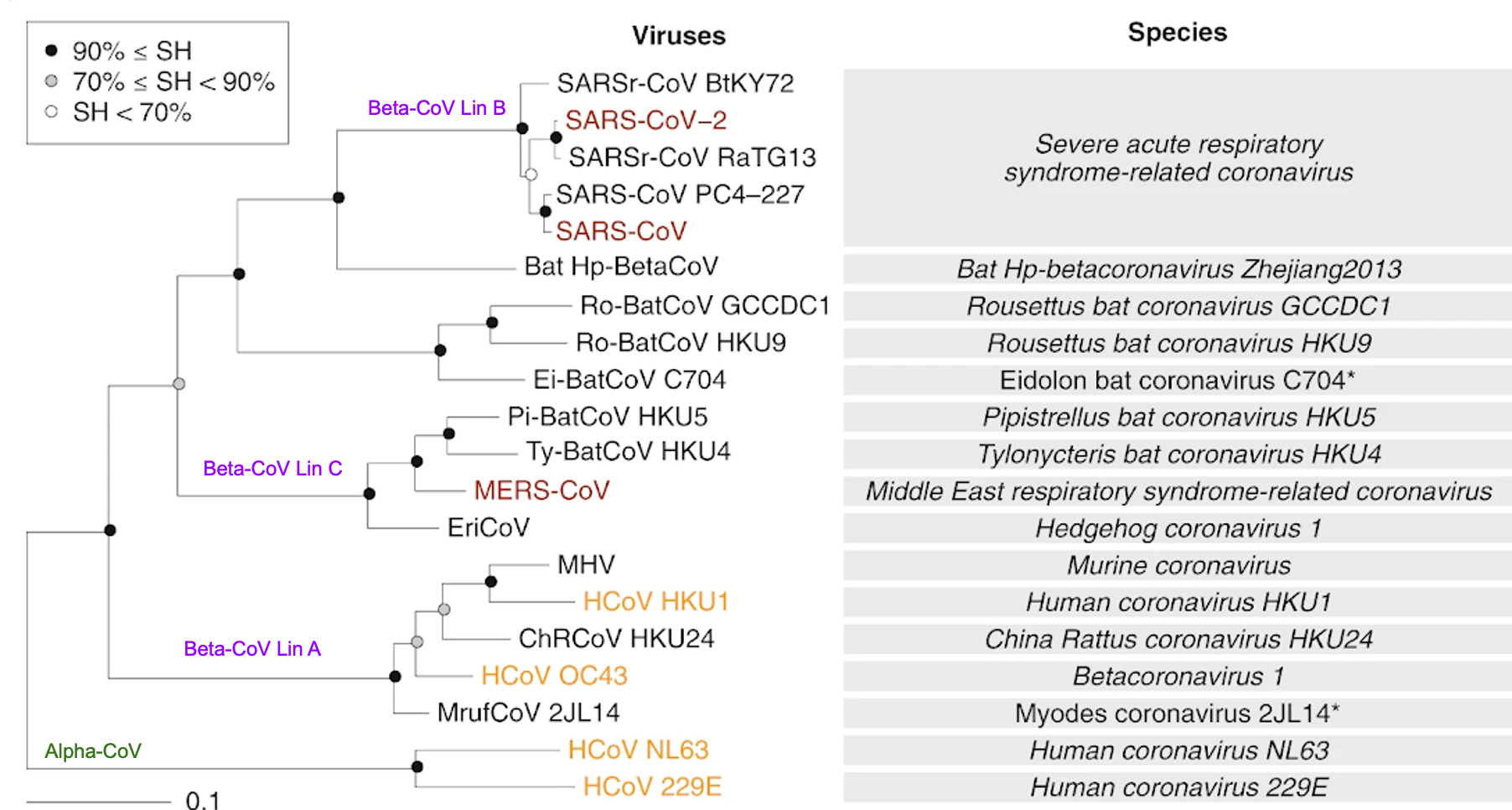
Evolution of Coronaviruses

## Learning Objectives: (a) To examine the evolutionary relationships between different Coronavirus strains, using the amino acid sequence of the Main Protease. (b) To compare the structures of the Main Protease from SARS and SARS-CoV-2

## Introduction

Evolutionary relationships between different types of coronavirus genomes can shed light on how close or distant their relationships are (in evolutionary scale). For example, the relationship between different species of coronavirus are seen in the following phylogenetic tree.



