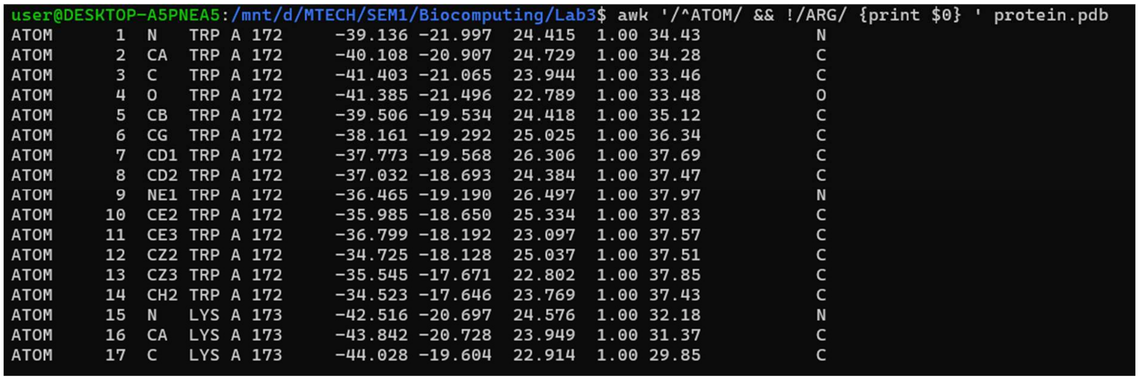
```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ awk '/^ATOM/ {res=substr($0,18,3);
 if(res ~ /E$/) print res}' protein.pdb
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
ILE
PHE
PHE
PHE
PHE
PHE
PHE
PHE
PHE
```

19)



```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ sed '/TER/d; /END/d
' protein.pdb > Edited_protein.pdb
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ less Edited_protein
.pdb
```

20)



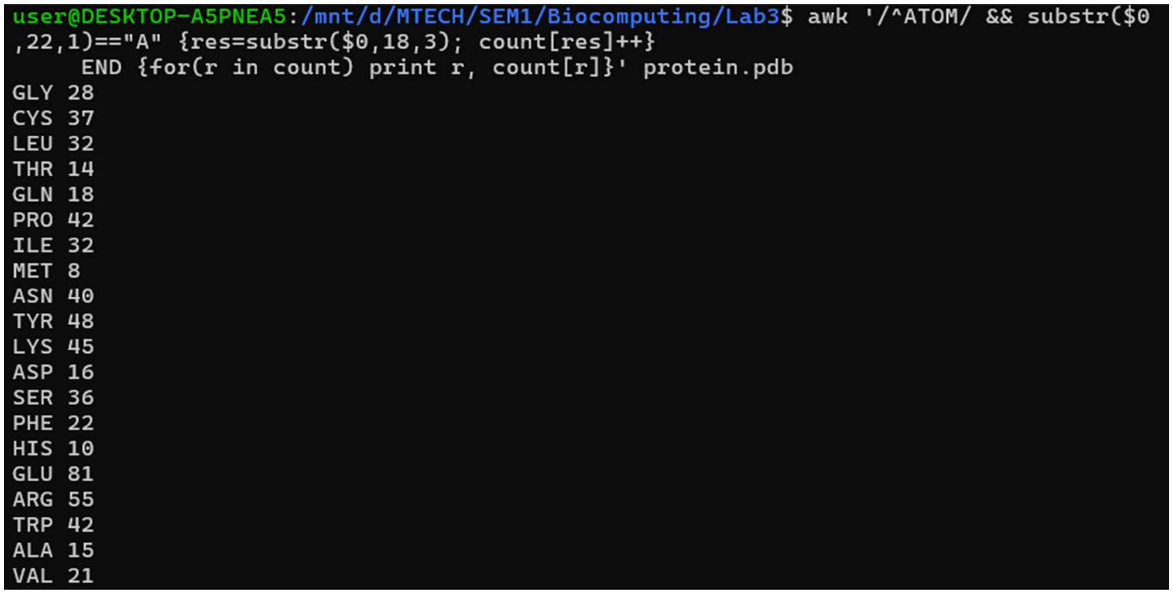
```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ awk '/^ATOM/ && !/ARG/ {print $0} ' protein.pdb
ATOM      1  N   TRP A 172     -39.136 -21.997  24.415  1.00 34.43           N
ATOM      2  CA  TRP A 172     -40.108 -20.907  24.729  1.00 34.28           C
ATOM      3  C   TRP A 172     -41.403 -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
```

