

Indian Institute of Technology Gandhinagar

BE623 Biocomputing

Sem1 2025-2026

Lab Assignment –2

Linux & Shell Scripting with Biological Data Files

### Part 1 - vi Basics & File Editing

1. Open a new file called notes.txt in vi.

- Insert exactly one line of text:

Have a nice day

(Make sure there is no trailing space at the end.)

- Save and exit.

- Verify that the file contains exactly one line and 15 characters.

A terminal window with a black background and white text. The prompt is 'khushi@DESKTOP-DE71G03: ~/Lab\_session2'. The user enters 'pwd' and the output is '/home/khushi'. Then the user enters 'vi notes.txt'. After exiting vi, the user enters 'wc -l notes.txt' and the output is '1 notes.txt'. Then the user enters 'wc -m notes.txt' and the output is '16 notes.txt'. Finally, the user enters 'cd Lab\_session2/'.

```
khushi@DESKTOP-DE71G03: ~/Lab_session2
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

khushi@DESKTOP-DE71G03:~$ pwd
/home/khushi
khushi@DESKTOP-DE71G03:~$ vi notes.txt
khushi@DESKTOP-DE71G03:~$ wc -l notes.txt
1 notes.txt
khushi@DESKTOP-DE71G03:~$ wc -m notes.txt
16 notes.txt
khushi@DESKTOP-DE71G03:~$ cd Lab_session2/
```

I tried many times but got the same results yet again for the number of characters as 16 only.

But then with the help of [chatgpt](#) and [google AI Mode search](#) I got one code which I tried and got the number of characters as 15

I got the reason behind it was the trailing space or new line character which by default gets counted

So we need an extra command to not let it get counted as a character

I used the command `head -c -1 filename | wc -m file name`

In this command they only consider all the characters except of the last one which was supposedly the new line character and I got result as 15.

```
khushi@DESKTOP-DE71G03: ~/Lab_session2
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

khushi@DESKTOP-DE71G03:~$ vi notes.txt
khushi@DESKTOP-DE71G03:~$ head -c -1 notes.txt | wc -m
15
khushi@DESKTOP-DE71G03:~$ cd Lab_session2
khushi@DESKTOP-DE71G03:~/Lab_session2$ ls
```

## Part 2 - Pattern Matching in FASTA Files

2. Display the last four lines of sequence.fasta without opening the file in an editor.

```
1 notes.txt
khushi@DESKTOP-DE71G03:~$ wc -m notes.txt
16 notes.txt
khushi@DESKTOP-DE71G03:~$ cd Lab_session2/
khushi@DESKTOP-DE71G03:~/Lab_session2$ pwd
/home/khushi/Lab_session2
khushi@DESKTOP-DE71G03:~/Lab_session2$ tail -n 4 sequence.fasta
TAAGTACTGATAAGTTACAAACTGTTTTCTATCTAAAGGGCAATACAGCCCTAGACTCTCCAGGTAT
TTGACTCCTCGAGCAAAAAGGGAAATTGAGGAAATAGAGCAAGCTATTTCAGAGGCAACTATATCACA
TAGACACCCCG

khushi@DESKTOP-DE71G03:~/Lab_session2$ grep ">" sequence5.fasta
>ahr
>clock
```

3. In sequence5.fasta, print all header lines (lines starting with >).

```
16 notes.txt
khushi@DESKTOP-DE71G03:~$ cd Lab_session2/
khushi@DESKTOP-DE71G03:~/Lab_session2$ pwd
/home/khushi/Lab_session2
khushi@DESKTOP-DE71G03:~/Lab_session2$ tail -n 4 sequence.fasta
TAAGTACTGATAAGTTACAAACTGTTTTCTATCTAAAGGGCAATACAGCCCTAGACTCTCCAGGTAT
TTGACTCCTCGAGCAAAAAGGGAAATTGAGGAAATAGAGCAAGCTATTTCAGAGGCAACTATATCACA
TAGACACCCCG

khushi@DESKTOP-DE71G03:~/Lab_session2$ grep ">" sequence5.fasta
>ahr
>clock
>hif1a
>hif2a
>hif3a
>nps1
>nps2
>nps3
>nps4
>sim1
>sim2
>arnt1
>bmal1
khushi@DESKTOP-DE71G03:~/Lab_session2$ grep "A.G" sequence5.fasta
IFRTKHKLDFTPIGCDAKGRIVLGYTEAELCTRGSGYQFIHAADMLYCAESHIRMIKTGESGHIVFRLLT
```

4. Find all matches in sequence5.fasta where A is followed by any single character and then G.

```
>arnt1
>bmal1
khushi@DESKTOP-DE71G03:~/Lab_session2$ grep "A.G" sequence5.fasta
IFRTKHKLDFTPIGCDAKGRIVLGYTEAELCTRGSGYQFIHAADMLYCAESHIRMIKTGESGHIVFRLLT
DAARSRRSQETEVLYQLAHTLPFARGVSAHLDKASIMRLTISYLRHRLCAAGEHNVQVAGGGEPLDACYL
KALEGFVMVLTAEQDMAYLSENVSKHLGLSQLELIGHSIFDFIHPCDQEEQLDQALTPPTERCFSLRMKST
KEKSRNAARSRRGKENLEFFELAKLLPLPGATSSQDKASIVRLSVTYLRLRRFAALGAPPWGLRAAGPP
AGLAPGRGPAALVSEVFQHLGGHILQSLDGFVFALNQEGKFLYSETVSIYGLSQVEMTGSSVFDYI
HPGDHSEVLQGLGVQERSFVRMKSTLTKRGLHVKAAGYKVIHVTGRLRALGLVALGHTLPPAPLAELP
MLQRAGGFVNLQSVATVAGSGKSPGEHVLVSHVLSQAEGGQT
GASKARRDQINAEIRNLKELLPLAEADKVLVSVLHMSLACIYTRKGVFFAGGTPLAGPTGLLSAQELED
IVAALPGFLVFTAEGLLYLSESVSEHLGHSHVDLVAAQDSYDIIDPADHLTVRQQLTLTORLFCRFR
EKSKMAARTTRREKSEFYLAKLLPLPSATTSQLDKASTIRLTTSYLRHVRVFFPEGLGEINRHSRTSP
ETIERSFFLRMKCVLAKRINAGLTCSGVKVIHCSGYLKIRIVGLVAVQSLPPSAITETKLYSNMFMRASL
EKSNAARTTRREKSEFYLAKLLPLPSATTSQLDKASTIRLTTSYLRHVRVFFPEGLGDANRQPSRAGP
ETIERSFFLRMKCVLAKRINAGLTCSGVKVIHCSGYLKIRIVGLVAVQSLPPSAITETKLYSNMFMRASL
ELKHLILEAADGFLFIVSCETGRVVVSDSVTPVLNQPSQSEWFGSTLYDQVHPDDVDKLRQLSTSRMCM
GSRRSFICRMRCGSSEPHFVVHCTGYIKAKFLVATIGRLQVTSPPNCTDMSHVCPTFEISRHNIEGIF
DELKHLILRAADGFLFVVGCDRGKILFVSESVFKILNYSQNDLIGQSLFDYLPKDKAKVKEQLSSRLC
SGARRSFFCRMKNRPRKSFCTIHSTGYLKSNSCLVAIGRLHSHVVPQVNGEIRVKSMEVYSRHAIDG
```

