**INTRODUCTION TO NUCLEOTIDE SEQUENCE AND PROTEIN ANALYSIS IN MEGA AND PyMol USING CORONAVIRUS SARS-Cov-2 AND OTHER RELATED PATHOGENS**

*Nicholas Lorusso, Maria Shumskaya, Kean University, 1000 Morris Ave, Union, NJ 07083*

**Disclaimer**: Due to the rapid emerging nature of data on the novel coronavirus 2019-nCoV, this exercise provides not the most current but rather represent predictions based on publications available on 3/30/2020. Nucleotide sequences were downloaded from NCBI GenBank and then trimmed for educational purposes. The authors will appreciate any modifications suggested by colleagues.

This exercise is designed to introduce a learner into how a variety of computational approaches can be used to answer some biological questions and is based around a specific example: coronavirus SARS-CoV-2, which caused 2019-2020 pandemic.

The exercise consists of two assignments with each given in one 2.5 hours standard lab periods.

During week 1, publicly available nucleotide sequences of SARS-CoV-2 and other viruses will be used to compare similarity and create hypotheses for the relationships between viral species. A freely available software, MEGA (Kumar et al. 2018), will be used to compare different RNA viruses within the *Coronaviridae*. The learner will 1) align sequences for the RNA-dependent RNA polymerase (RdRP) gene, 2) create a hypothetical phylogeny using maximum parsimony, and 3) create a second hypothetical phylogeny using maximum likelihood. We will then compare the produced predictions and consider the strengths and weaknesses of either approach.

During week 2, the learner will use a free for education PyMol ("The PyMOL Molecular Graphics System" 2010) software to visualize a receptor-binding domain of the spike protein of SARS-CoV-2 together with the host receptor ACE2, and label regions and amino acids important for binding to the receptor.

**Learning objectives**

After successful completion of this exercise, students will be able to:

* Develop abilities necessary for understandings about scientific inquiries.
* Identify appropriate computational approaches addressing biological questions.
* Evaluate data and graphical representation.
* Critically assess experimental results and draw conclusions based on the data.
* Apply bioinformatics methods to solve biological problems:
  + Find and interpret data from major online databases such as NCBI GenBank and PDB.
  + Use basic bioinformatics software such as MEGA and PyMol and analyze results provided by such software.

Table of Contents

[LAB 1 3](#_Toc36818414)

[Introduction to NCBI database and phylogeny in MEGA using SARS-CoV-2 and other pathogens 3](#_Toc36818415)

[Background 3](#_Toc36818416)

[1. Install MEGA and download data files 4](#_Toc36818417)

[2. Accessing the NCBI database 4](#_Toc36818418)

[3. Aligning sequences retrieved from available databases 5](#_Toc36818419)

[4. Creating predictive phylogenies using maximum likelihood and maximum parsimony. 7](#_Toc36818420)

[LAB 2 11](#_Toc36818421)

[Introduction to PyMol molecular modeling using SARS-CoV-2 S-protein with ACE2 receptor 11](#_Toc36818422)

[Background 11](#_Toc36818423)

