ASSIGNMENT TUTORIAL 1

- **ROMICA RAISINGHANI (2021101053)**

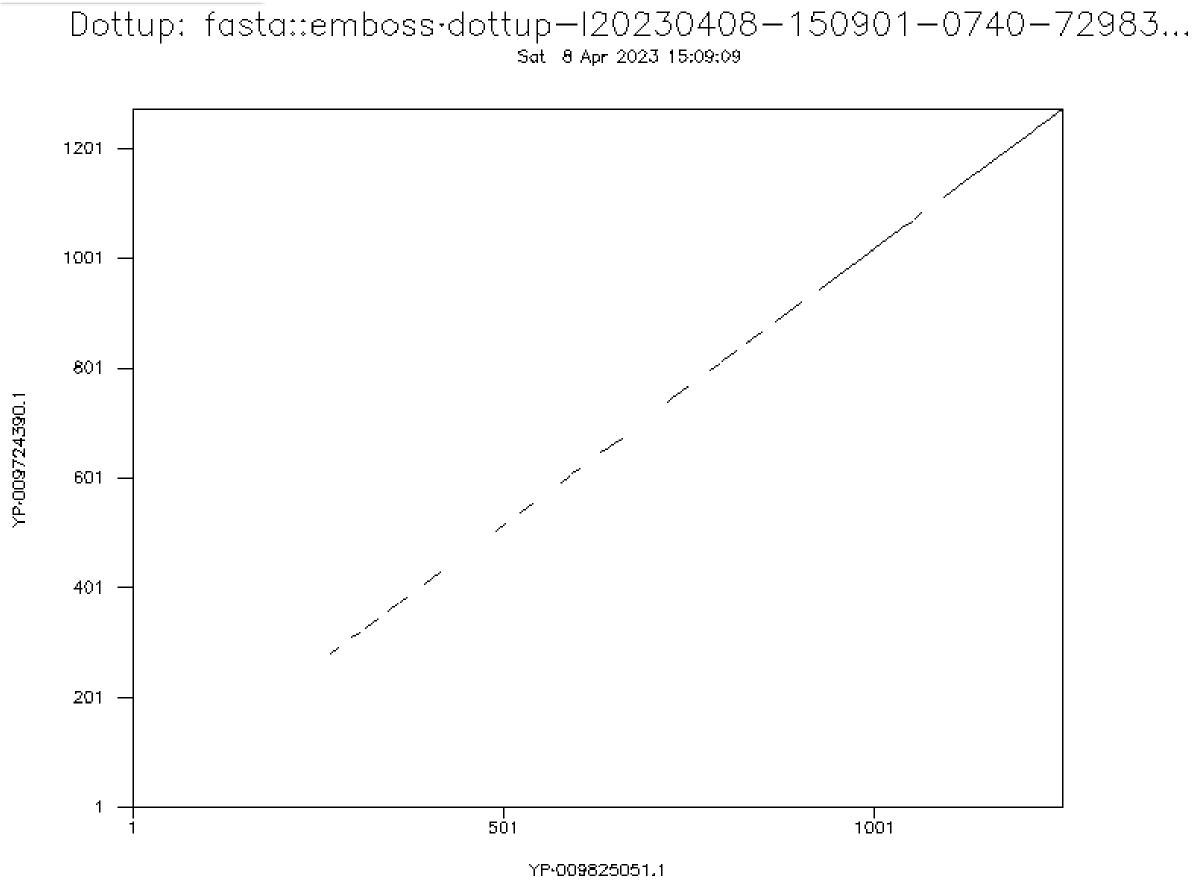
Ques1)

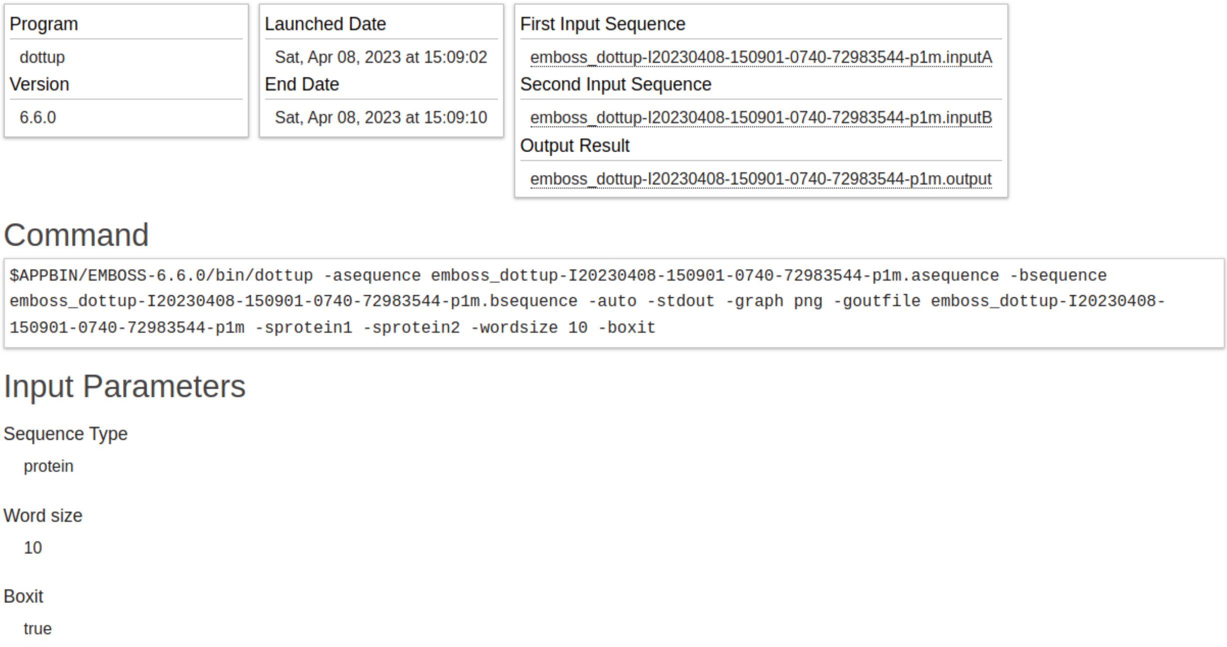
In Dotmatcher it compares amino acids even if there are not exactly same, whereas in Dottup its a binary value: 1- completely same, 0- completely distinct.

Dottup Plots(Proteins) :

1) Sars-Cov-2 vs Sars-Cov

K-tuple: 3





Program

dottup

Version

6.6.0

Launched Date

Sat, Apr 08, 2023 at 15:09:02

End Date

Sat, Apr 08, 2023 at 15:09:10

First Input Sequence

emboss_dottup-I20230408-150901-0740-72983544-p1m.inputA

Second Input Sequence

emboss_dottup-I20230408-150901-0740-72983544-p1m.inputB

Output Result

emboss_dottup-I20230408-150901-0740-72983544-p1m.output

Command

```
$APPBIN/EMBOSS-6.6.0/bin/dottup -asequence emboss_dottup-I20230408-150901-0740-72983544-p1m.asequence -bsequence
emboss_dottup-I20230408-150901-0740-72983544-p1m.bsequence -auto -stdout -graph png -goutfile emboss_dottup-I20230408-
150901-0740-72983544-p1m -sprotein1 -sprotein2 -wordsize 10 -boxit
```

Input Parameters

Sequence Type

protein

Word size

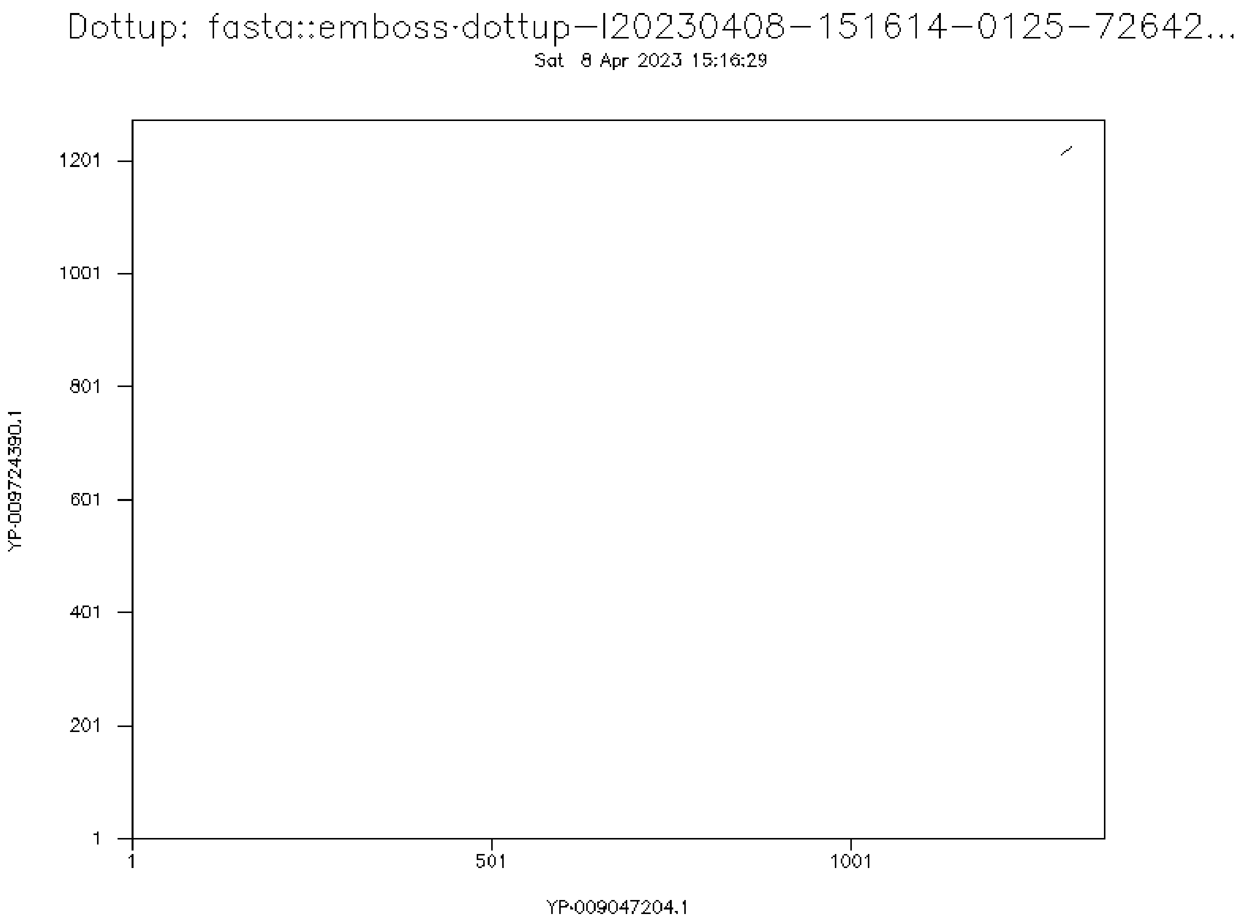
10

Boxit

true

2) Sars-Cov-2 vs Mers

K-tuple: 3



Program

dottup

Version

6.6.0

Launched Date

Sat, Apr 08, 2023 at 15:16:15

End Date

Sat, Apr 08, 2023 at 15:16:31

First Input Sequence

emboss_dottup-I20230408-151614-0125-72642090-p1m.inputA

Second Input Sequence

emboss_dottup-I20230408-151614-0125-72642090-p1m.inputB

Output Result

emboss_dottup-I20230408-151614-0125-72642090-p1m.output

Command

```
$APPBIN/EMBOSS-6.6.0/bin/dottup -asequence emboss_dottup-I20230408-151614-0125-72642090-p1m.asequence -bsequence
emboss_dottup-I20230408-151614-0125-72642090-p1m.bsequence -auto -stdout -graph png -goutfile emboss_dottup-I20230408-
151614-0125-72642090-p1m -sprotein1 -sprotein2 -wordsize 10 -boxit
```

