Mansuri Naushin Parveen Sagir Ahmed (24310041)

24310041

Lab Assignment - 2

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

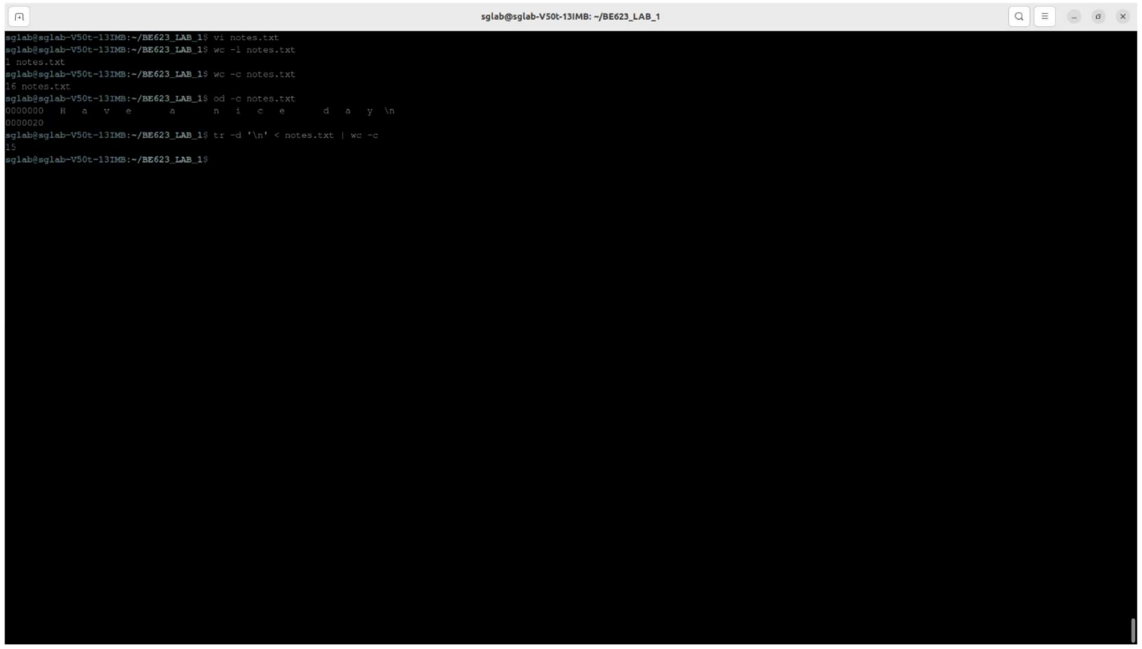
- Insert exactly one line of text:

Have a nice day

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

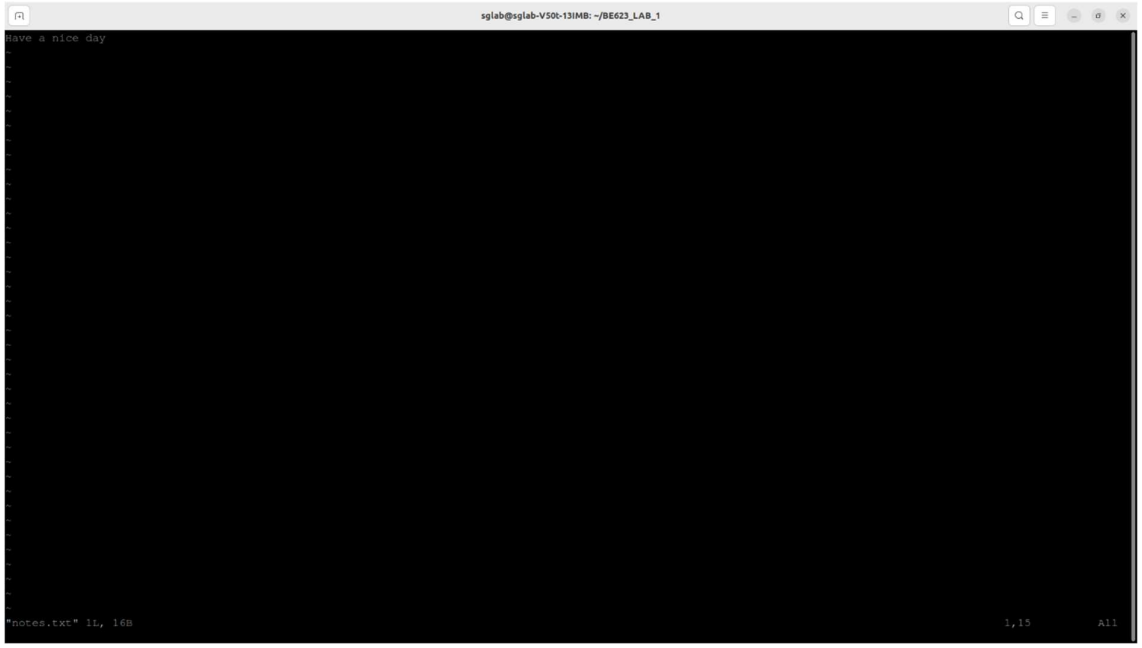
- Save and exit.

