

Bioinformatics **Project**

Kushagra Agarwal
2018113012

Q1)

The screenshot shows the NCBI BLAST search interface. The 'blastp' program is selected under the 'Program Selection' tab. The 'Choose Search Set' section shows the 'Database' set to 'Non-redundant protein sequences (nr)'. The 'Organism' field is set to 'SARS-CoV-2 (taxid:2697049)'. The 'Exclude' section is empty. The 'Enter Query Sequence' section shows a text input field with a sequence: YDAQPCSDKAYKIEELFYSYATHSDKFTDGVCLFWNCNVDRYPANSIVCRFDTRVLSNLPDCDGG SLY VNKHAFHTPAFDKSAFVNLKQLPFFYYSDSPCESHGKQVVSDIDYVPLKSATCITRCNLGGAVCRHH ANE YRLYLDAYNMMISAGFSLWYKQFDYTNLWNTFTRLQ. The 'Job Title' field is empty. The 'Align two or more sequences' checkbox is unchecked.

(i) Which BLAST program did you consider for the search?

I used the Blastp program for the search

(ii) Which database was it searched against and why?

I searched it against the Non Redundant Protein Sequences (nr) database. The nr database is compiled by the NCBI (National Center for Biotechnology Information) as a protein database for Blast searches. It contains non-identical sequences from GenBank CDS translations, PDB, Swiss-Prot, PIR, and PRF. We wanted to find our query sequence in all of these databases, hence I used this nr database.

iii) Based on the search what can you say about your sequence?

I performed the database search twice. Once excluding all hits from SARS-CoV-2 (to find matches from other organisms) and once including SARS-CoV-2.

The top 5 hits from the one excluding SARS-CoV-2 is:

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
✓	ORF1ab [synthetic construct]	1134	1134	100%	0.0	100.00%	QIG55856.1
✓	orf1ab polyprotein [Bat coronavirus RaTG13]	1129	1129	100%	0.0	99.24%	QHR63299.1
✓	orf1ab polyprotein [Pangolin coronavirus]	1120	1120	100%	0.0	98.67%	QIG55944.1
✓	orf1ab polyprotein [Pangolin coronavirus]	1108	1108	100%	0.0	96.96%	QIA48622.1
✓	orf1ab polyprotein [Pangolin coronavirus]	1108	1108	100%	0.0	96.96%	QIA48613.1

The top 5 hits from the one including SARS-CoV-2 is:

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
✓	ORF1ab polyprotein .partial [Severe acute respiratory syndrome coronavirus 2]	1134	1134	100%	0.0	100.00%	QNE12858.1
✓	ORF1ab polyprotein [Severe acute respiratory syndrome coronavirus 2]	1134	1134	100%	0.0	100.00%	QNO98541.1
✓	ORF1ab polyprotein [Severe acute respiratory syndrome coronavirus 2]	1134	1134	100%	0.0	100.00%	QJR86467.1
✓	ORF1ab polyprotein [Severe acute respiratory syndrome coronavirus 2]	1134	1134	100%	0.0	100.00%	QMT57234.1
✓	ORF1ab polyprotein [Severe acute respiratory syndrome coronavirus 2]	1134	1134	100%	0.0	100.00%	QJD23979.1

From both the results, it is evident that the query sequence is a part of the ORF1ab polyprotein found in SARS-CoV-2.

iv) Report the organism name, sequence coverage, %age identity, and the e-value of the top non-SARS-CoV-2 sequence.

Organism Name	Coverage	% Identity	E-value
Bat coronavirus RaTG13	100%	99.24%	0.0
Pangolin coronavirus	100%	98.67%	0.0
Bat SARS-like coronavirus	100%	95.83%	0.0

BtRs-BetaCoV/YN2013	100%	95.83%	0.0
Bat coronavirus	100%	95.64%	0.0

v) Find 5 hits to your sequence with identity in the range ~ 95% - 25% - list the accession id, organism name, query coverage, percentage identity, and e-value for the selected 5 sequences.

