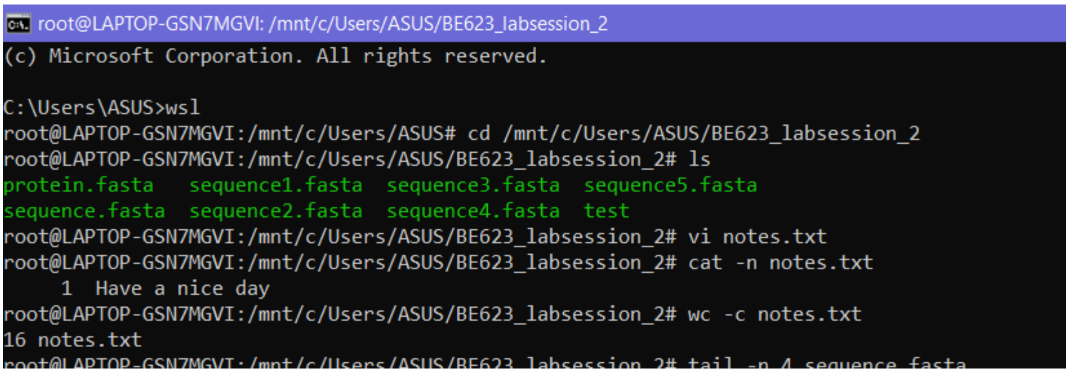
BE623 Biocomputing

Lab Assignment 2

Part 1 – vi Basics and file editing

1.