5. Find all matches in sequence5.fasta where P is followed by any character except A, then L.

```
DEKMFCEKRRRPFYVSDRRIKELVSEVFKELTSGHCEKQGEPTDFIPRDKRYREQSRRKE
SGARRSFFCRMKNRPRKSFCTIHSTGYLKSNSCLVAIGRLHSHVVPQVNGEIRVKSMEVYSRHAIDG
khushi@DESKTOP-DE71G03:~/Lab_session2$ grep "P[" sequence5.fasta
grep: Invalid regular expression
khushi@DESKTOP-DE71G03:~/Lab_session2$ grep "P[^\A]L" sequence5.fasta
QLHWQIPPPSPMLMERCFCICRLCLDNSSGFLAMNFQGLKLYLPQALFAIATPLQPPSILEIRTKNF
MRMKCTVTNRGRTVNLKSATWKLHCTGQVKVPELLSCLIMCEPIQHPSHIDPLDSKTFLSRHSNDM
LTSRGTLLNLKAATWVLNCSGHMRAVEPPIQCLVLCEAIPHPGSLPEPLGRGAFLSRHSMDKMFYCYD
FTQLMLEALDGFIIAVTDGSIYVSDSITPLIGHLPDSDMDQNLNLFPEQHSVEYKILSSEYLSKDS
ELKHLILEAADGFLFIVSCETGRVVVSDSVTPVLNQPSQSEWFGSTLYDQVHPDDVDKLRQLSTSRMCM
khushi@DESKTOP-DE71G03:~/Lab_session2$ grep -E "V" sequence5.fasta
DAVFERELNLQEGEFLQALNGEVLPITDALVYASSTIDVLEQDSRTHORVYVLTHTERDEFOR
```

6. Print all lines in sequence5.fasta that have exactly 2 consecutive Vs anywhere in the line.

```
khushi@DESKTOP-DE71G03: ~/Lab_session2$ grep -E "VV" sequence5.fasta
AAAFREGNLQEGFLLQALNGFLVVTDAALVFYASSTIQDYLGFQSDVHQSVELIHTEDRAEFQR
IWLQTHYIITYHQNRPFEIVCTHTVSYAEVRAE
TVIYNTKNSQPQCIVCNVYVSGIIQHDL
QMDNLYLKALEGFIADVITQGDIMFLSENISKFMGLTQVELTGHISIFDTHPCDHEIRENLSSTERDFF
KFTYCDRITTELIGYHPEELLGRSAYEFYHALDSENMTKSHQNLCTKGQVSGQYRLAKHGYYVLELQ
DRAEAVAGYSPDOLIGCSAYEFYHALDSDAVSKSIHTLLSKGQAVTGQYRFLARSGGYLLTQTATVSG
QTHYIITYHQNRPFEIVCTHTVSYAEVRAE
VDYHPGDHVEAEQGLTLEFSEFIRMKSTLTKRGVHKSSGKYVHTIGRLRLRMGLVVAHALPPPTI
TSEVLTYLGEERSELLCKSWGLLHPEDLAHASAQHYRLAESGDIQAEHVIQLAKTGGNAHYCYLLY
EKSNAARTREKENSEFELAKLLPLPSAITSQDKASTIIRLTTSYLKMRVFPFEGLEAMGHSSRTSP
LDNVGRELQSHLLQTLQDGFIFVAPDGKIMYISETASVHLGLSQVELTGNISIEYIHPADHDEMTAVLTA
LDGVAKELGSHLLQTLQDGFIFVAPDGKIMYISETASVHLGLSQVELTGNISIEYIHPADHDEMTAVLTA
SYATVHNSRSRPHCIVSVNVLTEIEYKEL
ELKHLILEADGFLFVSCETGRVVVSDVSTPVLNQPSSEWFGSTLYDQVHPDDVKLREQLSTRMCH
GSRRSFICRMRCGSSEPHFVVHCTGYIKAKFLVAIGRLQVTSNPCTDMSHVCQTEFISRHNIEGIF
TFVDHRCVATGVGPQQLLGNKIVEFCHPEDQQLRDSFQQVYKLGQVLSVMFRFSKNQEWLWRTSS
DELKHLILRAADGFLFVGCGRGKILFVSESVFKILNYSQNDLIGQSLFDYLHPKDIKAVKEQLSSRLC
SGARRSFRCMKCNRPKSFCTIHTGYLKSNSCLVAIGRLHSHVYQPVNGEIRVKSMYVSRHAIDG
RWFSPMHPMTKEVYIVSTHTVVL
```

