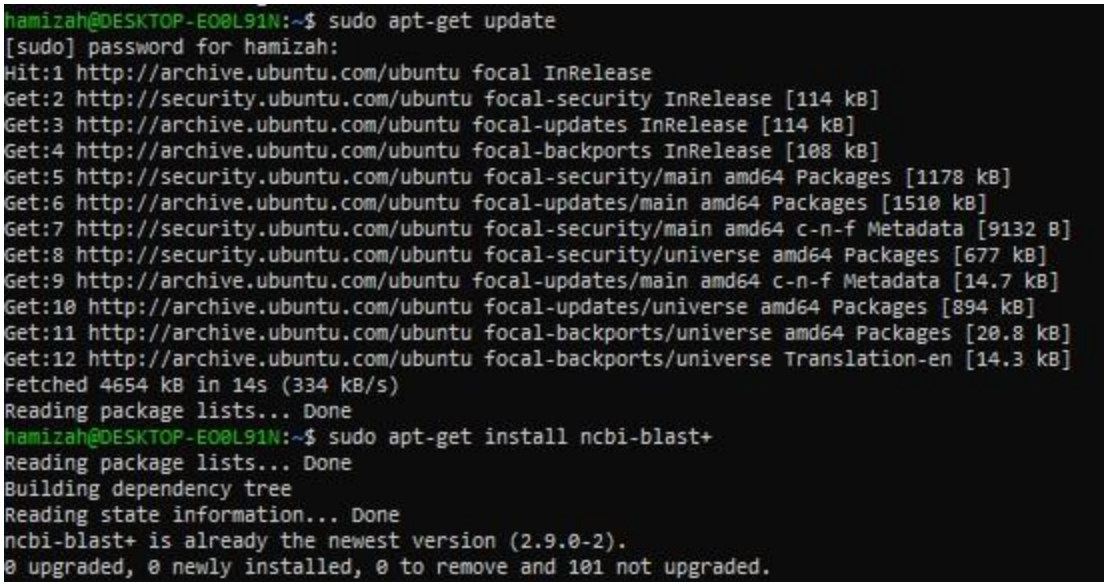
COMPUTER OPERATING SYSTEM – BIOINFORMATICS ANALYSIS

Compare protein sequences of non-pathogenic strain and pathogenic strain using BLAST in Linux using Ubuntu

- Install blast+



- Download NC_000913.fasta and NC_002655.fasta

```
hamizah@DESKTOP-EO0L91N:~$ wget --no-check-certificate 'https://docs.google.com/uc?export=download&id=1eHmlUi4m8pyolMw275NxD3zicdhJy4ab' -O NC_000913.fasta
--2022-01-28 04:32:23--  https://docs.google.com/uc?export=download&id=1eHmlUi4m8pyolMw275NxD3zicdhJy4ab
Resolving docs.google.com (docs.google.com)... 142.250.199.14, 2404:6800:4001:803::200e
Connecting to docs.google.com (docs.google.com)|142.250.199.14|:443... connected.
HTTP request sent, awaiting response... 302 Moved Temporarily
Location: https://doc-04-bg-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc7l7deffksulhg5h7mbp1/9oe4kpf71bdl9q49gj6tojp4mpkin83f/1643315475000/01839286
ing]
Warning: wildcards not supported in HTTP.
--2022-01-28 04:32:24--  https://doc-04-bg-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc7l7deffksulhg5h7mbp1/9oe4kpf71bdl9q49gj6tojp4mpkin83f/164331
ownload
Resolving doc-04-bg-docs.googleusercontent.com (doc-04-bg-docs.googleusercontent.com)... 172.217.26.65, 2404:6800:4001:810::2001
Connecting to doc-04-bg-docs.googleusercontent.com (doc-04-bg-docs.googleusercontent.com)|172.217.26.65|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 8760 (8.6K) [application/octet-stream]
Saving to: 'NC_000913.fasta'

NC_000913.fasta                     100%[=============================================================================================================>

2022-01-28 04:32:25 (5.62 MB/s) - 'NC_000913.fasta' saved [8760/8760]

hamizah@DESKTOP-EO0L91N:~$ wget --no-check-certificate 'https://docs.google.com/uc?export=download&id=1911cWfvL6c6kvX0yAAT2WGNb_S5wNz2g' -O NC_002655.fasta
--2022-01-28 04:33:05--  https://docs.google.com/uc?export=download&id=1911cWfvL6c6kvX0yAAT2WGNb_S5wNz2g
Resolving docs.google.com (docs.google.com)... 216.58.196.14, 2404:6800:4001:806::200e
Connecting to docs.google.com (docs.google.com)|216.58.196.14|:443... connected.
HTTP request sent, awaiting response... 302 Moved Temporarily
Location: https://doc-0o-bg-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc7l7deffksulhg5h7mbp1/gadga89g7auhkm8a20enuptr9bf125je/1643315550000/01839286
ing]
Warning: wildcards not supported in HTTP.
--2022-01-28 04:33:05--  https://doc-0o-bg-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc7l7deffksulhg5h7mbp1/gadga89g7auhkm8a20enuptr9bf125je/164331
ownload
Resolving doc-0o-bg-docs.googleusercontent.com (doc-0o-bg-docs.googleusercontent.com)... 216.58.196.33, 2404:6800:4001:807::2001
Connecting to doc-0o-bg-docs.googleusercontent.com (doc-0o-bg-docs.googleusercontent.com)|216.58.196.33|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 6088 (5.9K) [application/octet-stream]
Saving to: 'NC_002655.fasta'

NC_002655.fasta                     100%[=============================================================================================================>

2022-01-28 04:33:06 (3.66 MB/s) - 'NC_002655.fasta' saved [6088/6088]
```

