# Question 1

## (i)

SARS-CoV2 is similar to SARS-CoV1 rather than MERS-COV. This can be observed from the plots below.

## (ii)

It is easier to identify the similarity using protein sequences rather than DNA sequences. DNA only consists of 4 components, so the chance of accidentally being similar is much higher. On the other hand, there are 20 amino acids. Thus, only similar proteins will have similar protein sequences.

## (iii)

**Dottup Parameters**

- Word Size (k-tuple) = 10

**Dotmatcher Parameters**

- Window Size = 15
- Threshold Value = 50

**Plots**

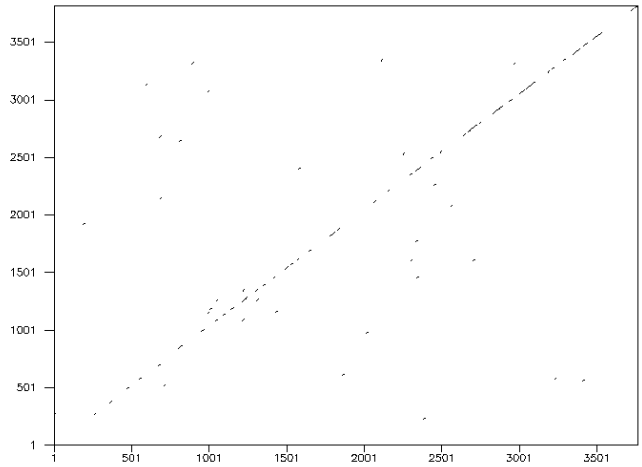


Figure 1: Dottup SARSCoV-SARSCoV2 DNA

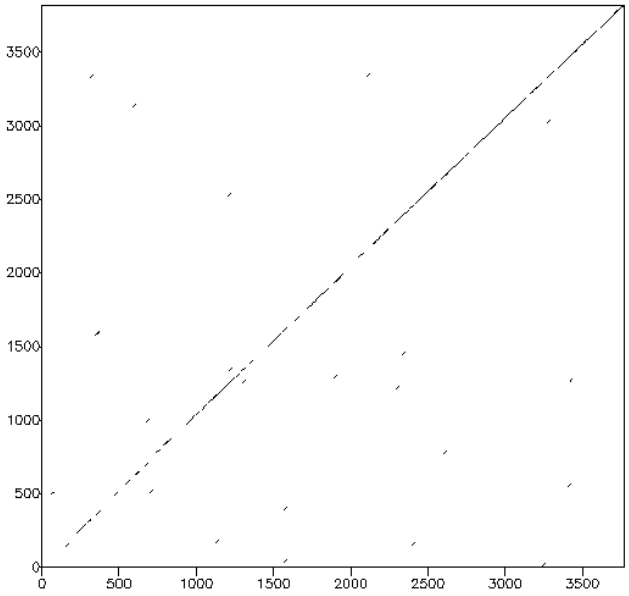


Figure 2: Dotmatcher SARSCoV-SARSCoV2 DNA

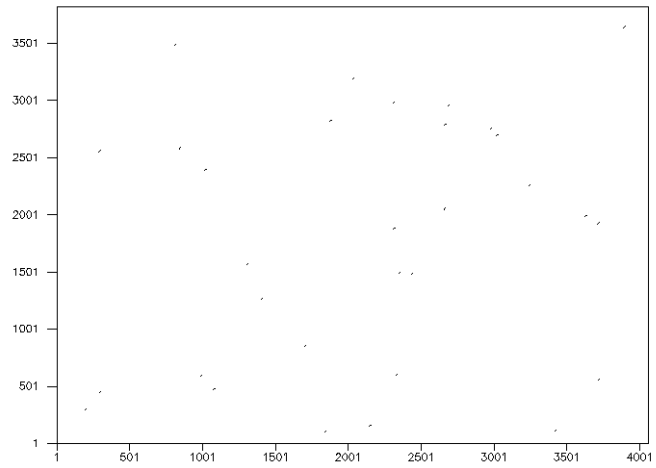


Figure 3: Dottup MERSCoV-SARSCoV2 DNA