Input Parameters

Sequence Type

protein

Word size

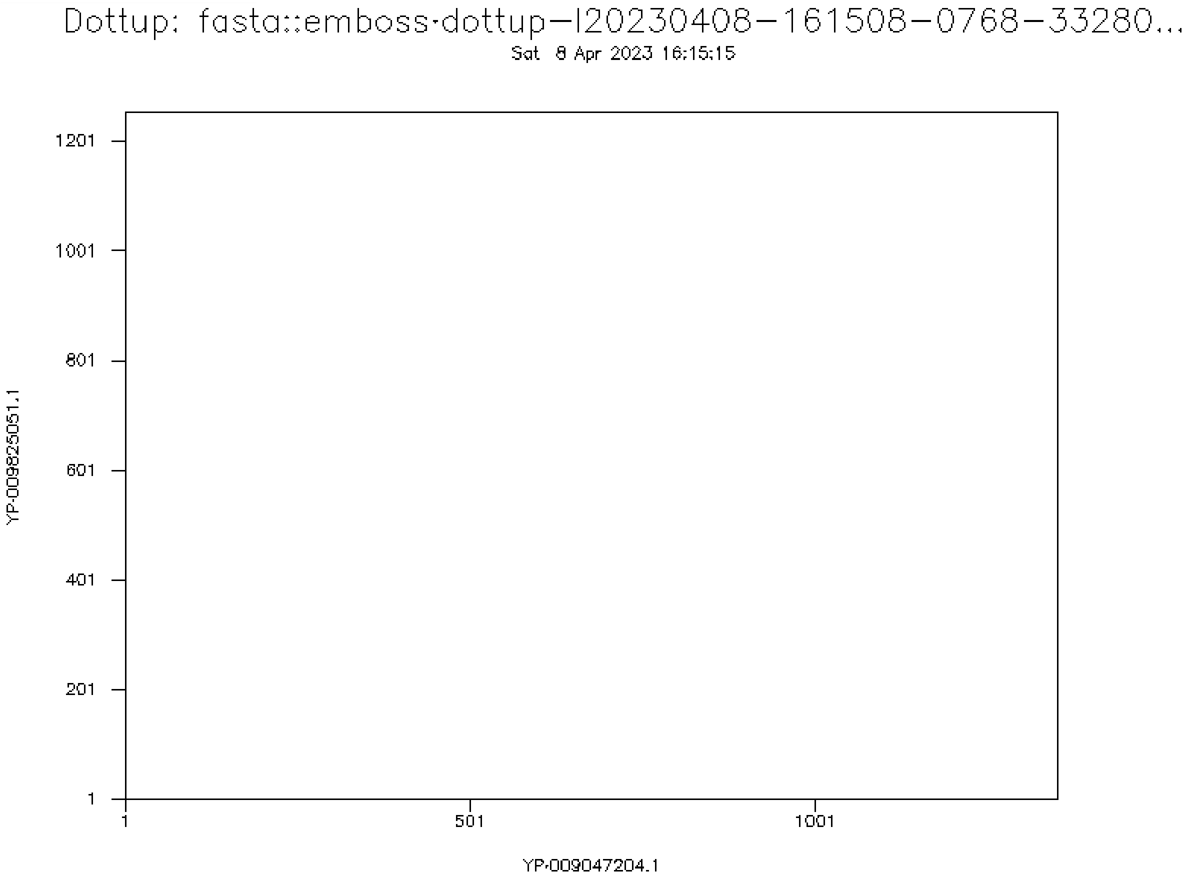
10

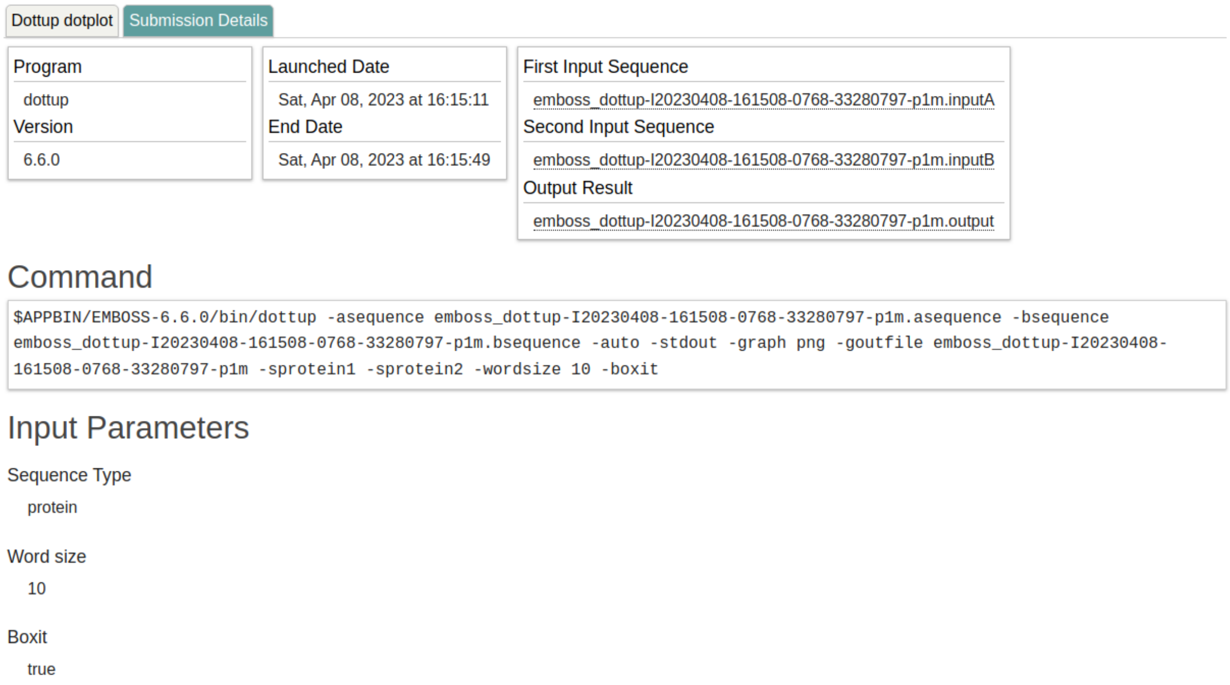
Boxit

true

3) Sars-Cov vs Mers

K-tuple: 3





Program

dottup

Version

6.6.0

Launched Date

Sat, Apr 08, 2023 at 16:15:11

End Date

Sat, Apr 08, 2023 at 16:15:49

First Input Sequence

emboss_dottup-I20230408-161508-0768-33280797-p1m.inputA

Second Input Sequence

emboss_dottup-I20230408-161508-0768-33280797-p1m.inputB

Output Result

emboss_dottup-I20230408-161508-0768-33280797-p1m.output

Command

```
$APPBIN/EMBOSS-6.6.0/bin/dottup -asequence emboss_dottup-I20230408-161508-0768-33280797-p1m.asequence -bsequence
emboss_dottup-I20230408-161508-0768-33280797-p1m.bsequence -auto -stdout -graph png -goutfile emboss_dottup-I20230408-
161508-0768-33280797-p1m -sprotein1 -sprotein2 -wordsize 10 -boxit
```

Input Parameters

Sequence Type

protein

Word size

10

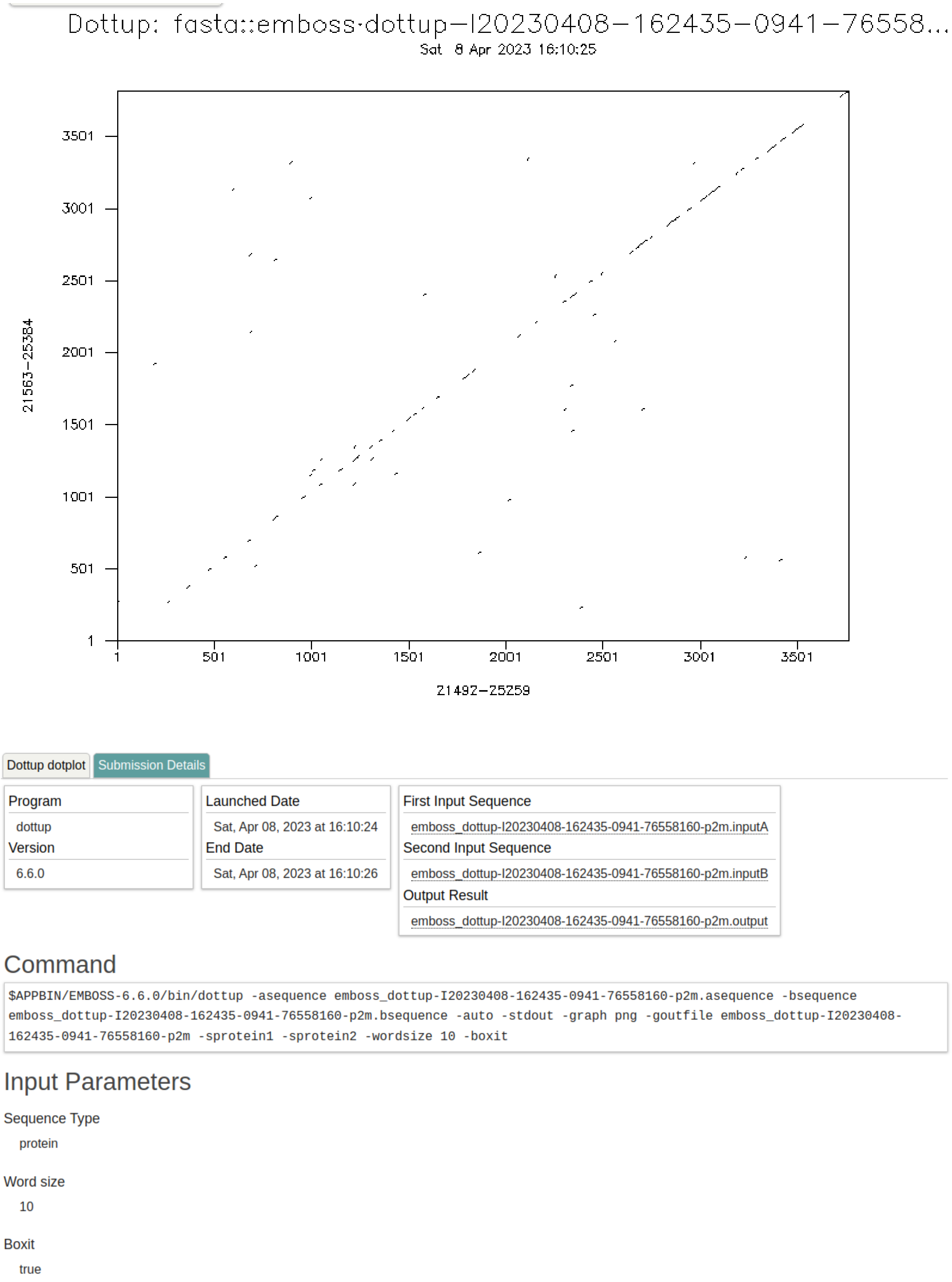
Boxit

true

Dottup Plots(DNA) :