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



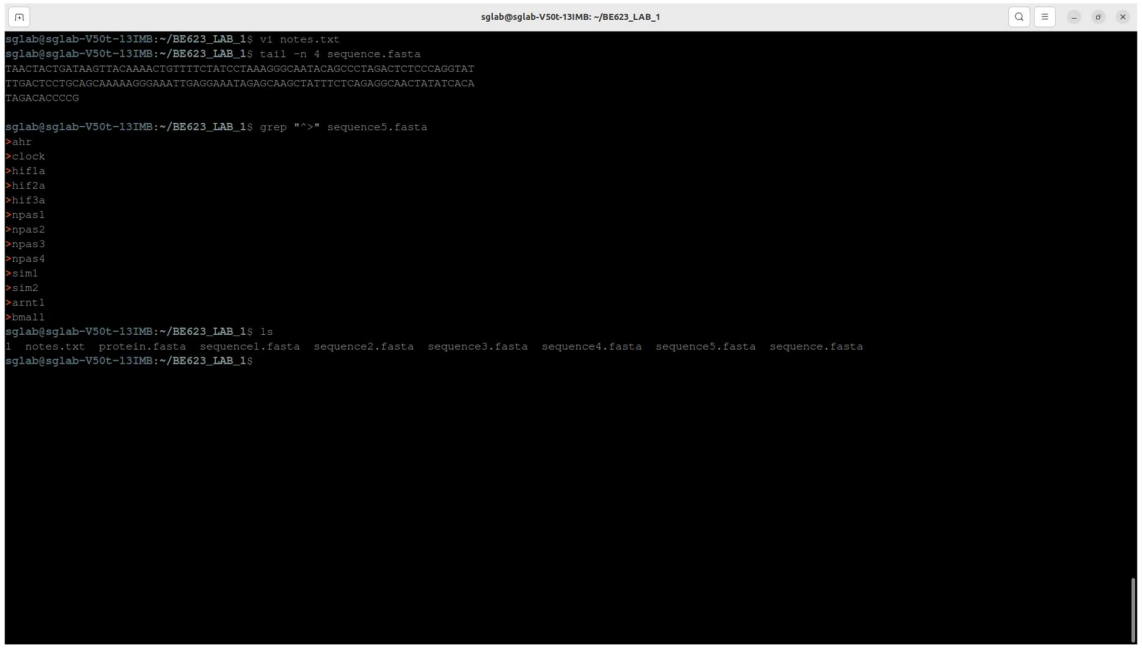
During Part 1, I faced an issue where wc -m showed 16 characters instead of 15. To solve this, I used two extra Linux commands (beyond the lab sheet):

od -c notes.txt → shows hidden characters like newline (\n) or carriage return (\r). This confirmed that the extra count was due to a newline, and then I used tr –d '\n' < notes.txt | wc –m to count characters without the newline (Ref: used chat GPT).



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

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



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

```
sglab@sglab-V50t-13IMB: ~/BE623_LAB_1
sglab@sglab-V50t-13IMB:~/BE623_LAB_1$ grep "A.G" sequence5.fasta
IFRTKHKLDFTPIGCDAKGRIVLGYTEAELCTRGSGYQFIHAADMLYCAESHIRMIKTGESGMIVFRLLT
DAARSRRSQETEVLYQLAHTLPFARGVSAHLDKASIMRLTISYLRMHRLCAAGEWNQVGAGGEPLDACYL
KALEGFVMVLTAEGDMAYLSENVSKHLGLSQLELIGHSIFDFIHPCDQEELQDALTPPTERCFSLRMKST
KEKSRNAARSRRGKENLEFFELAKLLPLPGAISSQLDKASIVRLSVTYLRLRRFAALGAPPWGLRAAGPP
AGLAPGRRGPAALVSEVFEQHLGGHILQSLDGFVFALNQEGKFLYISETVSIYLGLSQVEMTGSSVFDYI
HPGDHSEVLEQLGLVQERSFFVRMKSTLTKRGLHVKASGYKVIHVTGRLRALGLVALGHTLPPAPLAELP
WLQRAGGFVWLQSVATVAGSGKSPGEHHVLWVSHVLSQAEGGQT
GASKARRDQINAEIRNLKELLPLAEADKVRLSYLHIMSLACIYTRKGVFFAGGTPLAGPTGLLSAQELED
IVAALPGFLLVFTAEGKLLYLSESVSEHLGHSMVDLVAQGDSIYDIIDPADHLTVRQQLTLTDRLFRCRF
EKSKNAARTRREKENSEFYELAKLLPLPSAITSQLDKASIIRLTTSYLKMRVVFPEGLGEAWGHSSRTSP
EIERSFFLRMKCVLAKRNAGLTCGGYKVIHCSGYLKIRNVGLVAVGHSLPPSAVTEIKLHSNMFMFRASL
EKSKNAAKTRREKENGEFYELAKLLPLPSAITSQLDKASIIRLTTSYLKMRAVFPEGLGDAWGQPSRAGP
EIERSFFLRMKCVLAKRNAGLTCSGYKVIHCSGYLKIRIVGLVAVGQSLPPSAITEIKLYSNMFMFRASL
ELKHLILEAADGFLFIVSCETGRVVYVSDSVTPVLNQPQSEWFGSTLYDQVHPDDVDKLREQLSTSRMCM
GSRRSFICRMRCGSSEPHFVVVHCTGYIKAKFCLVAIGRLQVTSSPNCTDMSNVCQPTEFISRHNIEGIF
DELKHLILRAADGFLFVVGCDRGKILFVSESVFKILNYSQNDLIGQSLFDYLHPKDIAKVKEQLSSSRLC
SGARRSFFCRMKCNRPRKSFCTIHSTGYLKSNLSCLVAIGRLHSHVVPQPVNGEIRVKSMEYVSRHAIDG
sglab@sglab-V50t-13IMB:~/BE623_LAB_1$
```

