BE623 Biocomputing | Lab assignment 2 | Anchal Kaushik | 24410001

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.

Output:

```
anchal@hp:/mnt/c/Users/kaush$ vi notes.txt
anchal@hp:/mnt/c/Users/kaush$ cat notes.txt
Have a nice dayanchal@hp:/mnt/c/Users/kaush$ wc notes.txt
 0  4 15 notes.txt
anchal@hp:/mnt/c/Users/kaush$ wc -m notes.txt
15 notes.txt
anchal@hp:/mnt/c/Users/kaush$ wc -l notes.txt
0 notes.txt
anchal@hp:/mnt/c/Users/kaush$
```

Note: The file notes.txt contains the text *"Have a nice day"*, which is one visible line of text.

- **wc -m notes.txt showed 15 characters.**
- **wc -l notes.txt showed 0 lines, because the file was saved without a newline (\n) at the end.**
- **I used :set binary, :set noeol commands to remove new line and to get exactly 15 characters.**

Part 2 - Pattern Matching in FASTA Files

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

Output:

```
anchal@hp:/mnt/c/Users/kaush$ tail -n 4 sequence.fasta
TAACTACTGATAAGTTACAAAACTGTTTTCTATCCTAAAGGGCAATACAGCCCTAGACTCTCCCAGGTAT
TTGACTCCTGCAGCAAAAAGGGAAATTGAGGAAATAGAGCAAGCTATTTCTCAGAGGCAACTATATCACA
TAGACACCCCG
```

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

Output:

```
anchal@hp:/mnt/c/Users/kaush$ grep "^>" sequence5.fasta
>ahr
>clock
>hif1a
>hif2a
>hif3a
>npas1
>npas2
>npas3
>npas4
>sim1
>sim2
>arnt1
>bmal1
anchal@hp:/mnt/c/Users/kaush$
```

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

Output:

```
anchal@hp:/mnt/c/Users/kaush$ grep -E "A.G" sequence5.fasta
IFRTKHKLDFTPIGCDAKGRIVLGYTEAELCTRGSGYQFIHAADMLYCAESHIRMIKTGESGMIVFRLLT
DAARSRRSQETEVLYQLAHTLPFARGVSAHLDKASIMRLTISYLRMHRLCAAGEWNQVGAGGEPLDACYL
KALEGFVMVLTAEGDMAYLSENVSKHLGLSQLELIGHSIFDFIHPCDQEELQDALTPPTERCFSLRMKST
KEKSRNAARSRRGKENLEFFELAKLLPLPGAISSQLDKASIVRLSVTYLRLRRFAALGAPPWGLRAAGPP
AGLAPGRRGPAALVSEVFEQHLGGHILQSLDGFVFALNQEGKFLYISETVSIYLGLSQVEMTGSSVFDYI
HPGDHSEVLEQLGLVQERSFFVRMKSTLTKRGLHVKASGYKVIHVTGRLRALGLVALGHTLPPAPLAELP
WLQRAGGFVWLQSVATVAGSGKSPGEHHVLWVSHVLSQAEGGQT
GASKARRDQINAEIRNLKELLPLAEADKVRLSYLHIMSLACIYTRKGVFFAGGTPLAGPTGLLSAQELED
IVAALPGFLLVFTAEGKLLYLSESVSEHLGHSMVDLVAQGDSIYDIIDPADHLTVRQQLTLTDRLFRCRF
EKSKNAARTRREKENSEFYELAKLLPLPSAITSQLDKASIIRLTTSYLKMRVVFPEGLGEAWGHSSRTSP
EIERSFFLRMKCVLAKRNAGLTCGGYKVIHCSGYLKIRNVGLVAVGHSLPPSAVTEIKLHSNMFMFRASL
EKSKNAAKTRREKENGEFYELAKLLPLPSAITSQLDKASIIRLTTSYLKMRAVFPEGLGDAWGQPSRAGP
EIERSFFLRMKCVLAKRNAGLTCSGYKVIHCSGYLKIRIVGLVAVGQSLPPSAITEIKLYSNMFMFRASL
ELKHLILEAADGFLFIVSCETGRVVYVSDSVTPVLNQPQSEWFGSTLYDQVHPDDVDKLREQLSTSRMCM
GSRRSFICRMRCGSSEPHFVVVHCTGYIKAKFCLVAIGRLQVTSSPNCTDMSNVCQPTEFISRHNIEGIF
DELKHLILRAADGFLFVVGCDRGKILFVSESVFKILNYSQNDLIGQSLFDYLHPKDIAKVKEQLSSSRLC
SGARRSFFCRMKCNRPRKSFCTIHSTGYLKSNLSCLVAIGRLHSHVVPQPVNGEIRVKSMEYVSRHAIDG
anchal@hp:/mnt/c/Users/kaush$
```

