Tutorial Assignment 1

Science 2

Kishore Kumar | 2020101126

Question 1

DotPlot Analysis using

(a) Dottup – k-tuple search

(b) Dotmatcher – sliding window

Given are the sequences of spike glycoprotein (both DNA and protein) of the following:

1. SARS-CoV (2003)
2. MERS-COV (2012)
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.

Dottup analysis

Parameters

- Word-size: 10
- Box it: Yes

Figure 1: DNA: Sars-Cov-2 vs Sars-Cov

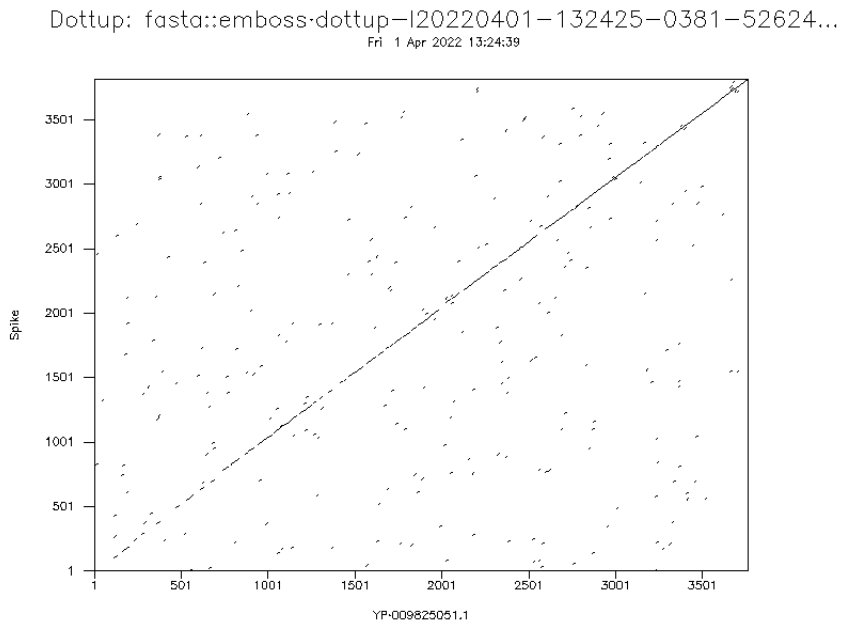


Figure 2: DNA: Sars-Cov-2 vs Mers-Cov

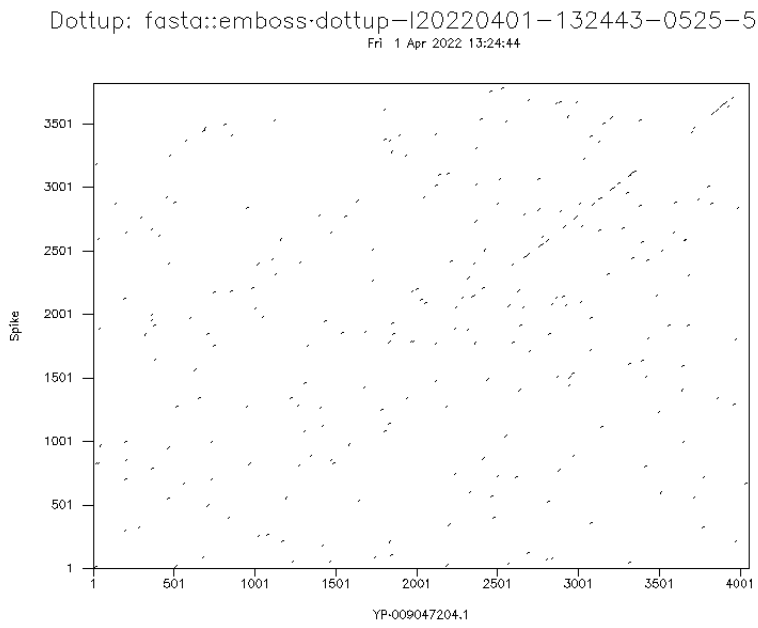


Figure 3: Protein: Sars-Cov-2 vs Sars-Cov

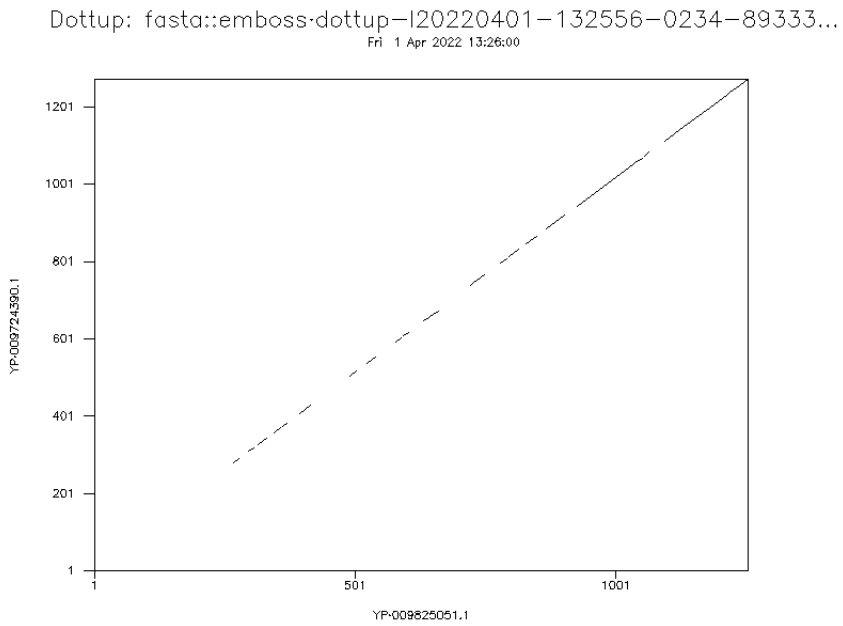
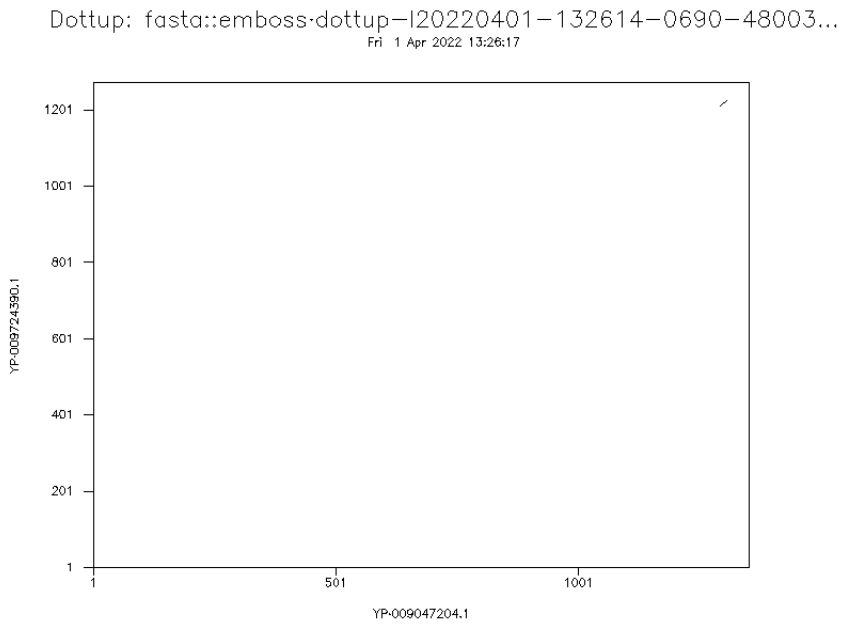