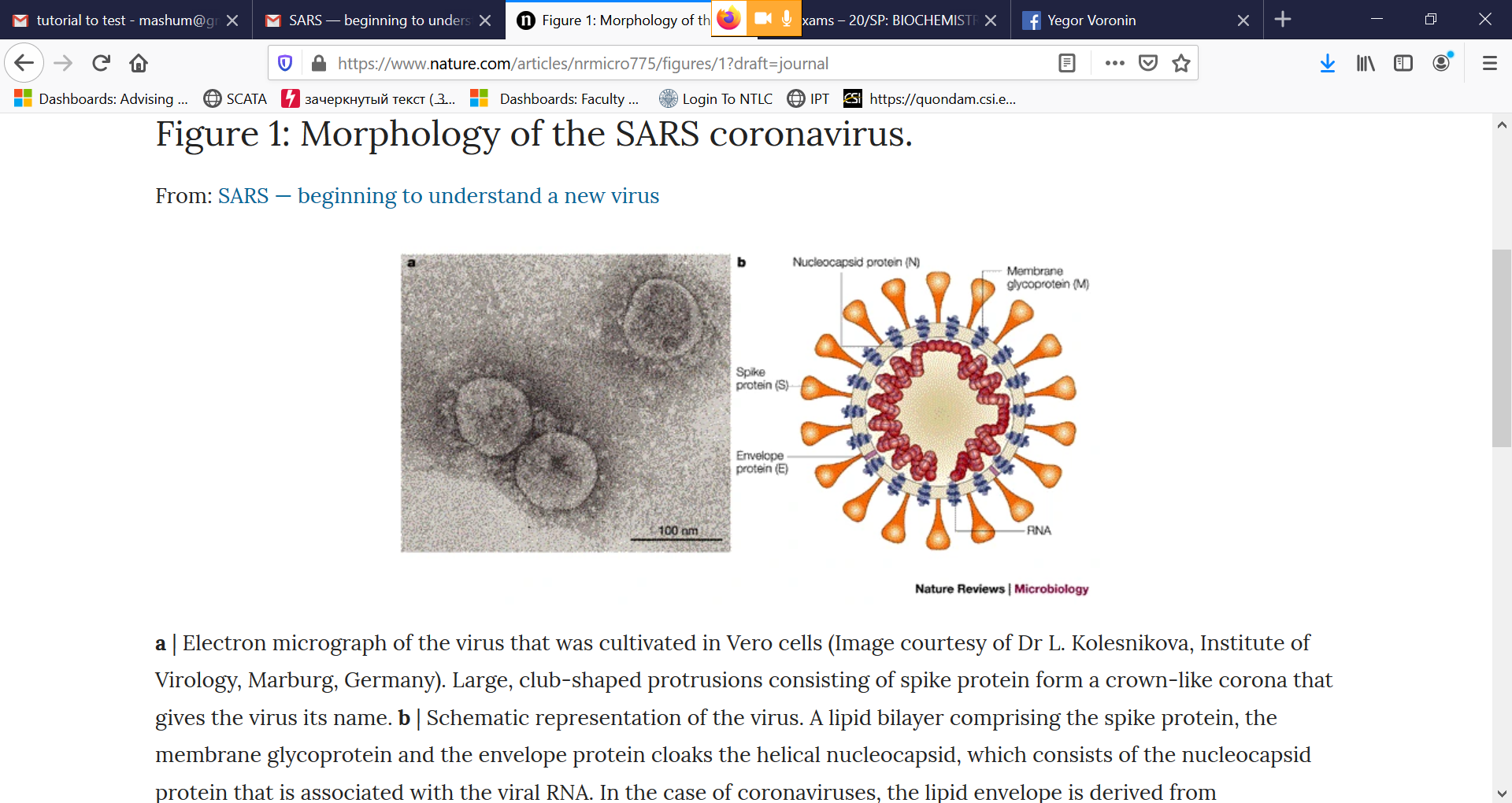
[1. Install PyMol 11](#_Toc36818424)

[2. Accessing the Protein Data Bank (PDB) 11](#_Toc36818425)

[3. Viewing PyMol and Basic Commands 13](#_Toc36818426)

[4. Important Menus in PyMol. 14](#_Toc36818427)

[5. 3D modeling of SARS-CoV-2 and ACE2 protein complex 16](#_Toc36818428)



**Morphology of the SARS coronavirus. a**: Electron micrograph of the SARS-CoV virus that was cultivated in Vero cells (Image courtesy of Dr L. Kolesnikova, Institute of Virology, Marburg, Germany). Large, club-shaped protrusions consisting of spike protein form a crown-like corona that gives the virus its name. **b** Schematic representation of the virus. From (Stadler et al. 2003).

