Tutorial Assignment 1

NAIMEESH NARAYAN TIWARI

2020101074

▼ Question 1

1. **DotPlot Analysis** using (a) Dottup – k-tuple search, and (b) Dotmatcher – sliding window

Show them first the example on the webserver (HBA_Human with HBB_Human proteins). Ask them to run these programs for different k-tuples (Dottup) and different window size and threshold (Dotmatcher).

Given are the sequences of spike glycoprotein (both DNA and protein) of the following: (1) SARS-CoV (2003), MERS-COV (2012), and (3) SARS-CoV2 (2019). Submit the results of Dottup and Dotmatcher and answer the following Qs:

(i) Identify SARS-CoV2 is similar to which of the earlier two viruses?
(ii) Is it easy to identify the similarity using DNA or protein sequences? Give reasons.
(iii) Submit the graphs and give the k-tuple values used, and window size and threshold values used.

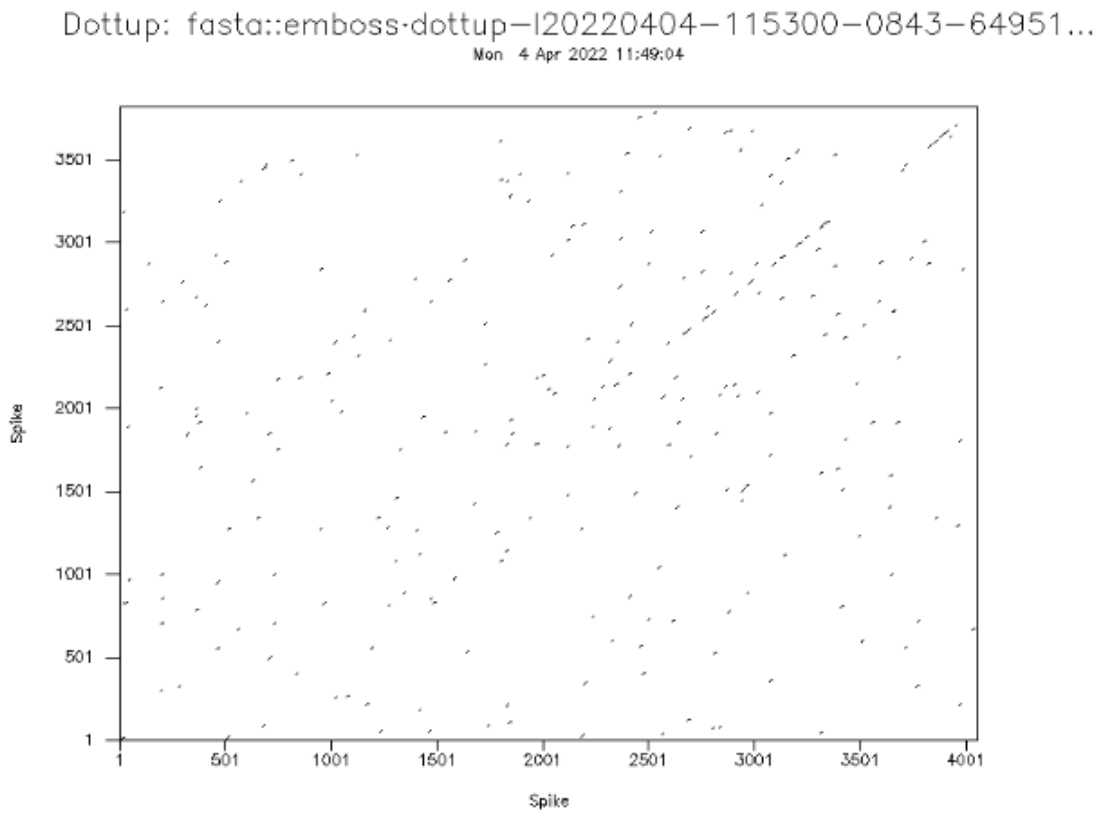
▼ Similarity Comparison

SARS-CoV (2003) has more similarities than MERS-COV(2012). We infer this from the DotPlot analysis graphs:

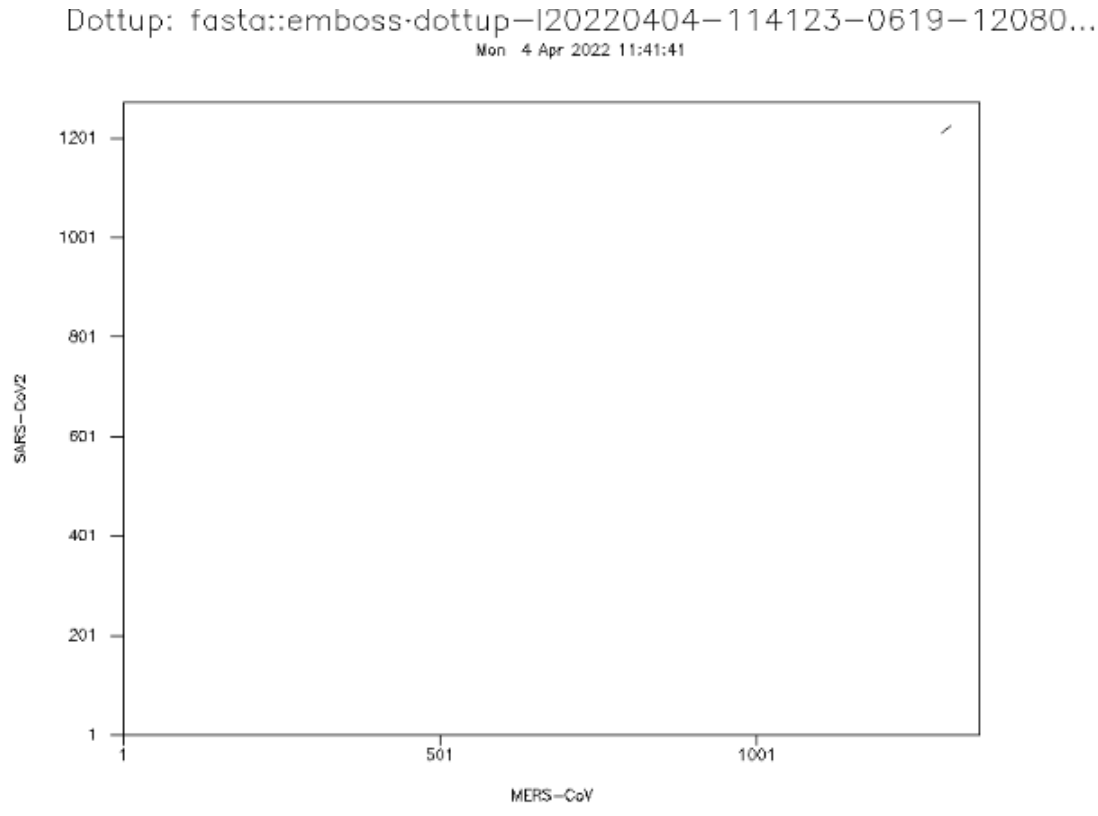
Analysis done on results from Dottup

DNA and Protein matches for MERS-CoV with SARS-CoV2 using Dottup :

DNA:


Dottup: fasta::emboss-dottup-I20220404-115300-0843-64951...
Mon 4 Apr 2022 11:49:04
Spike
Spike
1
501
1001
1501
2001
2501
3001
3501
4001