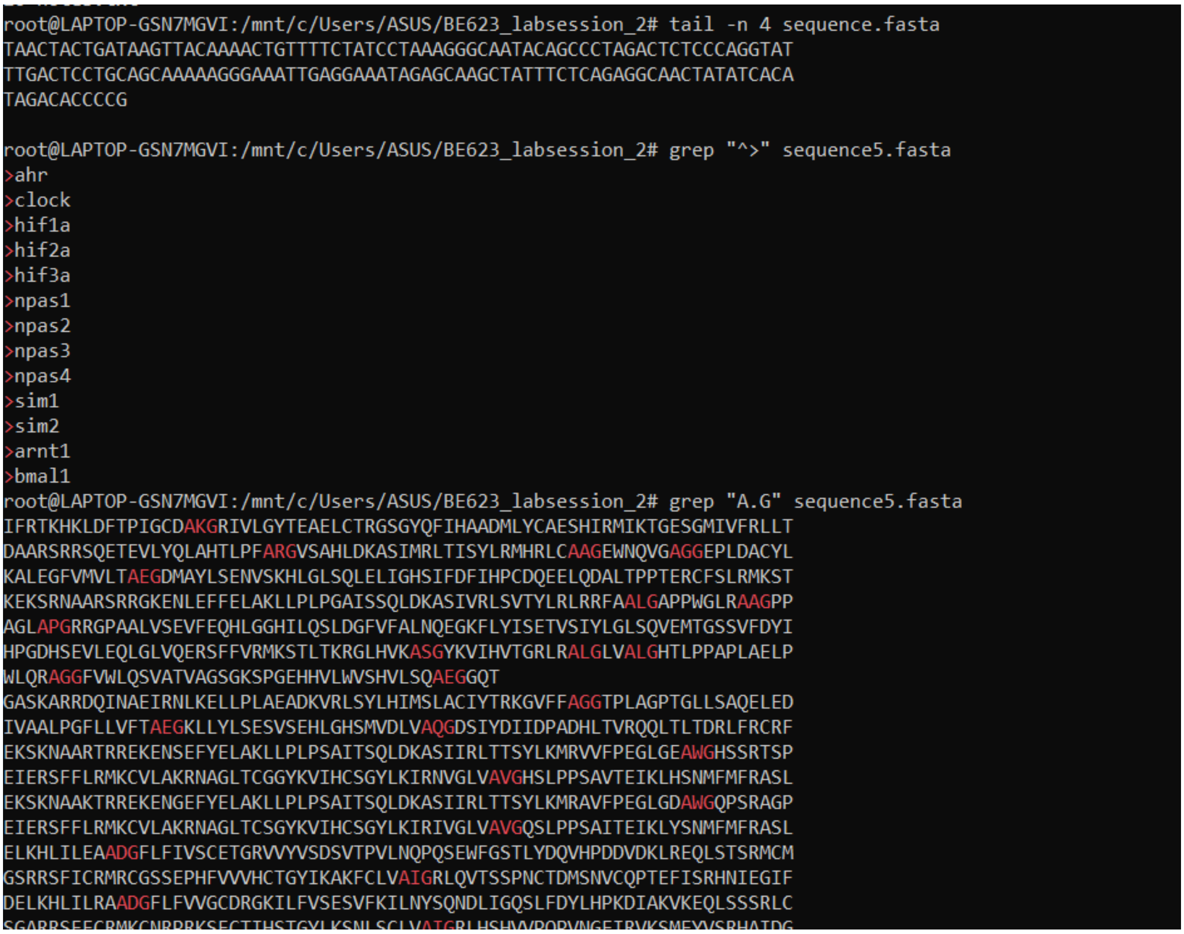


root@LAPTOP-GSN7MGVI: /mnt/c/Users/ASUS/BE623_labsession_2

```
(c) Microsoft Corporation. All rights reserved.

C:\Users\ASUS>wsl
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS# cd /mnt/c/Users/ASUS/BE623_labsession_2
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# ls
protein.fasta   sequence1.fasta  sequence3.fasta  sequence5.fasta
sequence.fasta  sequence2.fasta  sequence4.fasta  test
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# vi notes.txt
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# cat -n notes.txt
     1  Have a nice day
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# wc -c notes.txt
16 notes.txt
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# tail -n 4 sequence.fasta
```

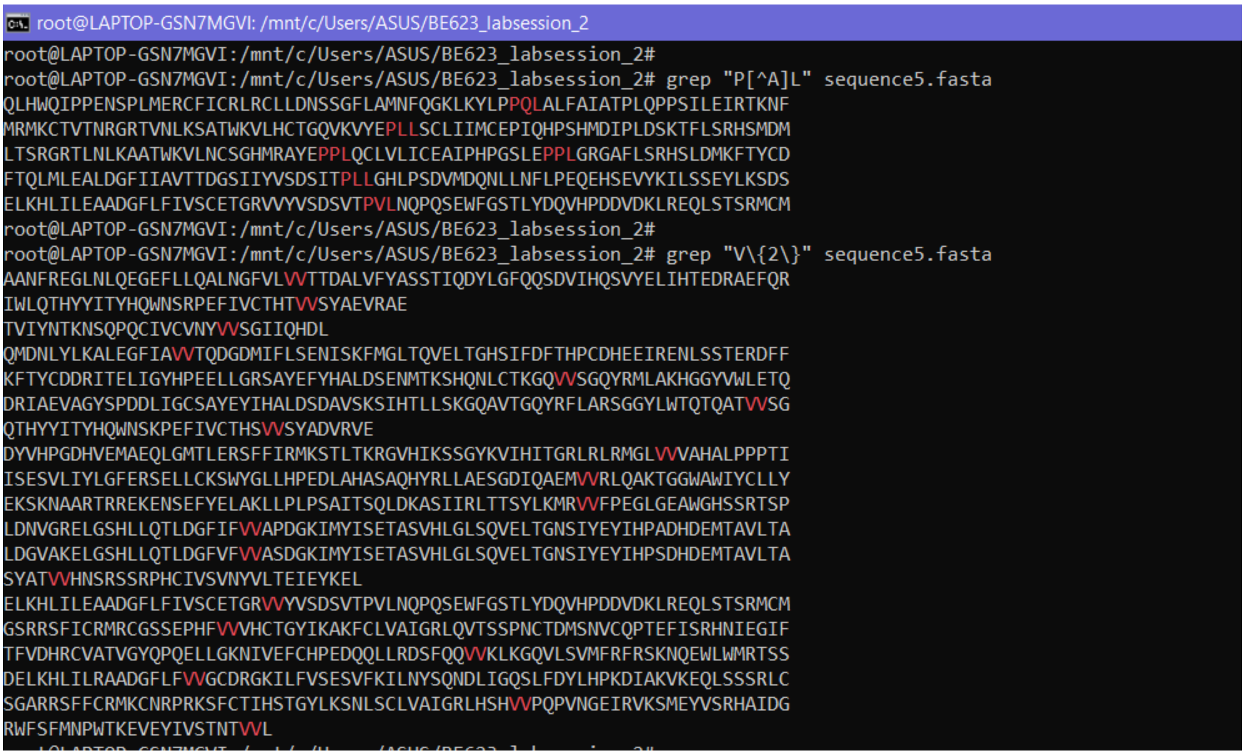
Part 2 – Pattern Matching in FASTA Files