# LAB 1

# Introduction to NCBI database and phylogeny in MEGA using SARS-CoV-2 and other pathogens

## Background

Viral pathogens pose a number of consequences for agriculture, biodiversity, and human health; as such, they are a constant focus in various fields of biological research. The fact that genomes of viral taxa are subject to a rapid rate of mutation (Domingo and Perales 2019) and as a result viruses can shift into novel hosts (Martinez et al. 2014) makes understanding how viral genomes change between related strains critically important. Recent outbreaks such as SARS-CoV-2 likely emerged from these types of shifts between animal hosts in China (Zhang et al. 2020), with a reservoir in bat bats such as *Rhinolophus affinis* (Andersen et al. 2020). Given these opportunistic shifts for viral pathogens, evaluating the relationships between potentially dangerous related taxa is one way of better predicting and understanding viruses with economic or medical consequences. One challenge, given their microscopic nature, is that viruses are most easily studied using phylogenetic comparison, which requires tailored questions and research methods.

Single-stranded RNA viruses in the family *Coronaviridae* have posed a number of health risks to humans since the turn of the 21st century. A number related conditions have resulted from members of this family such as sudden acute respiratory syndrome (SARS, caused by SARS-CoV), Middle East Respiratory Syndrome (MERS, caused by MERS-CoV), and most recently Coronavirus disease (COVID-19, caused by SARS-CoV-2) and have become a recent focus for research in the biomedical sciences, virology, epidemiology, and evolutionary biology. One major issue confronting this research, however, is that studying viruses can be difficult due to their small size and the resulting need to use specific technical approaches. Despite the challenges, establishing relationships between related viruses can make important connections for researchers attempting to target potential pathogens of concern or develop vaccines between related taxa.

This exercise is intended for anyone interested in better understanding how scientists determine the relationships between different viral taxa and strains. While there are a number of ways to accomplish this we will guide you through a specific phylogenetic approach meant to **compare and visualize relationships of Coronavirus-19 (COVID-19) to other pathogenic taxa**. To make this type of comparison we will use publicly available data of RNA sequences from the genomes of viral species of interest. We will process the data to determine similarity between the sequences and compare them visually using a phylogenetic tree, also called a **phylogeny** (or a **cladogram**), showing the evolutionary relationship of multiple species to each other (Figure 1). These types of trees depict hypotheses as to where species share common ancestors, and how a lineage has diversified over time, and how individuals can be grouped into **clades** (groups of organisms believed to have evolved from a common ancestor).

Then, we will evaluate the quality of the prediction of the relationship the phylogeny predicts. There are many ways to do evolutionary analysis, and ultimately you need to read many papers to decide what you think is the best way to do it. This exercise is a very basic one and can be considered a starting point for learning more about the topic.