Figure 4: Protein: Sars-Cov-2 vs Mers-Cov



Dotmatcher Analysis

Parameters

- Window-size: 10

- Threshold: 23
- Matrix: DNAfull / BLOSUM62

Figure 5: DNA: Sars-Cov-2 vs Sars-Cov

Figure 6: DNA: Sars-Cov-2 vs Mers-Cov

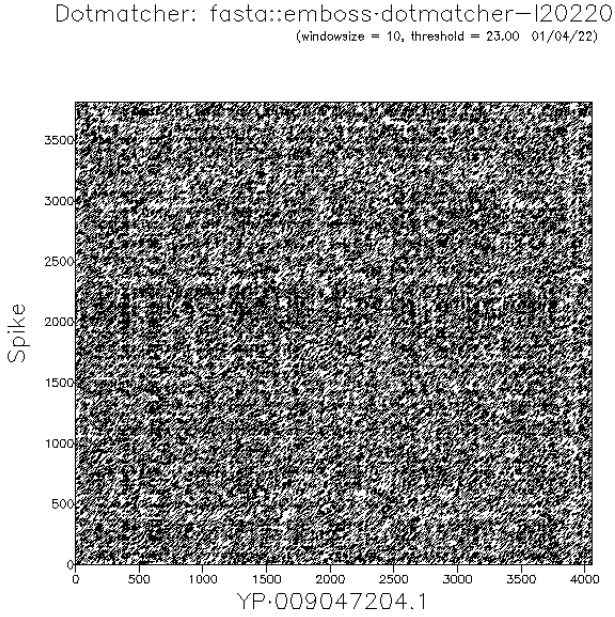


Figure 7: Protein: Sars-Cov-2 vs Sars-Cov

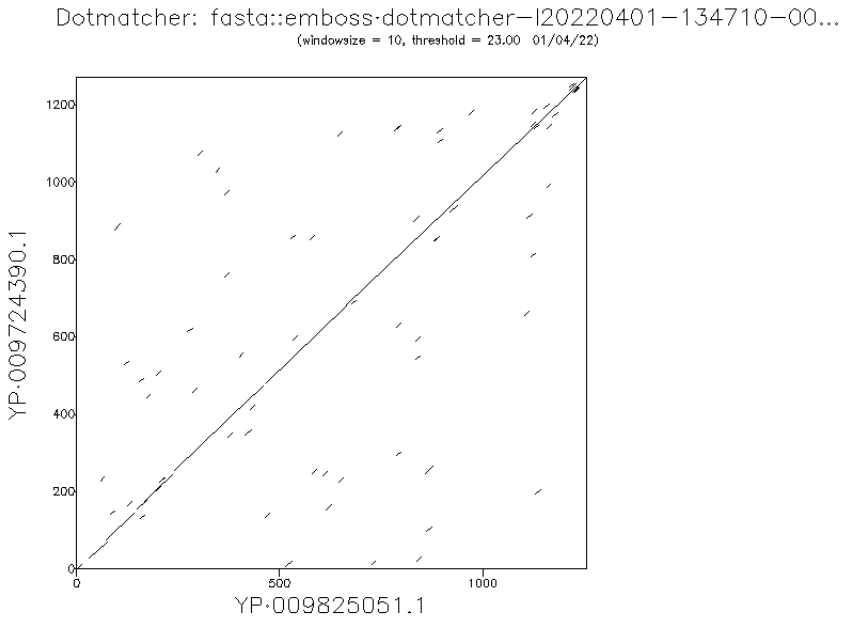
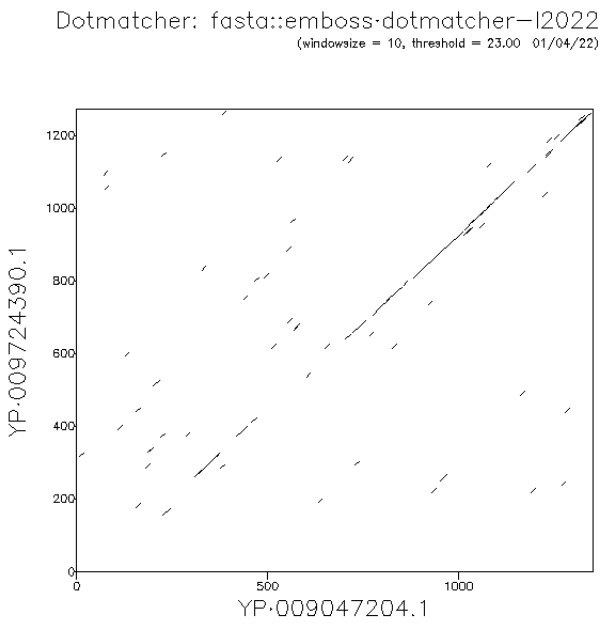


Figure 8: Protein: Sars-Cov-2 vs Mers-Cov



Conclusion

(i) **Identify SARS-CoV2 is similar to which of the earlier two viruses?** **Answer:** Observing the Dottup and Dotmatcher plots we generated, we can easily see

that SARS-CoV-2 has far more matches with SARS-CoV than it has with MERS-CoV. Hence we can say that SARS-CoV-2 is similar to SARS-CoV.