7. Print all lines in sequence5.fasta that contain either AA or DD.

```
khushi@DESKTOP-DE71G03: ~/Lab_session2$ grep -E "AA|DD" sequence5.fasta
LDGVAKELGSHLLQTLQDGFIFVAPDGKIMYISETASVHLGLSQVELTGNISIEYIHPADHDEMTAVLTA
SYATVHNSRSRPHCIVSVNVLTEIEYKEL
ELKHLILEADGFLFVSCETGRVVVSDVSTPVLNQPSSEWFGSTLYDQVHPDDVKLREQLSTRMCH
GSRRSFICRMRCGSSEPHFVVHCTGYIKAKFLVAIGRLQVTSNPCTDMSHVCQTEFISRHNIEGIF
TFVDHRCVATGVGPQQLLGNKIVEFCHPEDQQLRDSFQQVYKLGQVLSVMFRFSKNQEWLWRTSS
DELKHLILRAADGFLFVGCGRGKILFVSESVFKILNYSQNDLIGQSLFDYLHPKDIKAVKEQLSSRLC
SGARRSFRCMKCNRPKSFCTIHTGYLKSNSCLVAIGRLHSHVYQPVNGEIRVKSMYVSRHAIDG
RWFSPMHPMTKEVYIVSTHTVVL
khushi@DESKTOP-DE71G03: ~/Lab_session2$ grep -E "AA|DD" sequence5.fasta
AAAFREGNLQEGFLLQALNGFLVVTDAALVFYASSTIQDYLGFQSDVHQSVELIHTEDRAEFQR
IFRTKHKLDFTPIGCDAGRIVLGYTEAELCTRGSGYQFIHAAOMLYCAESHIRMIKTGESGMIVFRLLT
NCEYLKALDGFVHVLTDGDMYISDNVKNYMLTQFELTGHVSFDFTHPCDHEEMREMLTHNTQSFFL
KEKSRDAARCRRSKEFVYFELAHQLPLPHNVSHLDKASVHRLTISYLRVRKLLDAGOLDIEDDKAKQH
NCEYLKALDGFVHVLTDGDMYISDNVKNYMLTQFELTGHVSFDFTHPCDHEEMREMLTHNTQSFFL
KEKSRDAARCRRSKEFVYFELAHQLPLPHNVSHLDKASVHRLTISYLRVRKLLDAGOLDIEDDKAKQH
AGLAPGRGPAALVSEVFEQLGHILQSLDGFVVALNQEGKFLYISETSVIYGLSQVELTGNISIEYIHPADHDEMTAVLTA
LEWKFLLDHRAPPIITGYLPEFVLTSGDYVYHIDDELARHCHQLMQFGKGSYCYRFLTKGQGMWIL
SRDAARSRGKENFELAKLLPLPSAITSQDKASTIIRLTTSYLKMRDFANQDPPNLRMEGPPPTI
IVAALPGFLLVFTAEGKLLVSESVSEHLSHVDLVAQGSIDVIDPADHLLTVRQQLTLTDLFRFRF
EKSNAARTREKENSEFELAKLLPLPSAITSQDKASTIIRLTTSYLKMRVFPFEGLEAMGHSSRTSP
EKSNAARTREKENSEFELAKLLPLPSAITSQDKASTIIRLTTSYLKMRVFPFEGLEAMGHSSRTSP
ELKHLILEADGFLFVSCETGRVVVSDVSTPVLNQPSSEWFGSTLYDQVHPDDVKLREQLSTRMCH
DELKHLILRAADGFLFVGCGRGKILFVSESVFKILNYSQNDLIGQSLFDYLHPKDIKAVKEQLSSRLC
KPFVFDQRAITAILAYLPQELLGTSCYEFHQDIGHLAECRHQVLTREKITTNCYKFKIKDGSFITLRS
khushi@DESKTOP-DE71G03: ~/Lab_session2$ grep -v ">" sequence5.fasta | grep -E "P" sequence5.fasta
SNPSKRHRDRINTELDRLASLLPFPQDVINKDLKLSVLRISVYLRKSFDFVALKSSPTERNGGQONCR
QLPHQITPPENSLMERCFCICRLCLDNNSSGLAMNFQGLKYLPPQLALFAIATPLQPPSILEIRTKNF
IFRTKHKLDFTPIGCDAGRIVLGYTEAELCTRGSGYQFIHAAOMLYCAESHIRMIKTGESGMIVFRLLT
KNNRWTHVQSNARLLYKNGRPDIYIVTQRPLTDEEGTEHLR
VSRNKSEKKRRDQFNVLKELGSMPLGNARKMDKSTVLQKSIDFLRKHKETAQSDASEIRQDKPFTLS
NEEFTQLMLEALDGFLLAINTDGSIIYVSESVTSLLEHLPSDLVQSIINFPIEGEHSEVYKILSTEYK
SKHQLEFCCMHLRGITDKEPSTVEYVKFIGNFKSLYEDRVCFVATVRLATQFIKEMCTVEEINEEFTS
RHSLKWLFLDHRAPPIITGYLPEFVLTSGDYVYHIDDELARHCHQLMQFGKGSYCYRFLTKGQGMWIL
IWLQTHYIITYHQNRPFEIVCTHTVSYAEVRAE
KEKSRDAARSRRSKEFVYFELAHQLPLPHNVSHLDKASVHRLTISYLRVRKLLDAGOLDIEDDKAKQH
NCEYLKALDGFVHVLTDGDMYISDNVKNYMLTQFELTGHVSFDFTHPCDHEEMREMLTHNTQSFFL
RMKCTLSRGRTHNKSATKVLHCTGHIHYKPPMTCLVICEPIPHPSNIEIPLDSKTFLSRHSIDMK
FSYCDERITELMGYEPPEELLGRSIEYVYHALDSOHLTKTHDMFTKGQVTTGQYRLAKRGYVWETQA
TVIYNTKNSQPQCIVCNVYVSGIIQHDL
```