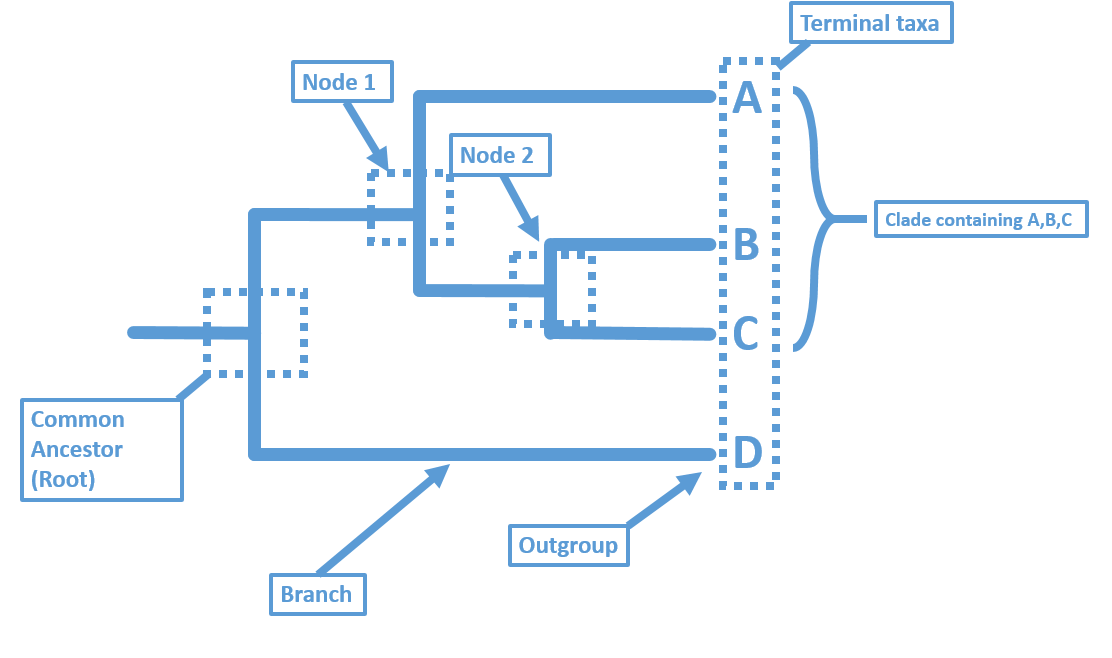


Figure 1. Phylogenetic tree. B and C – sister taxa, they are most closely related and share an ancestor at the node 2. Both B and C are equally related to A and share an ancestor at the node 1. D is a basal taxon and an outgroup for the clade containing A, B, and C.

In order to build a phylogenetic tree, we need to choose one gene to compare to each other. We have selected the RdRp gene as it is unique to RNA viruses and is universal within that group, since it allows for replication of the viral genome. Given that it is found in all RNA viruses, it is a good candidate gene for evaluating changes in nucleotide sequences (depending of the research question, other genes can be used as well).We will use MEGA 7 (or more modern version MEGA X) free program to construct such a tree for coronaviruses. This program can compare sets of RNA sequences and suggest a cladogram based on the analysis. Make sure you download the proper version for your operational system.

### Install MEGA and download data files

* 1. Go to <http://www.megasoftware.net/>. Choose your operating system and make sure you download a version with the GUI (graphic user interface). Install following the prompts after download completes.
  2. Download and save data file “coronavirusesRdRp.fas” to work with in MEGA. The file will be provided by your instructor.

### Accessing the NCBI database

Let’s retrieve a sequence for RdRP from an online data repository to visualize what the data going into the analysis looks like.

* 1. Go to <https://www.ncbi.nlm.nih.gov/>
  2. Change “All databases” to “Nucleotide”. In the search line, type “RdRp” and click “Search”.

**Question 1.** How many items show up in the results?

**Question 2.** Write the first sequence that appears in the search results. What is its accession number?

* 1. As you figured out, RdRp occurs in more than just a coronavirus. Let’s restrict the search to bat coronaviruses. In the right top corner, find “Results by taxon” and click on “More…” to display all available organisms. Click on “Bat coronavirus”.

**Question 3.** How many results appear in the search? What is the accession number of the first result in the search?

* 1. Click on the first entry to open it.

**Question 4.** Who are the authors of this DNA sequence?

* 1. Scroll down to see the nucleotide sequence.

**Question 5**. How many nucleotides are in it?

You can now observe a lot of information about the found gene, the evidence supporting its annotation, and potentially related sequences. We’re mostly focused on the **sequence** for the gene – so under the heading and GenBank number, select “**FASTA**” to view the sequence for the gene. This sequence is the type of input we will use to make our comparison. See Figure 2 for an example of such sequence from a different virus.