2, 3 and 4



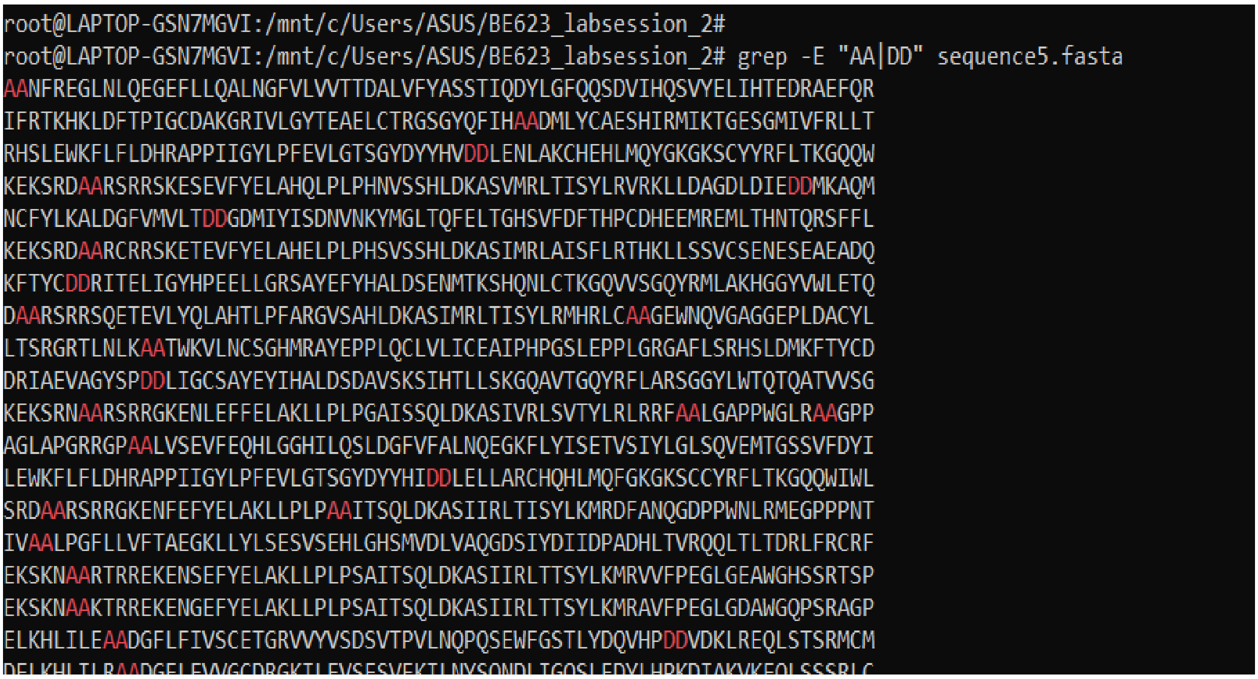
```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# tail -n 4 sequence.fasta
TAACTACTGATAAGTTACAAAACTGTTTTCTATCCTAAAGGGCAATACAGCCCTAGACTCTCCCAGGTAT
TTGACTCCTGCAGCAAAAAGGGAAATTGAGGAAATAGAGCAAGCTATTTCTCAGAGGCAACTATATCACA
TAGACACCCCG

root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# grep "^>" sequence5.fasta
>ahr
>clock
>hif1a
>hif2a
>hif3a
>npas1
>npas2
>npas3
>npas4
>sim1
>sim2
>arnt1
>bmal1
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# grep "A.G" sequence5.fasta
IFRTKHKLDFTPIGCDAKGRIVLGYTEAELCTRGSGYQFIHAADMLYCAESHIRMIKTGESGMIVFRLLT
DAARSRRSQETEVLYQLAHTLPFARGVSAHLDKASIMRLTISYLRMHRLCAAGEWNQVGAGGEPLDACYL
KALEGFVMVLTAEGDMAYLSENVSKHLGLSQLELIGHSIFDFIHPCDQEELQDALTPPTERCFSLRMKST
KEKSRNAARSRRGKENLEFFELAKLLPLPGAISSQLDKASIVRLSVTYLRLRRFAALGAPPWGLRAAGPP
AGLAPGRRGPAALVSEVFEQHLGGHILQSLDGFVFALNQEGKFLYISETVSIYLGLSQVEMTGSSVFDYI
HPGDHSEVLEQLGLVQERSFFVRMKSTLTKRGLHVKASGYKVIHVTGRLRALGLVALGHTLPPAPLAELP
WLQRAGGFVWLQSVATVAGSGKSPGEHHVLWVSHVLSQAEGGQT
GASKARRDQINAEIRNLKELLPLAEADKVRLSYLHIMSLACIYTRKGVFFAGGTPLAGPTGLLSAQELED
IVAALPGFLLVFTAEGKLLYLSESVSEHLGHSMVDLVAQGDSIYDIIDPADHLTVRQQLTLTDRLFRCRF
EKSKNAARTRREKENSEFYELAKLLPLPSAITSQLDKASIIRLTTSYLKMRVVFPEGLGEAWGHSSRTSP
EIERSFFLRMKCVLAKRNAGLTCGGYKVIHCSGYLKIRNVGLVAVGHSLPPSAVTEIKLHSNMFMFRASL
EKSKNAAKTRREKENGEFYELAKLLPLPSAITSQLDKASIIRLTTSYLKMRAVFPEGLGDAWGQPSRAGP
EIERSFFLRMKCVLAKRNAGLTCSGYKVIHCSGYLKIRIVGLVAVGQSLPPSAITEIKLYSNMFMFRASL
ELKHLILEAADGFLFIVSCETGRVVYVSDSVTPVLNQPQSEWFGSTLYDQVHPDDVDKLREQLSTSRMCM
GSRRSFICRMRCGSSEPHFVVVHCTGYIKAKFCLVAIGRLQVTSSPNCTDMSNVCQPTEFISRHNIEGIF
DELKHLILRAADGFLFVVGCDRGKILFVSESVFKILNYSQNDLIGQSLFDYLHPKDIAKVKEQLSSSRLC
SGARRSFFCRMKCNRRRKSFCTIHSTGYLKSNLSCLVAIGRLHSHVVRQRVNGEIRVKSMEYVSRHAIDG
```