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

Output:

```
anchal@hp:/mnt/c/Users/kaush$ grep -E "P[^A]L" sequence5.fasta
QLHWQIPPENSPLMERCFICRLRCLLDNSSGFLAMNFQGKLKYLPPQLALFAIATPLQPPSILEIRTKNF
MRMKCTVTNRGRTVNLKSATWKVLHCTGQVKVYEPLLSCLIIMCEPIQHPSHMDIPLDSKTFLSRHSMDM
LTSRGRTLNLKAATWKVLNCSGHMRAYEPPLQCLVLICEAIPHPGSLEPPLGRGAFLSRHSLDMKFTYCD
FTQLMLEALDGFIIAVTTDGSIIYVSDSITPLLGHLPSDVMDQNLLNFLPEQEHSEVYKILSSEYLKSDS
ELKHLILEAADGFLFIVSCETGRVVYVSDSVTPVLNQPQSEWFGSTLYDQVHPDDVDKLREQLSTSRMCM
anchal@hp:/mnt/c/Users/kaush$
```

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

Output:

```
anchal@hp:/mnt/c/Users/kaush$ grep -E "VV[^V]" sequence5.fasta
AANFREGLNLQEGEFLLQALNGFVLVVTTDALVFYASSTIQDYLGFQQSDVIHQSVYELIHTEDRAEFQR
IWLQTHYYITYHQWNSRPEFIVCTHTVVSYAEVRAE
TVIYNTKNSQPQCIVCVNYVVSGIIQHDL
QMDNLYLKALEGFIAVVTQDGDMIFLSENISKFMGLTQVELTGHSIFDFTHPCDHEEIRENLSSTERDFF
KFTYCDDRITELIGYHPEELLGRSAYEFYHALDSENMTKSHQNLCTKGQVVSGQYRMLAKHGGYVWLETQ
DRIAEVAGYSPDDLIGCSAYEYIHALDSDAVSKSIHTLLSKGQAVTGQYRFLARSGGYLWTQTQATVVSG
QTHYYITYHQWNSKPEFIVCTHSVVSYADVRVE
DYVHPGDHVEMAEQLGMTLERSFFIRMKSTLTKRGVHIKSSGYKVIHITGRLRLRMGLVVVAHALPPPTI
ISESVLIYLGFERSELLCKSWYGLLHPEDLAHASAQHYRLLAESGDIQAEMVVRLQAKTGGWAWIYCLLY
EKSKNAARTRREKENSEFYELAKLLPLPSAITSQLDKASIIRLTTSYLKMRVVFPEGLGEAWGHSSRTSP
LDNVGRELGSHLLQTLDGFIFVVAPDGKIMYISETASVHLGLSQVELTGNSIYEYIHPADHDEMTAVLTA
LDGVAKELGSHLLQTLDGFVFVVASDGKIMYISETASVHLGLSQVELTGNSIYEYIHPSDHDEMTAVLTA
SYATVVHNSRSSRPHCIVSVNYVLTEIEYKEL
ELKHLILEAADGFLFIVSCETGRVVYVSDSVTPVLNQPQSEWFGSTLYDQVHPDDVDKLREQLSTSRMCM
GSRRSFICRMRCGSSEPHFVVVHCTGYIKAKFCLVAIGRLQVTSSPNCTDMSNVCQPTEFISRHNIEGIF
TFVDHRCVATVGYQPQELLGKNIVEFCHPEDQQLLRDSFQQVVKLKGQVLSVMFRFRSKNQEWLWMRTSS
DELKHLILRAADGFLFVVGCDRGKILFVSESVFKILNYSQNDLIGQSLFDYLHPKDIAKVKEQLSSSRLC
SGARRSFFCRMKCNRPRKSFCTIHSTGYLKSNLSCLVAIGRLHSHVVPQPVNGEIRVKSMEYVSRHAIDG
RWFSFMNPWTKEVEYIVSTNTVVL
anchal@hp:/mnt/c/Users/kaush$
```