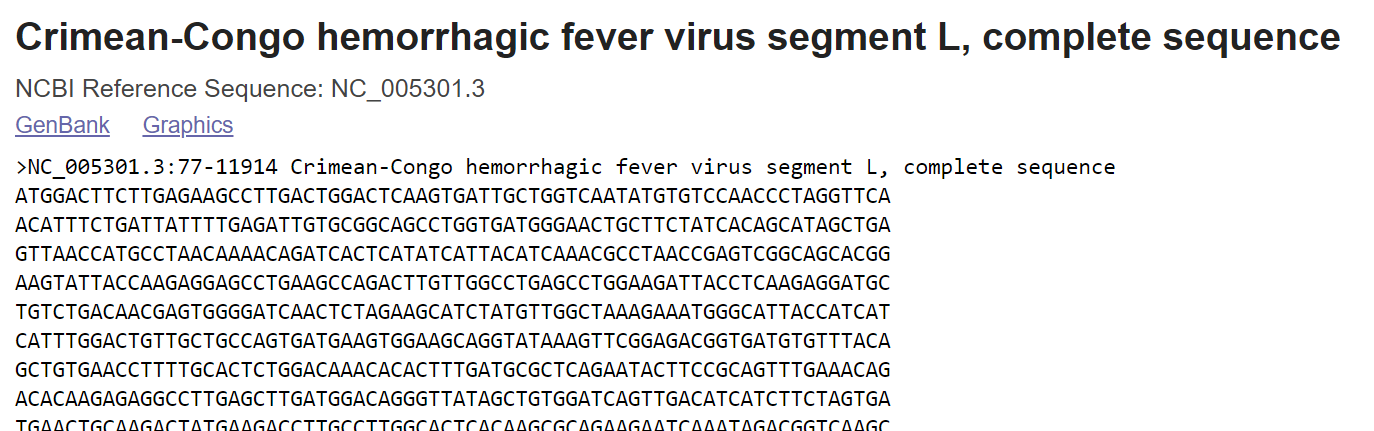


Figure 2. Excerpt of the RdRP gene from Crimean-Congo hemorrhagic fever virus

### Aligning sequences retrieved from available databases

To simplify the assignment, we have downloaded several available RdRp sequences for coronaviruses from NCBI and put them together in a document in ‘fasta’ format. You should have downloaded this file at step 1 of this exercise. ‘Fasta’ is a format that can be opened by Notepad (try to see what is inside!) and is used by many bioinformatics software. Luckily for researches of viral pathogens, the genomes of viruses are relatively short and obtaining reference sequence data for pandemics such as COVID-19 can happen relatively quickly (Wang et al. 2020). These data make it possible to compare with other related taxa to understand the origins of viral outbreaks and help create methods of control (e.g. vaccines).

* 1. Open MEGA and follow the following procedure to create an alignment file that can be used for comparison of the viral taxa under consideration.

Before you construct your phylogenetic tree, you need to **align** the sequences in your file. Alignment is always done first using a computer program to get you started. A good alignment is extremely important, as your later analysis and results will be dependent on it. A **sequence alignment** is a way of arranging the sequences to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships (*Wikipedia*). If sequences look similar, but are of different lengths, then **gaps** (empty spaces) need to be inserted to make the sequences line up (=align). Where such gaps should go can be hard to determine. Since each nucleotide position will be a character in the phylogenetic analysis, it is important that each nucleotide is put in a reasonable place in the alignment.

* 1. Open the “coronavirusesRdRp.fas” by navigating to “File” -> “Open a File/Session” and selecting the file from its location on your computer. You will be asked if you want to align or analyze – choose align.
  2. The new window that opens will contain sequences of 11 different viral taxa for alignment. To use MEGA to perform an automated alignment, navigate to the “Alignment” menu in the alignment editor, select “Align by MUSCLE”, and click “OK” (use default parameters).

In the resulting output the sequences should be aligned and you should see some columns with colored letters or areas where there are structural differences (e.g. insertions or deletions) between samples (shown as white dash marks). Each row is a taxon we are comparing. Your alignment will output similar to the image below.

**Question 6.** What is the color for adenine nucleotide (A)? For guanine (G)?

* 1. Scroll the far left and far right of the data. Notice aligned (conservative) regions of your sequences, and non-conservative regions substituted with gaps or other nucleotides in other sequences.
  2. Save the alignment by navigating to “Data” then “Export alignment” and select “MEGA” as the format. Name the file “covid\_alignment” and close the alignment window.