Adapted from Gorbalenya et al., 2020, Nature Microbiology, 5, 536–544

Although genomic (nucleic acid) sequences are commonly used to generate phylogenetic trees, protein sequences of homologous proteins can also be used.

This exercise focuses on comparing the amino acid sequences of the main protease from a number of different types of coronavirus.

## Download the amino acid sequence of the SARS-CoV-2 Main Protease

One source of these protein sequences is the Protein Data Bank.

* Start by saving the sequence of the SARS-CoV-2 Main Protease sequence as follows:

|  |  |
| --- | --- |
|  | In the top search bar on the RCSB PDB home page ([www.rcsb.org](http://www.rcsb.org)) type in the **PDB ID 6yb7** to open the structure summary page. |
|  | From the top right corner Download the FASTA  Save the file for use in a later section of the exercise |

### A1. Open the contents of the FASTA sequence of the SARS CoV-2 Main Protease and paste it below

|  |
| --- |
|  |

When proteins are translated, the first amino acid is commonly Methionine (or Met or M). However, the first amino acid here is a Serine (or Ser or S).

### A2. Why do you think that the first amino acid in this protein is not a Met?

## Compare sequences of 10 coronavirus main proteases

For this part of the exercise, using the same steps as in section A, the sequences of 9 other coronavirus Main proteases (MPro) were saved from the following sources

* SARS-CoV (severe acute respiratory syndrome coronavirus, PDB ID 1q2w);
* MERS (Middle East respiratory syndrome coronavirus, PDB ID 5c3n);
* PEDV (Porcine Epidemic Diarrhea Virus, PDB ID 4xfq);
* TGEV (Porcine Transmissible Gastroenteritis Virus, PDB ID 1p9u);
* HKU4 (Tylonycteris bat coronavirus HKU4, PDB ID 4yo9);
* A59 (Mouse hepatitis virus A59, PDB ID 6jij);
* HKU1 (Human coronavirus HKU1, PDB ID 3d23)
* FIPV (Feline infectious peritonitis virus, PDB ID 5eu8); and
* NL63 (Human coronavirus NL63, PDB ID 3tlo).

The file with all 10 MPro sequences is called All\_Unique\_MPro\_in\_PDB.fasta.txt Download it for use in this exercise.

* To examine the evolutionary relationships between these organisms (based on the sequences of the Main Protease) draw a phylogenetic tree

|  |  |
| --- | --- |
|  | Go to the Clustal Omega website at EMBL-EBI (<https://www.ebi.ac.uk/Tools/msa/clustalo/>). |
|  | Upload the file with all 10 Main Protease sequences All\_Unique\_Nsp5\_in\_PDB.fasta.txt as the input sequence file |
|  | Set Your Parameters to set ORDER to “Aligned” under “More Options”  Click on Submit. |

|  |
| --- |
| Exploring the Results:  Once the run is complete the results will be shown as an alignment of all the 10 sequences uploaded to the tool. Note that all the amino acid residues that are completely conserved are marked by a \* while those that have substitutions with similar amino acids are marked with a : or . symbol. Try coloring the sequences by clicking on the “Show Colors” button. |

* Among the various tabs on the result page click on the “Phylogenetic Tree” button to view the relationship between these sequences.
* Take a screenshot of the tree and import it into a powerpoint slide for labeling.

### B1. Mark the various branches with the labels of Beta-CoV (Lineage B, A, C) and Alpha-CoV. Does this match the Phylogenetic Tree in the beginning of the exercise.

|  |
| --- |
|  |

* Click on the Results Summary tab and Percent Identity Matrix to explore how closely related these sequences are.
* Read the row of numbers labeled SARS-CoV-2.