21)

CODE FROM CHATGPT:

awk '$1=="ATOM" && substr($0,22,1)=="A" {res=substr($0,18,3); count[res]++}

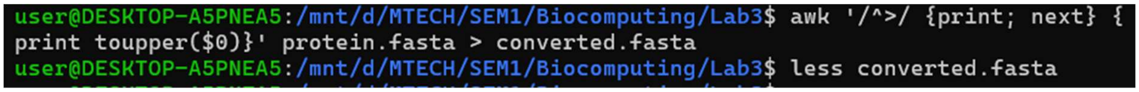
END {for(r in count) print r, count[r]}'



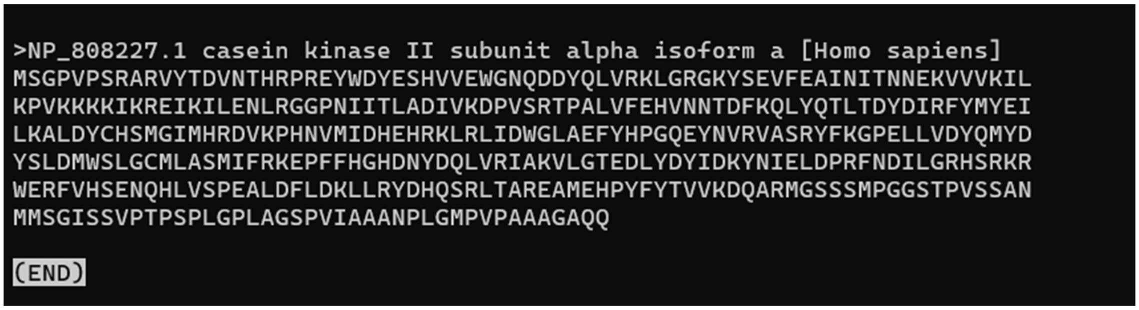
```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ awk '/^ATOM/ && substr($0
,22,1)=="A" {res=substr($0,18,3); count[res]++}
     END {for(r in count) print r, count[r]}' protein.pdb
GLY 28
CYS 37
LEU 32
THR 14
GLN 18
PRO 42
ILE 32
MET 8
ASN 40
TYR 48
LYS 45
ASP 16
SER 36
PHE 22
HIS 10
GLU 81
ARG 55
TRP 42
ALA 15
VAL 21
```

22)Toupper function to convert lower to uppercase was identified from ChatGPT

Usage of /^>/ {print; next} → if line is a header (starts with >), print as-is.(was got from ChatGPT)

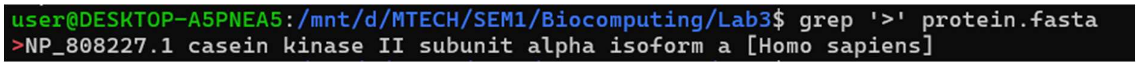


```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ awk '/^>/ {print; next} {
print toupper($0)}' protein.fasta > converted.fasta
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ less converted.fasta
```



```
>NP_808227.1 casein kinase II subunit alpha isoform a [Homo sapiens]
MSGPVPSRARVYTDVNTHRPREYWDYESHVVEWGNQDDYQLVRKLGRGKYSEVFEAINITNNEKVVVKIL
KPVKKKKIKREIKILENLRGGPNIITLADIVKDPVSRTPALVFEHVNNTDFKQLYQTLTDYDIRFYMYEI
LKALDYCHSMGIMHRDVKPHNVMIDHEHRKLRLIDWGLAEFYHPGQEYNVRVASRYFKGPELLVDYQMYD
YSLDMWSLGCMLASMIFRKEPFFHGHDNYDQLVRIAKVLGTEDLYDYIDKYNIELDPRFNDILGRHSRKR
WERFVHSENQHLVSPEALDFLDKLLRYDHQSRLTAREAMEHPYFYTVVKDQARMGSSSMPGGSTPVSSAN
MMSGISSVPTPSPLGPLAGSPVIAAANPLGMPVPAAAGAQQ

(END)
```