### Creating predictive phylogenies using maximum likelihood and maximum parsimony.

We will now take the alignment we created in the previous steps and use it to compare taxa by their sequence similarity. To do this – we will use both **Maximum Parsimony** and **Maximum Likelihood** methods, which both allow inference about hypothetical relationships in different ways.

**Maximum Parsimony** methods evaluate a large (ideally infinite) number of trees that could all be possible and then considers which of those trees requires the least number of genetic changes across the members. The tree with has the fewest number of changes to explain the data is considered most likely.

**Maximum Likelihood** is similar to parsimony methods in that it considers a large number of potential trees but instead of selecting which tree would require the fewest number of changes as best – this method uses statistical models which incorporate *how likely* a given changes (e.g. insertions, transversions, etc.) are and considers trees that fit those models as best.

#### Building a tree using Maximum Parsimony

#### In the main window for MEGA navigate to “Analysis” -> “Construct/test tree using Maximum Parsimony” and this should bring up a dialog window. You will need to use your ‘covid\_alignment.meg’ file. There are many options for this type of analysis but for the purposes of this exercise we recommend using the following be selected:

**Analysis Preferences** = Phylogeny reconstruction  
**Statistical Method** = Maximum parsimony  
**Test of phylogeny** = Bootstrap method  
**No. of Bootstrap Replications** = 100  
**MP Search Method** = Tree-Bisection-Reconnection (TBR)  
**No. of initial trees (Random addition)** = 10  
**MP Search Level** = 1  
**Max No. of Trees to Retain** = 100

Click **Run** and wait for the analysis to complete (This can take up to 10 minutes depending on your computer specifications). Each clade on the tree is an evolutionary hypothesis. It would be good to know which of these hypotheses are supported by more data, which are supported by less. While you are waiting, MEGA is performing a **Bootstrapping method** to generate a measure of support for the different branches on the evolutionary tree. After the software has considered the possible trees you should see something like Figure 3.