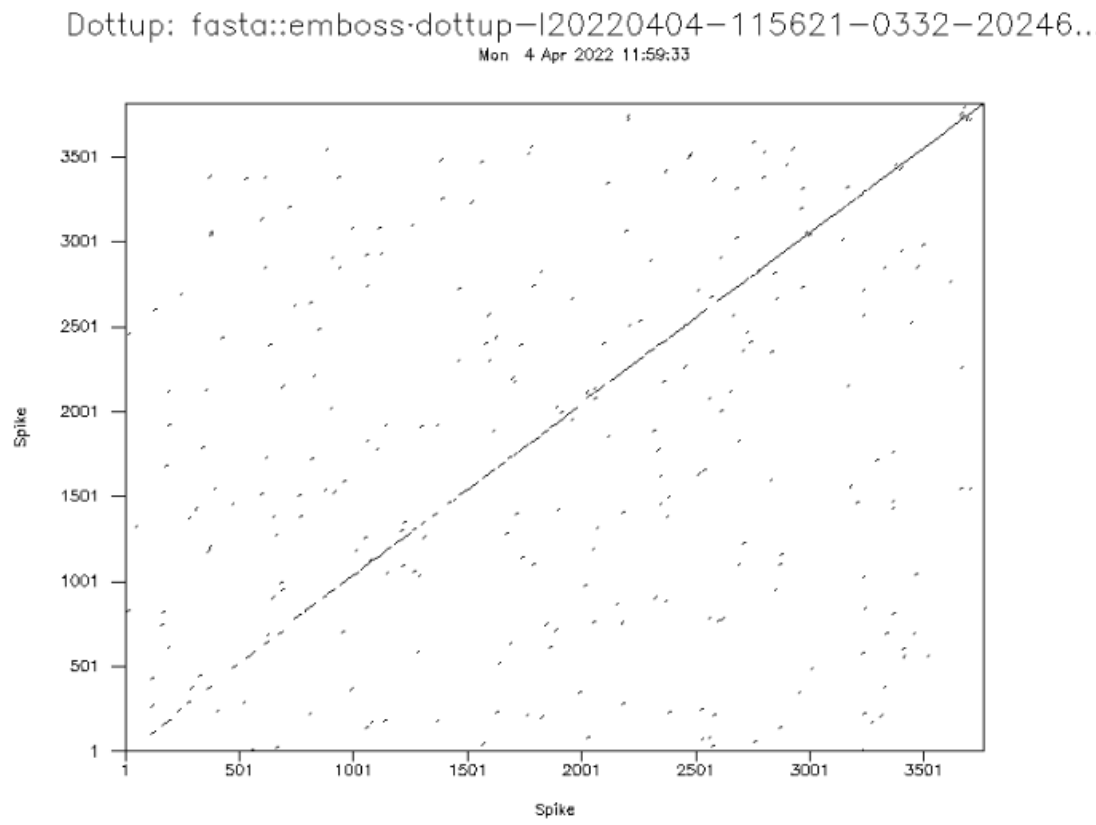
Protein:


Dottup: fasta::emboss-dottup-I20220404-114123-0619-12080...
Mon 4 Apr 2022 11:41:41
SARS-CoV2
MERS-CoV
1
201
401
601
801
1001
1201
501

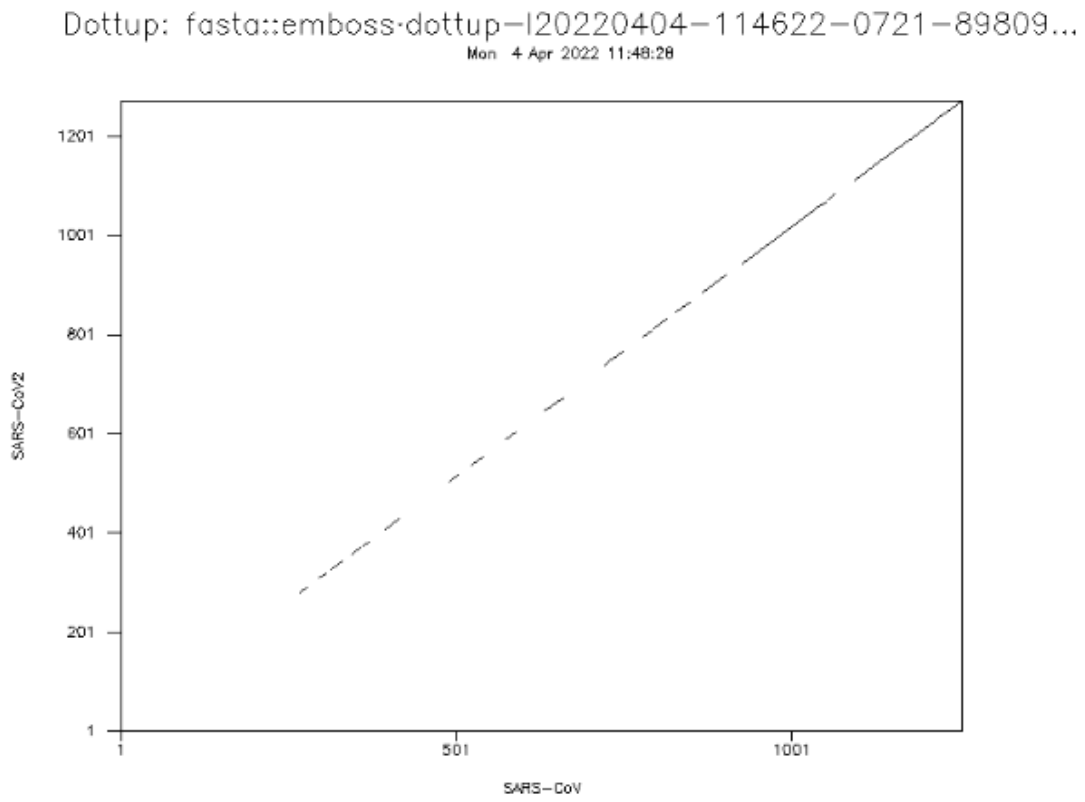
As we are unable to observe a line like plot in the above two we can say that these two viruses aren’t very similar

