

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)

I created the file then i did not understand the question , so i used :set number command in vi to just assign the numbers . i used cat and sed (search and replace as i understand)

```
usage: mv [-f | -i | -n] [-hv] source target      my times each residue appears (frequency of amino acids)
          mv [-f | -i | -n] [-v] source ... directory
[scifghost@sids-MacBook-Air lab_assignment_3 % mv ~/Downloads/biocomputing/BE623_labsession_3/Sed\ Cheatsheet.pdf .
[scifghost@sids-MacBook-Air lab_assignment_3 % ls
Awk-Cheatsheet.pdf      lab_assignment_3.pdf      protein.pdb          testline.txt
clock_gene.fasta         protein.fasta           Sed Cheatsheet.pdf
[scifghost@sids-MacBook-Air lab_assignment_3 % cat testline.txt | sed s/^$/g
hello

i am testing this      • sed '/^>/d' clock_gene.fasta | awk '{seqlen+=length($0)} END {print "Total length:", seqlen}'
for                   • seqlen'

the use of biocomputing lab
• sed 's/HIS/HSE/g' protein.pdb | awk '/^ATOM/ && $3=="CA"'
what is up ?           # In PDB, replace HIS with HSE (sed), then print only C-alpha lines (awk)

hack the planet         sed -n '/^ATOM/p' protein.pdb | awk '$3=="N" || $3=="CA" || $3=="C" || $3=="O"'
[scifghost@sids-MacBook-Air lab_assignment_3 % cat testline.txt | sed s/^$/d
sed: 1: "s/^$/d": bad flag in substitute command: 'd'
[scifghost@sids-MacBook-Air lab_assignment_3 % cat testline.txt | sed '/^$/d'; total+=length($0)} END {print hello
i am testing this     "GC% =", (gc/total)*100'
for                   # % OF GC content in sequence
the use of biocomputing lab
what is up ?
hack the planet
scifghost@sids-MacBook-Air lab_assignment_3 %
```

lab_assignment_3 — vim testline.txt — 157x47

```
hello
i am testing this
for
the use of biocomputing lab
what is up ?
hack the planet
~ "Every alternate line" means you write something on one line, leave the next line blank, then write again, leave the next blank, and so on.
~ So it looks like this:
~ scss
~ Line 1 text
~ <empty>
~ Line 3 text
~ <empty>
~ Line 5 text
~ <empty>
~ Line 7 text
```

what is every alternate line mean

"Every alternate line" means you write something on one line, leave the next line blank, then write again, leave the next blank, and so on.

So it looks like this:

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

I thought i can use vi to assign the numbers. Then i used google and found that i can use NR to assign the numbers to every line in vi :set number does that (i knew that one)