1) Sars-Cov-2 vs Sars-Cov

K-tuple: 1



Program

dottup

Version

6.6.0

Launched Date

Sat, Apr 08, 2023 at 16:10:24

End Date

Sat, Apr 08, 2023 at 16:10:26

First Input Sequence

emboss_dottup-I20230408-162435-0941-76558160-p2m.inputA

Second Input Sequence

emboss_dottup-I20230408-162435-0941-76558160-p2m.inputB

Output Result

emboss_dottup-I20230408-162435-0941-76558160-p2m.output

Command

```
$APPBIN/EMBOSS-6.6.0/bin/dottup -asequence emboss_dottup-I20230408-162435-0941-76558160-p2m.asequence -bsequence
emboss_dottup-I20230408-162435-0941-76558160-p2m.bsequence -auto -stdout -graph png -goutfile emboss_dottup-I20230408-
162435-0941-76558160-p2m -sprotein1 -sprotein2 -wordsize 10 -boxit
```

Input Parameters

Sequence Type

protein

Word size

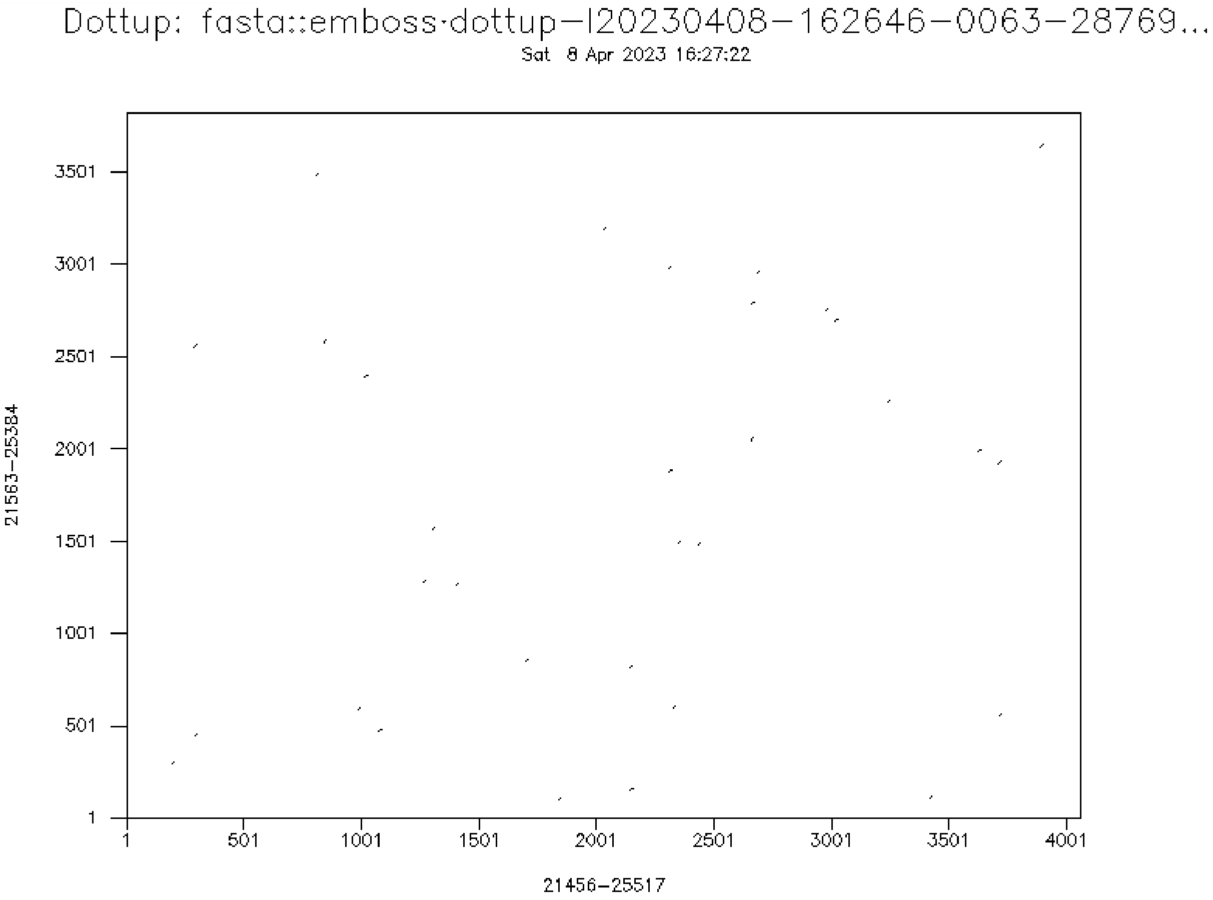
10

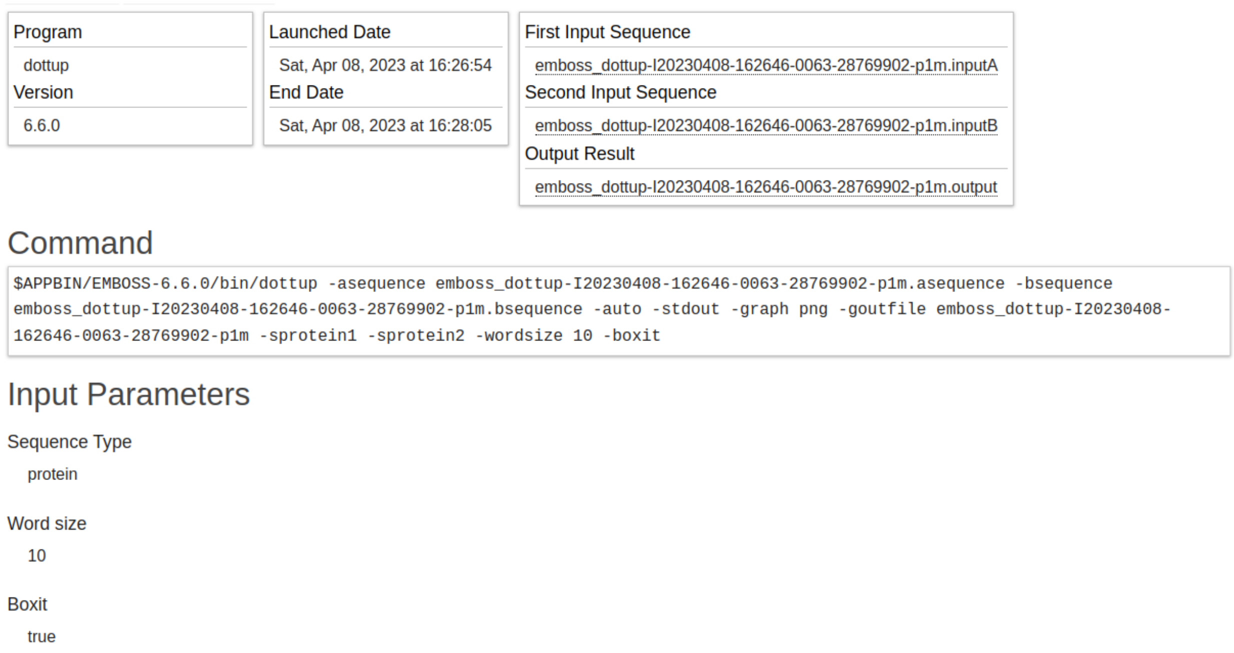
Boxit

true

2) Sars-Cov-2 vs Mers

K-tuple: 1





Program	Launched Date	First Input Sequence
dottup	Sat, Apr 08, 2023 at 16:26:54	emboss_dottup-I20230408-162646-0063-28769902-p1m.inputA
Version	**End Date**	**Second Input Sequence**
6.6.0	Sat, Apr 08, 2023 at 16:28:05	emboss_dottup-I20230408-162646-0063-28769902-p1m.inputB
		Output Result
		emboss_dottup-I20230408-162646-0063-28769902-p1m.output

Command

```
$APPBIN/EMBOSS-6.6.0/bin/dottup -asequence emboss_dottup-I20230408-162646-0063-28769902-p1m.asequence -bsequence
emboss_dottup-I20230408-162646-0063-28769902-p1m.bsequence -auto -stdout -graph png -goutfile emboss_dottup-I20230408-
162646-0063-28769902-p1m -sprotein1 -sprotein2 -wordsize 10 -boxit
```