5 and 6.



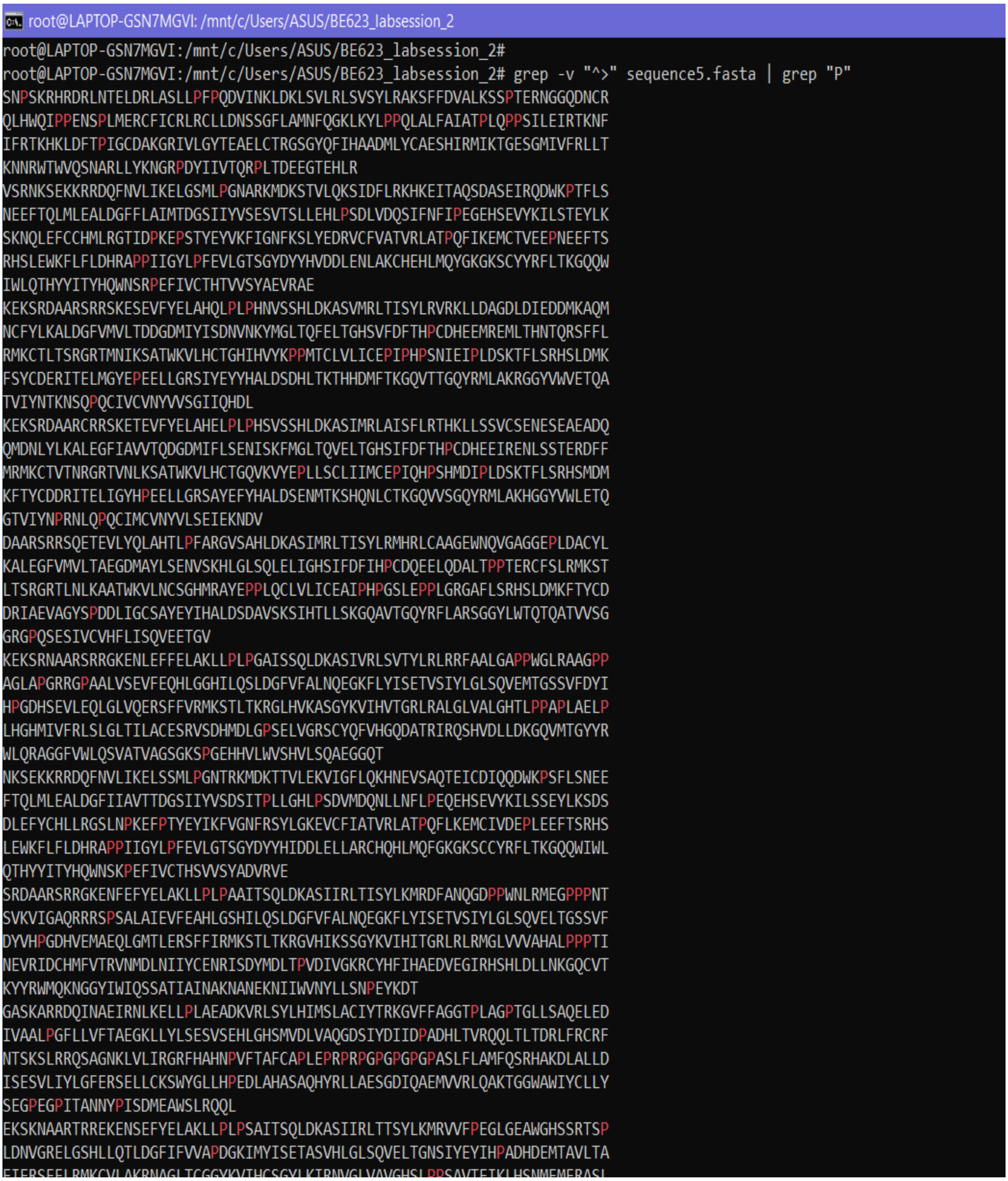
```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2#
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# grep "P[^A]L" sequence5.fasta
QLHWQIPPENSPLMERCFICRLRCLLDNSSGFLAMNFQGKLKYLPPQLALFAIATPLQPPSILEIRTKNF
MRMKCTVTNRGRTVNLKSATWKVLHCTGQVKVYEPLLSCLIIMCEPIQHPSHMDIPLDSKTFLSRHSMDM
LTSRGRTLNLKAATWKVLNCSGHMRAYEPPLQCLVLICEAIPHPGSLEPPLGRGAFLSRHSLDMKFTYCD
FTQLMLEALDGFIIAVTTDGSIIYVSDSITPLLGHLPSDVMDQNLLNFLPEQEHSEVYKILSSEYLKSDS
ELKHLILEAADGFLFIVSCETGRVVYVSDSVTPVLNQPQSEWFGSTLYDQVHPDDVDKLREQLSTSRMCM
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2#
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# grep "V\{2\}" sequence5.fasta
AANFREGLNLQEGEFLLQALNGFVLVVTTDALVFYASSTIQDYLGFQQSDVIHQSVYELIHTEDRAEFQR
IWLQTHYYITYHQWNSRPEFIVCTHTVVSYAEVRAE
TVIYNTKNSQPQCIVCVNYVVSGIIQHDL
QMDNLYLKALEGFIAVVTQDGDMIFLSENISKFMGLTQVELTGHSIFDFTHPCDHEEIRENLSSTERDFF
KFTYCDDRITELIGYHPEELLGRSAYEFYHALDSENMTKSHQNLCTKGQVVSGQYRMLAKHGGYVWLETQ