### B2. Which other coronavirus in this list is most closely related to SARS-CoV-2 and which is the least?

## BLAST Search to identify other sequences that match the SARS-CoV-2 Main Protease

In the previous section the exercise focused on the protein sequences of 10 different coronavirus Main Proteases that were submitted to the PDB. In the next section the SARS-CoV-2 Main Protease sequence will be used to search and find other matches.

The goal of this exercise is to identify other coronavirus that are really closely related to SARS-CoV-2. Perhaps this will shed light on where the virus came from.

* Go to the NCBI interface for running BLAST search and click on the Protein Blast box. Alternatively go directly to <https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome>

|  |  |
| --- | --- |
|  | Upload the FASTA format sequence of the SARS-CoV-2 Main Protease (from PDB ID 6yb7) |
|  | Keep all default options and click on the BLAST button |

Once the BLAST run is complete (after a few minutes) examine the results page and answer the following questions.

### C1. What is the organism and protein that is the most common match in the result list?

* Repeat the BLAST search with some modified parameters as follows:

|  |  |
| --- | --- |
|  | Click on the Edit Search box on the top of the page to return to the page for setting up options for the search |
|  | Note that the Query sequence is already filled in (from the FASTA file uploaded earlier) |
|  | Update the settings so that the results are reported after excluding all SARS-CoV-2 sequences - i.e., the results will show matches from other sources |
|  | Before starting the search, click on the Algorithm Parameters to select 50 in the Maximum target sequences. This is the number of results that will be returned.  Click on the BLAST button to start the search. |

### C2. Name 3 other virus names whose sequences matched the query with the revised settings.

Hint: Click on the Taxonomy tab in the results section to see the source(s) of the matched sequences.

## Compare the structures of SARS-CoV and SARS-CoV-2 Main Proteases

Of all the different viruses that closely matched SARS-CoV-2 the only structure that we have in the PDB (as of Summer 2020) is the SARS-CoV protease. This is the virus that caused the SARS epidemic in 2002-2003.

* In this section we will superpose the structures of the main protease from SARS-CoV (PDB ID 1q2w) and SARS-CoV-2 (PDB ID 6yb7)
* Go to the <https://www.rcsb.org/3d-view/>

|  |  |
| --- | --- |
|  | In the Download Structure options >> Type in the PDB IDs in the box next to PDB Id(s) >>  “6yb7, 1q2w”  Click on Apply |
|  | Both structures will show up in the 3D-Canvas  The 2 symmetry related chains of the SARS-CoV-2 protease in PDB Id 6yb7 (Chain A ASM\_1 and Chain A ASM\_2) are colored green  The 2 chains of SARS CoV in PDB Id 1q2w (Chains A and B) - are colored orange and purple |

* Select the chains from each of the structures for superposition as follows:
  + Activate the selection mode by clicking on the arrow icon A picture containing table

    Description automatically generated
  + Change the picking level to “Chain” by clicking on the word “Residue” to see options in the pulldown menu

The goal here is to align the chains labeled A from PDB ID 6yb7 and and PDB ID 1q2w.

* + Select Chain A in the PDB entry 6yb7 by clicking in the sequence panel
  + Change the PDB entry to 1q2w in the menus at the top of the sequence panel
  + Select chain A in this structure too
  + In the 3D canvas, Chain A of both the structures should be highlighted with a green halo

|  |  |
| --- | --- |
|  | In the Superposition section of the control panel click on “By Chains” >> you should see the chains that you selected - chain ID A in both structures listed here  Click on Superpose  Once the superposition is done click anywhere in the white space in the 3D Canvas to clear selection. |

* Note the rmsd (root mean square deviation) between atoms in these chains listed in the log panel (bottom of the 3D Canvas).

### D1. Save an image of the superposed structures and record the rmsd value below. Do you think that the two structures are similar or different? What does this mean in terms of evolutionary relationships between SARS-CoV and SARS-CoV-2?

|  |
| --- |
|  |

### D2. Compared to the SARS-CoV-2 Main Protease, do you think that the structures of pangolin and bat Main Proteases will be similar or different? Explain your answer.