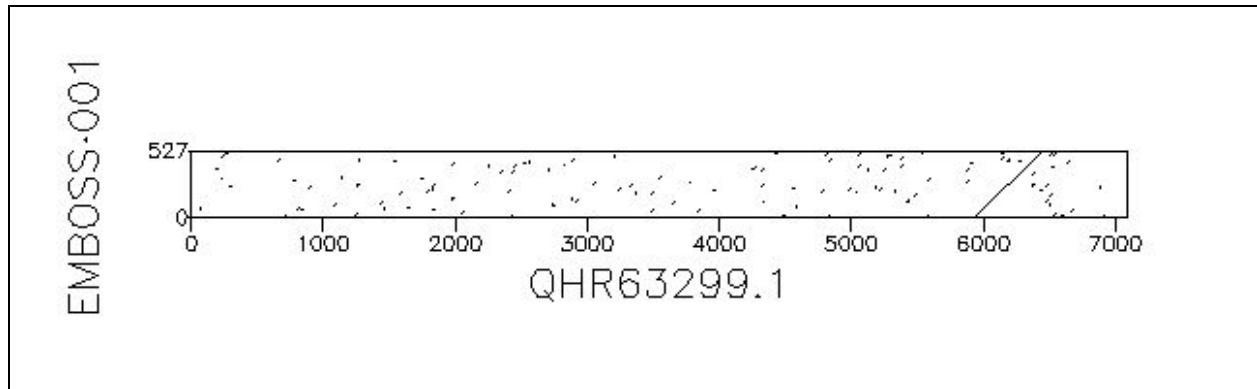
AccessionID	Organism Name	Coverage	% Identity	E-value
AAV91630.1	SARS coronavirus A022	100%	94.88%	0.0
ADY69164.1	Zaria bat coronavirus	100%	72.35%	0.0
ALB08320.1	Middle East respiratory syndrome-related coronavirus	100%	62.76%	0.0
BBM61484.1	Bovine coronavirus	99%	58.21%	0.0
QJS39707.1	Infectious bronchitis virus	99%	53.52%	0.0

vi) Give the dot plots of the query with the 1st and 5th sequences from the list in (iv).

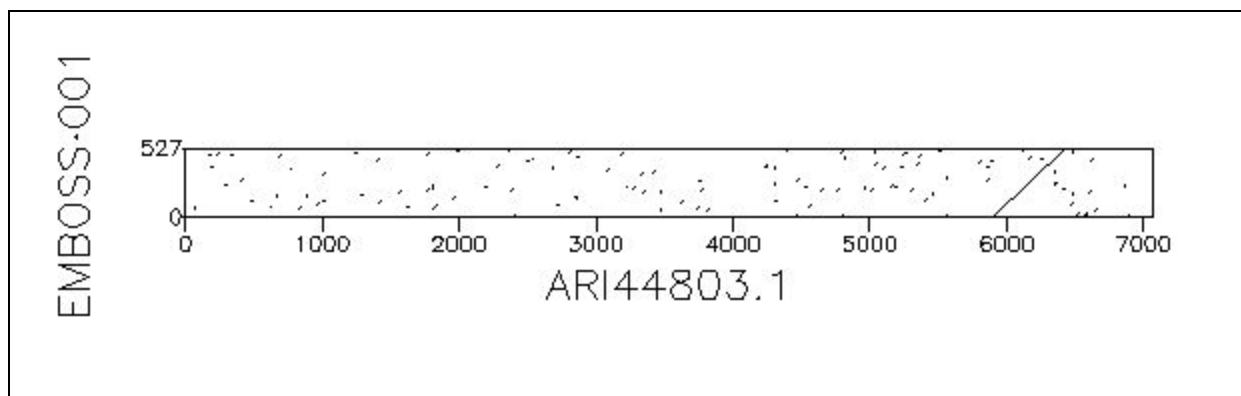
Dotmatcher Parameters:

STEP 2 - Set options		
WINDOW SIZE	THRESHOLD	MATRIX
10	23	BLOSUM62 ▼

Organism: Bat coronavirus RaTG13 Accession ID: QHR63299.1



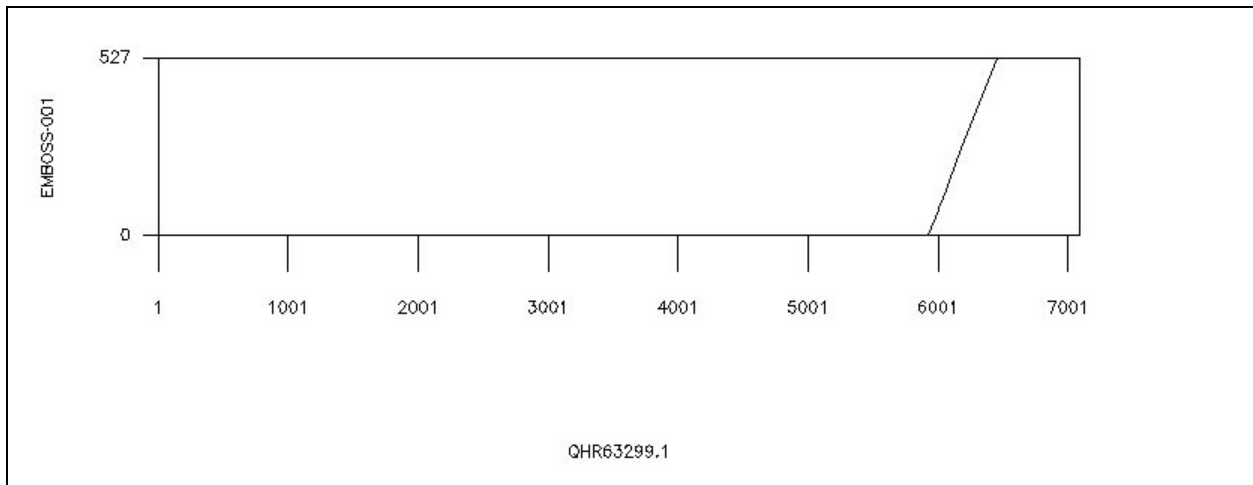
Organism: Bat coronavirus Accession ID: ARI44803.1



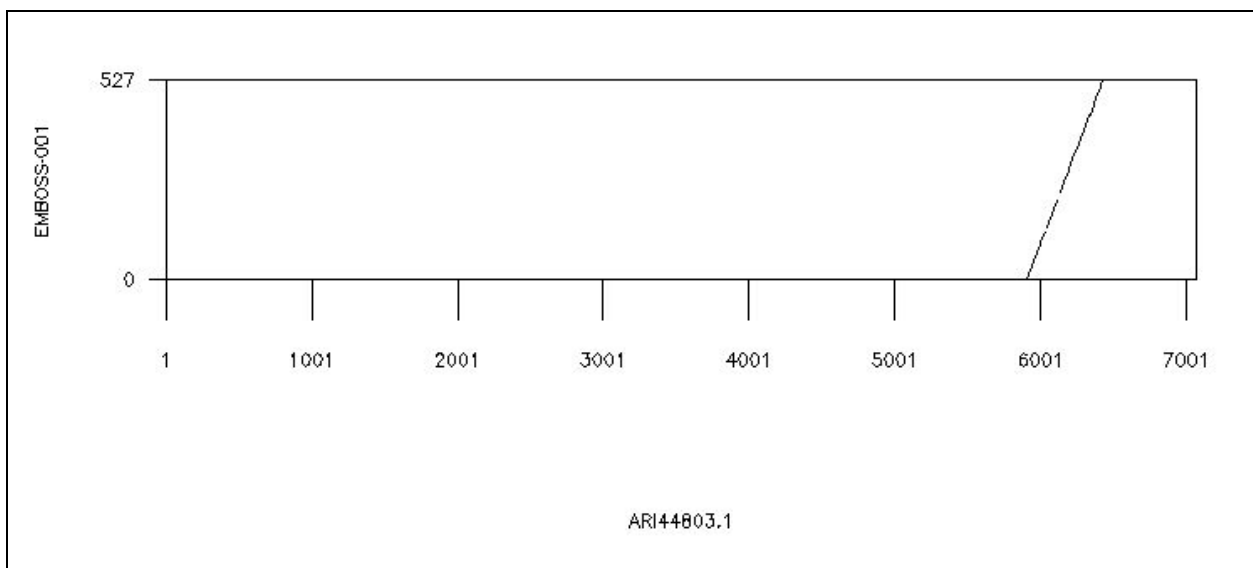
Dottup Parameters:

STEP 2 - Set options	
WORD SIZE	BOXIT
10	yes ▼

Organism: Bat coronavirus RaTG13 Accession ID: QHR63299.1



Organism: Bat coronavirus Accession ID: ARI44803.1



vii) Give the pairwise alignment by Needleman-Wunsch algorithm – is it similar to the alignment given by BLAST?

Used Emboss Needle to perform the alignments for all the 5 sequences mentioned in iv)

OUTPUT FORMAT					
pair					
MATRIX	GAP OPEN	GAP EXTEND	END GAP PENALTY	END GAP OPEN	END GAP EXTEND
BLOSUM62	10	0.5	false	10	0.5

i) Bat coronavirus RaTG13 [QHR63299.1](#)

https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=emboss_needle-I20201006-170842-0062-71717588-p1m

Also available in the file: Needle/QHR63299.1.txt

ii) Pangolin coronavirus [QIG55944.1](#)

https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=emboss_needle-I20201006-170909-0830-51969656-p1m

Also available in the file: Needle/QIG55944.1.txt

iii) Bat SARS-like coronavirus [ATO98118.1](#)

https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=emboss_needle-I20201006-170848-0039-65717786-p1m

Also available in the file: Needle/ATO98118.1.txt

iv) BtRs-BetaCoV/YN2013 [AIA62329.1](#)

https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=emboss_needle-I20201006-170857-0144-79261247-p1m

Also available in the file: Needle/AIA62329.1.txt

v) Bat coronavirus [ARI44803.1](#)

https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=emboss_needle-I20201006-170832-0764-6570675-p2m

Also available in the file: Needle/ARI44803.1.txt

Inference: The alignments for all the 5 sequences with the query sequence are the same as those obtained from the BLAST search tool.

I solved Q2 and Q3 for two different query sequences. The question asks us to perform the MSA using the query sequence which would mean the sequence given in the assignment question. But on creating the phylogenetic trees for this MSA I observed all the trees were different (even the bootstrap confidence was weak). This was because the query sequence was very small compared to the 5 hits in Q1) iv). Therefore I performed both Q2) and Q3) for both the query sequence given in the question and the protein sequence (ORF1ab in SARS-CoV-2) we found in Q4).

Q2) Perform multiple sequence alignment of your query sequence with the 5 'hits' selected from the database search in step-1 using CLUSTAL. Submit the alignment and score.

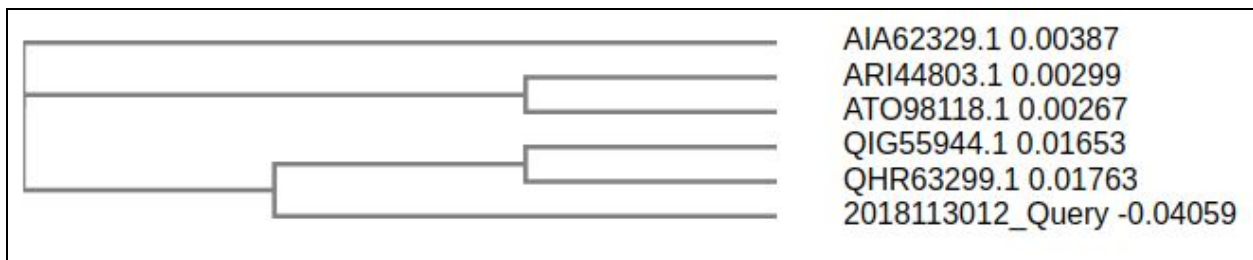
STEP 2 - Set your parameters			
OUTPUT FORMAT			
ClustalW with character counts			
DEALIGN INPUT SEQUENCES	MBED-LIKE CLUSTERING GUIDE-TREE	MBED-LIKE CLUSTERING ITERATION	NUMBER of COMBINED ITERATIONS
no	yes	yes	default(0)
MAX GUIDE TREE ITERATIONS	MAX HMM ITERATIONS	ORDER	
default	default	aligned	