DNA and Protein matches for SARS-CoV with SARS-CoV2 using Dottup :

DNA:



Protein:



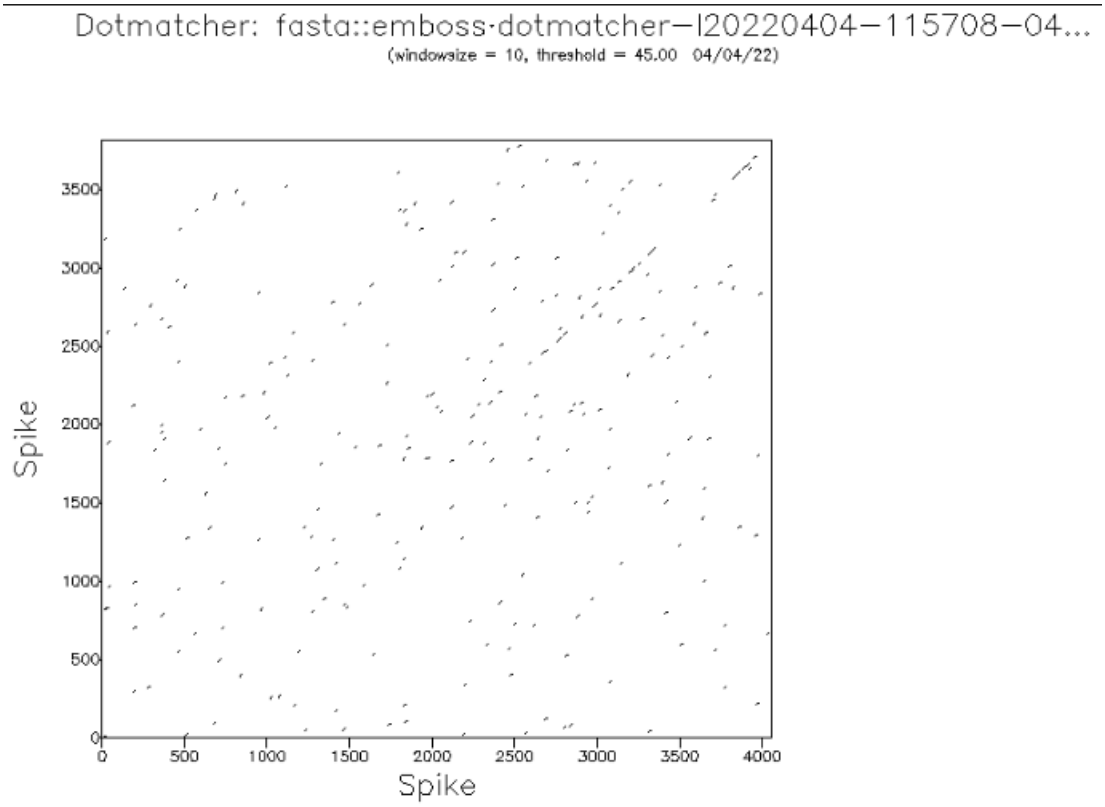
we observe straight line(for DNA match) and almost straight line for (Protein match)