Input Parameters

Sequence Type

protein

Word size

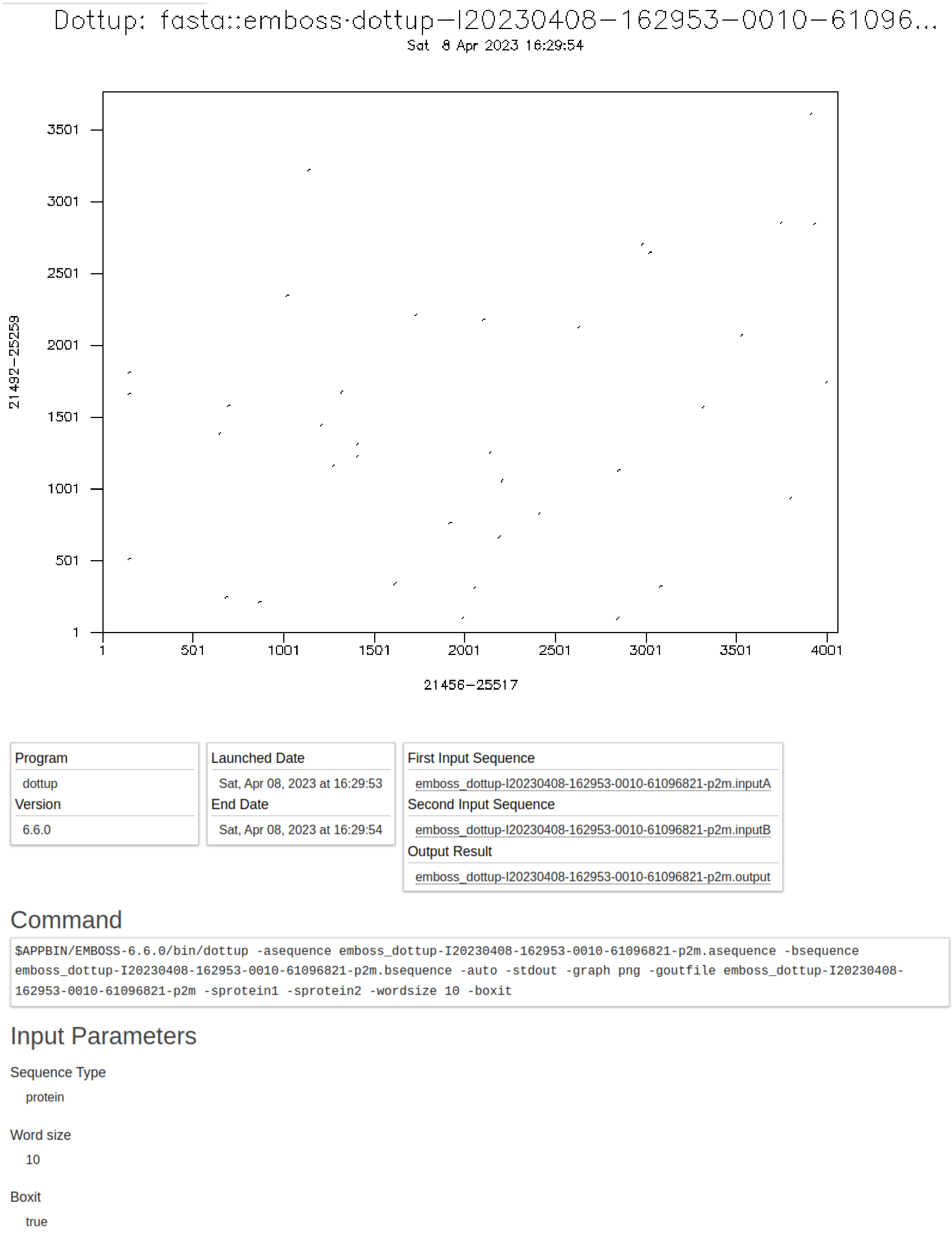
10

Boxit

true

3) Sars-Cov vs Mers

K-tuple: 1



Program	Launched Date	First Input Sequence
dottup	Sat, Apr 08, 2023 at 16:29:53	emboss_dottup-I20230408-162953-0010-61096821-p2m.inputA
Version	End Date	Second Input Sequence
6.6.0	Sat, Apr 08, 2023 at 16:29:54	emboss_dottup-I20230408-162953-0010-61096821-p2m.inputB
		Output Result
		emboss_dottup-I20230408-162953-0010-61096821-p2m.output

Command

```
$APPBIN/EMBOSS-6.6.0/bin/dottup -asequence emboss_dottup-I20230408-162953-0010-61096821-p2m.asequence -bsequence
emboss_dottup-I20230408-162953-0010-61096821-p2m.bsequence -auto -stdout -graph png -goutfile emboss_dottup-I20230408-
162953-0010-61096821-p2m -sprotein1 -sprotein2 -wordsize 10 -boxit
```

Input Parameters

Sequence Type

protein

Word size

10

Boxit

true

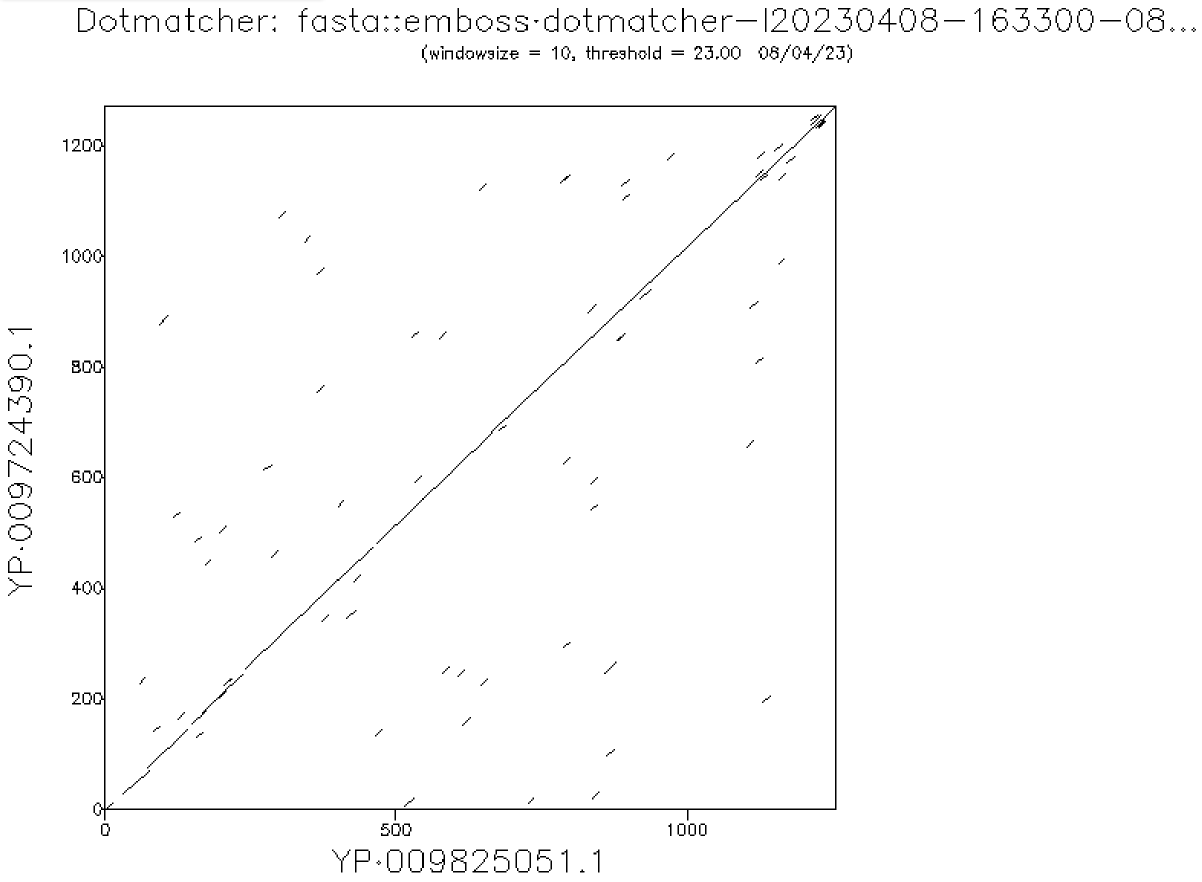
Dotmatcher Plots(Proteins) :

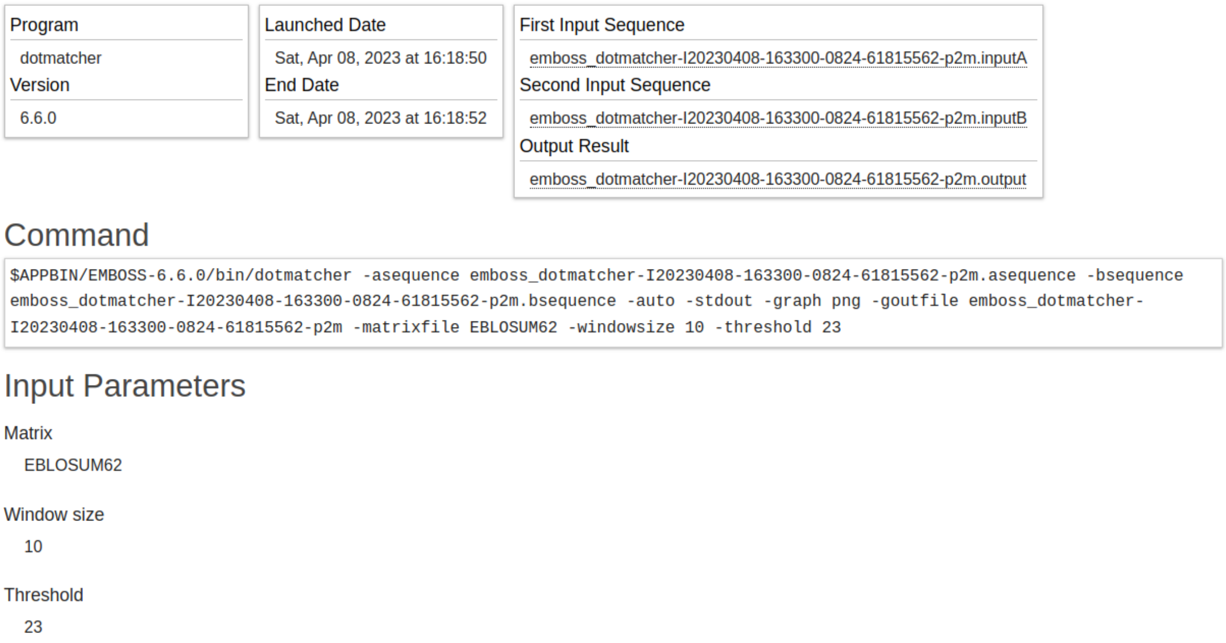
1) Sars-Cov-2 vs Sars-Cov

K-tuple: 3

Window-size: 10

Threshold value: 23





Program

dotmatcher

Version

6.6.0

Launched Date

Sat, Apr 08, 2023 at 16:18:50

End Date

Sat, Apr 08, 2023 at 16:18:52

First Input Sequence

emboss_dotmatcher-I20230408-163300-0824-61815562-p2m.inputA

Second Input Sequence

emboss_dotmatcher-I20230408-163300-0824-61815562-p2m.inputB

Output Result

emboss_dotmatcher-I20230408-163300-0824-61815562-p2m.output

Command

```
$APPBIN/EMBOSS-6.6.0/bin/dotmatcher -asequence emboss_dotmatcher-I20230408-163300-0824-61815562-p2m.asequence -bsequence
emboss_dotmatcher-I20230408-163300-0824-61815562-p2m.bsequence -auto -stdout -graph png -goutfile emboss_dotmatcher-
I20230408-163300-0824-61815562-p2m -matrixfile EBLOSUM62 -windowsize 10 -threshold 23
```