le add line number in front of every line using sed or awk

AI Mode All Images Videos Short videos Shopping Forums More options Tools

AI Overview

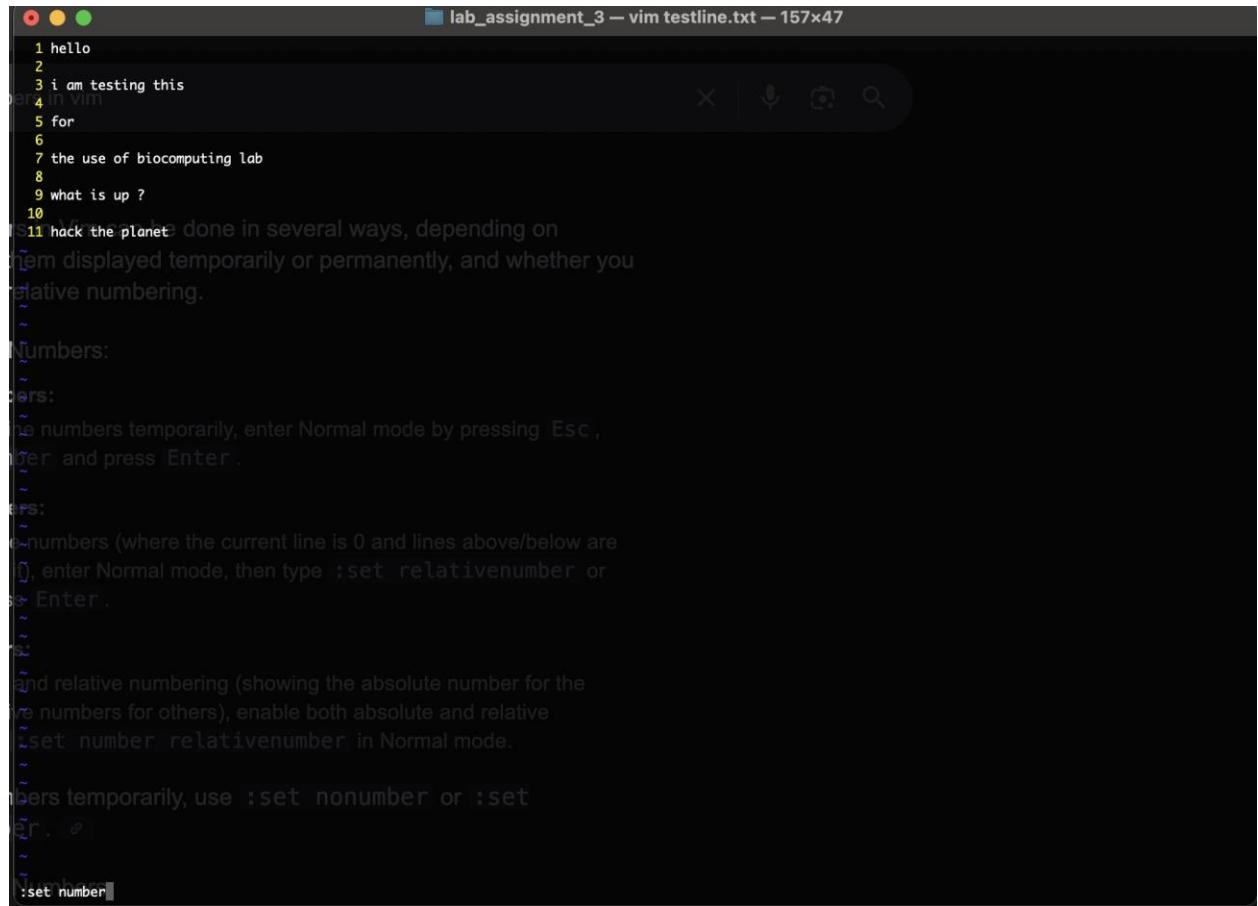
Listen

To add line numbers in front of every line of a file using `sed` or `awk`, the following commands can be used:

Using `awk`:

```
Code
awk '{print NR, $0}' filename.txt
```

```
[scifighost@sids-MacBook-Air lab_assignment_3 % cat newtestfile.txt | awk '{print NR, $0}'  
1 hello  
2 i am testing this  
3 for  
4 the use of biocomputing lab Al Mode  
5 what is up ?  
6 hack the planet  
scifighost@sids-MacBook-Air lab_assignment_3 %
```



```
1 hello  
2 i am testing this  
3 for  
4 the use of biocomputing lab  
5 what is up ?  
10  
11 hack the planet done in several ways, depending on  
them displayed temporarily or permanently, and whether you  
relative numbering.  
~  
Numbers:  
~  
bers:  
~  
he numbers temporarily, enter Normal mode by pressing Esc ,  
user and press Enter .  
~  
ers:  
~  
numbers (where the current line is 0 and lines above/below are  
0, enter Normal mode, then type :set relativenumber or  
Enter .  
~  
~  
and relative numbering (showing the absolute number for the  
ve numbers for others), enable both absolute and relative  
:set number relativenumber in Normal mode.  
~  
bers temporarily, use :set nonumber or :set  
~r. ~  
~  
:set number
```