MSA of top 5 hits for the query sequence:

Identity Matrix:

1: AIA62329.1	100.00	99.24	99.18	86.74	86.38	95.83
2: ARI44803.1	99.24	100.00	99.43	86.77	86.42	95.64
3: AT098118.1	99.18	99.43	100.00	86.79	86.50	95.83
4: QIG55944.1	86.74	86.77	86.79	100.00	96.58	98.67
5: QHR63299.1	86.38	86.42	86.50	96.58	100.00	99.24
6: 2018113012_Query	95.83	95.64	95.83	98.67	99.24	100.00

Phylogenetic Tree:

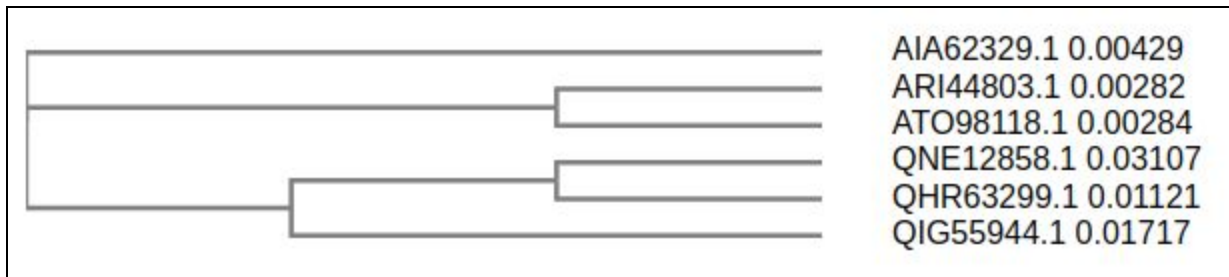


MSA of top 5 hits for protein (ORF1ab in SARS-CoV-2) in Q4:

Identity Matrix:

1: AIA62329.1	100.00	99.24	99.18	84.54	86.37	86.73
2: ARI44803.1	99.24	100.00	99.43	84.60	86.42	86.76
3: AT098118.1	99.18	99.43	100.00	84.70	86.50	86.77
4: QNE12858.1	84.54	84.60	84.70	100.00	95.77	94.08
5: QHR63299.1	86.37	86.42	86.50	95.77	100.00	96.58
6: QIG55944.1	86.73	86.76	86.77	94.08	96.58	100.00

Phylogenetic Tree:



For Q4Protein (ORF1ab in SARS-CoV-2)

<https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=clustalo-I20201006-180844-0580-74792240-p1m&analysis=alignments>

Can be also found in the file MSA/Q4Protein.clustal_num

For Query Sequence

<https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=clustalo-I20201006-180851-0447-84982777-p2m&analysis=alignments>

Can be also found in the file MSA/Query.clustal_num

Q3) Construct a phylogenetic tree using the MSA obtained in step-2. You may choose a method of your choice from the PHYLIP suite of programs.

- 1) Which method was used to construct the tree? Give reasons for choosing this method.

I used proml to create the trees. proml stands for Maximum Likelihood Tree for protein sequences. ML is considered the best method (slowest as well)

- Maximum likelihood is similar to maximum parsimony, in that analysis is performed for each column of the alignment, all possible trees are considered, and trees with the fewest changes are usually more likely
- ML allows corrections for variations in the mutation rates by considering explicit evolutionary models
- The method can be used to explore relationships among more diverse sequences

2) Give bootstrap values.