We can hence say that these two viruses are similar

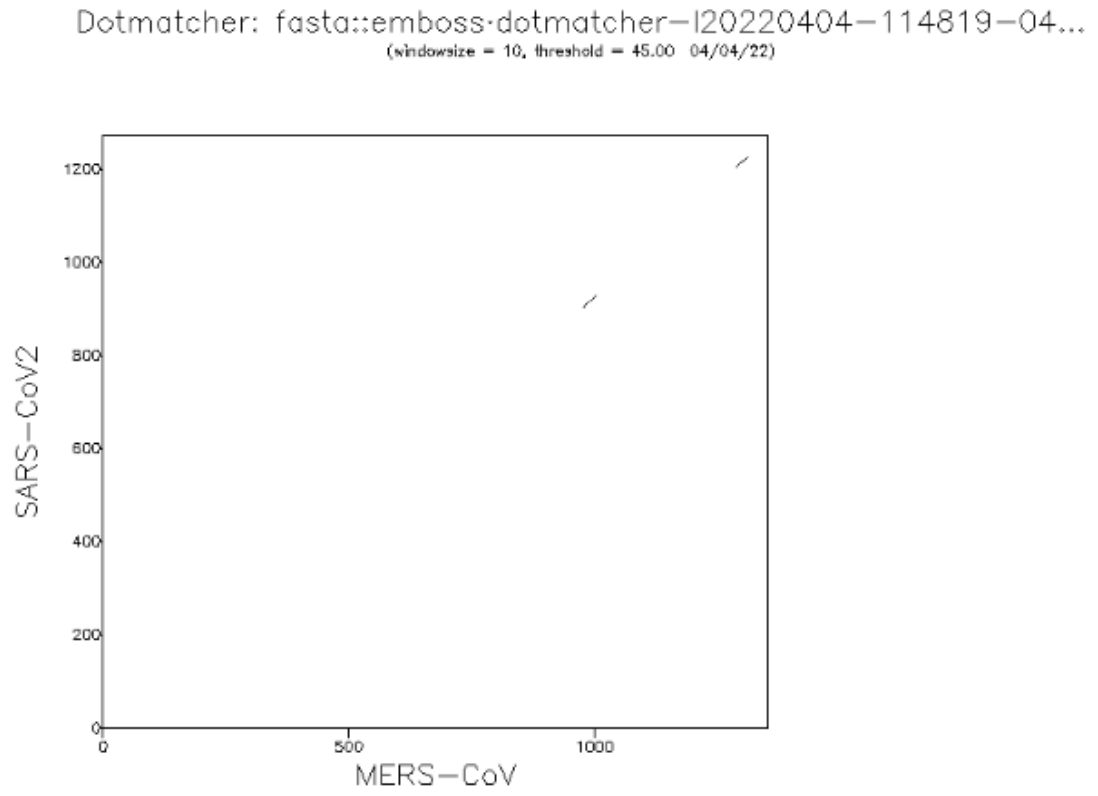
Analysis done on results from Dotmatcher

DNA and Protein matches for MERS-CoV and SARS-CoV2 using Dotmatcher :

DNA:


Dotmatcher: fasta::emboss-dotmatcher-I20220404-115708-04...
(windowsize = 10, threshold = 45.00 04/04/22)
Spike
Spike
0
500
1000
1500
2000
2500
3000
3500
4000