8. Print only the sequence lines (ignore headers) from sequence5.fasta that contain the letter P.

```
khushi@DESKTOP-DE71G03: ~/Lab_session2$ grep -v ">" sequence5.fasta | grep -E "P" sequence5.fasta
SNPSKRHRDRINTELDRLASLLPFPQDVINKDLKLSVLRISVYLRKSFDFVALKSSPTERNGGQONCR
QLPHQITPPENSLMERCFCICRLCLDNNSSGLAMNFQGLKYLPPQLALFAIATPLQPPSILEIRTKNF
IFRTKHKLDFTPIGCDAGRIVLGYTEAELCTRGSGYQFIHAAOMLYCAESHIRMIKTGESGMIVFRLLT
KNNRWTHVQSNARLLYKNGRPDIYIVTQRPLTDEEGTEHLR
VSRNKSEKKRRDQFNVLKELGSMPLGNARKMDKSTVLQKSIDFLRKHKETAQSDASEIRQDKPFTLS
NEEFTQLMLEALDGFLLAINTDGSIIYVSESVTSLLEHLPSDLVQSIINFPIEGEHSEVYKILSTEYK
SKHQLEFCCMHLRGITDKEPSTVEYVKFIGNFKSLYEDRVCFVATVRLATQFIKEMCTVEEINEEFTS
RHSLKWLFLDHRAPPIITGYLPEFVLTSGDYVYHIDDELARHCHQLMQFGKGSYCYRFLTKGQGMWIL
IWLQTHYIITYHQNRPFEIVCTHTVSYAEVRAE
KEKSRDAARSRRSKEFVYFELAHQLPLPHNVSHLDKASVHRLTISYLRVRKLLDAGOLDIEDDKAKQH
NCEYLKALDGFVHVLTDGDMYISDNVKNYMLTQFELTGHVSFDFTHPCDHEEMREMLTHNTQSFFL
RMKCTLSRGRTHNKSATKVLHCTGHIHYKPPMTCLVICEPIPHPSNIEIPLDSKTFLSRHSIDMK
FSYCDERITELMGYEPPEELLGRSIEYVYHALDSOHLTKTHDMFTKGQVTTGQYRLAKRGYVWETQA
TVIYNTKNSQPQCIVCNVYVSGIIQHDL
```

NKNQLEFCHMLRGZIDPKEPSYEVYKFIQGNFKSLYEDRVCFVATVRLATQFQIKEMCTVEPNEEFTS  
 RSHLEWKFLFLDHRAPDIIIGYLFPEVLGTSGDYGVYHDDLNLAKCHEHLMQYGGKSCCYRFLTKGQQW  
 IWLQTHYYITYHQWNSRPEFIVCTHTVVSVAEVRAS  
 KEKSRDAARSRRSKESEVFYELAHQLPLPHNVSSHLDKASIMRLTISYLRVRKLIDLDAOLDDEDMKAQM  
 NCYFKALGDGVVWLTDGDMITYSIDMNMKVMGLTOFELVSEVDFTHPCDHEENREILTHNTQSSFL  
 KQKCTLSSTHMLKASVYLHCTGHRVYKPPMTCLVLCEPTHPMSHTEFLDSKFTLSRSHLDK  
 VSYCDERTELTNGVEPEELLGRSTIYHYHALDSHDKTKTHDMFTKGQVTTGQYRMLAKRGGVVWVETQA  
 TVIYTNKNSQQTICVVCNVYVSGTIQHDL  
 KEKSRDAARCRRSKETEVFYEALHELPLPHSVSSHLDKASIMRLTISFLRTHKLSSVSCSENESEAEADQ  
 QMDMLYLKALGEFIAVVTQDGMITLSENISKPFMGLTOVELTGHISDFDTHPCDHEETRENLSSTERDFF  
 MRMKCTVTNNGRTVNILKSAWKVLHCTGQVKVYEPLLSCLIIIMCEPIQHSHMDIPLDKFTLSRSHMD  
 KFTYCDORITELIGVHEPELLGRSAYEYHALDSENMTKSHQNLCTKGQVVSQYRMLAKHGQVYMLETQ  
 GTVIYNPRNLQPCIMCVNYVLSIEKNDV  
 DAARSRRSQTEVLYQLAHTLPFARGVSAHLDKASIMRLTISYLRMHRLCAAGENVQVGAGGEPDACYL  
 KALGGVWMLTAGDMYAMISSENVSHMLKLSOLELTGHSITDFTHPCDHELODALTPPTNCSFLRMKST  
 LSTRTNLGLATMLNCEGSHRANRANGPPVQLVLTCEATHPGSELPPGLGRAGFLSRHSIMKFTYCD  
 DRTAEVAGVSPDOLTGSAYEYHALDSAVKSIHTLLSKGQAVTGYRFLARSGGVLHTQQTATVSG  
 WRGQSEISYCVHFLTQSVEETGV  
 KEKSMAARSRRGKENIEFFELAKLLPLPAITSSQLDKASIVRLSVTYLRLRRFAALGAPPMGLRAAGPP  
 AGLAPGRRGGPAAVSEVFEQHLGGHILQSLDGFVVALNQEGLFYSETSVIYGLSQVEMTSSVDFYI  
 HPDGHSEVLEQLGQVRSFFVRMKSTLTRKGLHVKASGVKYIHTVGRLLRGLVALGHTLPPAPLAEPL  
 LHHGMITVRLSLGLTILACESRSDHMDLGPSELVGRSCYQFVHGQDATIRQSHVDLLDKGQVMTGYR  
 MLRAGGQFVWLQSVATVAGVSGKSPGEHVLVWHSVLSAQEGQT  
 NKSEKKRRDQFNWLKELSSMLPGNTRKMDKTLEKIVGLQKHNEVSAQTEICDIQQDWKPSFLSNEE  
 FTQLMLLEALDGFIIATVTDGSIIVSDSTPLLGHLLPSDVMQDLNLFLEPQEHSEVYKLLSSEYLSKDS  
 LEVFCYHLRQSLNKEFPYEVYKFGHIFREYLGKEVCFIATVRLATQFQIKEMCTVEPNEEFTS  
 RSHLEWKFLFLDHRAPDIIIGYLFPEVLGTSGDYGVYHDDLNLAKCHEHLMQYGGKSCCYRFLTKGQQW  
 IWLQTHYYITYHQWNSRPEFIVCTHTVVSVAEVRAS  
 KEKSMAARSRRGKENIEFFELAKLLPLPAITSSQLDKASIMRLTISYLRMHRAFNQGDPPMLRMEGPPNT  
 SVKVGARRRRSPALATIEVFEAHLGSHILQSLDGFVVALNQEGLFYSETSVIYGLSQVELTGSSVF  
 DYVHPGDHVEAEQGLTLEFFIRMKSTLTRKGVHKSGGYKYIHTVGRLLRMLGVVAHALPPPTI  
 NEVRIDCHMFVTRVMDMLIYECNRSIDYMDLPDVIKGRCYVFIHAEDVEGIRHSHLDLKNKGQCVT  
 KYRRWQKNGGYIWTQSSATIAINAKNANEKNIWNYLLSNPEYKDT  
 GASKARRDQINAEIRNMLKELPLAEADKVRLSYLHMSLACTIRYRGVFFAGGTPLAGTGLLSAQELED  
 IVAALPGFLLVFTAGKLLYLSSEVSEHLGHSMDLVAGQDSYDIIDPADHTRVQQLTLDRFLRCFR  
 NTKSLRRQSGAGKLVLTRGFHANHVPFATCAPLEPRRPGPGPGGASFLAMFQSRHAKDLAL  
 RSHLEWKFLFLDHRAPDIIIGYLFPEVLGTSGDYGVYHDDLNLAKCHEHLMQYGGKSCCYRFLTKGQQW  
 IWLQTHYYITYHQWNSRPEFIVCTHTVVSVAEVRAS  
 KEKSMAARSRREKENSEFELAKLLPLSAITSSQLDKASIMRLTISYLMRNVFPEGLGEAGHGSRTSP  
 LDNVRGLGSHLQTLQDGFVVPADGKTHYISETASVHGLSQVELTGNISYIYTHADHDEMTAVLTA  
 ETERSFFLRMKCVLAKRNAGLTGGYKVTHCSGYLKIRNVGLVAVGHSLLPSAVEITLHSMFMFMRASL  
 DMKLTFLDSRVAELTGVEPQDLIEKTLYHHVHGCTFHLRCAHLLLVKGQVTTKYRFLAKHGGMVWQ  
 SYATIVHNSRSRSPHICIVSVVNYVTDTEYKGL  
 KEKSMAARSRREKENSEFELAKLLPLSAITSSQLDKASIMRLTISYLMRNVFPEGLGDAMGQPSRAGP