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

Output:

```
anchal@hp:/mnt/c/Users/kaush$ 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
anchal@hp:/mnt/c/Users/kaush$
```

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

Output:

```
anchal@hp:/mnt/c/Users/kaush$ grep -v "^>" sequence5.fasta | grep "P"
SNPSKRHRDRLNTELDRLASLLPFPQDVINKLDKLSVLRLSVSYLRAKSFFDVALKSSPTERNGGQDNCR
QLHWQIPPENSPLMERCFICRLRCLLDNSSGFLAMNFQGKLKYLPPQLALFAIATPLQPPSILEIRTKNF
IFRTKHKLDFTPIGCDAKGRIVLGYTEAELCTRGSGYQFIHAADMLYCAESHIRMIKTGESGMIVFRLLT
KNNRWTWVQSNARLLYKNGRPDYIIVTQRPLTDEEGTEHLR
VSRNKSEKKRRDQFNVLIKELGSMLPGNARKMDKSTVLQKSIDFLRKHKEITAQSDASEIRQDWKPTFLS
NEEFTQLMLEALDGFFLAIMTDGSIIYVSESVTSLLEHLPSDLVDQSIFNFIPEGEHSEVYKILSTEYLK
SKNQLEFCCHMLRGTIDPKEPSTYEYVKFIGNFKSLYEDRVCFVATVRLATPQFIKEMCTVEEPNEEFTS
RHSLEWKFLFLDHRAPPIIGYLPFEVLGTSGYDYYHVDDLENLAKCHEHLMQYGKGKSCYYRFLTKGQQW
IWLQTHYYITYHQWNSRPEFIVCTHTVVSYAEVRAE
KEKSRDAARSRRSKESEVFYELAHQLPLPHNVSSHLDKASVMRLTISYLRVRKLLDAGDLDIEDDMKAQM
NCFYLKALDGFVMVLTDDGDMIYISDNVNKYMGLTQFELTGHSVFDFTHPCDHEEMREMLTHNTQRSFFL
RMKCTLTSRGRTMNIKSATWKVLHCTGHIHVYKPPMTCLVLICEPIPHPSNIEIPLDSKTFLSRHSLDMK
FSYCDERITELMGYEPEELLGRSIYEYYHALDSDHLTKTHHDMFTKGQVTTGQYRMLAKRGGYVWVETQA
TVIYNTKNSQPQCIVCVNYVVSGIIQHDL
KEKSRDAARCRRSKETEVFYELAHELPLPHSVSSHLDKASIMRLAISFLRTHKLLSSVCSENESEAEADQ
QMDNLYLKALEGFIAVVTQDGDMIFLSENISKFMGLTQVELTGHSIFDFTHPCDHEEIRENLSSTERDFF
MRMKCTVTNRGRTVNLKSATWKVLHCTGQVKVYEPLLSCLIIMCEPIQHPSHMDIPLDSKTFLSRHSMDM
KFTYCDDRITELIGYHPEELLGRSAYEFYHALDSENMTKSHQNLCTKGQVVSGQYRMLAKHGGYVWLETQ
GTVIYNPRNLQPQCIMCVNYVLSEIEKNDV
DAARSRRSQETEVLYQLAHTLPFARGVSAHLDKASIMRLTISYLRMHRLCAAGEWNQVGAGGEPLDACYL
KALEGFVMVLTAEGDMAYLSENVSKHLGLSQLELIGHSIFDFIHPCDQEELQDALTPPTERCFSLRMKST
LTSRGRTLNLKAATWKVLNCSGHMRAYEPPLQCLVLICEAIPHPGSLEPPLGRGAFLSRHSLDMKFTYCD
DRIAEVAGYSPDDLIGCSAYEYIHALDSDAVSKSIHTLLSKGQAVTGQYRFLARSGGYLWTQTQATVVSG
GRGPQSESIVCVHFLISQVEETGV
KEKSRNAARSRRGKENLEFFELAKLLPLPGAISSQLDKASIVRLSVTYLRLRRFAALGAPPWGLRAAGPP
AGLAPGRRGPAALVSEVFEQHLGGHILQSLDGFVFALNQEGKFLYISETVSIYLGLSQVEMTGSSVFDYI
HPGDHSEVLEQLGLVQERSFFVRMKSTLTKRGLHVKASGYKVIHVTGRLRALGLVALGHTLPPAPLAELP
LHGHMIVFRLSLGLTILACESRVSDHMDLGPSELVGRSCYQFVHGQDATRIRQSHVDLLDKGQVMTGYYR
WLQRAGGFVWLQSVATVAGSGKSPGEHHVLWVSHVLSQAEGGQT
NKSEKKRRDQFNVLIKELSSMLPGNTRKMDKTTVLEKVIGFLQKHNEVSAQTEICDIQQDWKPSFLSNEE