a. Create a folder "os_project"

```
hamizah@DESKTOP-EO0L91N:~$ ls
NC_000913.fasta  NC_002655.fasta
hamizah@DESKTOP-EO0L91N:~$ mkdir os_project
hamizah@DESKTOP-EO0L91N:~$ ls
NC_000913.fasta  NC_002655.fasta  os_project
```

b. Move downloaded "NC_00913.fasta" and "NC002655.fasta" into the folder "os_project"

```
hamizah@DESKTOP-EO0L91N:~$ mv NC_000913.fasta NC_002655.fasta os_project
hamizah@DESKTOP-EO0L91N:~$ ls
os_project
hamizah@DESKTOP-EO0L91N:~$ cd os_project
hamizah@DESKTOP-EO0L91N:~/os_project$ ls
NC_000913.fasta  NC_002655.fasta
```

c. Change filename "NC_00913.fasta" to "ecoli-k12.fasta" and filename "NC_002655.fasta" to "ecoli-h7.fasta"

```
hamizah@DESKTOP-EO0L91N:~/os_project$ mv NC_000913.fasta ecoli-k12.fasta
hamizah@DESKTOP-EO0L91N:~/os_project$ mv NC_002655.fasta ecoli-h7.fasta
hamizah@DESKTOP-EO0L91N:~/os_project$ ls
ecoli-h7.fasta  ecoli-k12.fasta
```

d. Set up a database file for E.coli K12. The command to use is “makeblastdb” from the BLAST+ package

```
hamizah@DESKTOP-EO0L91N:~/os_project$ makeblastdb -in ecoli-k12.fasta -dbtype prot -title "E.coli K12 Database" -out ecoli-k12db -parse_seqids


Building a new DB, current time: 01/28/2022 06:08:38
New DB name:   /home/hamizah/os_project/ecoli-k12db
New DB title:  E.coli K12 Database
Sequence type: Protein
Keep MBits: T
Maximum file size: 1000000000B
Adding sequences from FASTA; added 20 sequences in 0.081609 seconds.
hamizah@DESKTOP-EO0L91N:~/os_project$ ls
ecoli-h7.fasta  ecoli-k12.fasta  ecoli-k12db.phr  ecoli-k12db.pin  ecoli-k12db.pog  ecoli-k12db.psd  ecoli-k12db.psi  ecoli-k12db.psq
```

e. BLAST “ecoli-h7.fasta” protein sequence against “ecoli-k12.fasta” using program “blastp” in blast+ package and export the query result as “h7vsk12.txt”. Please note that you should use e-value 0.00001 for your blast query.

```
hamizah@DESKTOP-EO0L91N:~/os_project$ blastp -db ecoli-k12db -query ecoli-h7.fasta -out h7vsk12.txt -evalue 0.00001
hamizah@DESKTOP-EO0L91N:~/os_project$ ls
ecoli-h7.fasta  ecoli-k12.fasta  ecoli-k12db.phr  ecoli-k12db.pin  ecoli-k12db.pog  ecoli-k12db.psd  ecoli-k12db.psi  ecoli-k12db.psq  h7vsk12.txt
```

f. Use “awk” and “egrep” to extract the first 20 annotated proteins.

```
hamizah@DESKTOP-EO0L91N:~/os_project$ awk '/Query=/ ||/No hits/{print}' h7vsk12.txt | awk '{i++;line[i]=$0; if($0~/No hits/){print line[i-1]}}' | egrep -v "([Uu]nknown) | [Pp]utative | [Hh]ypothetical" | head -20
Query= AAG59188.2 thiamin biosynthesis, probable sulfur donor [Escherichia
Query= AAG57587.2 lipoprotein-34 [Escherichia coli O157:H7 str. EDL933]
Query= AAG57537.2 pyridoxal/pyridoxine/pyridoxamine kinase [Escherichia
Query= AAG57281.2 cytidine/deoxycytidine deaminase [Escherichia coli
```

- Brief description of final output

```
Query= AAG59188.2 thiamin biosynthesis, probable sulfur donor [Escherichia
Query= AAG57587.2 lipoprotein-34 [Escherichia coli O157:H7 str. EDL933]
Query= AAG57537.2 pyridoxal/pyridoxine/pyridoxamine kinase [Escherichia
Query= AAG57281.2 cytidine/deoxycytidine deaminase [Escherichia coli
```

The final output displays 4 protein sequences in the first 20 annotated proteins that present in the pathogenic strain E.coli pathogenic strain E. coli O157:H7 but not in the nonpathogenic strain E. coli K12.