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

```
sglab@sglab-V50t-13IMB: ~/BE623_LAB_1
sglab@sglab-V50t-13IMB:~/BE623_LAB_1$ grep "P[^A]L" sequence5.fasta
QLHWQIPPENSPLMERCFICRLRCLLDNSSGFLAMNFQGKLKYLPPQLALFAIATPLQPPSILEIRTKNF
MRMKCTVTNRGRTVNLKSATWKVLHCTGQVKVYEPLLSCLIIMCEPIQHPSHMDIPLDSKTFLSRHSMDM
LTSRGRTLNLKAATWKVLNCSGHMRAYEPPLQCLVLICEAIPHPGSLEPPLGRGAFLSRHSLDMKFTYCD
FTQLMLEALDGFIIAVTTDGSIIYVSDSITPLLGHLPSDVMDQNLLNFLPEQEHSEVYKILSSEYLKSDS
ELKHLILEAADGFLFIVSCETGRVVYVSDSVTPVLNQPQSEWFGSTLYDQVHPDDVDKLREQLSTSRMCM
sglab@sglab-V50t-13IMB:~/BE623_LAB_1$
```

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

```
sglab@sglab-V50t-13IMB: ~/BE623_LAB_1
sglab@sglab-V50t-13IMB:~/BE623_LAB_1$ grep "VV" sequence5.fasta
AANFREGLNLQEGEFLLQALNGFVLVVTTDALVFYASSTIQDYLGFQQSDVIHQSVYELIHTEDRAEFQR
IWLQTHYYITYHQWNSRPEFIVCTHTVVSYAEVRAE
TVIYNTKNSQPQCIVCVNYVVSGIIQHDL
QMDNLYLKALEGFIAVVTQDGDMIFLSENISKFMGLTQVELTGHSIFDFTHPCDHEEIRENLSSTERDFF
KFTYCDDRITELIGYHPEELLGRSAYEFYHALDSENMTKSHQNLCTKGQVVSGQYRMLAKHGGYVWLETQ
DRIAEVAGYSPDDLIGCSAYEYIHALDSDAVSKSIHTLLSKGQAVTGQYRFLARSGGYLWTQTATVVSG
QTHYYITYHQWNSKPEFIVCTHSVVSYADVRVE
DYVHPGDHVEMAEQLGMTLERSFFIRMKSTLTKRGVHIKSSGYKVIHITGRLRLRMGLVVVAHALPPPTI
ISESVLIYLGFERSELLCKSWYGLLHPEDLAHASAQHYRLLAESGDIQAEMVVRLQAKTGGWAWIYCLLY
EKSKNAARTRREKENSEFYELAKLLPLPSAITSQLDKASIIRLTTSYLKMRVVFPEGLGEAWGHSSRTSP
LDNVGRELGSHLLQTLDGFIFVVAPDGKIMYISETASVHLGLSQVELTGNSIYEYIHPADHDEMTAVLTA
LDGVAKELGSHLLQTLDGFVFVVASDGKIMYISETASVHLGLSQVELTGNSIYEYIHPSDHDEMTAVLTA
SYATVVHNSRSSRPHCIVSVNYVLTEIEYKEL
ELKHLILEAADGFLFIVSCETGRVVYVSDSVTPVLNQPQSEWFGSTLYDQVHPDDVDKLREQLSTSRMCM
GSRRSFICRMRCGSSEPHFVVVHCTGYIKAKFCLVAIGRLQVTSSPNCTDMSNVCQPTEFISRHNIEGIF
TFVDHRCVATVGYQPQELLGKNIVEFCHPEDQQLLRDSFQQVVKLKGQVLSVMFRFRSKNQEWLWMRTSS
DELKHLILRAADGFLFVVGCDRGKILFVSESVFKILNYSQNDLIGQSLFDYLHPKDIAKVKEQLSSSRLC
SGARRSFFCRMKCNRPRKSFCTIHSTGYLKSNLSCLVAIGRLHSHVVPQPVNGEIRVKSMEYVSRHAIDG
RWFSFMNPWTKEVEYIVSTNTVVL
sglab@sglab-V50t-13IMB:~/BE623_LAB_1$
```