FTQLMLEALDGFIIAVTTDGSIIYVSDSITPLLGHLPSDVMDQNLLNFLPEQEHSEVYKILSSEYLKSDS
DLEFYCHLLRGSLNPKEFPTYEYIKFVGNFRSYLGKEVCFIATVRLATPQFLKEMCIVDEPLEEFTSRHS
LEWKFLFLDHRAPPIIGYLPFEVLGTSGYDYYHIDDLELLARCHQHLMQFGKGKSCCYRFLTKGQQWIWL
QTHYYITYHQWNSKPEFIVCTHSVVSYADVRVE
SRDAARSRRGKENFEFYELAKLLPLPAAITSQLDKASIIRLTISYLKMRDFANQGDPPWNLRMEGPPPNT
SVKVIGAQRRRSPSALAIEVFEAHLGSHILQSLDGFVFALNQEGKFLYISETVSIYLGLSQVELTGSSVF
DYVHPGDHVEMAEQLGMTLERSFFIRMKSTLTKRGVHIKSSGYKVIHITGRLRLRMGLVVVAHALPPPTI
NEVRIDCHMFVTRVNMDLNIIYCENRISDYMDLTPVDIVGKRCYHFIHAEDVEGIRHSHLDLLNKGQCVT
KYYRWMQKNGGYIWIQSSATIAINAKNANEKNIIWVNYLLSNPEYKDT
GASKARRDQINAEIRNLKELLPLAEADKVRLSYLHIMSLACIYTRKGVFFAGGTPLAGPTGLLSAQELED
```

```
IVAALPGFLLVFTAEGKLLYLSESVSEHLGHSMVDLVAQGDSIYDIIDPADHLTVRQQLTLTDRLFRCRF
NTSKSLRRQSAGNKLVLIRGRFHAHNPVFTAFCAPLEPRPRPGPGPGPGPASLFLAMFQSRHAKDLALLD
ISESVLIYLGFERSELLCKSWYGLLHPEDLAHASAQHYRLLAESGDIQAEMVVRLQAKTGGWAWIYCLLY
SEGPEGPITANNYPISDMEAWSLRQQL
EKSKNAARTRREKENSEFYELAKLLPLPSAITSQLDKASIIRLTTSYLKMRVVFPEGLGEAWGHSSRTSP
LDNVGRELGSHLLQTLDGFIFVVAPDGKIMYISETASVHLGLSQVELTGNSIYEYIHPADHDEMTAVLTA
EIERSFFLRMKCVLAKRNAGLTCGGYKVIHCSGYLKIRNVGLVAVGHSLPPSAVTEIKLHSNMFMFRASL
DMKLIFLDSRVAELTGYEPQDLIEKTLYHHVHGCDTFHLRCAHHLLLVKGQVTTKYYRFLAKHGGWVWVQ
SYATIVHNSRSSRPHCIVSVNYVLTDTEYKGL
EKSKNAAKTRREKENGEFYELAKLLPLPSAITSQLDKASIIRLTTSYLKMRAVFPEGLGDAWGQPSRAGP
LDGVAKELGSHLLQTLDGFVFVVASDGKIMYISETASVHLGLSQVELTGNSIYEYIHPSDHDEMTAVLTA
EIERSFFLRMKCVLAKRNAGLTCSGYKVIHCSGYLKIRIVGLVAVGQSLPPSAITEIKLYSNMFMFRASL
DLKLIFLDSRVTEVTGYEPQDLIEKTLYHHVHGCDVFHLRYAHHLLLVKGQVTTKYYRLLSKRGGWVWVQ
SYATVVHNSRSSRPHCIVSVNYVLTEIEYKEL
NHSEIERRRRNKMTAYITELSDMVPTCSALARKPDKLTILRMAVSHMKSLRGTGNTSTDGSYKPSFLTDQ
ELKHLILEAADGFLFIVSCETGRVVYVSDSVTPVLNQPQSEWFGSTLYDQVHPDDVDKLREQLSTSRMCM
GSRRSFICRMRCGSSEPHFVVVHCTGYIKAKFCLVAIGRLQVTSSPNCTDMSNVCQPTEFISRHNIEGIF
TFVDHRCVATVGYQPQELLGKNIVEFCHPEDQQLLRDSFQQVVKLKGQVLSVMFRFRSKNQEWLWMRTSS
FTFQNPYSDEIEYIICTNTNVK
EAHSQIEKRRRDKMNSFIDELASLVPTCNAMSRKLDKLTVLRMAVQHMKTLRGATNPYTEANYKPTFLSD
DELKHLILRAADGFLFVVGCDRGKILFVSESVFKILNYSQNDLIGQSLFDYLHPKDIAKVKEQLSSSRLC
SGARRSFFCRMKCNRPRKSFCTIHSTGYLKSNLSCLVAIGRLHSHVVPQPVNGEIRVKSMEYVSRHAIDG
KFVFVDQRATAILAYLPQELLGTSCYEYFHQDDIGHLAECHRQVLQTREKITTNCYKFKIKDGSFITLRS
RWFSFMNPWTKEVEYIVSTNTVVL
anchal@hp:/mnt/c/Users/kaush$
```