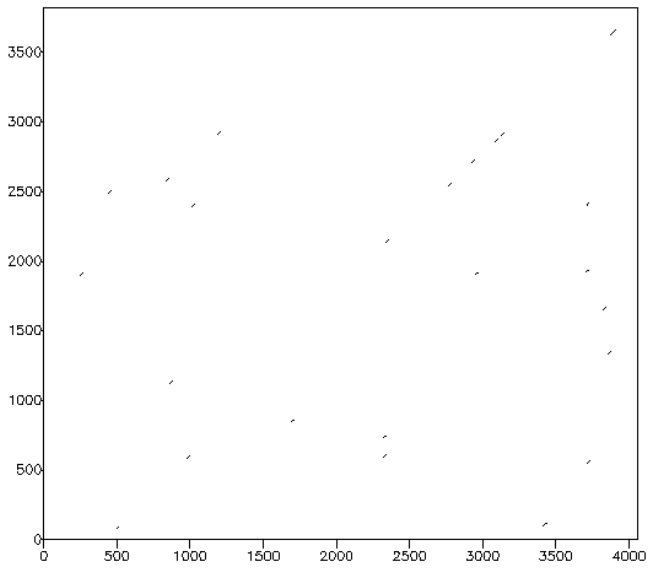


Figure 4: Dotmatcher MERSCoV-SARSCoV2 DNA

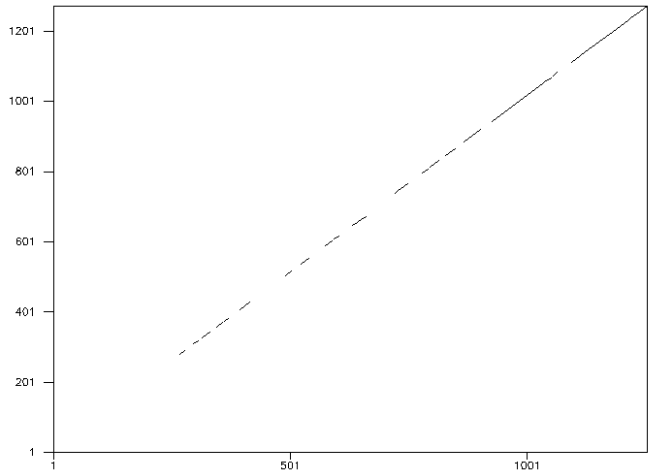


Figure 5: Dottup SARSCoV-SARSCoV2 Protein

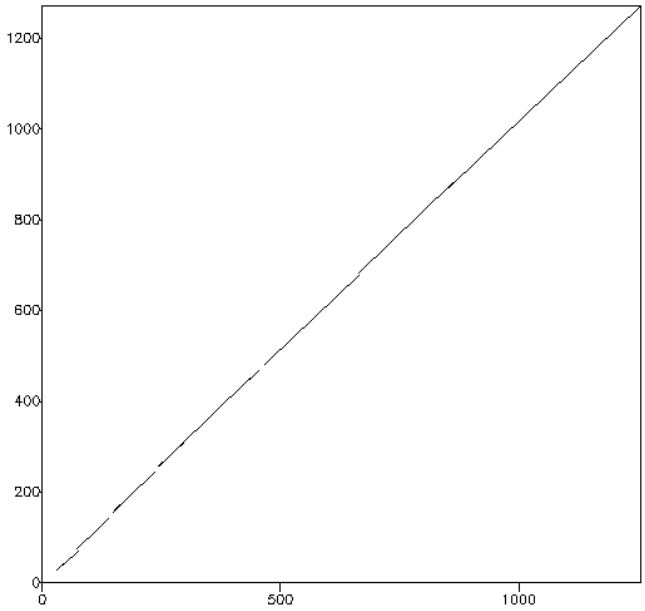


Figure 6: Dotmatcher SARSCoV-SARSCoV2 Protein

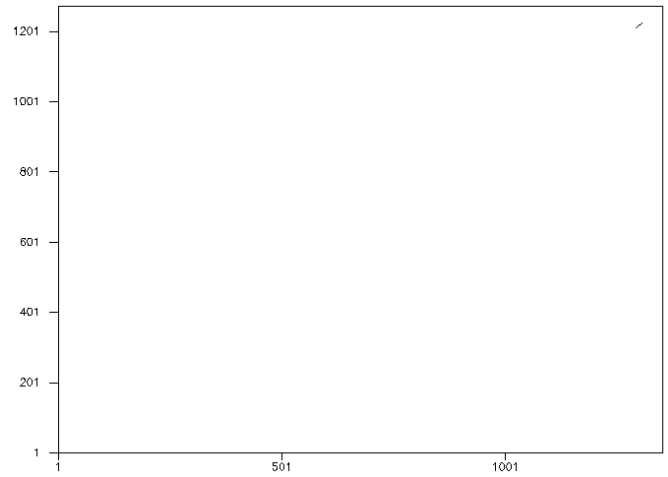


Figure 7: Dottup MERSCoV-SARSCoV2 Protein

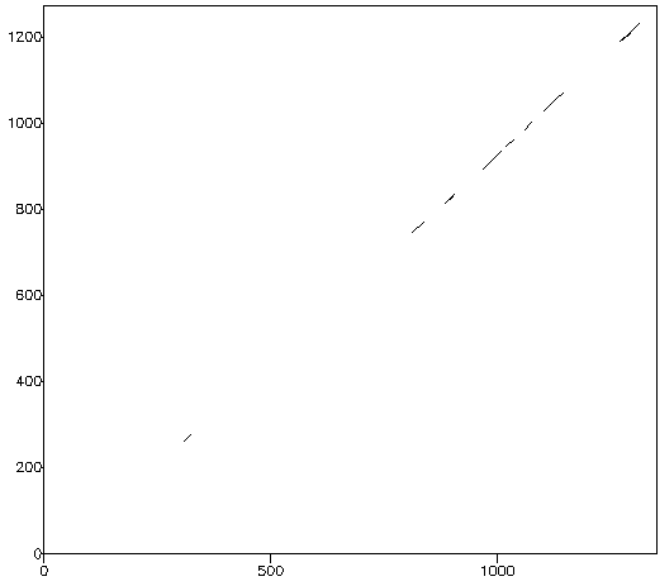


Figure 8: Dotmatcher MERSCoV-SARSCoV2 Protein

# Question 2

## (a) SARS-CoV vs SARS-CoV2

**(i)**

**Needle:**

| | **Identity** | **Similarity** |
|---|---|---|
| **DNA** | 72.8% | 72.8% |
| **Protein** | 76.4% | 87% |

**Water:**

| | **Identity** | **Similarity** |
|---|---|---|
| **DNA** | 73.3% | 73.3% |
| **Protein** | 76.4% | 87% |

The identity and similarity percentages of DNA are the same because the nucleotides either match, or they don't. Due to this, identity and similarity mean the same in this case.

BLOSUM62 assigns positive scores to conservative substitutions. Therefore the score for similarity is higher than the score for identity in case of proteins. It includes both substituions and exact matches.

**(ii)**

Identity refers to the exact matches between subsequences at a particular position, while similarity refers considers for exact matches as well as substitutions. Similarity accounts for conservative

substitutions, i.e, substitutions that don't significantly alter the functioning and structure of the protein, while identity does not account for it.

**(iii)**

Global alignment attempts to align the entire length of the sequence without any gaps at the ends, but local alignment attempts to identify subsequences of high similarity while allowing gaps at the ends.