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

I knew i can do it with sed but i forgot the 'p' . It was used in lab

```
6 hack the planet  
scifighost@sids-MacBook-Air lab_assignment_3 % sed -n '/^>/'  
sed: 1: "/^>/" command expected  
scifighost@sids-MacBook-Air lab_assignment_3 % sed -n '/^>/'p  
^C  
scifighost@sids-MacBook-Air lab_assignment_3 % sed -n '/^>/'p  
scifighost@sids-MacBook-Air lab_assignment_3 % cat clock_gene.fasta | sed -n '/^>/'  
sed: 1: "/^>/" command expected  
scifighost@sids-MacBook-Air lab_assignment_3 % cat clock_gene.fasta | sed -n '/^>/'p  
>NC_00004.12:c55546909-55427903 Homo sapiens chromosome 4, GRCh38.p14 Primary Assembly
```

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

```
scifghost@sids-MacBook-Air lab_assignment_3 % cat protein.fasta | grep 'CLOCK'  
>sp|015516|CLOCK_HUMAN Circadian locomoter output cycles protein kaput OS=Homo sapiens OX=9606 GN=CLOCK PE=1 SV=1
```

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

```
scifghost@sids-MacBook-Air lab_assignment_3 % cat protein.fasta | grep 'CC'  
TPEYLKSKNQLEFCCHMLRGTIIDPKEPSTYEYVKFIGNFKSLNSVSSSAHNGFEGTIQRT  
scifghost@sids-MacBook-Air lab_assignment_3 %
```

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

I do not know how to do it using sed or awk or i thought about using word count (wc)
but then i googled and found a good command called tr (i have also used it to
replace lower case letters to upper case in follow up questions .)

Then i thought there must be a way using grep , we can see man page of grep using
man grep command in terminal .

```
clock_gene.fasta      newtestfile.txt      protein.pab      testline.txt  
[scifghost@sids-MacBook-Air lab_assignment_3 % cat clock_gene.fasta | grep -o "G" | wc -l  
23472  
scifghost@sids-MacBook-Air lab_assignment_3 %
```

count character occurrences using bash from stdin

AI Overview

Listen

To count the occurrences of a specific character from standard input using Bash, one effective method involves using `tr` and `wc`.

- **Read from Standard Input and Filter Characters:**

The `tr -dc 'CHAR'` command reads from standard input and deletes all characters except the specified `CHAR`. This leaves only the occurrences of the desired character. Count Remaining Characters.

The output of `tr` is then piped to `wc -c`, which counts the number of bytes (which in this case, corresponds to the number of characters, as only the target character remains).

Example:

To count the occurrences of the character 'a' from standard input:

```
Code tr -dc 'a' | wc -c
```