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)

Output:

```
anchal@hp:/mnt/c/Users/kaush$ seq="sequence5.fasta"
anchal@hp:/mnt/c/Users/kaush$ grep -c "^>" "$seq"
13
anchal@hp:/mnt/c/Users/kaush$
```

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

Output:

```
anchal@hp:/mnt/c/Users/kaush$ test="G\{2,\}"
anchal@hp:/mnt/c/Users/kaush$ grep -v "^>" protein.fasta | grep "$test"
KPVKKKKIKREIKILENLRGGPNIITLADIVKDPVSRTPALVFEHVNNTDFKQLYQTLTDYDIRFYMYEI
WERFVHSENQHLVSPEALDFLDKLLRYDHQSRLTAREAMEHPYFYTVVKDQARMGSSSMPGGSTPVSSAN
anchal@hp:/mnt/c/Users/kaush$
```

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'

Output:

```
anchal@hp:/mnt/c/Users/kaush$ book="Biocomputing"
anchal@hp:/mnt/c/Users/kaush$ export book
anchal@hp:/mnt/c/Users/kaush$ bash -c 'echo $book'
Biocomputing
anchal@hp:/mnt/c/Users/kaush$
```

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"

Output:

```
anchal@hp:/mnt/c/Users/kaush$ #!/bin/bash
if [ -f sequence3.fasta ]; then
wc -l sequence3.fasta
else
echo "Missing file"
fi
19 sequence3.fasta
anchal@hp:/mnt/c/Users/kaush$
```

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

Output:

```
anchal@hp:/mnt/c/Users/kaush$ for file in *.fasta; do
    echo "Filename: $file"
    echo "Number of sequences: $(grep -c ">" "$file")"
    echo "File size (characters): $(wc -c < "$file")"
    echo "------------------------"
done
Filename: protein.fasta
Number of sequences: 1
File size (characters): 467
------------------------
Filename: sequence.fasta
Number of sequences: 1
File size (characters): 79551
------------------------
Filename: sequence1.fasta
Number of sequences: 1
File size (characters): 974
------------------------
Filename: sequence2.fasta
Number of sequences: 4
File size (characters): 1710
------------------------
Filename: sequence3.fasta
Number of sequences: 2
File size (characters): 1000
------------------------
Filename: sequence4.fasta
Number of sequences: 4
File size (characters): 2374
------------------------
Filename: sequence5.fasta
Number of sequences: 13
File size (characters): 4229
------------------------
anchal@hp:/mnt/c/Users/kaush$
```