[illegible]

## Part 3 - Using Variables

9. Store the filename sequence5.fasta in a variable called seq and print the number of sequences in it (headers count as sequences).

```
RWFSFMNIPWKEVEYIVSTNTVL
khushi@DESKTOP-DE71G03:~/Lab_session2$ seq="sequence5.fasta"
khushi@DESKTOP-DE71G03:~/Lab_session2$ grep -c ">" $seq
13
khushi@DESKTOP-DE71G03:~/Lab_session2$ pattern="G{2,}"
khushi@DESKTOP-DE71G03:~/Lab_session2$ grep -v ">" protein.fasta | grep "pattern" protein.fasta
khushi@DESKTOP-DE71G03:~/Lab_session2$ variable="Biocomputing"
khushi@DESKTOP-DE71G03:~/Lab_session2$ export variable
khushi@DESKTOP-DE71G03:~/Lab_session2$ bash -c 'echo $variable'
Biocomputing
khushi@DESKTOP-DE71G03:~/Lab_session2$ vi script.sh
```

10. Store the pattern `G{2,}` in a variable and search protein.fasta for sequence lines (ignore headers) with 2 or more consecutive Gs.

```
RWFSFMNIPWKEVEYIVSTNTVL
khushi@DESKTOP-DE71G03:~/Lab_session2$ seq="sequence5.fasta"
khushi@DESKTOP-DE71G03:~/Lab_session2$ grep -c ">" $seq
13
khushi@DESKTOP-DE71G03:~/Lab_session2$ pattern="G{2,}"
khushi@DESKTOP-DE71G03:~/Lab_session2$ grep -v ">" protein.fasta | grep "pattern" protein.fasta
khushi@DESKTOP-DE71G03:~/Lab_session2$ variable="Biocomputing"
khushi@DESKTOP-DE71G03:~/Lab_session2$ export variable
khushi@DESKTOP-DE71G03:~/Lab_session2$ bash -c 'echo $variable'
Biocomputing
khushi@DESKTOP-DE71G03:~/Lab_session2$ vi script.sh
```

11. Store "Biocomputing" in a variable, export it, and verify that it is available inside a new shell started using:

```
bash -c 'echo $VARIABLE_NAME'
```

```
RWFSFMNIPWKEVEYIVSTNTVL
khushi@DESKTOP-DE71G03:~/Lab_session2$ seq="sequence5.fasta"
khushi@DESKTOP-DE71G03:~/Lab_session2$ grep -c ">" $seq
13
khushi@DESKTOP-DE71G03:~/Lab_session2$ pattern="G{2,}"
khushi@DESKTOP-DE71G03:~/Lab_session2$ grep -v ">" protein.fasta | grep "pattern" protein.fasta
khushi@DESKTOP-DE71G03:~/Lab_session2$ variable="Biocomputing"
khushi@DESKTOP-DE71G03:~/Lab_session2$ export variable
khushi@DESKTOP-DE71G03:~/Lab_session2$ bash -c 'echo $variable'
Biocomputing
khushi@DESKTOP-DE71G03:~/Lab_session2$ vi script.sh
```

## Part 4 - File Existence & Loops

12. Write a shell script that checks if sequence3.fasta exists in the current folder. If yes, print the number of lines. If no, print "Missing file".

13. Using a for loop, go through all .fasta files in the current directory and print: filename, number of sequences, and file size in characters.