Got bootstrap values using phylip consense

Files for Query:

- with_bs/Query/Query_consense_outfile
- with_bs/Query/Query_consense_outtree

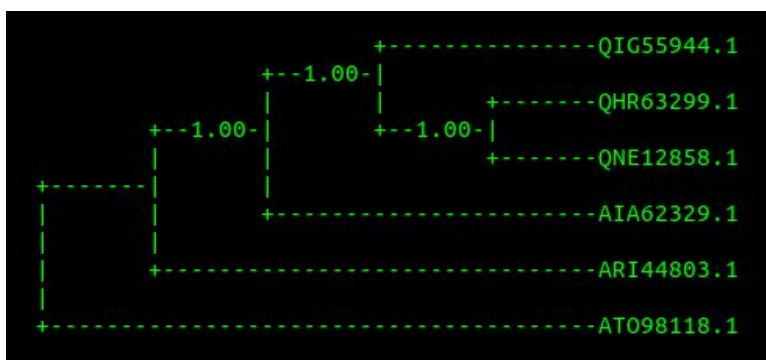
Consense Tree:



Files for Q4Protein:

- with_bs/Q4Protein/Q4Protein_consense_outfile
- with_bs/Q4Protein/Q4Protein_consense_outtree

Consense Tree:



- 3) Submit the tree and inferences drawn from this tree about the relatedness of the species considered?

For Query Sequence:

WITHOUT BOOTSTRAP

Files:

- without_bs/Query/Query_outfile
- without_bs/Query/Query_outtree

```
Jones-Taylor-Thornton model of amino acid change

      +2018113012
      +---4
+-----3 +QHR63299.1
|         |
|         +QIG55944.1
|
| +ARI44803.1
2--1
| +AT098118.1
|
+AIA62329.1
```

```
Ln Likelihood = -27517.99121

Between      And      Length      Approx. Confidence Limits
-----      ---      -
2            AIA62329.1    0.00373    ( 0.00220, 0.00526) **
2            3          0.13387    ( 0.12482, 0.14293) **
3            4          0.00882    ( 0.00520, 0.01245) **
4            2018113012 0.00010    ( zero, infinity)
4            QHR63299.1 0.01275    ( 0.00907, 0.01641) **
3            QIG55944.1 0.01333    ( 0.01020, 0.01645) **
2            1          0.00152    ( 0.00044, 0.00262) **
1            ARI44803.1 0.00246    ( 0.00129, 0.00363) **
1            AT098118.1 0.00320    ( 0.00187, 0.00453) **

* = significantly positive, P < 0.05
** = significantly positive, P < 0.01
```

WITH BOOTSTRAP

Run phylip seqboot (random seed 11) and gave Query.phylip as input to get Query_bs_input as output. Ran proml on this as input

Files:

- with_bs/Query/Query_outfile
- with_bs/Query/Query_outtree

Jones-Taylor-Thornton model of amino acid change

```
+ARI44803.1
+- -1
| +AT098118.1
|
|           +2018113012
|           +--4
2-----3 +QHR63299.1
|         |
|         +QIG55944.1
|
+AIA62329.1
```

Ln Likelihood = -27498.22782

Between	And	Length	Approx. Confidence Limits
-----	---	-----	-----
2	AIA62329.1	0.00424	(0.00258, 0.00589) **
2	1	0.00214	(0.00085, 0.00345) **
1	ARI44803.1	0.00367	(0.00226, 0.00509) **
1	AT098118.1	0.00293	(0.00166, 0.00420) **
2	3	0.13325	(0.12424, 0.14227) **
3	4	0.00868	(0.00513, 0.01222) **
4	2018113012	0.00010	(zero, infinity)
4	QHR63299.1	0.01243	(0.00882, 0.01602) **
3	QIG55944.1	0.01136	(0.00844, 0.01426) **