7. Print only lines 5 to 28 from clock.gene.fasta.

I used google to search for the answer.

```

scifighost@sids-MacBook-Air lab_assignment_3 % cat clock_gene.fasta | sed -n '5,28'p
GTGGAGGAGGGAAAGGAAGGGAGGGAGGGAGCTGGCACAGGAGCGCGAATTGGGGGGTG
GGTGGGGGGCGCCACTCACAGCCCAGGTGCTGCTGGAGGTGGAGCCGCGCCTCTGGACACAGGC
GGGTAGTGGTCCAGTCACCGCAGCAGGGAGACCTGGTGGGGAGGGAGAAGCCGGAGCCGCCAA
GCCACACGGTAGGGCGCGGGAAAGGGAGGGAGCAGGGGGCGCGTGTGTGGGGCGGGGGCGCGC
CAAGGGTGGGAAGGGCGGAGCTGAAGCCCAAGTTGGCGTGTCTAGTGTCTTCCCGGACT
TCGGGCGAGGCCCGCCCTGCGTGAAGGCCCTCTGGGCGAGCTGGGTTACCTGCGGGGCAGGGCGG
AGTGGGTGACGGCGGGCGGGCGTGAAGGGCGCCGGAGCTGCGCGGATTCAGCAGCTGGAG
GCGGGAAAGACGGGACCGGGTGCAGAGAGAGCTTCGCTGGGGACCCGCTAGGCCGTGACCCACTT
TATTCTGTACCAACTCGGGCACTTGGAGCAGCGCCAAATGGGGCGCCGGGCCAGCTCCCGG
GGAACCCCGCCCTCCGGCGCCCGTGCAGTCGAGCTCGAACGCCGCGTTGCCGGC
CGCGGCTGGTCTGGTAGTGGTGGTCCGGGTTCCGGCTTCCAGCGCAGCGCCGGCTATTAGCGTC
TGACTCCAGCGACCGCGGGTGCAGAGAGAGCTTCGCTGGGGACCCGCTGGTTCTCCGTCCACC
CGCGCTTCCGCTTCCCGTACCGCGCTGGCGTTGTAGATTCTCCGAGCTGAAGCTGGGTTTC
TGGAGTTGGCTCTGGCGCTCTGGCCCTGGAGTGTAAATTCTACAGCAGGCCAGAGTTATTC
TTTGAAGTGTAGTTGTAGCTTGAGAGTCCTGTGATGGTAGGTGAGCCTAATTCTGAAGATAAA
AGCCTAGTCTGACCTGCAGATCAAATCAGATTGTGGTTCTGTCTATTAGATGCCGTG
TATTAGACTTTAAGGCTTTAGCCTTAAAAAAATTTACAGTGGAAAGAAAGCACAA
GAAGTAAACTTTACAGTCGTGATTGACTATAACGCTGATCCCCAAATCAAAGGTATTCACTT
GAAGATTGCGTCTGATTGTAGCTTAAGCGATTAGAGAAAATTGCAATTCCCTCTACCTGTTT
GAAAATAAACATTCTAAAGGATGTAATTAGATAATGAATTGCTTCTGAAACTTATCCCTGGGA
CACCTAAATCTGATGGTTCAAGCTGGGGAGGAAATTAACTCTGTGATAAGTGGTTGGCTGA
ACAGAATGCTGAAAGGTTAGCTGAGCTAGGAGAAATACCTAATGTAACGACGAGTTAATGGGT
GCAGCACACCGCTGGCACATTATACATATGTGACAAACCTGCAGTTGCACTGTACCCTAGAAC
TTAAAGTATAATAAAAAAAAGTAAAAAAAGTTAGCCTGAAGAAAGCAGACTGAAAATGTTCT
scifighost@sids-MacBook-Air lab_assignment_3 % cat clock_gene.fasta | sed -n '5,28'p | wc
    24      24     1704
scifighost@sids-MacBook-Air lab_assignment_3 %

```

how to print range of lines from stdin in linux

AI Mode All Videos Images Short videos Forums Shopping More options Tools

AI Overview

Listen

To print a range of lines from standard input (stdin) in Linux, you can use various command-line tools like `sed`, `awk`, or a combination of `head` and `tail`.

1. Using `sed`:

The `sed` command is powerful for text manipulation and can easily extract line ranges.

```
Code
cat your_input_file | sed -n 'START_LINE,END_LINEp'
```

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

```
[scifighost@sids-MacBook-Air lab_assignment_3 % cat protein.fasta
>sp|015516|CLOCK_HUMAN Circadian locomoter output cycles protein kaput OS=Homo sapiens OX=9606 GN=CLOCK PE=1 SV=1
MLFTVSCSKMSSIVDRDDSSIFDGLVEEDDKDKAKRVSRNKSEKKRRDQFNVLIKELGSM
LPGNARKMDKSTVLQKSIDFLRKHKITAQSDASEIRQDWKPTFLSNEEFTQLMLEALDG
FFLAIMTDGSIYYVSESVTSLEEHPLSDLVDQSIFNFIPEGEHSEVYKLSTHLLESDSL
TPEYLSKSNQLEFCCHMLRGTDIPKEPSTYEYVKFIGNFKSLNSVSSSAHNGFEGTIQRT
HRPSYEDRVCVFATVRLATPQFIKEMCTVEEPNEEFTSRHSLEWKFLFLDHRAPPIGYL
PEFVELGTSGYDYHHVDDLENLAKCHEHLMQYGGKSCYRFLTKQGQWILQTHYYITYH
QWNRSRPEFIVCTHTVVSAYEVRARELGLIEESLPTAAADKSQDSGSNDRINTVSLKEA
LERFDHSPTPSASSRRKSHTAVSDPSSPTKIKPTDTSTPPROHLPAHEKMVQRSSF
SSQSINSQSVGSSLTOPVMSQATNLPIPGMSQFQFSAQOLGAMQHLKDQLEQRTRMIEAN
IHRQQEELRKIQLQEQQLQMVHGQGLQMFLQQSNPGLNFGSVQLSSGNSSNIQQLAPINMQQQ
VVPNTQIQSGMTNGHIGTTQHMIQQQTLOSTSTQSQQNVLSGHSQTSPLSQQTQSTLTAP
LYNTMVISQSPAAGSMVQIPSSMPQNSTDQIRFSQGQQLVTKLVTPAVAC
GAVMVPSTMQLMGQVVTAYPEFTAQQQQTLTSVQQQQQQSEQQLTSVQOPSQALQTQ
PPQQFLQTSRLLHGNPSTQLILSAAPLQLQSTFPQSHHQHSQQQQQLSRLHRTDSLDPD
[SKVQPQ]

scifighost@sids-MacBook-Air lab_assignment_3 % cat protein.fasta | awk -F ' ' '{print $2}'
015516
```