14. Modify the above loop so that it only prints files with more than 3 sequences.





## Part 5 - Applied Data Extraction

15. From sequence5.fasta, extract only the sequence lines (no headers) that contain 3 or more cysteines (C). Save the output to a file named cys\_rich.txt. Ensure the output file contains no empty lines.

```
khushi@DESKTOP-DE71G03:~/Lab_session2$ cat sequence5.fasta
Sequences:13
File size (characters): 4229
khushi@DESKTOP-DE71G03:~/Lab_session2$ grep -v "^>" sequence5.fasta | grep "C.*C" sequence5.fasta | grep -v "^$" sequence5.fasta > cys_rich.txt
khushi@DESKTOP-DE71G03:~/Lab_session2$ less cys_rich.txt
khushi@DESKTOP-DE71G03:~/Lab_session2$ man sort
khushi@DESKTOP-DE71G03:~/Lab_session2$ grep -c "^>" *.fasta | sort -k2 -nr | head -n 1
sequence5.fasta:13
khushi@DESKTOP-DE71G03:~/Lab_session2$ _
```

```
khushi@DESKTOP-DE71GO3: ~/Lab_session2
>ahr
SNPSKRRHRRDLNTELDRLASLLPFPQDVINKDLKLSVLRISVSYLRAKSFDFVALKSSPTERNGGQDNCR
AANFREGNLQEGEFLQALNGFVLVTTDALVFYASSTIQDVLGFQSDVHQSVELIHTEDRAEFQR
QLHMQIPPENSPLMERCFICRLRCLLNDSSGFLAMNFQGLKLYLPPQLALFAIATPLQPPSILEIRTKNF
IFRTKHLDFTPIGCDAGRIVLGYTEALCTRGSGYQFIAHADMLYCAESHIRMIKTGESGMIVFRLLT
KNNRWTVQSNARLLYKNGRDPYIIVTQRPLDDEEGTEHLR
>c1ock
VSRNKSEKKRRDQFNVLKELGSHLPGNARKMDKSTVLQKSTDFLRKHKEITAQSDASEIRQDKKPTFLS
NEEFQLMLEALDGGFLAINTDGSIIYVSESVTSLLEHLPSDLVQSTFNFIPGEHSEVYKLLSTEVYK
SKNQLFCGHLRGITIDPKEPSTYEVYKFTGNFKSLYEDRVCFVATVRLATPQFIKEMCTVEEPNEEFTS
RHSLEWFLFLDHRAPPYIIGVLPFEVLGTSGDYVYHVDLENLAKCHEHLNQYKGGKSCYYRFLTKGQQW
IWLQTHYIITYHQWNSRPEFIVCTHTVVSVAEVRAE
>hif1a
KEKSRDAARSRRSKSEVFYELAHQLPLPHNVSSHLDKASVMRLTISYLRVRKLLDAGDLTIEDDMKAQM
NCFYLKALDGFVMVLTDGDMITYSDINVKYMGLTQFELTGHVSFDFTHPCDHEEMREMLTHNTQSFFL
RMKCTLTSRGRTHMIKSAWKVLHCTGHIHYKPPMTCLVLICEPIPHPSNIEIPLDSKTFLSRHSMDM
FSYCDERITELMGYEPEELLGRSIEYHYHALDSHDLTKTHDMFTKGQVTTGQYRMLAKRGYVWVETQA
TVIYNTKNSQPQCIVCNVYVSGIIQHDL
>hif2a
KEKSRDAARSRRSKSEVFYELAHQLPLPHNVSSHLDKASIMRLATSFLRTHKLLSSVCSENESEAEADQ
QNDNLYLKALGFIIVTQDGMIFLSENIKFMGLTQVELTGHISFDFTHPCDHEEIRENLSTTERDFF
MRMKCTVYNTNRGRTVNLKSATWKVLHCTGQVKVVEPLLSCLTIKCEPIQHPSHMDIPLDSKTFLSRHSMDM
KFTYCDRITTELIGYHPEELLGRSAYEFYHALDSENMTKSHQNLCTKGQVVSQYRMLAKHGQVWILETQ
GTVIYNPNRLQPQCIMCVVYVSEIEKNV
>hif3a
DAARSRRSQETEVYQLAHTLPFARGVSAHLDKASIMRLTISYLRMRHLCAAGENVQVAGGEPDACYL
KALEGFVMVLTAEGDMAYLSENVSKHLGSQLLEIGHISFDFIHPDQDEELQDALTPPTERCFSLRMKST
LTSRGRITLNLKAATWKVLNCSGHRAYEPPLQCLVLICEAIPHPSLEPPLGRGAFLSRHSMDMKFTYCD
DRIAEVAGVSPDDLIGCSAYEYHALDSAVSKSIHTLLSKGQAVTGQYRFLARSGGYVLTQTATVVSQ
GRGPQSEIIVCVHFLTSQVEETGV
>npsa1
KEKSRDAARSRRKENLEFFELAKLLPLPGATSSQLDKASIVRLSVTYLRLRRFAALGAPPWGLRAAGPP
AGLAPRRGPAALVSEVFQHLGGHILQSLDGFVFNALNQGKFLYISETVSYLGLSQVEMTGSSVFDYI
HPGDHSEVLQGLVQERSFVRMKSTLTKRGLHVKASGYKVIHVTGRLRALGLVALGHTLPPAPLAEPL
LHGHHMIVFRLSLGLTILACESRVSDHMDLPSELVGRSCYQFVHGQDATRIRQSHVDLLDKGQVMTGYR
WLRAGGFVWLQSVATVAGSGKSPGEHVLWVSHVLSQAEGGT
>npsa2
NKSEKKRRDQFNVLKELSSMLPGNTRKMDKTTVLKVIKGLQKHNEVSAQTEICDIQDQWKPFLSNEE
FTQLMLEALDGFIIIVTQDGSIIYVSDSITPLLGHLPDVMQNLNLFLEQEHSEVYKILSSEYLSKDS
DLFEYCHLRLGSLNPKFPTYEYKFGVNFRLVKEVCFIATVRLATPQFLKENCIVDEPLEEFTSRHS
LEWKFLFLDHRAPPYIIGVLPFEVLGTSGDYVYHIDLELLARHQHLMQFQGGKSCCYRFLTKGQQWIML
QTHYIITYHQWNSKPEFIVCTHSVVSADVRVE
>npsa3
cys_r1ch.txt
```

```
khushi@DESKTOP-DE71GO3: ~/Lab_session2
>hif2a
KEKSRDAARSRRSKSEVFYELAHQLPLPHNVSSHLDKASIMRLATSFLRTHKLLSSVCSENESEAEADQ
QNDNLYLKALGFIIVTQDGMIFLSENIKFMGLTQVELTGHISFDFTHPCDHEEIRENLSTTERDFF
MRMKCTVYNTNRGRTVNLKSATWKVLHCTGQVKVVEPLLSCLTIKCEPIQHPSHMDIPLDSKTFLSRHSMDM
KFTYCDRITTELIGYHPEELLGRSAYEFYHALDSENMTKSHQNLCTKGQVVSQYRMLAKHGQVWILETQ
GTVIYNPNRLQPQCIMCVVYVSEIEKNV
>hif3a
DAARSRRSQETEVYQLAHTLPFARGVSAHLDKASIMRLTISYLRMRHLCAAGENVQVAGGEPDACYL
KALEGFVMVLTAEGDMAYLSENVSKHLGSQLLEIGHISFDFIHPDQDEELQDALTPPTERCFSLRMKST
LTSRGRITLNLKAATWKVLNCSGHRAYEPPLQCLVLICEAIPHPSLEPPLGRGAFLSRHSMDMKFTYCD
DRIAEVAGVSPDDLIGCSAYEYHALDSAVSKSIHTLLSKGQAVTGQYRFLARSGGYVLTQTATVVSQ
GRGPQSEIIVCVHFLTSQVEETGV
>npsa1
KEKSRDAARSRRKENLEFFELAKLLPLPGATSSQLDKASIVRLSVTYLRLRRFAALGAPPWGLRAAGPP
AGLAPRRGPAALVSEVFQHLGGHILQSLDGFVFNALNQGKFLYISETVSYLGLSQVEMTGSSVFDYI