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

```
sglab@sglab-V50t-13IMB: ~/BE623_LAB_1
sglab@sglab-V50t-13IMB:~/BE623_LAB_1$ grep -E "AA|DD" sequence5.fasta
AANFREGLNLQEGEFLLQALNGFVLVVTTDALVFYASSTIQDYLGFQQSDVIHQSVYELIHTEDRAEFQR
IFRTKHKLDFTPIGCDAKGRIVLGYTEAELCTRGSGYQFIHAADMLYCAESHIRMIKTGESGMIVFRLLT
RHSLEWKFLFLDHRAPPIIGYLPFEVLGTSGYDYYHVDDLENLAKCHEHLMQYGKGKSCYYRFLTKGQQW
KEKSRDAARSRRSKESEVFYELAHQLPLPHNVSSHLDKASVMRLTISYLRVRKLLDAGDLDIEDDMKAQM
NCFYLKALDGFVMVLTDDGDMIYISDNVNKYMGLTQFELTGHSVFDFTHPCDHEEMREMLTHNTQRSFFL
KEKSRDAARCRRSKETEVFYELAHELPLPHSVSSHLDKASIMRLAISFLRTHKLLSSVCSENESEAEADQ
KFTYCDDRITELIGYHPEELLGRSAYEFYHALDSENMTKSHQNLCTKGQVVSGQYRMLAKHGGYVWLETQ
DAARSRRSQETEVLYQLAHTLPFARGVSAHLDKASIMRLTISYLRMHRLCAAGEWNQVGAGGEPLDACYL
LTSRGRTLNLKAATWKVLNCSGHMRAYEPPLQCLVLICEAIPHPGSLEPPLGRGAFLSRHSLDMKFTYCD
DRIAEVAGYSPDDLIGCSAYEYIHALDSDAVSKSIHTLLSKGQAVTGQYRFLARSGGYLWTQTQATVVSG
KEKSRNAARSRRGKENLEFFELAKLLPLPGAISSQLDKASIVRLSVTYLRLRRFAALGAPPWGLRAAGPP
AGLAPGRRGPAALVSEVFEQHLGGHILQSLDGFVFALNQEGKFLYISETVSIYLGLSQVEMTGSSVFDYI
LEWKFLFLDHRAPPIIGYLPFEVLGTSGYDYYHIDDLELLARCHQHLMQFGKGKSCCYRFLTKGQQWIWL
SRDAARSRRGKENFEFYELAKLLPLPAAITSQLDKASIIRLTISYLKMRDFANQGDPPWNLRMEGPPPNT
IVAALPGFLLVFTAEGKLLYLSESVSEHLGHSMVDLVAQGDSIYDIIDPADHLTVRQQLTLTDRLFRCRF
EKSKNAARTRREKENSEFYELAKLLPLPSAITSQLDKASIIRLTTSYLKMRVVFPEGLGEAWGHSSRTSP
EKSKNAAKTRREKENGEFYELAKLLPLPSAITSQLDKASIIRLTTSYLKMRAVFPEGLGDAWGQPSRAGP
ELKHLILEAADGFLFIVSCETGRVVYVSDSVTPVLNQPQSEWFGSTLYDQVHPDDVDKLREQLSTSRMCM
DELKHLILRAADGFLFVVGCDRGKILFVSESVFKILNYSQNDLIGQSLFDYLHPKDIAKVKEQLSSSRLC
KFVFVDQRATAILAYLPQELLGTSCYEYFHQDDIGHLAECHRQVLQTREKITTNCYKFKIKDGSFITLRS
sglab@sglab-V50t-13IMB:~/BE623_LAB_1$
```