In context of SARS-CoV and SARS-CoV2, for the case of DNA sequence, the global and local alignments are slightly different. In case of protein sequence, the global and local alignments are the same.

**(iv)**

**Parameters:**

- Matrix: BLOSUM62
- Gap Open: 10
- Gap Extend: 0.5

```
########################################
# Program: needle
# Rundate: Fri  5 Apr 2024 14:11:39
# Commandline: needle
#    -auto
#    -stdout
#    -asequence emboss_needle-I20240405-141132-0407-36772968-p1m.asequence
#    -bsequence emboss_needle-I20240405-141132-0407-36772968-p1m.bsequence
#    -datafile EDNAFULL
#    -gapopen 10.0
#    -gapextend 0.5
#    -endopen 10.0
#    -endextend 0.5
#    -aformat3 pair
#    -snucleotide1
#    -snucleotide2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: DQ231462.2
# 2: 21563-25384
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 3951
# Identity:    2876/3951 (72.8%)
# Similarity:  2876/3951 (72.8%)
# Gaps:         312/3951 ( 7.9%)
# Score: 10412.0
#
#
#=======================================
```

Figure 9: Needle SARSCoV-SARSCoV2 DNA

```
########################################
# Program: water
# Rundate: Fri  5 Apr 2024 14:16:06
# Commandline: water
#    -auto
#    -stdout
#    -asequence emboss_water-I20240405-141522-0943-76901741-p1m.asequence
#    -bsequence emboss_water-I20240405-141522-0943-76901741-p1m.bsequence
#    -datafile EDNAFULL
#    -gapopen 10.0
#    -gapextend 0.5
#    -aformat3 pair
#    -snucleotide1
#    -snucleotide2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: DQ231462.2
# 2: 21563-25384
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 3933
# Identity:    2882/3933 (73.3%)
# Similarity:  2882/3933 (73.3%)
# Gaps:         276/3933 ( 7.0%)
# Score: 10412.0
#
#
#=======================================
```

Figure 10: Water SARSCoV-SARSCoV2 DNA

```
#######################################
# Program: needle
# Rundate: Fri  5 Apr 2024 14:12:41
# Commandline: needle
#    -auto
#    -stdout
#    -asequence emboss_needle-I20240405-141217-0898-48581873-p1m.asequence
#    -bsequence emboss_needle-I20240405-141217-0898-48581873-p1m.bsequence
#    -datafile EDNAFULL
#    -gapopen 10.0
#    -gapextend 0.5
#    -endopen 10.0
#    -endextend 0.5
#    -aformat3 pair
#    -snucleotide1
#    -snucleotide2
# Align_format: pair
# Report_file: stdout
#######################################

#=======================================
#
# Aligned_sequences: 2
# 1: 21431-25492
# 2: 21563-25384
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 4852
# Identity:    2340/4852 (48.2%)
# Similarity:  2340/4852 (48.2%)
# Gaps:        1820/4852 (37.5%)
# Score: 4469.0
#
#
#=======================================
```

Figure 11: Needle MERSCoV-SARSCoV2 DNA

```
#######################################
# Program: water
# Rundate: Fri  5 Apr 2024 14:17:28
# Commandline: water
#    -auto
#    -stdout
#    -asequence emboss_water-I20240405-141704-0972-73858133-p1m.asequence
#    -bsequence emboss_water-I20240405-141704-0972-73858133-p1m.bsequence
#    -datafile EDNAFULL
#    -gapopen 10.0
#    -gapextend 0.5
#    -aformat3 pair
#    -snucleotide1
#    -snucleotide2
# Align_format: pair
# Report_file: stdout
#######################################

#=======================================
#
# Aligned_sequences: 2
# 1: 21431-25492
# 2: 21563-25384
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 4833
# Identity:    2340/4833 (48.4%)
# Similarity:  2340/4833 (48.4%)
# Gaps:        1796/4833 (37.2%)
# Score: 4470.5
#
#
#=======================================
```