23)

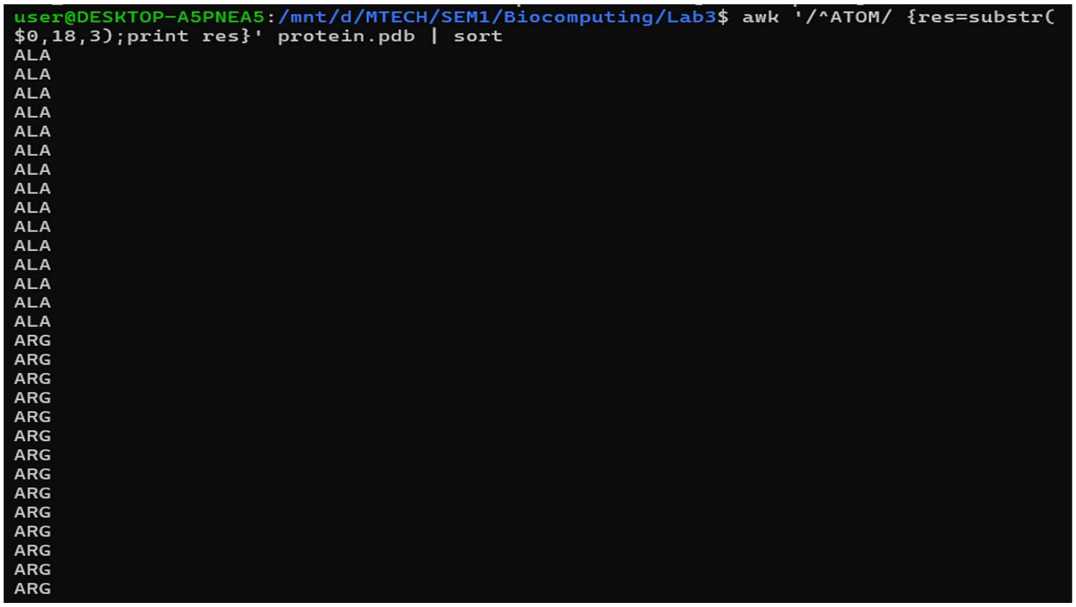


```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ grep '>' protein.fasta
>NP_808227.1 casein kinase II subunit alpha isoform a [Homo sapiens]
```

There is only one sequence in this file, hence it is the one with maximum length(I have solved this logically without code)

24)

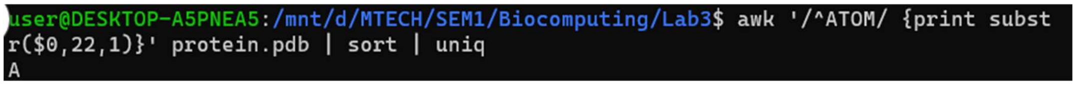
awk '$1=="ATOM" {res=substr($0,18,3); print res}' protein.pdb | sort | uniq (code from GPT and I have referred the sort usage from it)



```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ awk '/^ATOM/ {res=substr(
$0,18,3);print res}' protein.pdb | sort
ALA
ALA
ALA
ALA
ALA
ALA
ALA
ALA
ALA
ALA
ALA
ALA
ALA
ALA
ALA
ARG
ARG
ARG
ARG
ARG
ARG
ARG
ARG
ARG
ARG
ARG
ARG
ARG
ARG
```

NAME: ALAMELU
ROLL NUMBER: 25210013
LAB-ASSIGNMENT-3
Text processing (sed and awk)

25)

