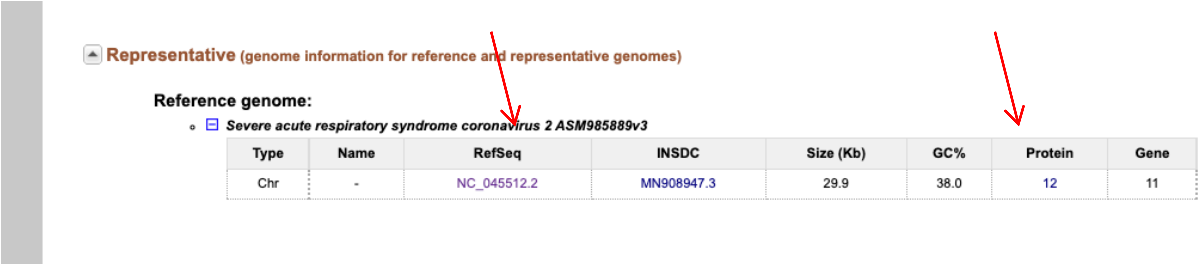
Algoritmos Bioinformática /Bioinformática 2021/2022

Assignment 3

In this assignment, you will analyze the genomic sequence of the SARS-CoV2 virus to identify the putative proteins. The goal is to identify all possible ORFs.

A. Fetch the data

- Download the Genome sequence from Genbank:
  - Go to: https://www.ncbi.nlm.nih.gov/genome/?term=MT072688



Representative (genome information for reference and representative genomes)

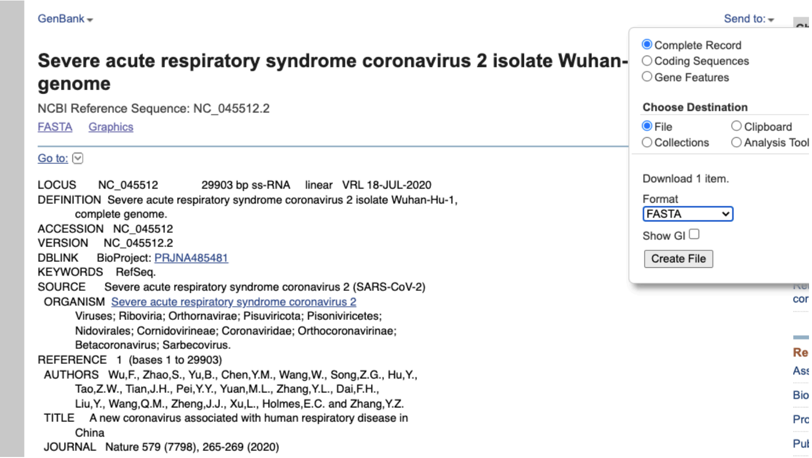
Reference genome:

Severe acute respiratory syndrome coronavirus 2 ASM985889v3

Type	Name	RefSeq	INSDC	Size (Kb)	GC%	Protein	Gene
Chr	-	NC_045512.2	MN908947.3	29.9	38.0	12	11

  - Click on the RefSeq link https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2

  - In option "Send to", select file, Format Fasta, Create File.



Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-
genome

NCBI Reference Sequence: NC_045512.2

```
LOCUS       NC_045512           29903 bp ss-RNA     linear   VRL 18-JUL-2020
DEFINITION  Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1,
            complete genome.
ACCESSION   NC_045512
VERSION     NC_045512.2
DBLINK      BioProject: PRJNA485481
KEYWORDS    RefSeq.
SOURCE      Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)
  ORGANISM  Severe acute respiratory syndrome coronavirus 2
            Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes;
            Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae;
            Betacoronavirus; Sarbecovirus.
REFERENCE   1  (bases 1 to 29903)
  AUTHORS   Wu,F., Zhao,S., Yu,B., Chen,Y.M., Wang,W., Song,Z.G., Hu,Y.,
            Tao,Z.W., Tian,J.H., Pei,Y.Y., Yuan,M.L., Zhang,Y.L., Dai,F.H.,
            Liu,Y., Wang,Q.M., Zheng,J.J., Xu,L., Holmes,E.C. and Zhang,Y.Z.
  TITLE     A new coronavirus associated with human respiratory disease in
            China
  JOURNAL   Nature 579 (7798), 265-269 (2020)
```

● Complete Record
○ Coding Sequences
○ Gene Features

Choose Destination

● File ○ Clipboard
○ Collections ○ Analysis Tool

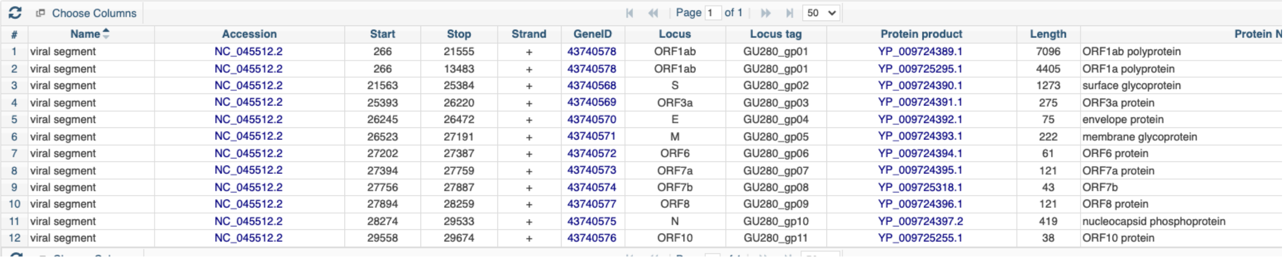
Download 1 item.