Figure 12: Water MERSCoV-SARSCoV2 DNA

```
#######################################
# Program: needle
# Rundate: Fri  5 Apr 2024 14:13:43
# Commandline: needle
#    -auto
#    -stdout
#    -asequence emboss_needle-I20240405-141336-0042-78864172-p1m.asequence
#    -bsequence emboss_needle-I20240405-141336-0042-78864172-p1m.bsequence
#    -datafile EBLOSUM62
#    -gapopen 10.0
#    -gapextend 0.5
#    -endopen 10.0
#    -endextend 0.5
#    -aformat3 pair
#    -sprotein1
#    -sprotein2
# Align_format: pair
# Report_file: stdout
#######################################

#=======================================
#
# Aligned_sequences: 2
# 1: ABB29898.2
# 2: YP_009724390.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 1277
# Identity:     975/1277 (76.4%)
# Similarity:  1111/1277 (87.0%)
# Gaps:          26/1277 ( 2.0%)
# Score: 5230.0
#
#
#=======================================
```

Figure 13: Needle SARSCoV-SARSCoV2 Protein

```
#######################################
# Program: water
# Rundate: Fri  5 Apr 2024 14:18:55
# Commandline: water
#    -auto
#    -stdout
#    -asequence emboss_water-I20240405-141842-0589-24054666-p1m.asequence
#    -bsequence emboss_water-I20240405-141842-0589-24054666-p1m.bsequence
#    -datafile EBLOSUM62
#    -gapopen 10.0
#    -gapextend 0.5
#    -aformat3 pair
#    -sprotein1
#    -sprotein2
# Align_format: pair
# Report_file: stdout
#######################################

#=======================================
#
# Aligned_sequences: 2
# 1: ABB29898.2
# 2: YP_009724390.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 1277
# Identity:     975/1277 (76.4%)
# Similarity:  1111/1277 (87.0%)
# Gaps:          26/1277 ( 2.0%)
# Score: 5230.0
#
#
#=======================================
```

Figure 14: Water SARSCoV-SARSCoV2 Protein

```
########################################
# Program: needle
# Rundate: Fri  5 Apr 2024 14:14:37
# Commandline: needle
#    -auto
#    -stdout
#    -asequence emboss_needle-I20240405-141435-0836-45938778-p1m.asequence
#    -bsequence emboss_needle-I20240405-141435-0836-45938778-p1m.bsequence
#    -datafile EBLOSUM62
#    -gapopen 10.0
#    -gapextend 0.5
#    -endopen 10.0
#    -endextend 0.5
#    -aformat3 pair
#    -sprotein1
#    -sprotein2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: AXP07355.1
# 2: YP_009724390.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 1455
# Identity:     436/1455 (30.0%)
# Similarity:   664/1455 (45.6%)
# Gaps:         284/1455 (19.5%)
# Score: 1567.5
#
#
#=======================================
```

Figure 15: Needle MERSCoV-SARSCoV2 Protein

```
########################################
# Program: water
# Rundate: Fri  5 Apr 2024 14:20:10
# Commandline: water
#    -auto
#    -stdout
#    -asequence emboss_water-I20240405-141949-0845-42971440-p1m.asequence
#    -bsequence emboss_water-I20240405-141949-0845-42971440-p1m.bsequence
#    -datafile EBLOSUM62
#    -gapopen 10.0
#    -gapextend 0.5
#    -aformat3 pair
#    -sprotein1
#    -sprotein2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: AXP07355.1
# 2: YP_009724390.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 1391
# Identity:     425/1391 (30.6%)
# Similarity:   646/1391 (46.4%)
# Gaps:         260/1391 (18.7%)
# Score: 1570.5
#
#
#=======================================
```

Figure 16: Water MERSCoV-SARSCoV2 Protein

## (b) MERS-COV vs SARS-CoV2

**Needle:**

| | **Identity** | **Similarity** |
|---|---|---|
| **DNA** | 48.2% | 48.2% |
| **Protein** | 30% | 45.6% |

**Water:**

| | **Identity** | **Similarity** |
|---|---|---|
| **DNA** | 48.4% | 48.4% |
| **Protein** | 30.6% | 46.4% |

**(i)**

No, the two proteins are not homologs. This is because MERS-COV and SARS-CoV2 have a relatively low percentage of identity and similarity, as compared to that between SARS-CoV and SARS-CoV2, which are homologs.