Input Parameters

Matrix

EBLOSUM62

Window size

10

Threshold

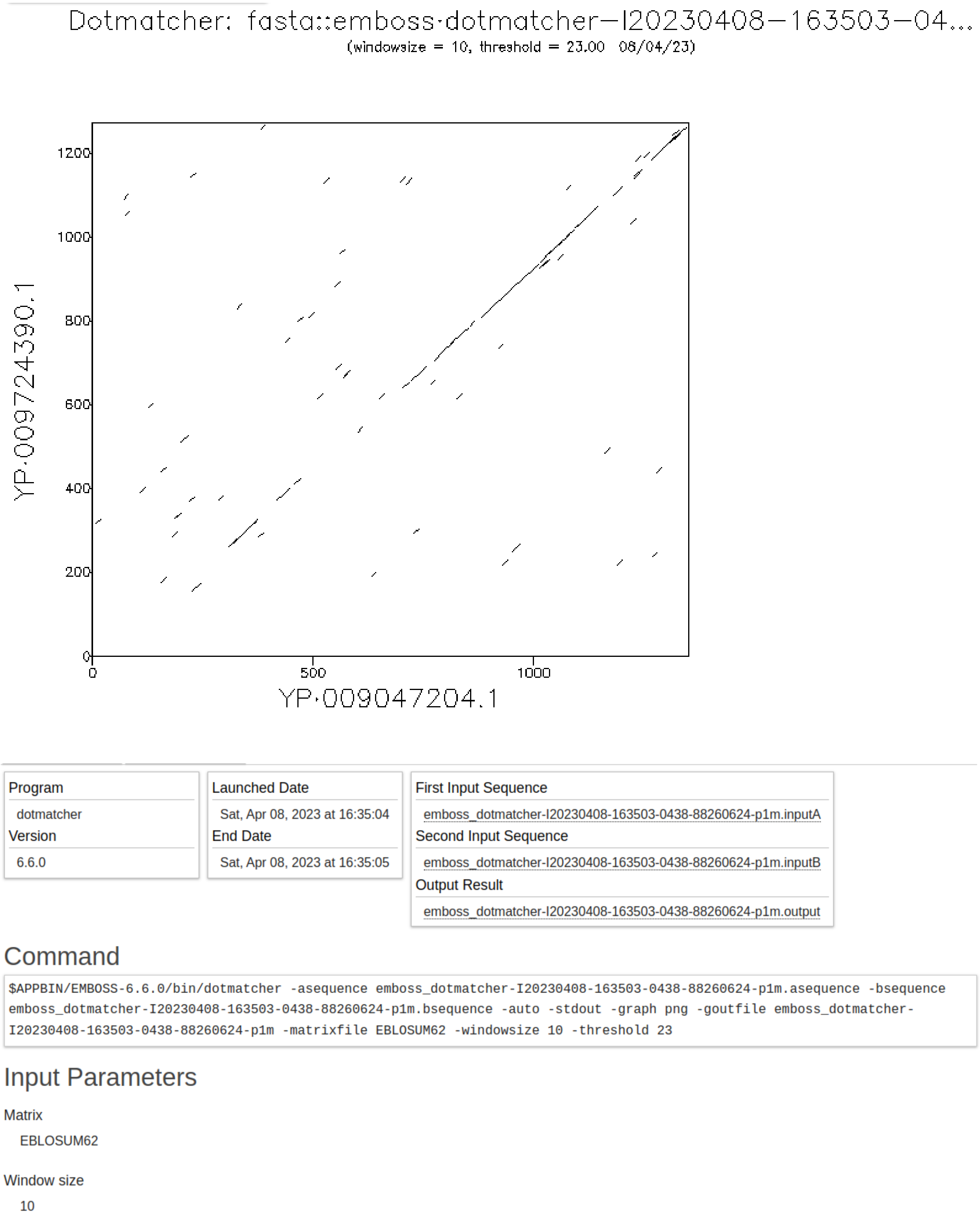
23

2) Sars-Cov-2 vs Mers

K-tuple: 3

Window-size: 10

Threshold value:


Dotmatcher: fasta::emboss·dotmatcher—I20230408—163503—04...
(windowsize = 10, threshold = 23.00 08/04/23)
YP·009724390.1
1200
1000
800
600
400
200
0
0
500
1000
YP·009047204.1

Program	Launched Date	First Input Sequence
dotmatcher	Sat, Apr 08, 2023 at 16:35:04	emboss_dotmatcher-I20230408-163503-0438-88260624-p1m.inputA
Version	End Date	Second Input Sequence
6.6.0	Sat, Apr 08, 2023 at 16:35:05	emboss_dotmatcher-I20230408-163503-0438-88260624-p1m.inputB
		Output Result
		emboss_dotmatcher-I20230408-163503-0438-88260624-p1m.output

Command

```
$APPBIN/EMBOSS-6.6.0/bin/dotmatcher -asequence emboss_dotmatcher-I20230408-163503-0438-88260624-p1m.asequence -bsequence
emboss_dotmatcher-I20230408-163503-0438-88260624-p1m.bsequence -auto -stdout -graph png -goutfile emboss_dotmatcher-
I20230408-163503-0438-88260624-p1m -matrixfile EBLOSUM62 -windowsize 10 -threshold 23
```

Input Parameters

Matrix

EBLOSUM62

Window size

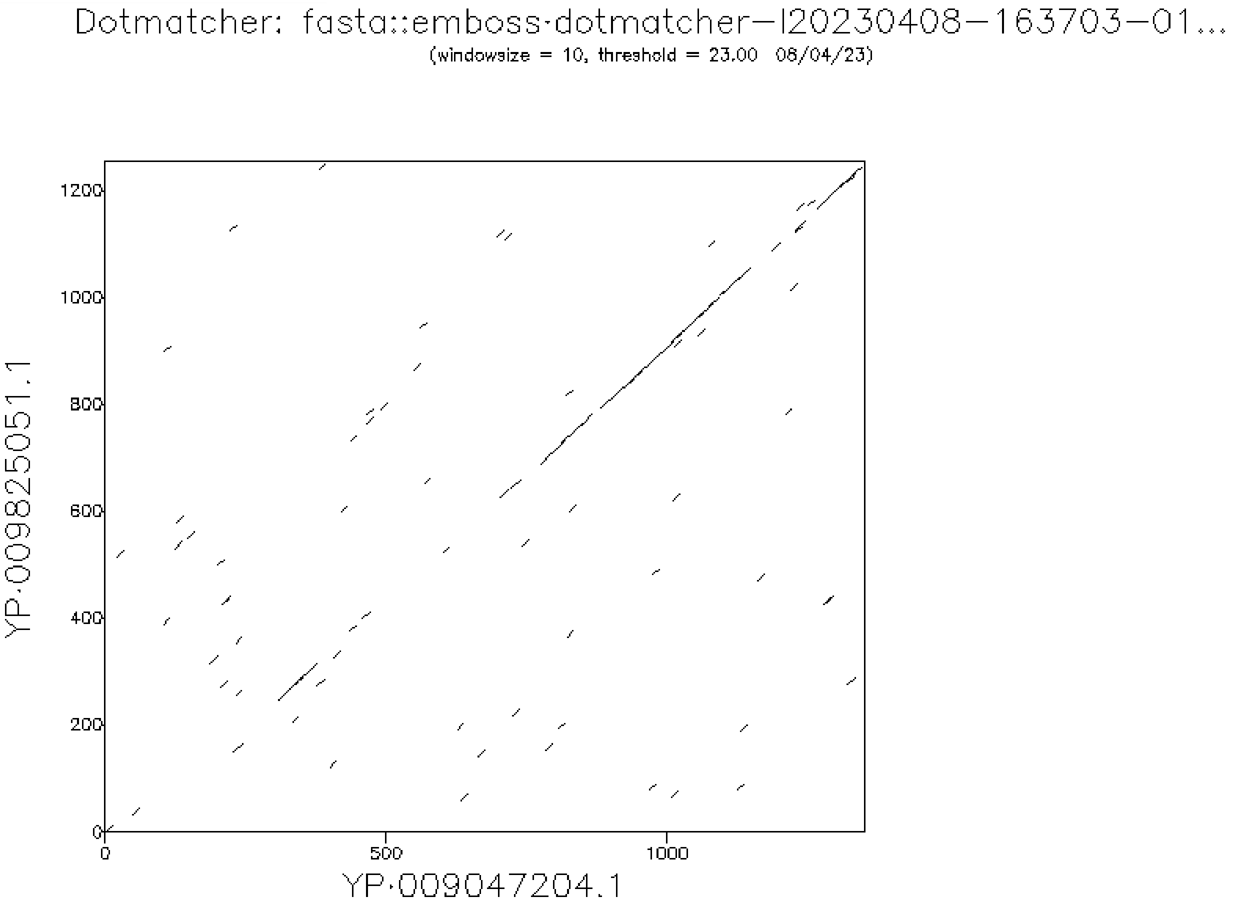
10

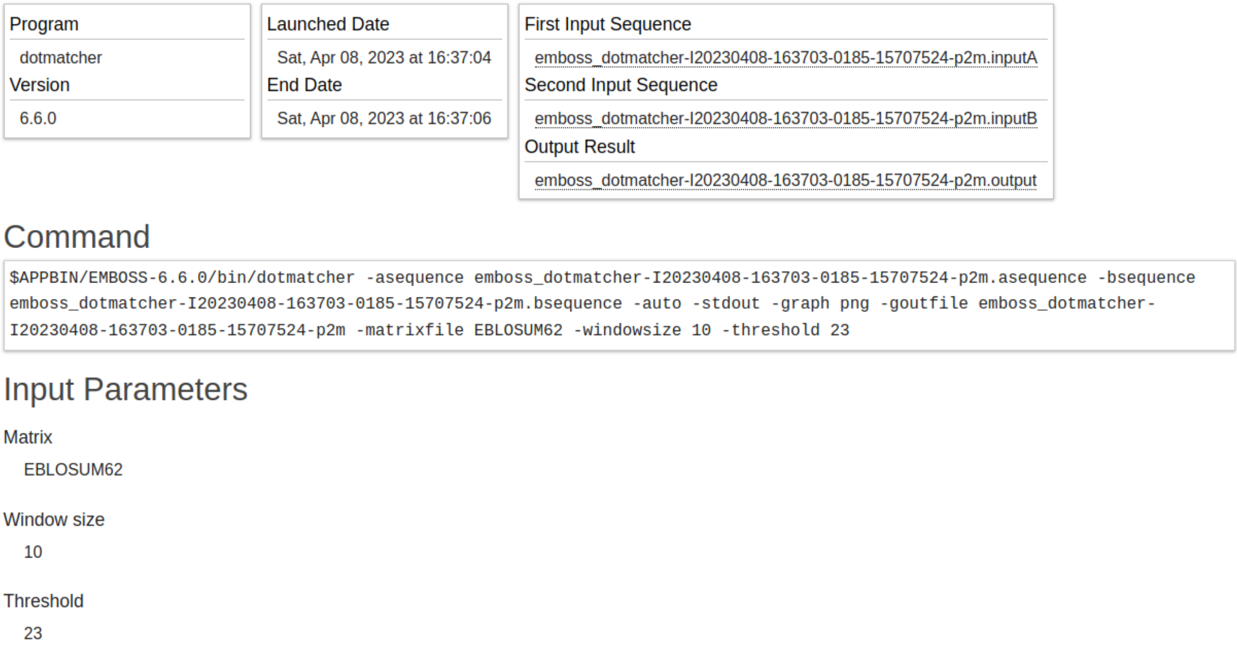
3) Sars-Cov vs Mers

K-tuple: 3

Window-size: 10

Threshold value: 23





Program	Launched Date	First Input Sequence
dotmatcher	Sat, Apr 08, 2023 at 16:37:04	emboss_dotmatcher-I20230408-163703-0185-15707524-p2m.inputA
Version	End Date	Second Input Sequence
6.6.0	Sat, Apr 08, 2023 at 16:37:06	emboss_dotmatcher-I20230408-163703-0185-15707524-p2m.inputB
		Output Result
		emboss_dotmatcher-I20230408-163703-0185-15707524-p2m.output