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

ll ATOM lines from protein.pdb that belong to chain A only.
t all ATOM lines for residues LYS or ARG in protein.pdb.
e every occurrence of LYS with ARG in protein.pdb.
nly the z-coordinate (third number in coordinates) for each atom from

```
[scifighost@sids-MacBook-Air lab_assignment_3 % ]
```

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

```
scifighost@sids-MacBook-Air lab_assignment_3 % cat protein.fasta | grep '^M.*Q$'
scifighost@sids-MacBook-Air lab_assignment_3 % ]
```

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

It was provided in pdf i just copy pasted it .

```
scifighost@sids-MacBook-Air lab_assignment_3 % awk '/>/&&!{if (seqlen){print header, seqlen}; header=$0; seqlen=0;
[next]}{seqlen+=length($0)}END{print header, seqlen}' protein.fasta
>sp|015516|CLOCK_HUMAN Circadian locomoter output cycles protein kaput OS=Homo sapiens OX=9606 GN=CLOCK PE=1 SV=1 846
scifighost@sids-MacBook-Air lab_assignment_3 % ]
```

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

```
[scifighost@sids-MacBook-Air lab_assignment_3 % cat protein.pdb | grep '^ATOM' | grep 'A'
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          C
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 -16.205  19.866  1.00 19.59          C
```

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

```
scifighost@sids-MacBook-Air lab_assignment_3 % cat protein.pdb | grep '^ATOM' | grep -E 'LYS|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
```

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

```
scifighost@sids-MacBook-Air lab_assignment_3 % cat protein.pdb | grep '^ATOM' | grep -E 'LYS|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
```

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

```
[scifghost@sids-MacBook-Air lab_assignment_3 % cat protein.pdb | grep '^ATOM' | awk '{print $9}'  
24.415  
24.729  
23.944  
22.789  
24.418  
25.025  
26.306  
24.384  
26.497  
25.334  
23.097  
25.037  
22.802  
23.769  
24.576  
23.949  
22.914  
21.976  
25.024  
24.519  
25.479  
24.901  
25.806  
23.090  
22.191  
21.728  
22.138  
22.913  
23.359  
24.206  
23.639  
25.445  
20.872  
20.564  
19.866  
20.142  
19.632  
18.612
```

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

```
[scifghost@sids-MacBook-Air lab_assignment_3 % cat protein.pdb | grep 'GLY' | wc  
33 421 2673  
scifghost@sids-MacBook-Air lab_assignment_3 % ]
```

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

I do not know how to remove first 2 lines . I should have just grep for atom and then pipe stdout to other commands .

```
scifghost@sids-MacBook-Air lab_assignment_3 % cat protein.pdb | grep -E 'ALA|GLY' | grep 'CA'  
AUTHOR A.SALA,S.CAPALDI,M.CAMPAGNOLI,B.FAGGION,S.LABO,M.PERDUCA,A.ROMANO,  
JRNL AUTH A.SALA,S.CAPALDI,M.CAMPAGNOLI,B.FAGGION,S.LABO,M.PERDUCA,  
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  
scifghost@sids-MacBook-Air lab_assignment_3 %
```