DRIAEVAGYSPDDLIGCSAYEYIHALDSDAVSKSIHTLLSKGQAVTGQYRFLARSGGYLWTQTQATVVSG
QTHYYITYHQWNSKPEFIVCTHSVVSYADVRVE
DYVHPGDHVEMAEQLGMTLERSFFIRMKSTLTKRGVHIKSSGYKVIHITGRLRLRMGLVVVAHALPPPTI
ISESVLIYLGFERSELLCKSWYGLLHPEDLAHASAQHYRLLAESGDIQAEMVVRLQAKTGGWAWIYCLLY
EKSKNAARTRREKENSEFYELAKLLPLPSAITSQLDKASIIRLTTSYLKMRVVFPEGLGEAWGHSSRTSP
LDNVGRELGSHLLQTLDGFIFVVAPDGKIMYISETASVHLGLSQVELTGNSIYEYIHPADHDEMTAVLTA
LDGVAKELGSHLLQTLDGFVFVVASDGKIMYISETASVHLGLSQVELTGNSIYEYIHPSDHDEMTAVLTA
SYATVVHNSRSSRPHCIVSVNYVLTEIEYKEL
ELKHLILEAADGFLFIVSCETGRVVYVSDSVTPVLNQPQSEWFGSTLYDQVHPDDVDKLREQLSTSRMCM
GSRRSFICRMRCGSSEPHFVVVHCTGYIKAKFCLVAIGRLQVTSSPNCTDMSNVCQPTEFISRHNIEGIF
TFVDHRCVATVGYQPQELLGKNIVEFCHPEDQQLLRDSFQQVVKLKGQVLSVMFRFRSKNQEWLWMRTSS
DELKHLILRAADGFLFVVGCDRGKILFVSESVFKILNYSQNDLIGQSLFDYLHPKDIAKVKEQLSSSRLC
SGARRSFFCRMKCNRPRKSFCTIHSTGYLKSNLSCLVAIGRLHSHVVPQPVNGEIRVKSMEYVSRHAIDG
RWFSFMNPWTKEVEYIVSTNTVVL
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2#
```