Command

```
$APPBIN/EMBOSS-6.6.0/bin/dotmatcher -asequence emboss_dotmatcher-I20230408-163703-0185-15707524-p2m.asequence -bsequence
emboss_dotmatcher-I20230408-163703-0185-15707524-p2m.bsequence -auto -stdout -graph png -goutfile emboss_dotmatcher-
I20230408-163703-0185-15707524-p2m -matrixfile EBLOSUM62 -windowsize 10 -threshold 23
```

Input Parameters

Matrix

EBLOSUM62

Window size

10

Threshold

23

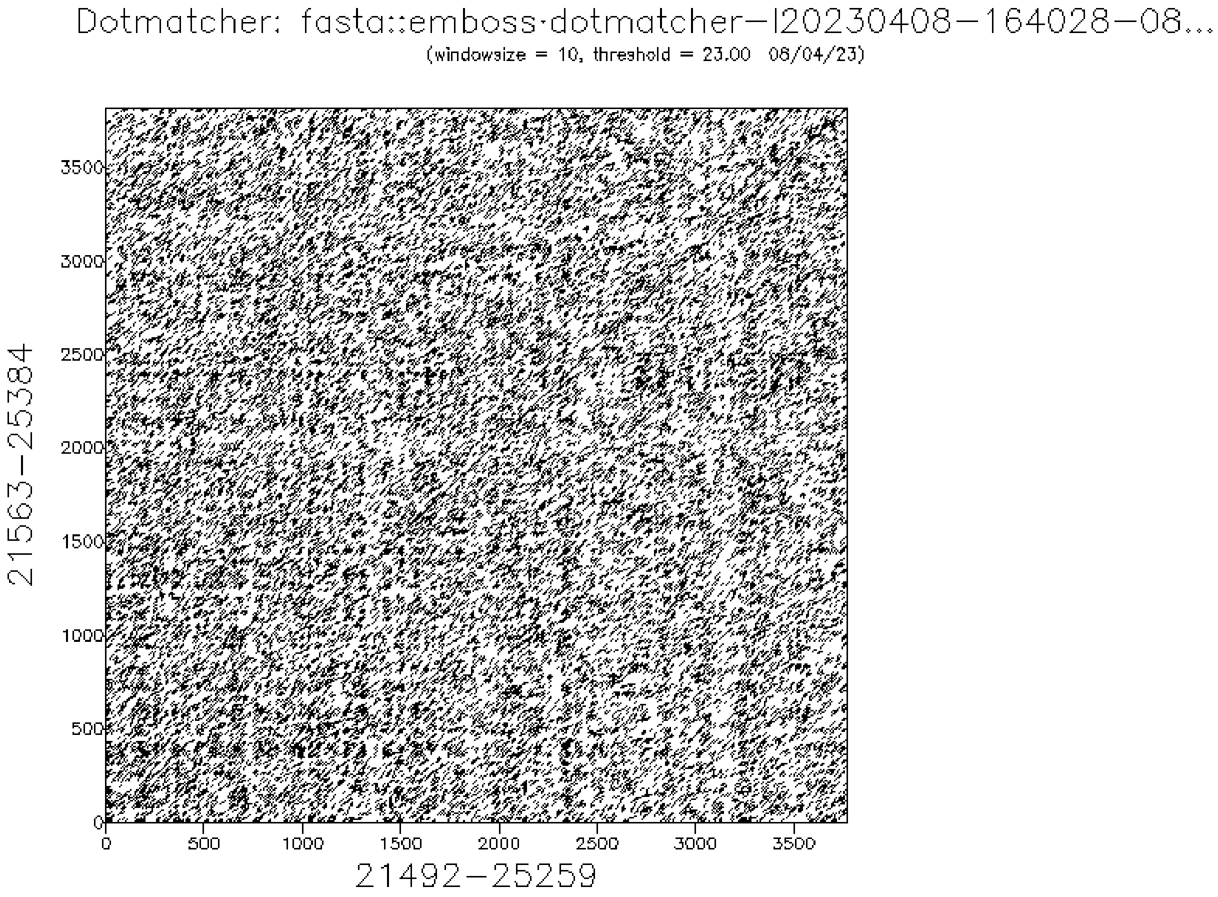
Dotmatcher Plots(DNA) :

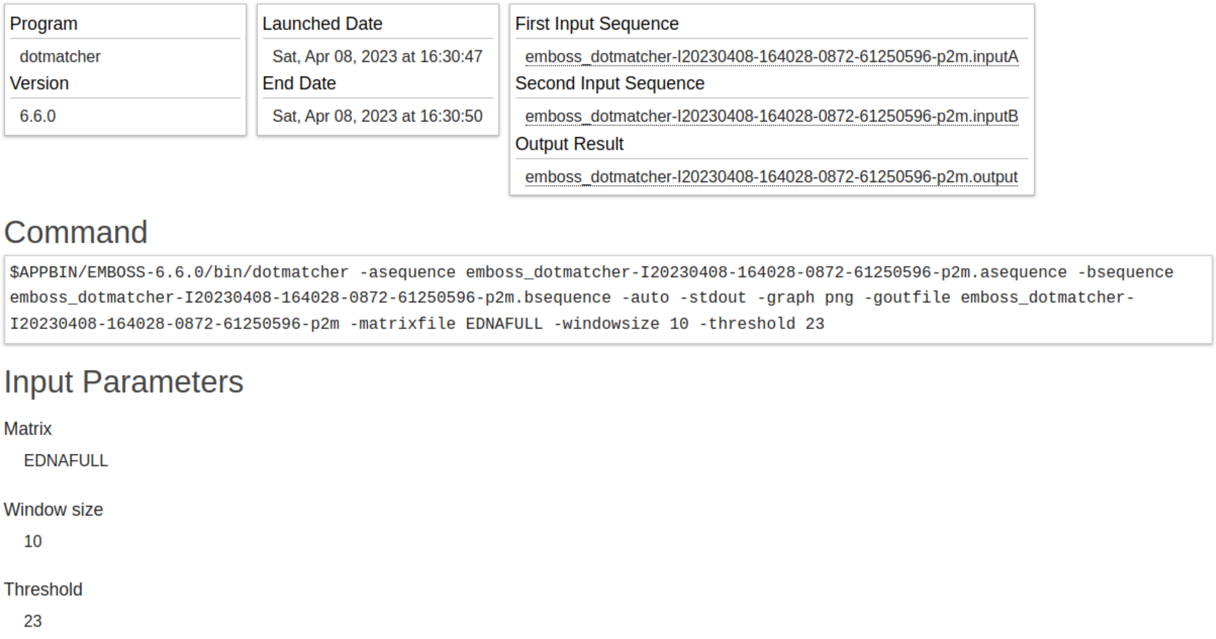
1) Sars-Cov-2 vs Sars-Cov

K-tuple: 1

Window-size: 10

Threshold value: 23





Program	Launched Date	First Input Sequence
dotmatcher	Sat, Apr 08, 2023 at 16:30:47	emboss_dotmatcher-I20230408-164028-0872-61250596-p2m.inputA
Version	End Date	Second Input Sequence
6.6.0	Sat, Apr 08, 2023 at 16:30:50	emboss_dotmatcher-I20230408-164028-0872-61250596-p2m.inputB
		Output Result
		emboss_dotmatcher-I20230408-164028-0872-61250596-p2m.output

Command

```
$APPBIN/EMBOSS-6.6.0/bin/dotmatcher -asequence emboss_dotmatcher-I20230408-164028-0872-61250596-p2m.asequence -bsequence
emboss_dotmatcher-I20230408-164028-0872-61250596-p2m.bsequence -auto -stdout -graph png -goutfile emboss_dotmatcher-
I20230408-164028-0872-61250596-p2m -matrixfile EDNAFULL -windowsize 10 -threshold 23
```