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

```
scifghost@sids-MacBook-Air lab_assignment_3 % grep '^ATOM' protein.pdb | awk '{print $12}' | grep 'C' | wc  
401 401 802  
scifghost@sids-MacBook-Air lab_assignment_3 %
```

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

```
[scifghost@sids-MacBook-Air lab_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
```

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

```
[scifghost@sids-MacBook-Air lab_assignment_3 % grep '^ATOM' protein.pdb | awk '{print $4}' | grep 'E$' | sort -u
ILE
PHE
scifghost@sids-MacBook-Air lab_assignment_3 %
```

```
scifighost@sids-MacBook-Air lab_assignment_3 % grep '^ATOM' protein.pdb | awk '{print $4}' | grep 'E$'  
ILE Example output:  
ILE  
ILE nginx  
ILE  
ILE ILE  
ILE PHE  
ILE GLU  
ILE  
ILE  
ILE  
ILE  
PHE ➡ Do you want me to also extend it so it prints residue name + residue number (e  
PHE of just names?  
PHE  
PHE  
PHE  
PHE  
PHE  
PHE  
PHE
```

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

```
scifighost@sids-MacBook-Air lab_assignment_3 % cat protein.pdb | grep -vE 'TER|END' | head  
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; Plus gives you higher limits, smarter models, and Sora for video.  
SOURCE 2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;  
SOURCE 3 ORGANISM_COMMON: HUMAN;  
scifighost@sids-MacBook-Air lab_assignment_3 %
```

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

```

ATOM      612   CYS A 251    -13.811  -17.173   17.525  1.00 25.715
scifighost@sids-MacBook-Air lab_assignment_3 % grep '^ATOM' protein.pdb | grep -v 'ARG'
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

```

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

```

scifighost@sids-MacBook-Air lab_assignment_3 % awk '/^ATOM/ {res[$4]++} END {for (r in res) print r, res[r]}' protein.pdb | sort
ALA 15 # Print atom serial number and residue name
ARG 55
ASN 40 awk '/^ATOM/ {res[$4]++} END {for (r in res) print r, res[r]}' protein.pdb
ASP 16
CYS 37 # Count how many times each residue appears (frequency of amino acids)
GLN 18 awk '/^ATOM/ && $3=="CA" {print $0}' protein.pdb
GLU 81
GLY 28 # Extract only C-alpha atoms
HIS 10
ILE 32
LEU 32 lined sed + awk
LYS 45
MET 8
PHE 22>/d' clock_gene.fasta | awk '{seqlen+=length($0)} END {print "Total length:", seqlen}'
PRO 42
SER 36
THR 14 lve headers from FASTA (sed), then count sequence lengths with awk
TRP 42 HIS/HSE/g' protein.pdb | awk '/^ATOM/ && $3=="CA"'
TYR 48
VAL 21 DB, replace HIS with HSE (sed), then print only C-alpha lines (awk)
scifighost@sids-MacBook-Air lab_assignment_3 %

```

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

I had to google how to set delimiter using awk because i knew how to print name, residue name and chain ID.

```
scifighost@sids-MacBook-Air lab_assignment_3 % cat protein.pdb | grep '^ATOM' | awk '{OFS=","; print $3,$4,$5}'
```

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 MET,A
CZ2,TRP,A
CZ3,TRP,A MET,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

Unlock more with Plus

Do you want me to also refine it so it removes the extra spaces in atom names like " CA" → "C"?

Get Plus

ChatGPT Plus gives you higher limits, smarter models, and Sora for video.