HPGDHSEVLQGLVQERSFVRMKSTLTKRGLHVKASGYKVIHVTGRLRALGLVALGHTLPPAPLAEPL
LHGHHMIVFRLSLGLTILACESRVSDHMDLPSELVGRSCYQFVHGQDATRIRQSHVDLLDKGQVMTGYR
WLRAGGFVWLQSVATVAGSGKSPGEHVLWVSHVLSQAEGGT
>npsa2
NKSEKKRRDQFNVLKELSSMLPGNTRKMDKTTVLKVIKGLQKHNEVSAQTEICDIQDQWKPFLSNEE
FTQLMLEALDGFIIIVTQDGSIIYVSDSITPLLGHLPDVMQNLNLFLEQEHSEVYKILSSEYLSKDS
DLFEYCHLRLGSLNPKFPTYEYKFGVNFRLVKEVCFIATVRLATPQFLKENCIVDEPLEEFTSRHS
LEWKFLFLDHRAPPYIIGVLPFEVLGTSGDYVYHIDLELLARHQHLMQFQGGKSCCYRFLTKGQQWIML
QTHYIITYHQWNSKPEFIVCTHSVVSADVRVE
>npsa3
SRDAARSRRKENLEFFELAKLLPLPAAITSSQLDKASIIRLTISYLMRDFANQDPPWNLRMGPPNPT
SVKVIQAQRRRSPSALAEVFEAHLGSHILQSLDGFVFNALNQGKFLYISETVSYLGLSQVELTGSSVF
DYVHPGDHVEMAEQGLHTLEFSFFIRMKSTLTKRGVHKSSGYKVIHVTGRLRLRMLGVVAHALPPPTI
NEVRIDCHMFVTRVMDLNIIVCENRISDYMIDTPVDIVGKRCYHFIAEDVEGIRHSHLDLLNKGQCVT
KYVRWQKNGGYIWIQSSATIAINAKNANEKNIIWVNYLLSNPEYKDT
>npsa4
GASKARRDQINAEIRNLKELLPLAEADKVRVSVLHMSLACIYTRKGVFFAGGTPLAGPTGLLSAQELED
IIVAALPGFLLVFTAEGKLLYLSSEVSEHLGHSMDLVAQDGSYDIDPADHLTVRQQLTLTDRLFRCRF
NTSKSLRQASAGNKLVLIRGRFAHNVPVTFACAPLEPRPRPGPGPGPASFLAMFQSRHAKDLALLD
TSESVLYTLGFERSELLCKSNYGLLHPEDLAHASAQHYRLLAESGDIQAEHVVRLQAKTGGNAIYCLLY
SEGPEGITANNYPIISDHEAMSRLQQL
>s1m1
EKSNAARTBREKENSEFYEALKLLPLPSAITSQDKASIIRLTISYLMRVVFPEGLGEAWGHSSRTSP
LDNVGRELGSHLLQTLQDGFIVVAPDGKIMYISETASVHLGLSQVELTGNISYIEYIHPADHDENTAVLA
ETERSFFLRMKCVLAKRNAGLTCGGYKVIHCSGYLKIRNVGLVAVGHSLLPSSAVTEIKLHSNMFMFRASL
DMKLIFLDSRVAELTGVEPQDLIEKTLYHHVHGCDTFHLRCAHLLLVKGQVTKYRYRFLAKHGGWVWVQ
SYATTIVHNSRSRPHCIVSVNYVLTDTYEKGL
>s1m2
```



```
khushi@DESKTOP-DE71GO3: ~/Lab_session2
HPGDHSEVLQGLVQERSFVRMKSTLTKRGLHVKASGYKVIHVTGRLRALGLVALGHTLPPAPLAEPL
LHGHIIVFRLSLGLTILACESRVSDHMDLGPSEVLRSCYQFVHGQDATRIRQSHVDL LDKGQVMTGYVR
MLQRAGGFVWLQSVATVAGSGKSPGEHVLWVSHVLSQAEGGQT
>nps2
NKSEKKRRDQFNVLKELSSMLPGNTRKMDKTTVLKVIIGFLQKHNEVSAQTEICDIQQDWKPSFLSNEE
FTQLMLEALDGFIIAVTDTGSIYVSDSITPLLGHLPDOVMDQNLNLFLEQEHSEVYKILSSEYKLSDS
DLFEYCHLRGSLNPKFPTIYKIFVGNFRSLGKEVCFIATVRLATPQFLKEMCIVDEPLEEFTSRHS
LEWKFLDLHRAPPIIGVLPFVLTSGVYDYHIDDLLELARCHQHLMQFGKGKSCCYRFLT KGQQQWJL
QTHYITTYHQNSKPEFIVCTHSVSVYADVRE
>nps3
SRDAARSRRKENFEFVELAKLLPLAAITSQLDKASIIRLTISYLKMRDFANQGDPPWNLRMGPPNPT
SVKVTGAQRRRSALAEVFEAHLGSHLQSLDGFVVALNQEGKFLYISETVSYLGLSQVELTGSSVF
DYVHPGDHVEAEQGLGHTLERSFIRMKSTLTKRGVHIKSSGYKVITHITGRLRLRMLVVAHALPPPTI
NEVRIDCHMFVTRVMDLNIIVCENRISDYMDLTPVDIVGKRCYHF IHAEDVEGIRHSHLDLLNKGQCVT
KYYRMMQKNGGYIWIQSSATIAINAKNANEKNIWVNYLLSNIPEYKDT
>nps4
GASKARRDQINAEIRNLKELLPLAEADKVLRYLHMSLACTYTRKGVFFAGGTPLAGPTGLLSAQELED
IVAALPGFLLVFTAEGKLLYLSSESVSEHLGHSMDLVAGQDSIYDIIDPADHLTVRQQLTLTDLRFRCRF
NTSKSLRQASAGNKLVLIRGRFAHNPFVTAFCAPLEPRPRPGPGGPGPASLFLAMFQSRHAKDALLD
ISESVLYLGFERSELLCKSWYGLLHPEDLAHASAQHYRLAESGDIOQAEHVVRLQAKTGGNAIYCLLY
SEGPEGITANNYPISDMEAWSLRQQL
>sim1
EKSNAARTREKENSEFVELAKLLPLPSAITSQDKASIIRLTTSYLKMRVVFPEGLGEAMGHSRTSP
LDNVGRELGSHLLQTLGDFIFVVPADGKIMYISETASVHLGLSQVELTGNSIYEHIPADHDEMTAVLT
EIERSSFFLRMKCVLAKRNAGLTCGYKVITHSGYLKIRIVGLVAGQSLPPSAITEIKLYSNMFMRASL
DKLFLDLSRVTEVTGVEPQDLIEKTLVHHVHGCDVFLRYAHLLLVKGQVTTYKYRFLSKRGGNVWVQ
SVATVHNSRSRPHCIVSVNYVLTDEYKGL
>sim2
EKSNAARTREKENSEFVELAKLLPLPSAITSQDKASIIRLTTSYLKMRVVFPEGLGEAMGQPSRAGP
LDGVAKELGSHLLQTLGDFIFVVPADGKIMYISETASVHLGLSQVELTGNSIYEHIPADHDEMTAVLT
EIERSSFFLRMKCVLAKRNAGLTCGYKVITHSGYLKIRIVGLVAGQSLPPSAITEIKLYSNMFMRASL
DKLFLDLSRVTEVTGVEPQDLIEKTLVHHVHGCDVFLRYAHLLLVKGQVTTYKYRFLSKRGGNVWVQ
SVATVHNSRSRPHCIVSVNYVLTDEYKGL
>ant1
NHSEIERRRRNKMTAYITELSDMVPTCSALARKPKLTIIRMAVSHMKSRLGTGNTSDGSKPSFLTDQ
LEKHLILEADGFLFVSCETGRVVYVSDSVTPVLNQPSSEWFGSTLYDQVHPDDVKLREQLSTSRMCM
GSRRSFICRMRCGSEPHFVWHCTGYIKAKFLVATGRLQVTSPPNCTDMSHVCQPTFISRHNIETGIF
TFVDHRCVATVGVQPQELLGKNIIEFCHPEDQQLLRDSFQQVVKLGQVLSVMFRFSKNQEWLWMTSS
FTQNPYSDEIEYICTNTNVK
>bma11
EAHSQIEKRRRDKMNSFIDELASLVPTCNAMSRKLDKLTVLRMAVQHMKTLRGATNPYTEANYKPTFLSD
DELKHLILRAADGFLFVVGCDRGKILFVSESVFKILNYSQNDLIGQSLFDYLHPKDIKAKVEQLSSSLR
SGARRSFFCRMKNCRPKSFCTIHTSYLKSNSLCLVAIGRLHSHVVPQVNGEIRVKSMEYVSRHAIDG
KVFVDQRATAILAYLPQELLGTSCYEFHQDIGHLAECHRQVLQTREKITTNCYKFKIKDGSFITLRS
RWFSFMMNPWKEVEYIVSTNTVVL
(END)
```