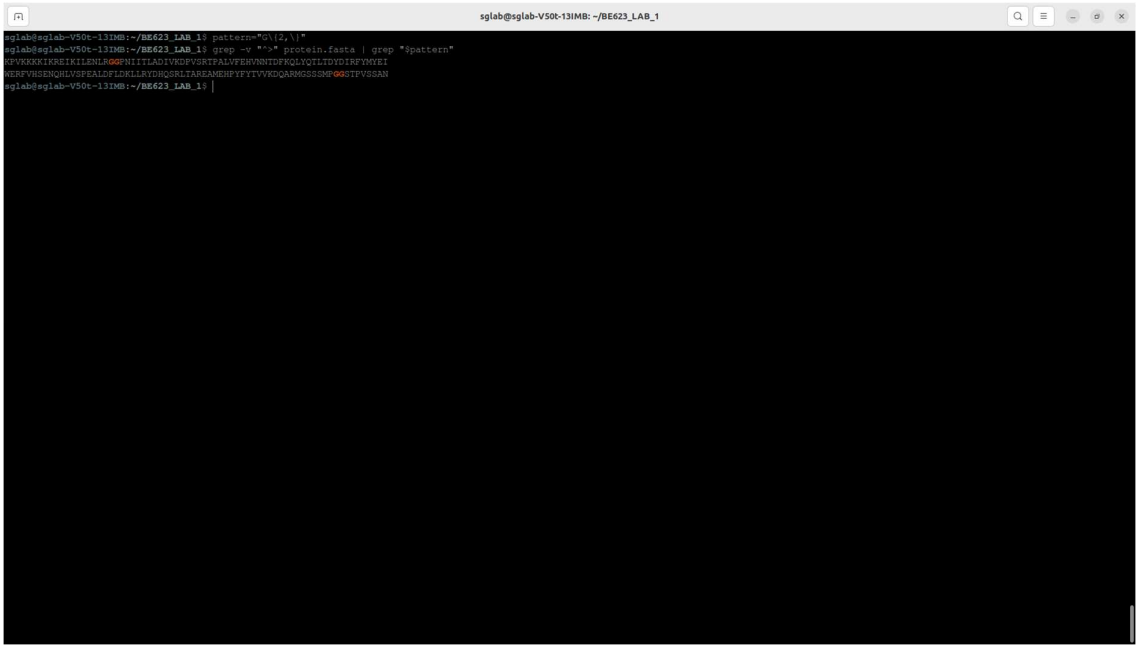
8. Print only the sequence lines (ignore headers) from sequence5.fasta that contain the letter P.

```
sglab@sglab-V50t-13IMB: ~/BE623_LAB_1
sglab@sglab-V50t-13IMB:~/BE623_LAB_1$ grep -v ">" sequence5.fasta | grep "P"
NSNPSKRHRDRLNTELDRLASLLPFPQDVINKLDKLSVLRLSVSYLRAKSFFDVALKSSPTERNGGQGNCR
QLHWQIPPENSPLMERCFICRLRCLLDNSSGFLAMNFQGKLKYLPPQLALFAIATPLQPPSILEIRTKNF
IFRTKHKLDFTPIGCDAKGRIVLGYTEAELCTRGSGYQFIHAADMLYCAESHIRMIKTGESGMIVFRLLT
KNNRWTWVQSNARLLYKNGRPDYIIVTQRPLTDEEGTEHLR
VSRNKSEKKRRDQFNVLIKELGSMLPGNARKMDKSTVLQKSIDFLRKHKEITAQSDASEIRQDWKPTFLS
NEEFTQLMLEALDGFFLAIMTDGSITYVSESVTSLLEHLPSDLVDQSIFNFIPEGEHSEVYKILSTEYLK
EKNQLEFCCHMLRGTIDPKEPGTYEVVKTIGNFKSLYEDRVCFVATVRLATPQFIKEMCTVEEPNEEFTS
RHSLEWKFLFLDHRAPPIIGYLPFEVLGTSGYDYYHVDDLENLAKCHEHLMQYGKGKSCYYRFLTKGQQW
IWLQTHYYITYHQWNSRPEFIVCTHTVVSYAEVRAE
KEKSRDAARSRRSKESEVFYELAHQLPLPHNVSSHLDKASVMRLTISYLRVRKLLDAGDLDIEDDMKAQM
WCFYLKALDGFVMVLTDDGDMIYISDNVNKYMGLTQFELTGHSVFDFTHPCDHEEMREMLTHNTQRSFFL
RMKCTLTSRGRTMNIKSATWKVLHCTGHIHVYKPPMTCLVLICEPIPHPSNIEIPLDSKTFLSRHSLDMK
FSYCDERITELMGYEPEELLGRSIYEYYHALDSDHLTKTHHDMFTKGQVTTGQYRMLAKRGGYVWVETQA
TVIYNTKNSQPQCIVCVNYVVSGIIQHDL
KEKSRDAARCRRGKETEVFYELAHELPLPHSVSSHLDKASIMRLAISFLRTHKLLSSGVCSENESEAEADQ
QMDNLYLKALEGFIAVVTQDGDMIFLSENISKFMGLTQVELTGHSIFDFTHPCDHEEIRENLSSTERDFF
MRMKCTVTNRGRTVNLKSATWKVLHCTGQVKVYEPLLSCLIIMCEPIQHPSHMDIPLDSKTFLSRHSMDM
KFTYCDDRITELIGYHPEELLGRSAYEFYHALDSENMTKSHQNLCTKGQVVSGQYRMLAKHGGYVWLETQ
GTVIYNPRNLQPQCIMCVNYVLSEIEKNDV
DAARSRRSQETEVLYQLAHTLPFARGVSAHLDKASIMRLTISYLRMHRLCAAGEWNQVGAGGEPLDACYL
KALEGFVMVLTAEGDMAYLSENVSKHLGLSQLELIGHSIFDFIHPCDQEELQDALTPPTERCFSLRMKST
LTSRGRTLNLKAATWKVLNCSGHMRAYEPPLQCLVLICEAIPHPGSLEPPLGRGAFLSRHSLDMKFTYCD
DRIAEVAGYSPDDLIGCSAYEYIHALDSDAVSKSIHTLLSKGQAVTGQYRFLARSGGYLWTQTQATVVSG
GRGPQSESIVCVHFLISQVEETGV
KEKSRNAARSRRGKENLEFFEIAKLLPLPGAISSQLDKASIVRLSVTYLRLRRFAALGAPPWGLRAAGPP
AGLAPGRRGPAALVSEVFEQHLGGHILQSLDGFVFALNQEGKFLYISETVSIYLGLSQVEMTGSSVFDYI
HPGDHSEVLEQLGLVQERSFFVRMKSTLTKRGLHVKASGYKVIHVTGRLRALGLVALGHTLPPAPLAELP
LHGHMIVFRLSLGLTILACESRVSDHMDLGPSELVGRSCYQFVHGQDATRIRQSHVDLLDKGQVMTGYYR
WLQRAGGFVWLQSVATVAGSGKSPGEHHVLWVSHVLSQAEGGQT
MKSEKKRRDQFNVLIKELSSMLPGNTRKMDKTTVLEKVIGFLQKHNEVSAQTEICDIQQDWKPSFLSNEE
FTQLMLEALDGFIIAVTTDGSITYVSDSITPILGHLPSDVMDQNLLNFLPEQEHSEVYKILSSEYLKSDS
DLEFYCHLLRGSLNPKEFPTYEYIKFVGNFRSYLGKEVCFIATVRLATPQFLKEMCIVDEPLEEFTSRHS
LEWKFLFLDHRAPPIIGYLPFEVLGTSGYDYYHIDDLELLARCHQHLMQFGKGKSCCYRFLTKGQQWIWL
QTHYYITYHQWNSKPEFIVCTHSVVSYADVRVE
SRDAARSRRGKENFEFYELAKLLPLPAAITSQLDKASIIRLTISYLKMRDFANQGDPPWNLRMEGPPPNT
SVKVIGAQRRRSPSALAIEVFEAHLGSHILQSLDGFVFALNQEGKFLYISETVSIYLGLSQVELTGSSVF
DYVHPGDHVEMAEQLGMTLERSFFIRMKSTLTKRGVHIKSSGYKVIHITGRLRLRMGLVVVAHALPPPTI
NEVRIDCRMSVTRVNMDLNIIYCENRISDYMDLTPVDIVGKRCYHFIHREDVEGIRNSHLDLLNMGQCVT
KYYRWMQKNGGYIWIQSSATIAINAKNANEKNIIWVNYLLSNPEYKDT
GASKARRDQINAEIRNLKELLPLAEADKVRLSYLHIMSLACIYTRKGVFFAGGTPLAGPTGLLSAQELED
IVAALPGFLLVFTAEGKLLYLSESVSEHLGHSMVDLVAQGDSIYDIIDPADHLTVRQQLTLTDRLFRCRF
NTSKSLRRQSAGNKLVLIRGRFHAHNPVFTAFCAPLEPRPRPGPGPGPGPASLFLAMFQSRHAKDLALLD
ISESVLIYLGFERSELLCKSWYGLLHPEDLAHASAQHYRLLAESGDIQAEMVVRLQAKTGGWAWIYCLLY
SEGPEGPITANNYPISDMEAWSLRQQL
EKSRNAARTRREKENSEFYELAKLLPLPSAITSQLDKASIIRLTTSYLKMRVVFPEGLGEAWGHSSRTSP
LDNVGRELGSHLLQTLDGFIFVVAPGKKMYISETASVRLGLSQVELTGNSYEYIHPADHEEMTAVLTA
EIERSFFLRMKCVLAKRNAGLTCGGYKVIHCSGYLKIRNVGLVAVGHSLPPSAVTEIKLHSNMFMFRASL
DMKLIFLDSRVAELTGYEPQDLIEKTLYHHVHGCDTFHLRCAHHLLLVKGQVTTKYYRFLAKHGGWVWVQ
GYATIVHNSRSSRPHCIVSVNYVLTDTEYKGL
```