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

```
scifighost@sids-MacBook-Air lab_assignment_3 % cat protein.fasta | tr 'a-z' 'A-Z'
```

>SP|0155161|CLOCK_HUMAN CIRCADIAN LOCOMOTER OUTPUT CYCLES PROTEIN KAPUT OS=HOMO SAPIENS OX=9606 GN=CLOCK PE=1 SV=1
MLFTVSCMSSIVDRDDSSIFDGLVEEDDKDKAKRVSRNKSEKKRRQFNVLIKELGSM
LPGNARKMDKSTVLQKSIDFLRKHKEITAQSDASEIRQDWKPTFLSNEEFTQLMLEALDG
FFLAJMTDGSIIYVSESVTSLLEHLPSDLVDQSIFNFPIPEGEHESEVYKILSTHLESDSL
TPEYLSKNQLEFCCHMLRGTDIPKEPSTYEVVKFIGFKSLNSVSSAHNGFEGTIQRT
HRPSYEDRVCFVATVRLATPOFIKEMCTVVEEPNEEFTSRHSLEWKFLFLDRAPPILGYL
PFEVLGTSGDYYYHVDDLENLAKCHEHLMQYGKGKSCYRFLTKGQQWIWLQTHYYITYH
QWNSRPEFIVCTHTVSYAEVRAERRRELGEIEESLPTAAKSDQSGSDNRINTVSKEA
LERFDHSPTPSASSRSRKSSHTAVSDPSSTPTKIPTDSTPPRQHLPAHEKMVQRSSF
SSQSINQSPVGSSLTQPVMSQATNLPIPQGMSQFQFSALQGAMQHLKDQLEQRTRMIEAN
IHRQQEELRKIQEQQLQMVHGOGLQMFLQQSNGPGLNFGSVQLSSGNSSNIQQLAPINMQGQ
VVPTNQIQLQGMNTGHIGTTQHMIQQQTQLSTSTQSQQNVLSGHSQQTSLPSQTQSTLTAP
LYNTMVISQPAAGSMVQIPSSMPQNQSTOSAAVTTFTQDRQIRFSQGQQLVTKLVTA
GAVMVPSTMLMGVVTAYPTFATQQQQSQTLSVTQQQQQSQSEQQQLTSVQQPSQAQLTQ
PPQQFLQTSRLLHGNPSTQLILSAAFPLQGSTFPQSHHQHQSQQQQQLSRHRTDSLPDP
SKVQPOcat protein.fasta | tr 'a-z' 'A-Z'

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

I have used this command and modified it a bit but do not know if the answer is right or wrong.

```
scifighost@sids-MacBook-Air lab_assignment_3 % awk '!/^>/ {if (seqlen){print seqlen}; seqlen=0; next}{seqlen+=length($0)}END{print seqlen}' protein.fasta
```

113 all residues and their frequencies from chain A.
0

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

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

```
scifighost@sids-MacBook-Air lab_assignment_3 % cat protein.pdb | grep '^ATOM' | awk '{print $4}' | sort -u
ALA ✓ awk '{print $4}' → extract residue name column
ARG
ASN ✓ sort -u → unique + alphabetical
ASP
CYS
GLN
GLU Do you want me to also include residue counts alongside names?
GLY
HIS
ILE ⌂ ⌃ ⌄ ⌅ ⌆ ⌇
LEU
LYS
MET
PHE
PRO
SER
THR
TRP
TYR
VAL      Unlock more with Plus
scifighost@sids-MacBook-Air lab_assignment_3 % smarter models, and Sora for video. Get Plus ×
```

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

```
scifighost@sids-MacBook-Air lab_assignment_3 % cat protein.pdb | grep '^ATOM' | awk '{print $5}' | sort -u
A      Unlock more with Plus
scifighost@sids-MacBook-Air lab_assignment_3 % smarter models, and Sora for video. Get Plus ×
```

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

Fold command breaks lines to parts like “SID” to “S\nI\nD\n” new line char.

```
scifighost@sids-MacBook-Air lab_assignment_3 % cat clock_gene.fasta | grep -v '^>' | fold -w1 | sort | grep -E '[ATGC]' | uniq -c
Would you like me to also give you a shorter awk one-liner for the same?
35332 A
21007 C
23471 G
39197 T      Unlock more with Plus
scifighost@sids-MacBook-Air lab_assignment_3 % smarter models, and Sora for video. Get Plus ×
```