7.



```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2#
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# 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
```

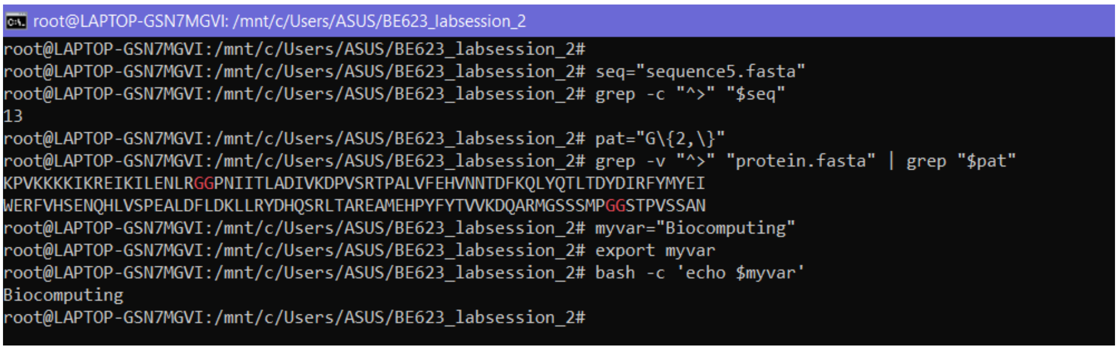
8.



```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2#
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# 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
EIERSEELRMKCVLAKRNAGLTCGGYKVIHCSGYLKIRNVGLVAVGHSLPPSAVTEIKLHSNMEMERASL
```