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

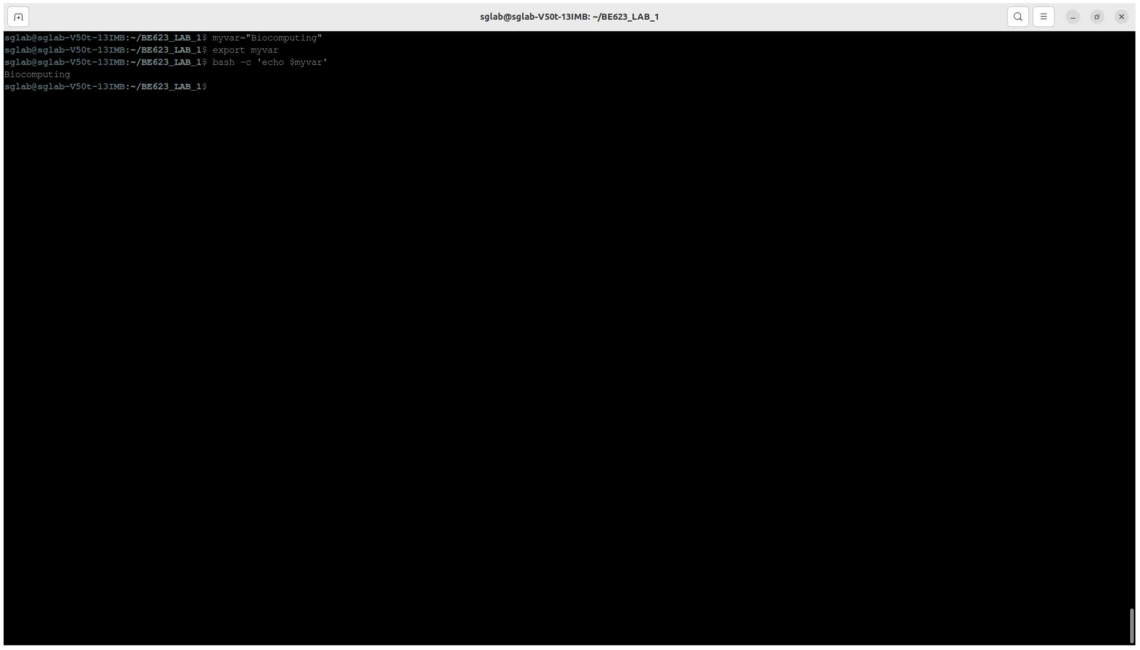
```
sglab@sglab-V50t-13IMB: ~/BE623_LAB_1
sglab@sglab-V50t-13IMB:~/BE623_LAB_1$ seq="sequence5.fasta"
sglab@sglab-V50t-13IMB:~/BE623_LAB_1$ echo "Number of sequences in $seq:"
Number of sequences in sequence5.fasta:
sglab@sglab-V50t-13IMB:~/BE623_LAB_1$ grep -c ">" $seq
13
sglab@sglab-V50t-13IMB:~/BE623_LAB_1$
```