## Extra Challenge (Optional)

Write a single shell command that finds the file in the current directory with the largest number of sequences (by header count) and prints:

<filename> has <count> sequences

Hint: You will likely need wc, grep, sort, and head.

```
filename:sequences.fasta
Sequences:13
File size (characters): 4229
khushi@DESKTOP-DE71G03:~/Lab_session2$ grep -v "^>" sequence5.fasta | grep "C.*C.*C" sequence5.fasta | grep -v "^$" sequence5.fasta > cys_rich.txt
khushi@DESKTOP-DE71G03:~/Lab_session2$ less cys_rich.txt
khushi@DESKTOP-DE71G03:~/Lab_session2$ man sort
khushi@DESKTOP-DE71G03:~/Lab_session2$ grep -c "^>" *.fasta | sort -k2 -nr | head -n 1
sequences5.fasta:13
khushi@DESKTOP-DE71G03:~/Lab_session2$ _
```

I used [google AI Mode search](#) to understand the sort command use in this question then I used command **sort -k2 -nr filename** from which I understood we use sort to segregate or arrange data. In this command by giving flag -k2 I specified the field which it has to look for to arrange data and by 2 I meant it has to look for 2<sup>nd</sup> column which has the number of sequences to find the file in the current directory with the largest number of sequences (by header count) then flag -nr I specified to sort numerical data and in reverse/descending order.