Input Parameters

Matrix

EDNAFULL

Window size

10

Threshold

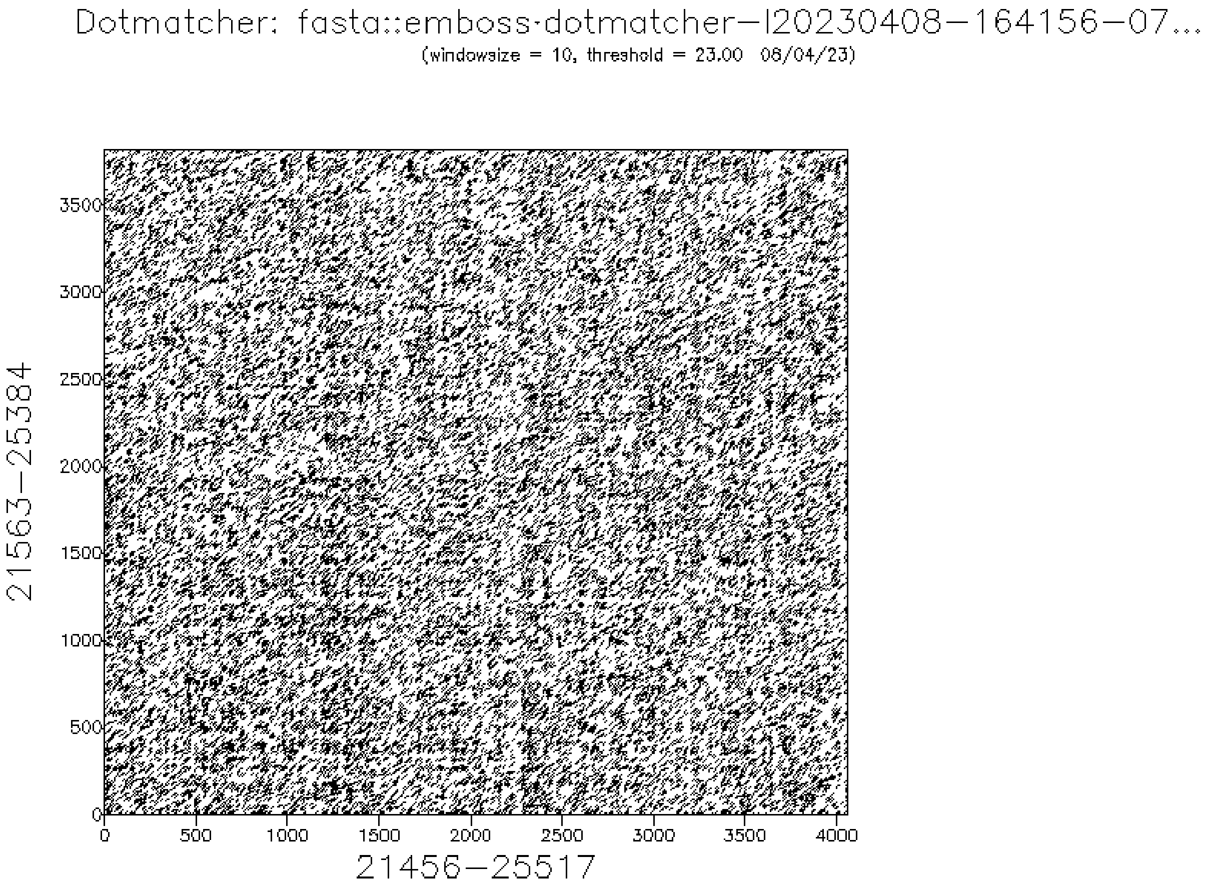
23

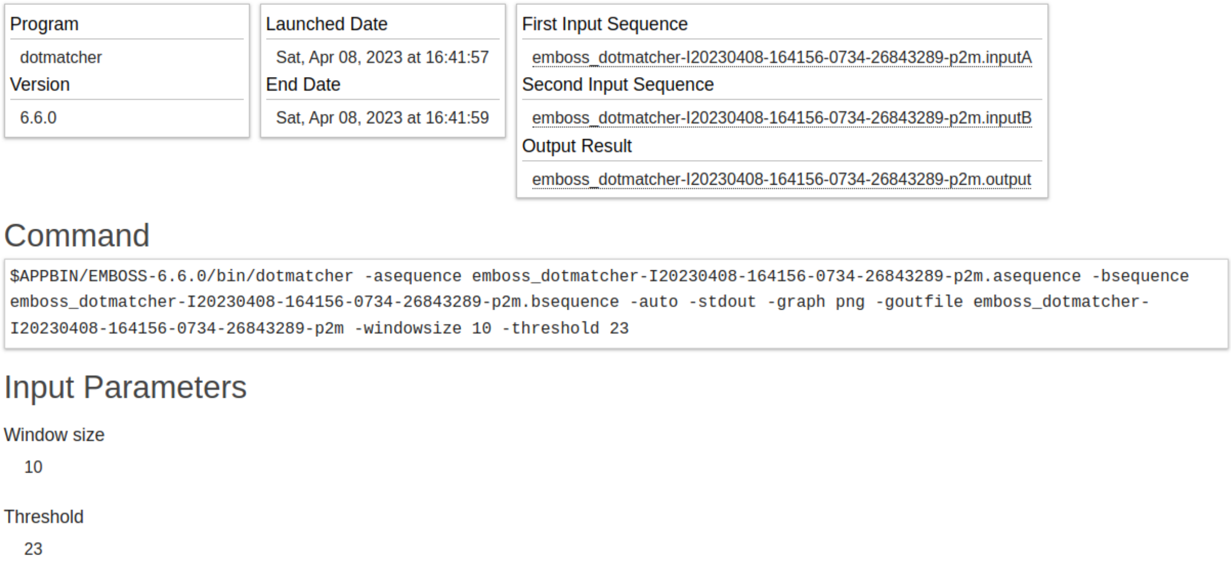
2) Sars-Cov-2 vs Mers

K-tuple: 1

Window-size: 10

Threshold value: 23





Program	Launched Date	First Input Sequence
dotmatcher	Sat, Apr 08, 2023 at 16:41:57	emboss_dotmatcher-I20230408-164156-0734-26843289-p2m.inputA
Version	End Date	Second Input Sequence
6.6.0	Sat, Apr 08, 2023 at 16:41:59	emboss_dotmatcher-I20230408-164156-0734-26843289-p2m.inputB
		Output Result
		emboss_dotmatcher-I20230408-164156-0734-26843289-p2m.output

Command

```
$APPBIN/EMBOSS-6.6.0/bin/dotmatcher -asequence emboss_dotmatcher-I20230408-164156-0734-26843289-p2m.asequence -bsequence
emboss_dotmatcher-I20230408-164156-0734-26843289-p2m.bsequence -auto -stdout -graph png -goutfile emboss_dotmatcher-
I20230408-164156-0734-26843289-p2m -windowsize 10 -threshold 23
```

Input Parameters

Window size

10

Threshold

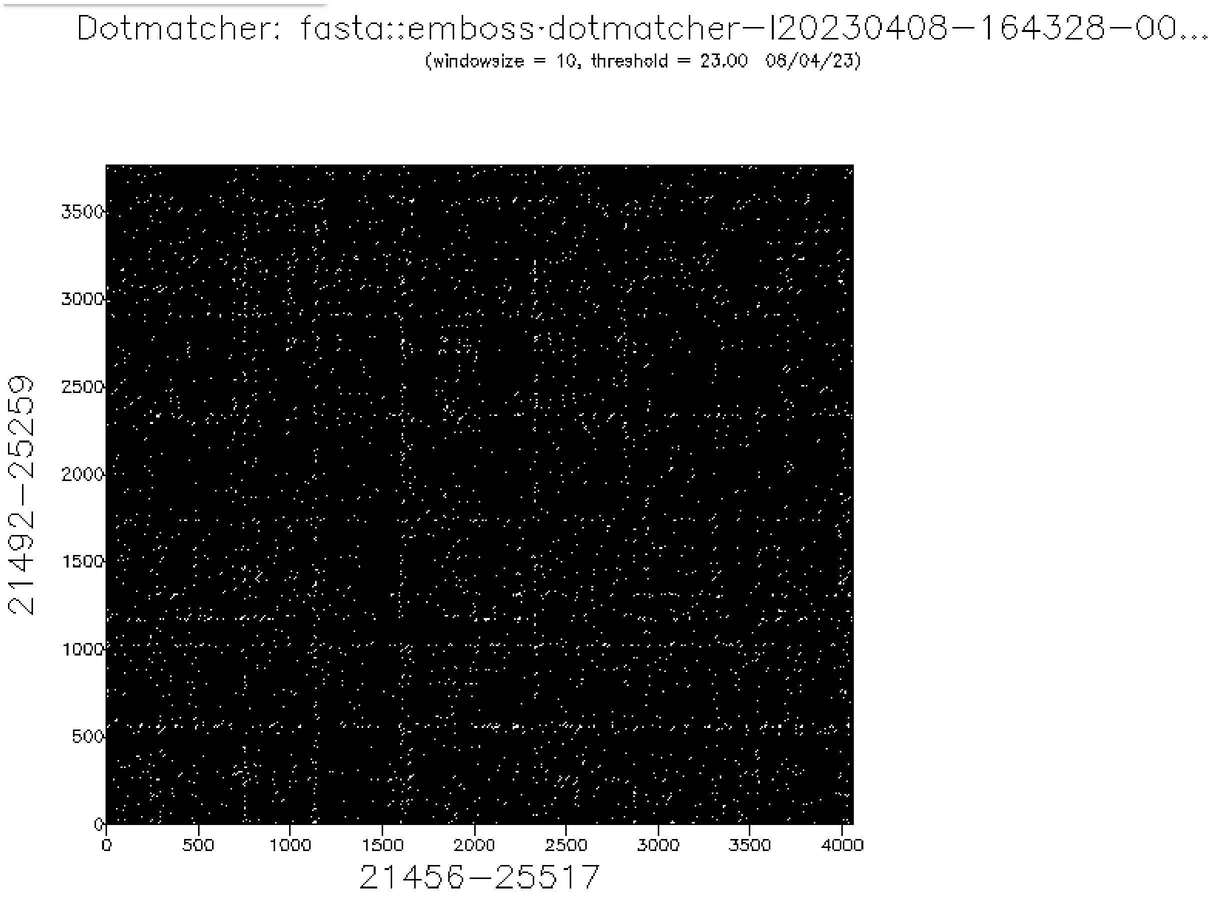
23

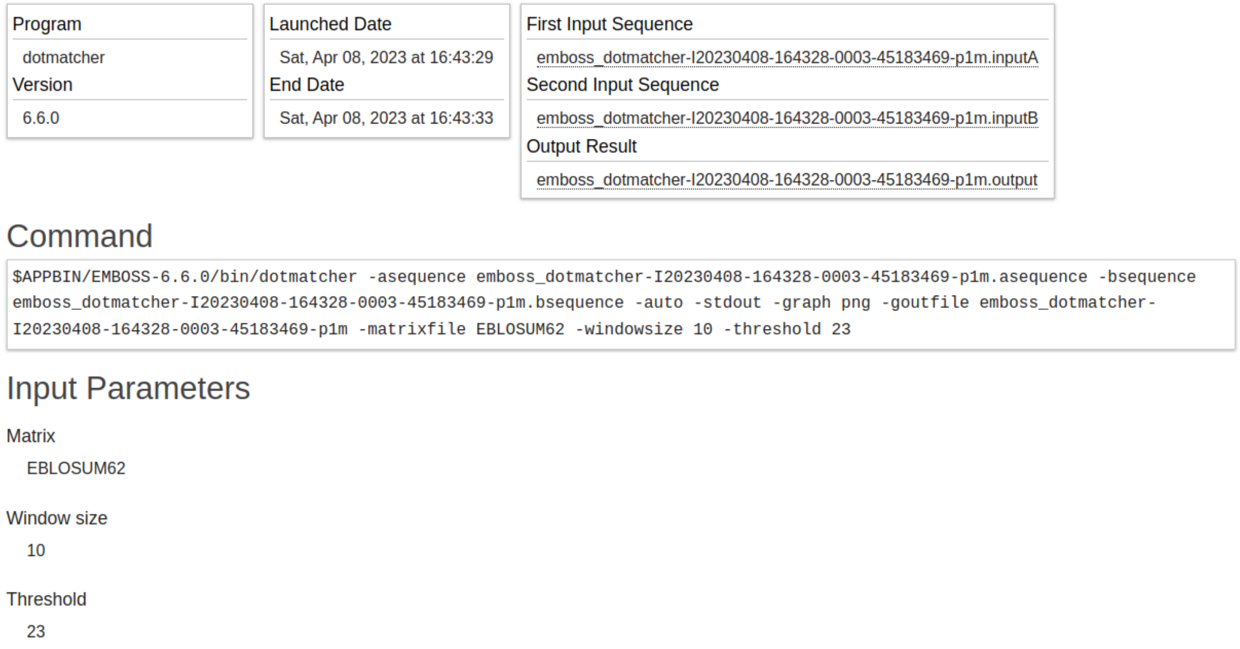
3) Sars-Cov vs Mers

K-tuple: 1

Window-size: 10

Threshold value: 23





Program

dotmatcher

Version

6.6.0

Launched Date

Sat, Apr 08, 2023 at 16:43:29

End Date

Sat, Apr 08, 2023 at 16:43:33

First Input Sequence

emboss_dotmatcher-I20230408-164328-0003-45183469-p1m.inputA

Second Input Sequence

emboss_dotmatcher-I20230408-164328-0003-45183469-p1m.inputB

Output Result

emboss_dotmatcher-I20230408-164328-0003-45183469-p1m.output

Command

```
$APPBIN/EMBOSS-6.6.0/bin/dotmatcher -asequence emboss_dotmatcher-I20230408-164328-0003-45183469-p1m.asequence -bsequence
emboss_dotmatcher-I20230408-164328-0003-45183469-p1m.bsequence -auto -stdout -graph png -goutfile emboss_dotmatcher-
I20230408-164328-0003-45183469-p1m -matrixfile EBLOSUM62 -windowsize 10 -threshold 23
```