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

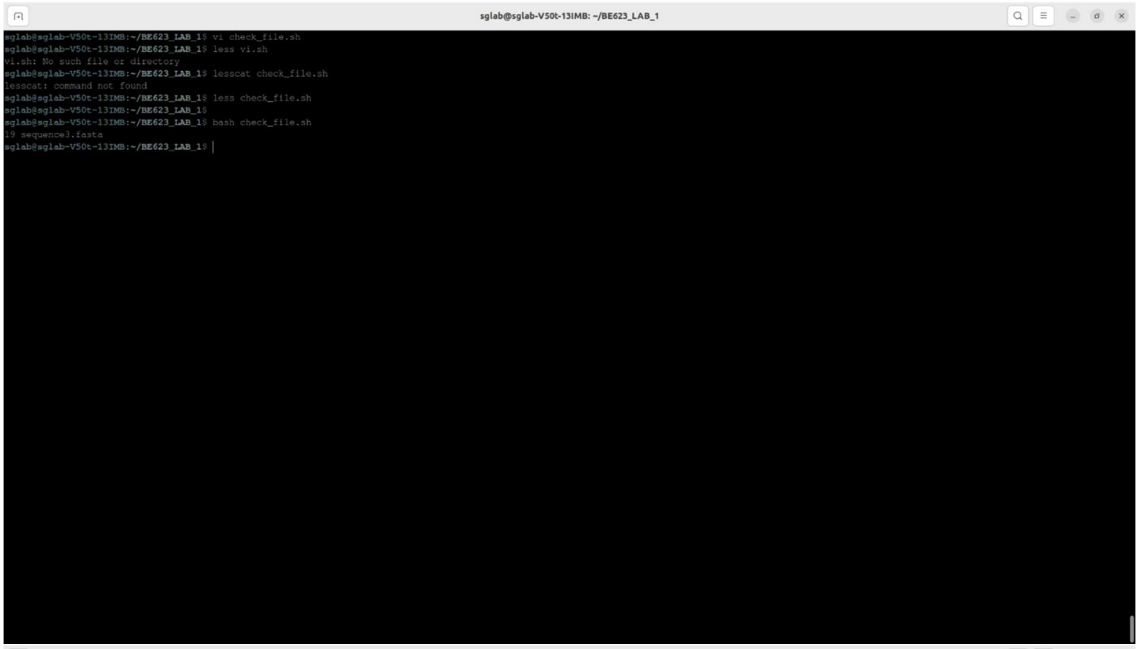


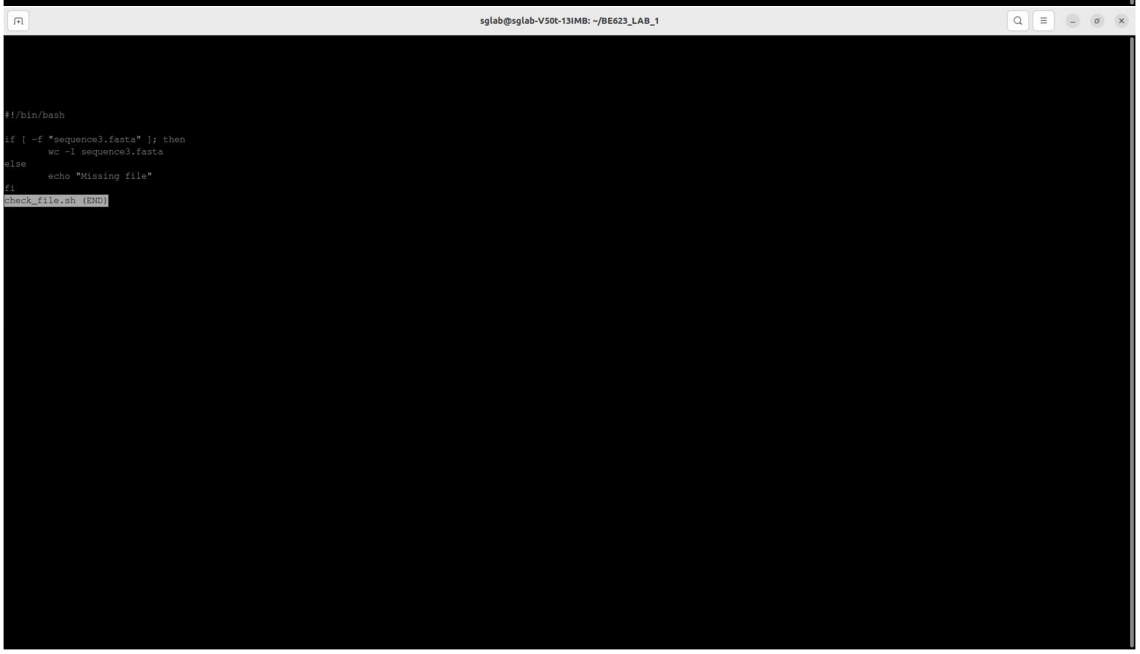
11. Store "Biocomputing" in a variable, export it, and verify that it is available inside a new shell started using:

bash -c 'echo $VARIABLE_NAME'