* = significantly positive, P < 0.05

** = significantly positive, P < 0.01

For Q4 Protein Sequence:

WITHOUT BOOTSTRAP

Files:

- without_bs/Q4Protein/Q4Protein_outfile
- without_bs/Q4Protein/Q4Protein_outtree

```
Jones-Taylor-Thornton model of amino acid change

+ARI44803.1
|
|           +QIG55944.1
|   +-----4
|   |           |   +QHR63299.1
1--2   +--3
|           |   +QNE12858.1
|           |
|   +AT098118.1
|
+AIA62329.1
```

```
Ln Likelihood = -28057.09193

Between      And      Length      Approx. Confidence Limits
-----      ---      -
1            AIA62329.1      0.00522      ( 0.00353, 0.00691) **
1            ARI44803.1      0.00246      ( 0.00129, 0.00363) **
1            2            0.00136      ( 0.00039, 0.00233) **
2            4            0.13410      ( 0.12505, 0.14317) **
4            QIG55944.1      0.01438      ( 0.01120, 0.01755) **
4            3            0.01344      ( 0.01033, 0.01656) **
3            QHR63299.1      0.00852      ( 0.00624, 0.01080) **
3            QNE12858.1      0.00716      ( 0.00507, 0.00925) **
2            AT098118.1      0.00186      ( 0.00077, 0.00296) **

* = significantly positive, P < 0.05
** = significantly positive, P < 0.01
```


WITH BOOTSTRAP

Run phylip seqboot (random seed 11) and gave Q4Protein.phylip as input to get Q4Protein_bs_input as output. Ran proml on this as input

Files:

- with_bs/Q4Protein/Q4Protein_outfile
- with_bs/Q4Protein/Q4Protein_outtree

Jones-Taylor-Thornton model of amino acid change

```
+AT098118.1
+- -1
| +ARI44803.1
|
| +QIG55944.1
2-----4
| | +QNE12858.1
| | +- -3
| | +QHR63299.1
|
+AIA62329.1
```

Ln Likelihood = -28150.10608

Between	And	Length	Approx. Confidence Limits
-----	---	-----	-----
2	AIA62329.1	0.00448	(0.00279, 0.00617) **
2	1	0.00191	(0.00065, 0.00316) **
1	AT098118.1	0.00293	(0.00166, 0.00420) **
1	ARI44803.1	0.00367	(0.00226, 0.00509) **
2	4	0.13307	(0.12407, 0.14208) **
4	QIG55944.1	0.01268	(0.00966, 0.01569) **
4	3	0.01309	(0.01003, 0.01616) **
3	QNE12858.1	0.00948	(0.00710, 0.01187) **
3	QHR63299.1	0.00825	(0.00601, 0.01049) **

* = significantly positive, P < 0.05

** = significantly positive, P < 0.01

iii) From the trees found we can conclude that the Q4Protein sequence (QNE12858.1) is the closest to Bat coronavirus RaTG13 (QHR63299.1) In all the three trees the topology is conserved and a value of 1 for the bootstrap consense tree confirms that the phylogeny inferred from the tree is correct.

For the trees for Query Sequence, we can see that 2018113012 (Query sequence) is also the closest to Bat coronavirus RaTG13 (QHR63299.1) In all the three trees the topology is conserved and a value of 1 for the bootstrap consense tree confirms that the phylogeny inferred from the tree is correct.

Therefore using both the methods, we have further confirmed our initial hypothesis that Bat coronavirus RaTG13 (QHR63299.1) is indeed the closest sequence to both our Query sequence and the Q4Protein (ORF1ab from SARS-CoV-2).


Q4) Give the name and accession id of your protein. Give a brief description of your protein and its function (based on the latest literature).

ORF1ab polyprotein Severe acute respiratory syndrome coronavirus 2
Accession ID: QNE12858.1

Percentage Identity	Coverage	E-value
100%	100%	0.0


As the percentage identity and coverage is 100% and the e-value is 0 (proves that this match is not by chance), we can say that the sequence given to us is from this ORF1ab polyprotein in SARS-CoV-2.