Input Parameters

Matrix

EBLOSUM62

Window size

10

Threshold

23

1. **Identify SARS-CoV-2 is similar to which of the earlier two viruses?**

Ans1. In the Dottup plots for comparing the viruses, we can clearly observe that the SARS-CoV-2 is more similar to SARS-CoV than MERS-CoV. There are regions where the diagonal lines representing the comparison of the SARS-CoV-2 sequence to the SARS-CoV sequence are longer and more continuous than the lines representing the comparison to the MERS-CoV sequence. This leads us to conclude that SARS-CoV-2 is more similar to SARS-CoV.

2. **Is it easy to identify the similarity using DNA or protein sequences? Give reasons.**

Ans2. There is no fixed method that says that similarities are easily identified by DNA or protein sequences. These are certain factors that should be kept in mind:

- Degree of similarity affects ease of identification; highly similar sequences are easier to compare and identify.
- Length of sequences can also affect ease of identification; longer sequences provide more information and can be easier to compare.
- The nature of the sequences (DNA or protein) can also influence ease of identification.

- Proteins are more complex than DNA and can be more challenging to compare due to a wider range of chemical properties and folding patterns.
- DNA sequences can be more prone to errors due to mutations or sequencing errors.

Overall, highly similar and longer sequences are easier to compare and identify, but more advanced computational methods may be required for sequences with low levels of similarity or complex properties.

Ques2)

PART A:

Needle Method (Global Alignment):

DNA Level:

```
Identity:    2833/3859 (73.4%)
Similarity:  2833/3859 (73.4%)
```

Protein Level:

```
Identity:     974/1277 (76.3%)
Similarity:  1111/1277 (87.0%)
```

Water Method (Local Alignment):

DNA Level:

```
Identity:    2833/3859 (73.4%)
Similarity:  2833/3859 (73.4%)
```

Protein Level:

```
Identity:      974/1277 (76.3%)
Similarity:   1111/1277 (87.0%)
```

(i) The % identity and similarity for protein level alignment are greater than for DNA level alignment. This is due to the fact that **multiple codons might encode the same amino acid**. As a result, the protein sequence is more flexible than the DNA sequence, resulting in a higher degree of conservation.

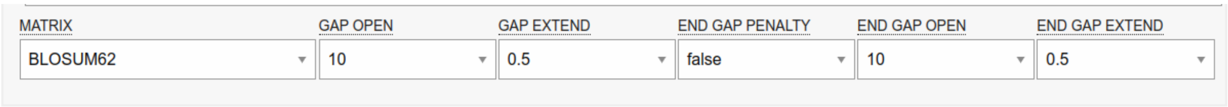
(ii) Identity: It refers to the number of alignment places where the two sequences share the same nucleotide or amino acid. It denotes the precise match of two sequences.

Similarity: It refers to the number of locations in the alignment where the two sequences have comparable or equivalent nucleotides or amino acids. It quantifies the degree of conservation between two sequences, taking into consideration both identical and comparable residues.

(iii) For the specified sequences, both local and global alignment are identical. This might be because the sequences are so brief.

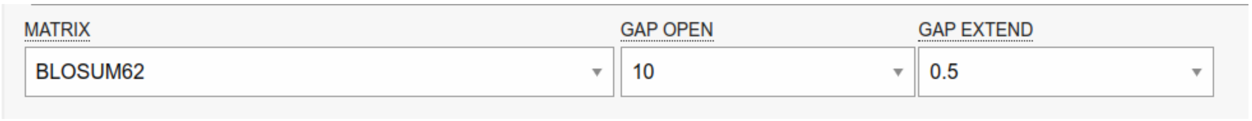
Nevertheless, for longer sequences, the local and global alignments would differ.

(iv) Needle Method:



MATRIX	GAP OPEN	GAP EXTEND	END GAP PENALTY	END GAP OPEN	END GAP EXTEND
BLOSUM62	10	0.5	false	10	0.5

Water Method:



MATRIX	GAP OPEN	GAP EXTEND
BLOSUM62	10	0.5

PART B:

Needle Method:

DNA:

```
Identity:    2194/4284 (51.2%)
Similarity:  2194/4284 (51.2%)
```

Protein:

```
Identity:     434/1454 (29.8%)
Similarity:   658/1454 (45.3%)
```

Water Method:

DNA:

```
Identity:    2194/4283 (51.2%)
Similarity:  2194/4283 (51.2%)
```

Protein:

```
Identity:     433/1440 (30.1%)
Similarity:   655/1440 (45.5%)
```

(i) Both sequences can be called homologous as they are identical more than 30% of the way through their protein alignments.

(ii) Since proteins constitute the functional units of the cell, this conclusion is based on protein sequence alignment.

Because genetic coding is degenerate, multiple nucleotide sequences may have the same amino acid sequences, and DNA sequences undergo more faster evolutionary changes, rendering them less conserved throughout evolutionary time, DNA is not commonly employed.

Ques3)

(i) SARS-CoV is the query sequence's closest homolog. This is true for DNA and protein database searches.

(ii)

For DNA database search -

Score - 6889

Percentage Identity - 99.97%

Percentage similarity -

Length of the alignment - 29903

Expect or e-value - 0.0

For protein database search -

Score - 2637

Percentage Identity - 99.92%

Percentage similarity -

Length of the alignment - 1310

Expect or e-value - 0.0

(iii) The spike glycoprotein of SARS-CoV is, in fact, the first non-synthetic hit. The alignment achieved using 'water' does not match the % identity and similarity. We have the following when we use 'water' for Sars-Cov:

Water Method (Local Alignment):

DNA:

Identity: 2833/3859 (73.4%)
Similarity: 2833/3859 (73.4%)

Protein:

Identity: 974/1277 (76.3%)
Similarity: 1111/1277 (87.0%)

The alignment is significant because a higher percentage of identity or similarity between the query sequence and a hit sequence indicates a tighter evolutionary link

between the two sequences.
It is crucial to remember, however, that a high proportion of identity or resemblance does not always imply functional or structural similarity.

(iv) Yes, in fact the bat SARS coronavirus spike glycoprotein very closely resembles sars-cov2.

Identity - Bat coronavirus Taxonomy ID: 1508220

For DNA database search -

Score - 5971

Percentage Identity - 94.61%

Percentage similarity -

Length of the alignment - 29838

Expect or e-value - 0.0

For protein database search -

Score - 2593

Percentage Identity - 98.43%

Percentage similarity -

Length of the alignment - 1269

Expect or e-value - 0.0

Ques4)

(i) The UniProt database contains approximately **200 million protein sequences**, while the GenBank database contains **over 350 billion nucleotide bases** across all organisms.

- To perform a search in these databases using dynamic programming (DP), we would need to compute a large number of matrix cells, which depends on the length of the query sequence, the length of the database sequences, and the size of the scoring matrix used for the DP algorithm.

- Assuming a query sequence length of 1000 bases, the number of matrix cells to be computed using DP for performing a search in UniProt or GenBank would be extremely large. For example, to search UniProt using a scoring matrix such as BLOSUM62 or PAM250, which typically have dimensions of 20x20 (for amino acids), we would need to compute approximately 3.6 x 10^18 cells. Similarly, to search GenBank using a scoring matrix such as NUC44, which typically has dimensions of 4x4 (for nucleotides), we would need to compute approximately 1.4 x 10^19 cells. These are extremely large numbers, and performing a search on these databases using DP would be computationally infeasible.

- Assuming a query sequence length of 1000 bases, the number of matrix cells to be computed using DP for performing a search in UniProt or GenBank would be extremely large. For example, to search UniProt using a scoring matrix such as BLOSUM62 or PAM250, which typically have dimensions of 20x20 (for amino acids), we would need to compute approximately 3.6 x 10^18 cells. Similarly, to search GenBank using a scoring matrix such as NUC44, which typically has dimensions of 4x4 (for nucleotides), we would need to compute approximately 1.4 x 10^19 cells. These are extremely large numbers, and performing a search on these databases using DP would be computationally infeasible.

Therefore, more efficient search algorithms such as `BLAST` or `FASTA` are typically used for searching large databases like UniProt and GenBank, which can handle such large datasets.

(ii) When comparing a query sequence of 1000 bases with the entire length of `Human Chromosome 1 (~249Mbp)` and `Mouse Chromosome 1 (~195Mbp)` using dynamic programming (DP), the memory or space requirement would be different for the two cases.

Assuming a scoring matrix such as NUC44, which typically has dimensions of 4x4 (for nucleotides), the number of cells to be computed using DP can be estimated as follows:

For Human Chromosome 1:
Number of cells = 1000 x 249,000,000 x 1000 x 4 = 9.96 x 10^14 cells

For Mouse Chromosome 1:
Number of cells = 1000 x 195,000,000 x 1000 x 4 = 7.8 x 10^14 cells

Therefore, the memory or space requirement for comparing the query sequence with Human Chromosome 1 using DP would be **slightly higher** than that for Mouse Chromosome 1, due to its larger size.

However, comparing a single query sequence with an entire chromosome using DP is not a common approach in bioinformatics, as it would be computationally intensive and time-consuming. Instead, more efficient algorithms such as `BLAST` or `FASTA` are typically used for comparing sequences against large databases, which are designed to handle such large datasets and can provide results in a reasonable amount of time.