Part 3 - Using Variables

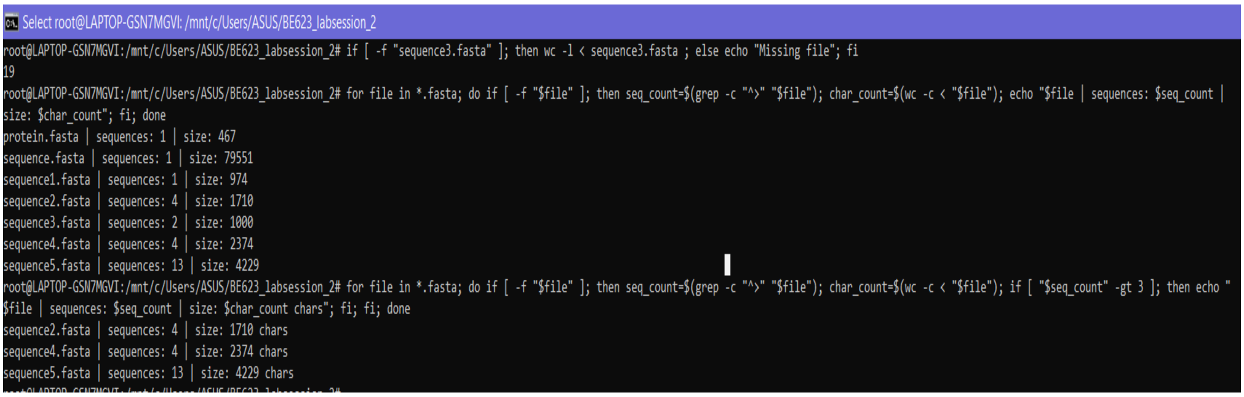
9, 10 and 11.



```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2#
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# seq="sequence5.fasta"
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# grep -c "^>" "$seq"
13
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# pat="G\{2,\}"
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# grep -v "^>" "protein.fasta" | grep "$pat"
KPVKKKKIKREIKILENLRGGPNIITLADIVKDPVSRTPALVFEHVNNTDFKQLYQTLTDYDIRFYMYEI
WERFVHSENQHLVSPEALDFLDKLLRYDHQSRLTAREAMEHPYFYTVVKDQARMGSSSMPGGSTPVSSAN
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# myvar="Biocomputing"
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# export myvar
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# bash -c 'echo $myvar'
Biocomputing
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2#
```

Part 4 - File existence and loops

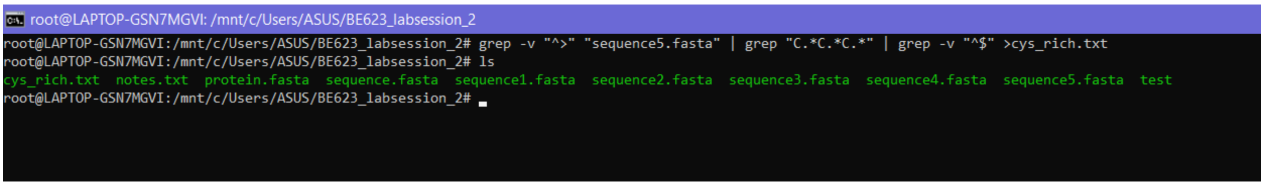
12,13 and 14.


Select root@LAPTOP-GSN7MGVI: /mnt/c/Users/ASUS/BE623_labsession_2

```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# if [ -f "sequence3.fasta" ]; then wc -l < sequence3.fasta ; else echo "Missing file"; fi
19
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# for file in *.fasta; do if [ -f "$file" ]; then seq_count=$(grep -c "^>" "$file"); char_count=$(wc -c < "$file"); echo "$file | sequences: $seq_count |
size: $char_count"; fi; done
protein.fasta | sequences: 1 | size: 467
sequence.fasta | sequences: 1 | size: 79551
sequence1.fasta | sequences: 1 | size: 974
sequence2.fasta | sequences: 4 | size: 1710
sequence3.fasta | sequences: 2 | size: 1000
sequence4.fasta | sequences: 4 | size: 2374
sequence5.fasta | sequences: 13 | size: 4229
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# for file in *.fasta; do if [ -f "$file" ]; then seq_count=$(grep -c "^>" "$file"); char_count=$(wc -c < "$file"); if [ "$seq_count" -gt 3 ]; then echo "
$file | sequences: $seq_count | size: $char_count chars"; fi; fi; done
sequence2.fasta | sequences: 4 | size: 1710 chars
sequence4.fasta | sequences: 4 | size: 2374 chars
sequence5.fasta | sequences: 13 | size: 4229 chars
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2#
```

Part 5 -Applied data extraction

15.



```
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# grep -v "^>" "sequence5.fasta" | grep "C.*C.*C.*" | grep -v "^$" >cys_rich.txt
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# ls
cys_rich.txt  notes.txt  protein.fasta  sequence.fasta  sequence1.fasta  sequence2.fasta  sequence3.fasta  sequence4.fasta  sequence5.fasta  test
root@LAPTOP-GSN7MGVI:/mnt/c/Users/ASUS/BE623_labsession_2# 
```