Format: FASTA

Show GI □

- Obtain the coordinates of the annotated proteins to compare to your results.
  - Go to: https://www.ncbi.nlm.nih.gov/genome/?term=MT072688
  - In the reference genome table, click on the link "12" for the column Protein. You should see a table as the one below. In the Download button get this

coordinates in a tabular format. The file is named *proteins_86693_757732.csv*



#	Name	Accession	Start	Stop	Strand	GeneID	Locus	Locus tag	Protein product	Length	Protein N
1	viral segment	NC_045512.2	266	21555	+	43740578	ORF1ab	GU280_gp01	YP_009724389.1	7096	ORF1ab polyprotein
2	viral segment	NC_045512.2	266	13483	+	43740578	ORF1ab	GU280_gp01	YP_009725295.1	4405	ORF1a polyprotein
3	viral segment	NC_045512.2	21563	25384	+	43740568	S	GU280_gp02	YP_009724390.1	1273	surface glycoprotein
4	viral segment	NC_045512.2	25393	26220	+	43740569	ORF3a	GU280_gp03	YP_009724391.1	275	ORF3a protein
5	viral segment	NC_045512.2	26245	26472	+	43740570	E	GU280_gp04	YP_009724392.1	75	envelope protein
6	viral segment	NC_045512.2	26523	27191	+	43740571	M	GU280_gp05	YP_009724393.1	222	membrane glycoprotein
7	viral segment	NC_045512.2	27202	27387	+	43740572	ORF6	GU280_gp06	YP_009724394.1	61	ORF6 protein
8	viral segment	NC_045512.2	27394	27759	+	43740573	ORF7a	GU280_gp07	YP_009724395.1	121	ORF7a protein
9	viral segment	NC_045512.2	27756	27887	+	43740574	ORF7b	GU280_gp08	YP_009725318.1	43	ORF7b
10	viral segment	NC_045512.2	27894	28259	+	43740577	ORF8	GU280_gp09	YP_009724396.1	121	ORF8 protein
11	viral segment	NC_045512.2	28274	29533	+	43740575	N	GU280_gp10	YP_009724397.2	419	nucleocapsid phosphoprotein
12	viral segment	NC_045512.2	29558	29674	+	43740576	ORF10	GU280_gp11	YP_009725255.1	38	ORF10 protein

Table 1: Coordinates of know proteins in the SARS-CoV2 genome.

B. Get statistics

The genomic sequence should be read only in the positive sense, i.e. from 5’ to 3’. Read the genomic sequence and obtain the following statistics:

1. Length of the sequence.
2. Frequency (in %) of A, C, G, T.
3. GC content.
4. Number of Start (AUG) codons found.
5. Number of Stop Codons (UAA, UAG, UGA).
6. Most and less frequent codon.

C. Get ORFs

Identify all potential ORFs. Using the complete genome sequence as input, locate all the potential ORFs in the positive sense.

- An ORF is defined as the region that starts with the start codon (AUG) and ends with the stop codon (UAA, UAG, UGA).
- For a given region, if alternative start codons are found, select the longest ORF.
- Select all ORFs with a minimum length of 120 nucleotides (40 amino acids).

In this step, you should output the following information:

7. A **file** with all the protein sequences named ***all_potential_proteins.txt***, with a sequence per line.
8. A **file** with the genomic coordinates of all the ORFs, named ***orf_coordinates.txt****.* The genomic coordinates correspond the start and end position in the genome in the format:

    Start1, End1, ORF1
    Start2, End2, ORF2
    …..
    StartN, EndN, ORFN

D. Overlap with annotation

Compare the results you obtained with those from the annotation in file proteins_86693_757732.csv (see Table 1). This file contains the genomic coordinates of the ORFs that code for the different proteins. For instance, the ORFs that code for the Spike Protein (S) is (start=)21563 (end=)25384. This represents the start and end of the genomic coordinates of this ORF. Check the overlap of your ORFs with those in this table.

The overlap between sequences A and B can be calculated as:

Overlap(A, B) = $\frac{|A \cap B|}{\min(A,B)}$, where the intersection region between A and B is defined by the length of the genomic region in common between A and B. For this exercise, we will define a slightly different version of the overlap, where A is one of the annotated ORF in Table 1 and B is an ORF from your list. Thus, the overlap is defined taking in account the length of the ORF in Table 1.

Overlap (A, B) = $\frac{|A \cap B|}{|A|}$

9. For each of the ORFs in table 1 you should output the longest overlap obtained with an ORF in your list. Use the identifiers in the column "Locus" to identify the ORF, e.g.

ORF1ab	72%
S	93%
….	

So, in the case above ORF1ab has an overlap of 72% with one of your ORFs, being 72 the percentage corresponding to the longest overlap. Hint: Define a function that compares the overlap between two sequences as defined above and test for each ORF in Table 1 the overlap with each of the ORFs in your list.

10. Output

For this exercise, you should submit a file name a python script named **run.py**. Note that all submissions in different format will not be considered! The file should be run as:

Python run.py sequence.fasta

Points **1 to 6 and 9**, should be written to **output**, with **the one item per line.**
Points **7 and 8** should save the results in **files** with the names as indicated above.