**(ii)**

As discussed in question 1 part (ii), protein sequences present a better measure of similarity rather than DNA sequences, although ideally both the sequences should be considered. Thus, majorly on the basis of protein sequences, I have made the inference that MERS-COV and SARS-CoV2 are not homologs.

# Question 3

DNA Results and Protein Results

### (i)
The closest homolog of the query sequence is *Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) (SARS-CoV)*

### (ii)
- Score: 5204
- Percentage Identity: 76%
- Percentage Similarity: 87%
- Length: 1255
- E-Value: 0

### (iii)
Yes, SARS-CoV is one of the hits, and the percentage identity and percentage similarity results do match with the alignments obtained using `water`.

The significance of this is that both Smith-Waterman and BLAST utilize local alignment algorithms, hence outputting almost the same values for similarities.

### (iv)
*Bat coronavirus HKU3 (BtCoV) (SARS-like coronavirus HKU3)* Spike glycoprotein is also a homolog.
- Score: 5081
- Percentage Identity: 76%
- Percentage Similarity: NA
- Length: 1242
- E-Value: 0

## Question 4

### Protein Database
Source
- 571282 reviewed proteins and 248234451 unreviewed proteins are present in UniProt.
- Thus, the total size is 248805733 proteins in UniProt
- 87,574,368,369 amino acids

### Nucleotide Database
Source
- 249060436 sequences, 2570711588044 bases

### (i)
For search in protein database, we multiply the number of amino acids in the database with the number of amino acids in the query sequence.

$$87574368369 * \frac{1000}{3} = 29191456123000$$

For search in nucleotide database, we multiply the number of nucleotides in the database with the number of nucleotides in the query sequence.

$$249060436 * 1000 = 249060436000$$

Assuming $10^7$ cells are computed per second, the time required is $33.78$ days and $6.91$ hours respectively.

### (ii)

For Human Chr 1 and a query sequence of 1000 bases, the number of matrix cells required is

$$2.49 * 10^8 * 10^3 = 2.49 * 10^{11}$$

For Human Chr 1 and Mouse Chr 1, the number of matrix cells required is

$$2.49 * 10^8 * 1.95 * 10^8 = 4.8555 * 10^{16}$$

Therefore, assuming the memory required for each cell is 1 unit, the memory required for these matrices respectively is $2.49 * 10^{11}$ units and $4.8555 * 10^{16}$ units.

## Question 5

```
import requests


def find_protein_by_taxonomy_id(taxonomy_id: int) -> None:
    url1 = f"https://rest.uniprot.org/uniprotkb/search?query=%28protein_name%3AInsulin%29+AND+%
28taxonomy_id%3A{taxonomy_id}%29"

    resp1 = requests.get(url1)
    results = resp1.json()

    print(f"Total number of protein entries: {resp1.headers.get('X-Total-Results')}")

    accession_ids = [entry["primaryAccession"] for entry in results["results"]]
    print("Accession IDs:", accession_ids)

    first_id = accession_ids[0]
    url2 = f"https://www.uniprot.org/uniprot/{first_id}.fasta"
    resp2 = requests.get(url2)
    with open(f"{first_id}.fasta", "wb") as fasta_file:
        fasta_file.write(resp2.content)
    print("Fasta downloaded!")


if __name__ == "__main__":
    find_protein_by_taxonomy_id(9606)
```