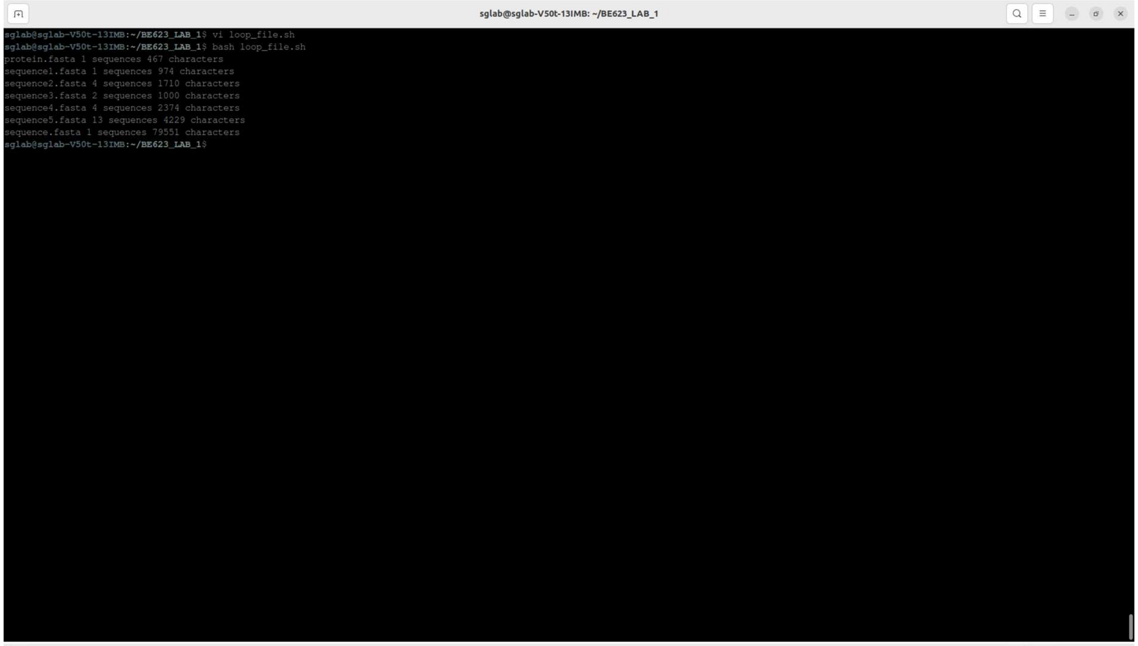


12. Write a shell script that checks if sequence3.fasta exists in the current folder. If yes, print the number of lines. If no, print "Missing file".

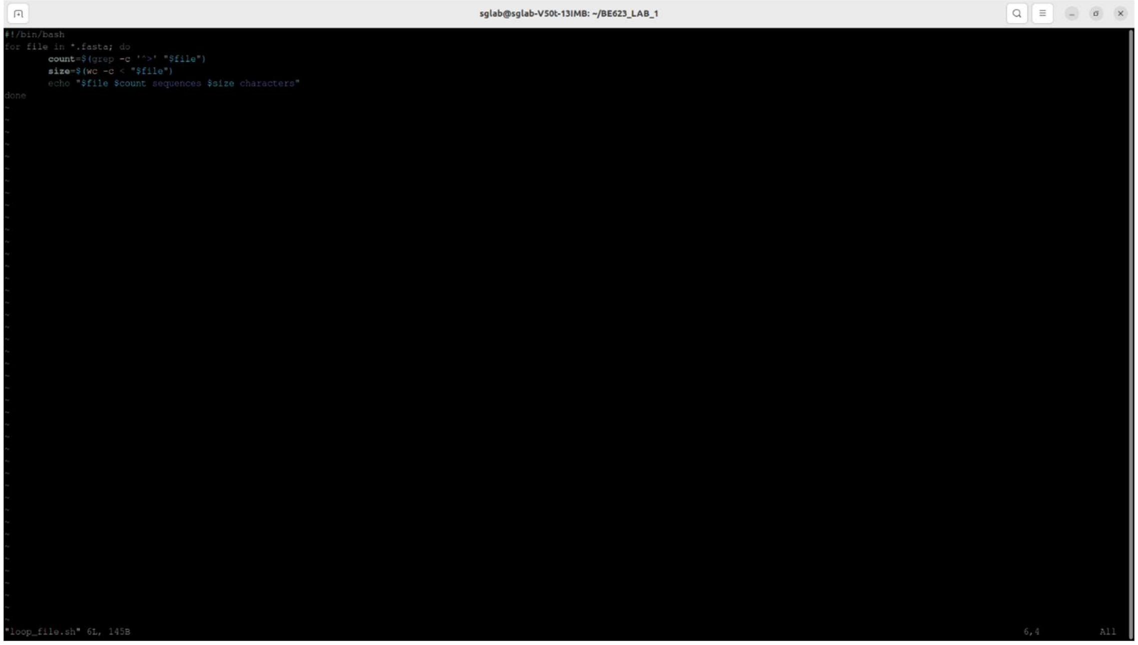




13. Using a for loop, go through all .fasta files in the current directory and print: filename, number of sequences, and file size in characters.



```
sglab@sglab-V50t-13IMB:~/BE623_LAB_1$ vi loop_file.sh
sglab@sglab-V50t-13IMB:~/BE623_LAB_1$ bash loop_file.sh
protein.fasta 1 sequences 467 characters
sequence1.fasta 1 sequences 974 characters
sequence2.fasta 4 sequences 1710 characters
sequence3.fasta 2 sequences 1000 characters
sequence4.fasta 4 sequences 2374 characters
sequence5.fasta 13 sequences 4229 characters
sequence.fasta 1 sequences 79551 characters
sglab@sglab-V50t-13IMB:~/BE623_LAB_1$
```



```
#!/bin/bash
for file in *.fasta; do
        count=$(grep -c '^>' "$file")
        size=$(wc -c < "$file")
        echo "$file $count sequences $size characters"
done
"loop_file.sh" 6L, 145B                                                    6,4           All
```

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

```
sglab@sglab-V50t-13IMB: ~/BE623_LAB_1
sglab@sglab-V50t-13IMB:~/BE623_LAB_1$ vi loop_modified.sh
sglab@sglab-V50t-13IMB:~/BE623_LAB_1$ bash loop_modified
bash: loop_modified: No such file or directory
sglab@sglab-V50t-13IMB:~/BE623_LAB_1$ bash loop_modified.sh
sequence2.fasta 4 sequences 1710 characters
sequence4.fasta 4 sequences 2374 characters
sequence5.fasta 13 sequences 4229 characters
sglab@sglab-V50t-13IMB:~/BE623_LAB_1$
```

```
sglab@sglab-V50t-13IMB: ~/BE623_LAB_1
#!/bin/bash
for file in *.fasta; do
        count=$(grep -c '^>' "$file")
        if [ "$count" -gt 3 ]; then
                size=$(wc -c < "$file")
                echo "$file $count sequences $size characters"
        fi
done
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
"loop_modified.sh" 9L, 181B                                    6,2-16        All
```

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.