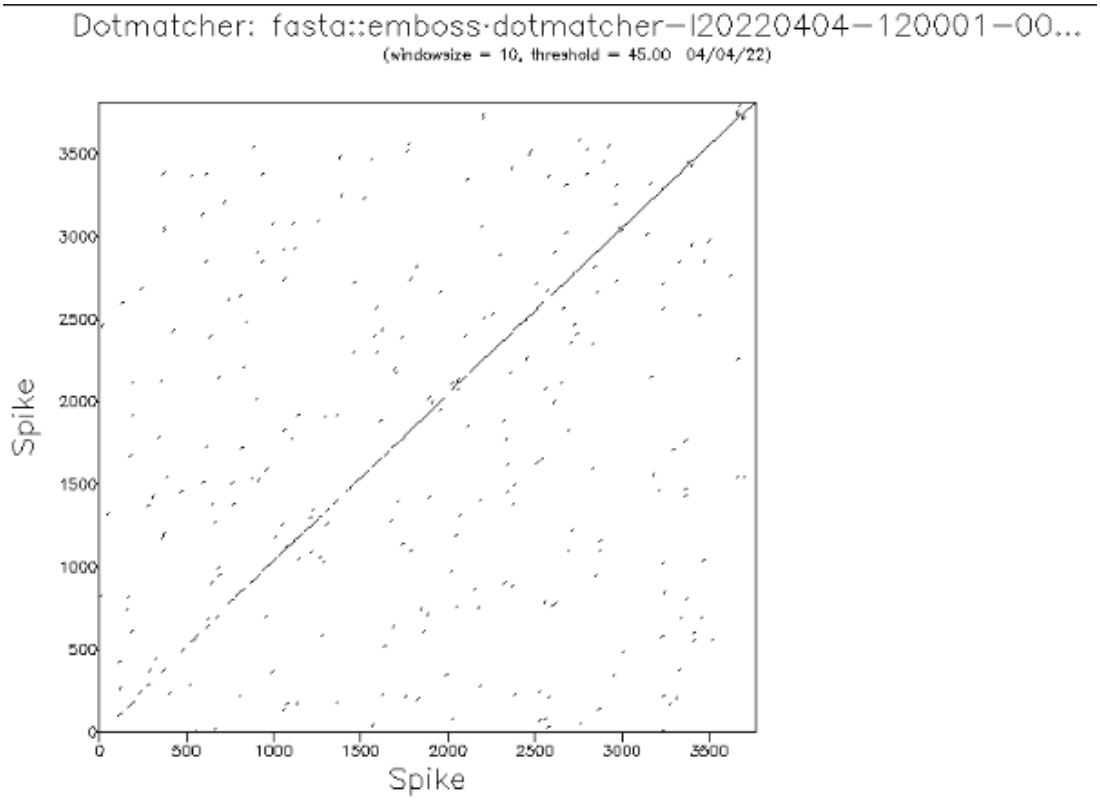
Protein:



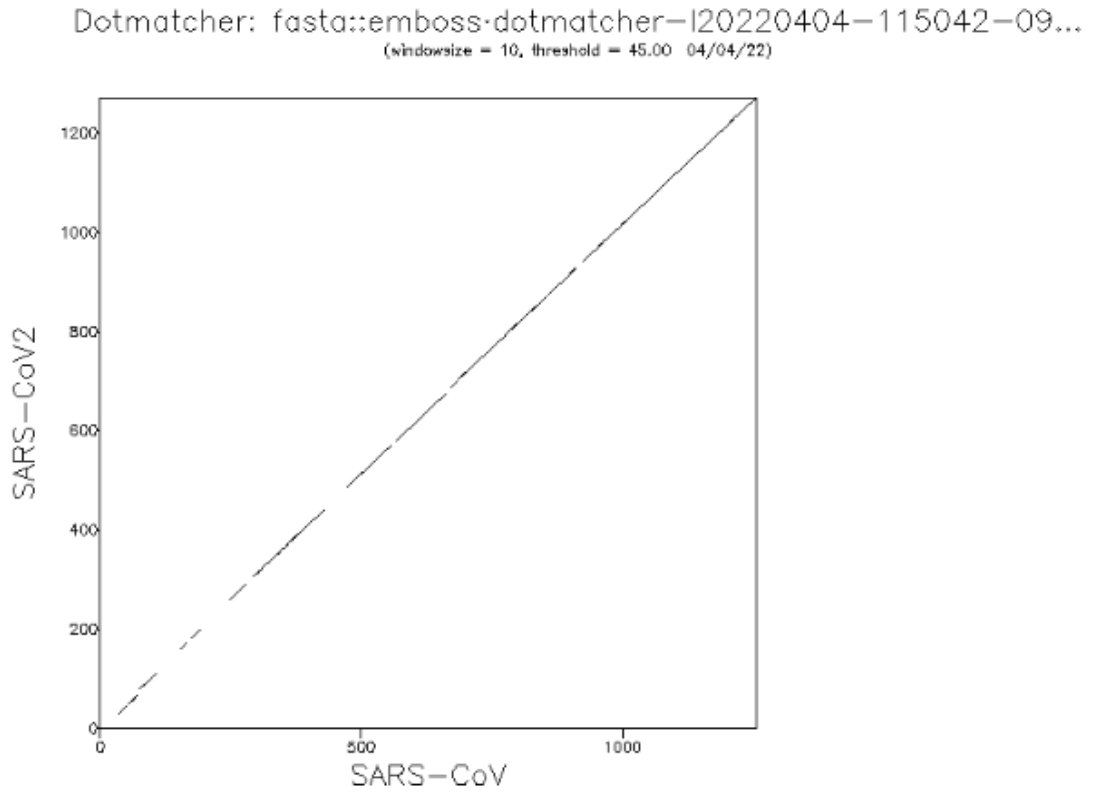
Except for the noise, these plots are very similar to the dottup plots hence pretty much same observation noted