About ORF1ab polyprotein:



Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is an enveloped, positive-sense, single-stranded RNA virus that causes coronavirus disease 2019 (COVID-19). Virus particles include the RNA genetic material and structural proteins needed for the invasion of host cells. Once inside the cell the infecting RNA is used to encode structural proteins that make up virus particles, nonstructural proteins that direct virus assembly, transcription, replication, and host control and accessory proteins whose function has not been determined.[~] ORF1ab, the largest gene, contains overlapping open reading frames that encode polyproteins PP1ab and PP1a. The polyproteins are cleaved to yield 16 nonstructural proteins, NSP1-16. Production of the longer (PP1ab) or shorter protein (PP1a) depends on a -1 ribosomal frameshifting events. The proteins, based on similarity to other coronaviruses, include the papain-like proteinase protein (NSP3), 3C-like proteinase (NSP5), RNA-dependent RNA polymerase (NSP12, RdRp), helicase (NSP13, HEL), endoRNAse (NSP15), 2'-O-Ribose-Methyltransferase (NSP16) and other nonstructural proteins. SARS-CoV-2 nonstructural proteins are responsible for viral transcription, replication, proteolytic processing, suppression of host immune responses, and suppression of host gene expression. The RNA-dependent RNA polymerase is a target of antiviral therapies.


1. Host translation inhibitor (nsp1) QHD43415_1
2. Non-structural protein 2 (nsp2) QHD43415_2
3. Papain-like proteinase (PLpro) QHD43415_3
4. Non-structural protein 4 (nsp4) QHD43415_4
5. Main proteinase 3CL-PRO QHD43415_5
6. Non-structural protein 6 (nsp6) QHD43415_6
7. Non-structural protein 7 (nsp7) QHD43415_7
8. Non-structural protein 8 (nsp8) QHD43415_8
9. Non-structural protein 9 (nsp9) QHD43415_9

- 
10. Replicase Polyprotein 1ab QHD43415_11
 11. Helicase (Hel) QHD43415_12
 12. Guanine-N7 methyltransferase (ExoN) / Proofreading exoribonuclease QHD43415_13
 13. Uridylate-specific endoribonuclease (NendoU) QHD43415_14

Used COVID3D (<http://biosig.unimelb.edu.au/covid3d/list>)

<https://jvi.asm.org/content/71/12/9313.short>

This means that the ORF1a protein can be cleaved into eight processing end products: nsp1 to nsp8. By micro sequence analysis of the nsp5 and nsp7 N termini, we have now formally confirmed the specificity of the SP for Glu / (Gly/Ser) substrates. Importantly, our studies revealed that the C-terminal half of the ORF1a protein (nsp3-8) can be processed by the SP following two alternative pathways, which appear to be mutually exclusive. In the majority of the nsp3-8 precursors, the SP cleaves the nsp4/5 site, yielding nsp3-4 and nsp5-8. Subsequently, the latter product is cleaved at the nsp7/8 site only, whereas the newly identified nsp5/6 and nsp6/7 sites appear to be inaccessible to the protease. In the alternative proteolytic cascade, which is used at a low but significant level in infected cells, it is the nsp4/5 site which remains uncleaved, while the nsp5/6 and nsp6/7 sites are processed to yield a set of previously unnoticed processing products. Coexpression studies revealed that nsp3-8 has to interact with cleaved nsp2 to allow processing of the nsp4/5 junction, the first step of the major processing pathway. When the nsp2 cofactor is absent, the nsp4/5 site cannot be processed and nsp3-8 is processed following the alternative, minor pathway.



ORF1a which comprises nsp's helps in coping with cellular stresses and maintains the functional integrity of the cellular components along with the pivotal roles in viral replication. ORF1b encodes viral RNA-dependent RNA polymerase (nsp 12), helicase (nsp 13), exonuclease (nsp14), a polyU (Uridylate) specific endonuclease (nsp15), and methyltransferase (nsp16).