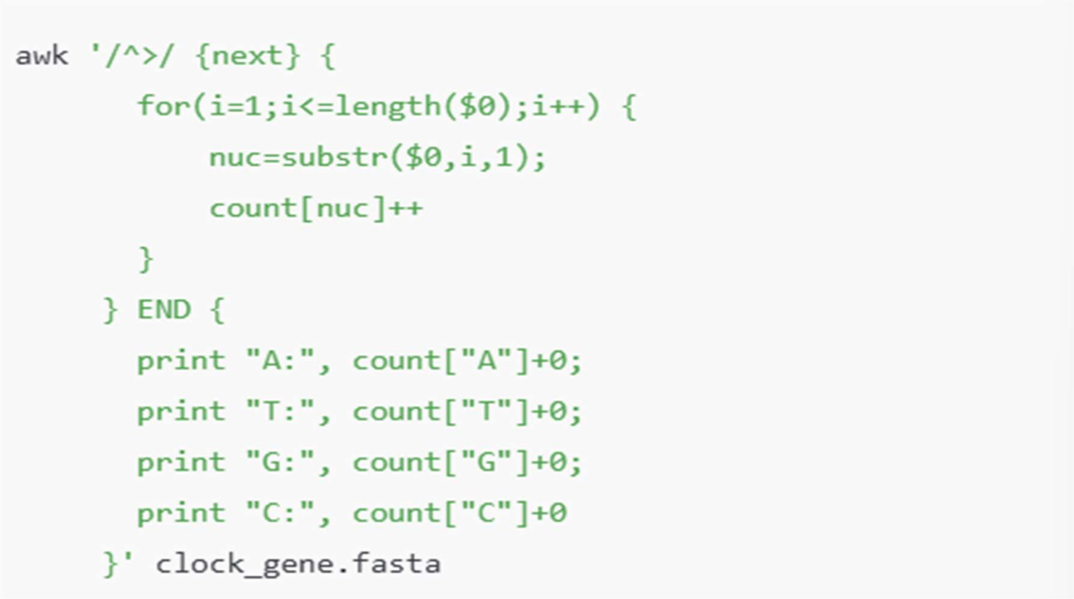


```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ awk '/^ATOM/ {print subst
r($0,22,1)}' protein.pdb | sort | uniq
A
```

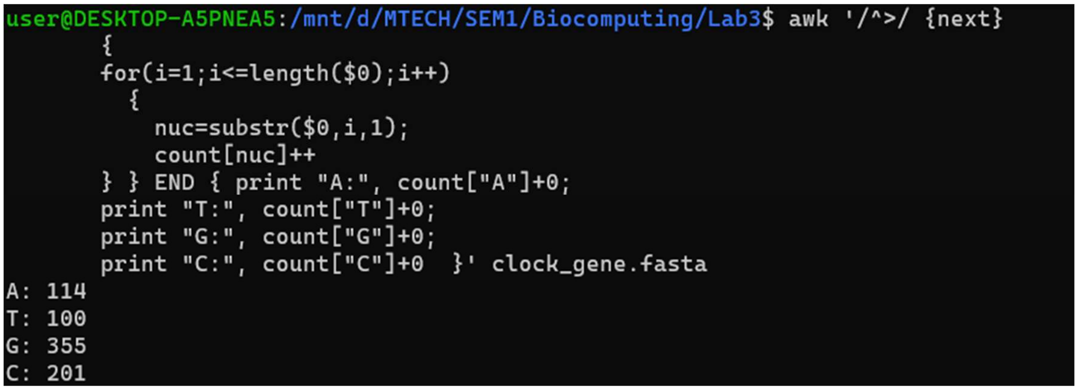
INFERENCE: Only A chain is present in the protein.pdb file

26)

The code was complicated, so I got the complete code from CHATGPT, but I’ve understood the logic of the code



```
awk '/^>/ {next} {
        for(i=1;i<=length($0);i++) {
            nuc=substr($0,i,1);
            count[nuc]++
        }
     } END {
        print "A:", count["A"]+0;
        print "T:", count["T"]+0;
        print "G:", count["G"]+0;
        print "C:", count["C"]+0
     }' clock_gene.fasta
```



```
user@DESKTOP-A5PNEA5:/mnt/d/MTECH/SEM1/Biocomputing/Lab3$ awk '/^>/ {next}
       {
       for(i=1;i<=length($0);i++)
         {
           nuc=substr($0,i,1);
           count[nuc]++
       } } END { print "A:", count["A"]+0;
       print "T:", count["T"]+0;
       print "G:", count["G"]+0;
       print "C:", count["C"]+0  }' clock_gene.fasta
A: 114
T: 100
G: 355
C: 201
```