DNA and Protein matches for SARS-CoV and SARS-CoV2 using Dotmatcher :

DNA:


Dotmatcher: fasta::emboss·dotmatcher−I20220404−120001−00...
(windowsize = 10, threshold = 45.00 04/04/22)
Spike
Spike
0
500
1000
1500
2000
2500
3000
3500

Protein:


Dotmatcher: fasta::emboss·dotmatcher−I20220404−115042−09...
(windowsize = 10, threshold = 45.00 04/04/22)
SARS−CoV2
SARS−CoV
0
200
400
600
800
1000
1200
500

Again similar to dottup

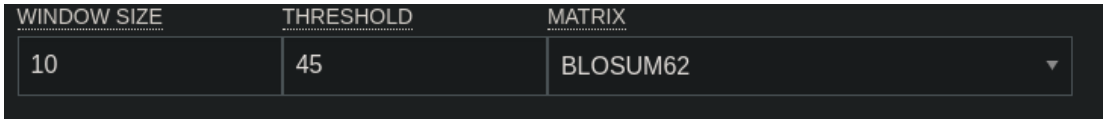
▼ Part 2

Protein sequences are easier to use because the DNA plots are noisier compared to their Protein counterparts for both SARS-Cov and MERS-CoV.

▼ Part 3

Dottup Protein and DNA:

Dotmatcher Protein:



Dotmatcher DNA:

▼ Question 2a

2. (a) **Pairwise Alignment:** Perform pairwise alignment of spike glycoprotein of SARS-CoV2 with that of SARS-CoV, both at the DNA and protein level, using programs 'needle' and 'water'. Answer the following Qs.
   (i) What is percentage identity and percentage similarity at DNA level and protein level? Which is larger and why, give reasons.
   (ii) What is the difference between the identity and similarity?
   (iii) Is there any difference in the global and local alignments of these two sequences?
   (iv) Submit the alignment giving the scoring scheme and gap penalties used.

▼ Part 1: SARS-CoV and SARS-CoV2

▼ Needle

DNA Level:

Identity: 3338/3902 (85.5%)

Similarity: 3338/3902 (85.5%)

Protein Level:

Identity: 974/1277 (76.3%)

Similarity: 1110/1277 (86.9%)

▼ Water

DNA Level:

Identity: 3339/3899 (85.6%)

Similarity: 3339/3899 (85.6%)

Protein Level:

Identity: 974/1277 (76.3%)

Similarity: 1110/1277 (86.9%)

- At DNA levels similarity and identity are similar
- At protein level, similarity is greater than identity
- Percentage Identity searches for exact matches between two strings, whereas Percentage Similarity searches for similarities.

- Protein sequences share a common ancestor and, as a result, share comparable charcteristics.
- The sequence's actual elements may differ, resulting in a lower identity %.

▼ Part 2: Difference between identity and similarity

- Identity is the number of characters that match exactly between two different sequences. Gaps are not considered here.
- similarity is the degree of resemblance between any 2 sequences on comparison. It depicts the length to which the sequences are aligned

▼ Part 3: difference in global and local alignments

Global Alignment is Needle
Local Alignment is Water

- Here, local alignment and global alignment are approximately the same at both DNA and protein level.

▼ Part 4: Scoring Scheme and gap penalties

Scoring Scheme : BLOSUM62

Gap penalty: 10.0

For more details, I have also uploaded the different analysis on needle and water with protein and DNA

▼ Question 2b

(b) Pairwise Alignment: Perform pairwise alignment of spike glycoprotein of SARS-CoV2 with that of MERS-COV virus, both at the DNA and protein level, using programs 'needle' and 'water'. Answer the following Qs.
(i) Based on the sequence alignment, can you say that the two proteins are homologs, i.e., related?
(ii) Are you able to make this inference from alignment of DNA sequences or protein sequences?

from Needle:

DNA:

```
# Identity:    2618/4543 (57.6%)
# Similarity:  2618/4543 (57.6%)
```

Protein:

```
# Identity:     434/1440 (30.1%)
# Similarity:   659/1440 (45.8%)
```

From Water:

DNA:

```
# Identity:    2619/4534 (57.8%)
# Similarity:  2619/4534 (57.8%)
```

Protein:

```
# Identity:     433/1426 (30.4%)
# Similarity:   656/1426 (46.0%)
```

a common rule of thumb is that two sequences are homologous if they are **more than 30% identical over their entire lengths**

Here we can see the percentage identity to be just over 30% hence we can declare them homolog.