(ii) **Is it easy to identify the similarity using DNA or protein sequences? Give reasons.**
Answer: It is much easier to identify similarity using protein sequences instead of DNA sequences. There are multiple reasons for this, such as:

(a) DNA sequences require us to have an exact one-to-one match whereas where comparing protein sequences we can rely on using more sensitive matrices like BLOSUM

(b) This becomes more exaggerated when we consider the fact that multiple DNA sequences map to the same protein. Hence we find much more random matches when comparing DNA sequences than proteins.

(c) Proteins, on the whole, do not change during evolution. As a consequence, protein search shows distant evolutionary ties as well.

(iii) **Submit the graphs and give the k-tuple values used, and window size and threshold values used.**
Answer: The graphs and parameters used have been attached above.

Question 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.

(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?

(a) (i) **What is percentage identity and percentage similarity at DNA level and protein level? Which is larger and why, give reasons.**

Answer: Both percentage identity and percentage similarity is larger at the DNA level than at the protein level. The DNA sequence is longer than the amino acid sequence, in-fact, three nucleotides are equal to one amino acid. There are four nucleotides in a DNA sequence, hence there is a 25

(ii) **What is the difference between the identity and similarity?**
Answer: When we say identity we mean the **exact same** amino acid or nucleotide. On the other hand, similarity captures those amino acids or nucleotides which may not be the exact same but might be very similar.

(iii) **Is there any difference in the global and local alignments of these two sequences?**
Answer: The global and local alignments of these two sequences are nearly identical. This is due to the fact that the query and search string lengths are nearly identical.

(iv) **Submit the alignment giving the scoring scheme and gap penalties used.**
Answer: The necessary files have been uploaded here: Secret Github Gist

(b) (i) **Based on the sequence alignment, can you say that the two proteins are homologs, i.e., related?**
Answer: The proteins aren't homologous. This is due to the low percentages of identity and likeness that we observe.

(ii) **Are you able to make this inference from alignment of DNA sequences or protein sequences?**
Answer: Because the percent identity and similarity are so low, we can clearly conclude this more so from protein alignment than dna alignment.

Question 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?

(i) **Which is the closest homolog of the query sequence?**

Answer: Closest homolog: Spike glycoprotein of SARS-CoV.

(ii) **Give the score, percentage identity, percentage similarity, length of the alignment, and the expect or e-value.**
Answer:

(a) Protein search:

(i) Score: 2009.2

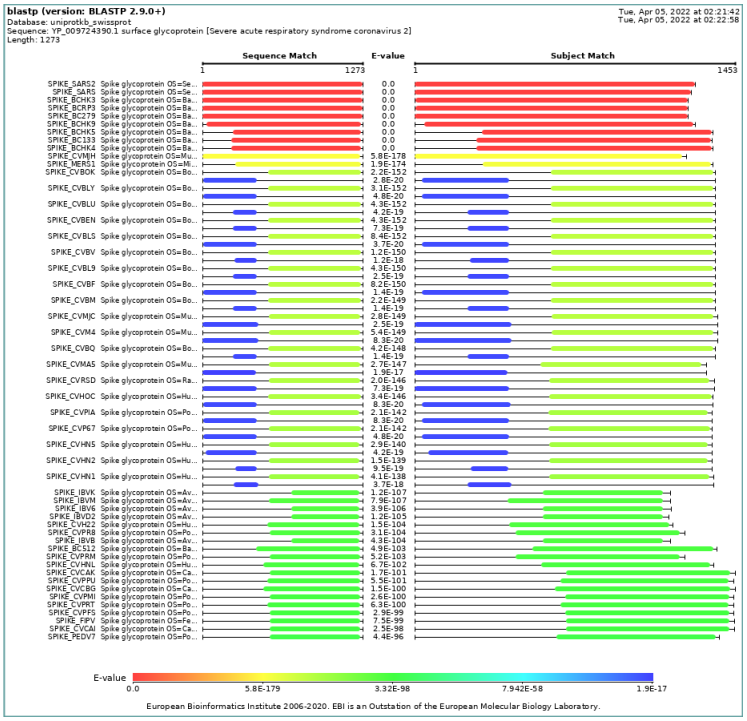
(ii) Percentage identity: 76.0

(iii) Percentage similarity: 86.8

(iv) Length of alignment: 1255

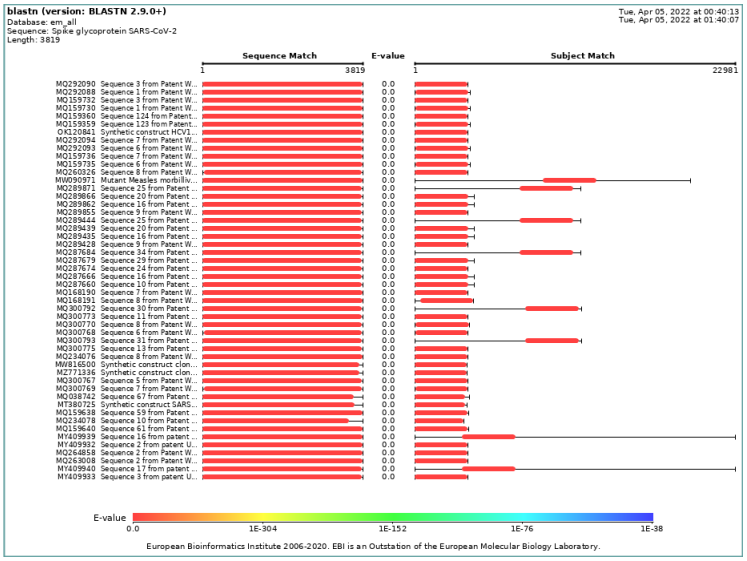
(v) E-value: 0.0

Figure 9: BLASTp query



(b) DNA search: Only showed results of vaccine patents.

Figure 10: BLASTn query



(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?**

Answer: Yes, the spike glycoprotein of SARS-CoV is in fact the best hit (excluding SARS-CoV-2 obviously). Yes, the percentage identity and percentage similarity results of our Blastp search does match with the results we obtained by running the water program on it.

(a) BLASTp query

(i) Percentage Identity: 76.0

(ii) Percentage Similarity: 86.8

(b) Water results

(i) Percentage Identity: 76.3

(ii) Percentage Similarity: 87.0

We can conclude that SARS-CoV-2 and SARS-CoV are very near homologs since there is significant alignment between the two viruses.

(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?**

Answer: Yes I do. In fact, in our database query, the spike glycoprotein of Bat Coronavirus (RaTG13) is the second closest homolog match.

(i) Percentage identity: 76.0

(ii) Percentage similarity: 85.4

(iii) Length of alignment: 1242

(iv) Score: 1961.8

(v) E-value: 0.0

Question 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

(i) **Find out the size of protein database, UniProt, and nucleotide database, GenBank.**
Answer:

(a) UniProt (As of Jan 19th 2022): 80596791741 amino acids.

(b) GenBank (As of Feb 2022): 236338284 sequences and 1173984081721 nucleotides.

(ii) **Performing search in protein database, UniProt, and nucleotide database, GenBank and the time required assuming query sequence of length 1000 bases.**
Answer:

(a) UniProt

- Required matrix size: 80596791741 $\times 1000$
- Assumed calcs per second possible: 10^8
- Time required: 805967 seconds $\implies \approx 9.3$ days

(b) GenBank

- Required matrix size: 1173984081721 $\times 1000$
- Assumed calcs per second possible: 10^8
- Time required: 11739840 seconds $\implies \approx 135.8$ days

(iii) **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**
Answer:

(a) Human chromosome: Requires matrix of size $249 \times 10^6 \times 1000 = 2.49 \times 10^{11}$

(b) Mouse chromosome: Requires matrix of size $195 \times 10^6 \times 1000 = 1.95 \times 10^{11}$