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

Output:

```
anchal@hp:/mnt/c/Users/kaush$ for file in *.fasta; do
    seq_count=$(grep -c ">" "$file")
    if [ "$seq_count" -gt 3 ]; then
        echo "Processing: $file"
        echo "Sequences: $seq_count"
        echo "File size (characters): $(wc -c < "$file")"
        echo "------------------------"
    fi
done
Processing: sequence2.fasta
Sequences: 4
File size (characters): 1710
------------------------
Processing: sequence4.fasta
Sequences: 4
File size (characters): 2374
------------------------
Processing: sequence5.fasta
Sequences: 13
File size (characters): 4229
------------------------
anchal@hp:/mnt/c/Users/kaush$
```

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

Output:

```
anchal@hp:/mnt/c/Users/kaush$ grep -v "^>" sequence5.fasta | grep -v "^$" | grep -E "(C.*){3,}" > cys_rich.txt
anchal@hp:/mnt/c/Users/kaush$ cat cys_rich.txt
QLHWQIPPENSPLMERCFICRLRCLLDNSSGFLAMNFQGKLKYLPPQLALFAIATPLQPPSILEIRTKNF
IFRTKHKLDFTPIGCDAKGRIVLGYTEAELCTRGSGYQFIHAADMLYCAESHIRMIKTGESGMIVFRLLT
SKNQLEFCCHMLRGTIDPKEPSTYEYVKFIGNFKSLYEDRVCFVATVRLATPQFIKEMCTVEEPNEEFTS
RMKCTLTSRGRTMNIKSATWKVLHCTGHIHVYKPPMTCLVLICEPIPHPSNIEIPLDSKTFLSRHSLDMK
MRMKCTVTNRGRTVNLKSATWKVLHCTGQVKVYEPLLSCLIIMCEPIQHPSHMDIPLDSKTFLSRHSMDM
LTSRGRTLNLKAATWKVLNCSGHMRAYEPPLQCLVLICEAIPHPGSLEPPLGRGAFLSRHSLDMKFTYCD
DLEFYCHLLRGSLNPKEFPTYEYIKFVGNFRSYLGKEVCFIATVRLATPQFLKEMCIVDEPLEEFTSRHS
LEWKFLFLDHRAPPIIGYLPFEVLGTSGYDYYHIDDLELLARCHQHLMQFGKGKSCCYRFLTKGQQWIWL
NEVRIDCHMFVTRVNMDLNIIYCENRISDYMDLTPVDIVGKRCYHFIHAEDVEGIRHSHLDLLNKGQCVT
EIERSFFLRMKCVLAKRNAGLTCGGYKVIHCSGYLKIRNVGLVAVGHSLPPSAVTEIKLHSNMFMFRASL
EIERSFFLRMKCVLAKRNAGLTCSGYKVIHCSGYLKIRIVGLVAVGQSLPPSAITEIKLYSNMFMFRASL
GSRRSFICRMRCGSSEPHFVVVHCTGYIKAKFCLVAIGRLQVTSSPNCTDMSNVCQPTEFISRHNIEGIF
SGARRSFFCRMKCNRPRKSFCTIHSTGYLKSNLSCLVAIGRLHSHVVPQPVNGEIRVKSMEYVSRHAIDG
KFVFVDQRATAILAYLPQELLGTSCYEYFHQDDIGHLAECHRQVLQTREKITTNCYKFKIKDGSFITLRS
anchal@hp:/mnt/c/Users/kaush$
```

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

Output:

```
anchal@hp:/mnt/c/Users/kaush$ grep -c "^>" *.fasta | sort -t: -k2,2nr | head
 -n 1 | awk -F: '{print $1 " has " $2 " sequences"}'
sequence5.fasta has 13 sequences
anchal@hp:/mnt/c/Users/kaush$
```