Two proteins are homolog if they have a common ancestor

Since percentages are lower in protein, we can say the above results are from protein.

▼ Question 3

3. Database search: Perform DNA and protein database search using spike glycoprotein of SARS-CoV2 as query and answer the following Qs:
   (i) Which is the closest homolog of the query sequence?

(ii) Give the score, percentage identity, percentage similarity, length of the alignment, and the expect or e-value.
(iii) Do you find the spike glycoprotein of SARS-CoV as one of the hits? Does the percentage identity and percentage similarity results match with the alignment obtained using 'water'? What is the significance of this alignment?
(iv) It was speculated that SARS-CoV2 has come from bat. Do you find any relation of spike glycoprotein of SARS-CoV2 with that of bat SARS coronavirus spike glycoprotein? What is identity, similarity, length of alignment, score and e-value?

▼ Part 1

Spike Glycoprotein- Bat coronavirus RaTG13

▼ Part 2

- Score: 2565
- Percentage identity: 97%
- Percentage similarity: 98%
- Length of alignment: 1269
- E-value: 0.0

▼ Part 3

Yes, We did find spike glycoprotein of SARS-Cov and related nearby variants in the blastp.

There are multiple instances of Sars Coronavirus in the search

The values do match with that of EMBOSS water analysis.

The significance is that we can verify that SARS-CoV2 and SARS-CoV are homolog. There is an evolutionary relationship between them. and SARS-CoV2 is the nearest variant of SARS-CoV which infects humans.

▼ Part 4

Bat Coronavirus was found to be the closest homolog in Part 1

Hence values are:

- Score: 2565

- Percentage identity: 97%
- Percentage similarity: 98%
- Length of alignment: 1269
- E-value: 0.0

▼ Question 4

4. Find out the size of protein database, UniProt, and nucleotide database, GenBank. Compute No. of matrix cells to be computed using DP for:
   (i) Performing search in protein database, UniProt, and nucleotide database, GenBank and the time required assuming query sequence of length 1000 bases.
   (ii) Comparing Human Chr 1 ~249Mbp with a query sequence of 1000 bases using DP, and comparing it with Chr 1 of Mouse (~195Mbp)? What is the memory or space requirement in the two cases?

▼ Part 1

Size of Uniprot = `230328648 sequences = 80596791741 amino acids`

(as of 19 Jan 2022)

source: https://www.ebi.ac.uk/uniprot/TrEMBLstats

Size of GenBank = 236338284 sequences = 1173984081721 bases

(as of Feb 2022)

source: https://www.ncbi.nlm.nih.gov/genbank/statistics/

Length of query sequence = 1000 bases

No. of matrix cells to be computed for comparing a protein sequence:

= Total number of acids in UniProt * Length of the query sequence
= 80596791741 * 1000
= 80596791741000

No. of matrix cells to be computed for comparing a DNA sequence:

= Total number of bases in GenBank * Length of the query sequence
= 1173984081721 * 1000
= 1173984081721000

to get time required,

Number of iterations that would be made for comparing protein sequence:

= Total number of acids in UniProt * Length of the query sequence
= 80596791741 * 1000
= 80596791741000

Assume 10^7 iterations in 1 second,

total time taken: 8059679.1741

Number of iterations that would be made for comparing a DNA sequence:
= Total number of bases in GenBank * Length of the query sequence
= 1173984081721 * 1000
= 1173984081721000

Assume 10^7 iterations in 1 second,
Total time taken = 117398408.1721

▼ Part 2

For Human Chr 1:

m = 1000
n = 249Mbp = 249 * 10^6 * 2 = 498 * 10^6
Memory complexity = m * n = 498 * 10^9
Time complexity = m * n = 498 * 10^9

For Mouse Chr 1:

m = 1000
n = 195Mbp = 195 * 10^6 * 2 = 390 * 10^6
Memory complexity = m * n = 390 * 10^9
Time complexity = m * n = 390 * 10^9