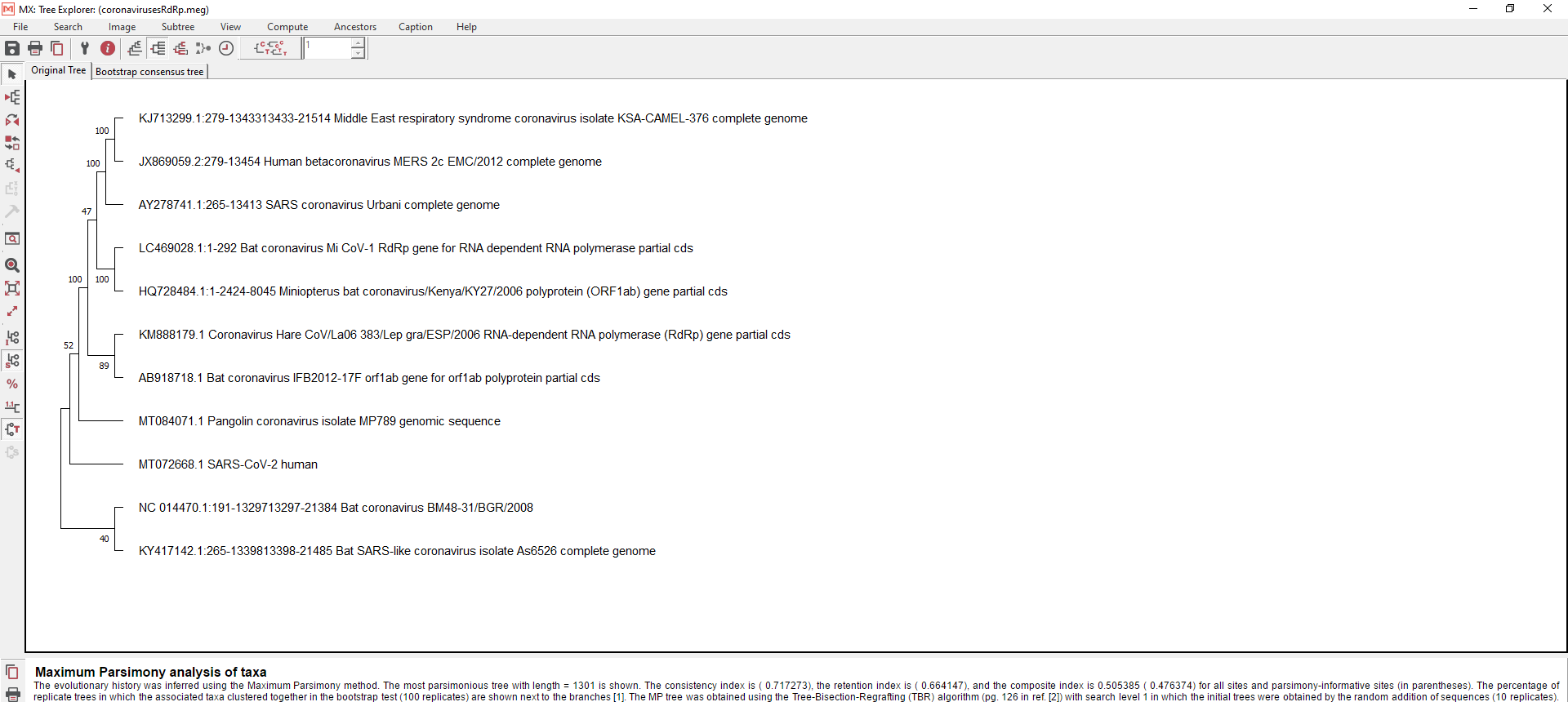


Figure 3. Maximum parismony tree. In this tree the number on the branches represents the likelihood of the relationship shown occurring out of the total number of trees considered (closer to 100 is more likely).

* + 1. Before going further, you should learn how to designate an **outgroup**, and **roo**t the tree to make sense of it. An outgroup should be an isolate, or small group of isolates, that are known to be more distantly related than all the other isolates being analyzed, but similar enough that the RNA can still be aligned. A good choice of outgroup in our case is an isolate from *Botryotinia fuckeliana* totivirus. Select an outgroup for the tree by clicking the button on the left hand pane of the tree viewer with the red arrow () and then selecting “*Botryotinia fuckeliana* totivirus” as the outgroup. This will root the tree and will examine the relationships between the remaining taxa relative to that taxa. There should only be one tree.

*Alternatively, given the role of bats as reservoirs for CoV-2 in China, you can root the tree by selecting “*Bat coronavirus IFB2012…”*– though this will produce a different presentation of the relationships.*

* + 1. Be sure that in the tree view window you are in the “Original Tree” tab.

**Question 7.** Save this tree for future comparison using the Copy button in the tool bar to save it in a word document. Write a caption for this tree.

* + 1. Close the tree after saving to proceed to the next exercise.

**Question 8.** Observing the relationships shown on the resulting tree, how confident are you in the relationship between SARS-CoV, SARS-CoV-2, SARS-like coronavirus from pangolin and Bat SARS-like coronavirus?

**Question 9.** Based on this tree, and only this tree, would you expect MERS or SARS-CoV to be more similar to SARS-CoV-2?

#### Building a Tree using Maximum Likelihood

Since Maximum Likelihood relies on the selection of a statistical model to choose a best tree, we **first need to select an appropriate model for MEGA to use**. MEGA can actually compare models based off of model fit scores (e.g. AICc or BIC) to help select the appropriate model.

* + 1. Click “Models” in the MEGA tool bar and select “Find Best DNA/Protein Models (ML)” from the dropdown menu. **If prompted to reselect your alignment file, choose the file we made in the first part of the exercise from your computer.** Click OK.
    2. Observe the results in the table produced by MEGA. It might seem like a daunting amount of information but all we are interested in is the row at the very top showing the “best” model based on our data (which has the lowest fit scores). **Record the model description from the first column and whether it says “+G” or “+I”.**

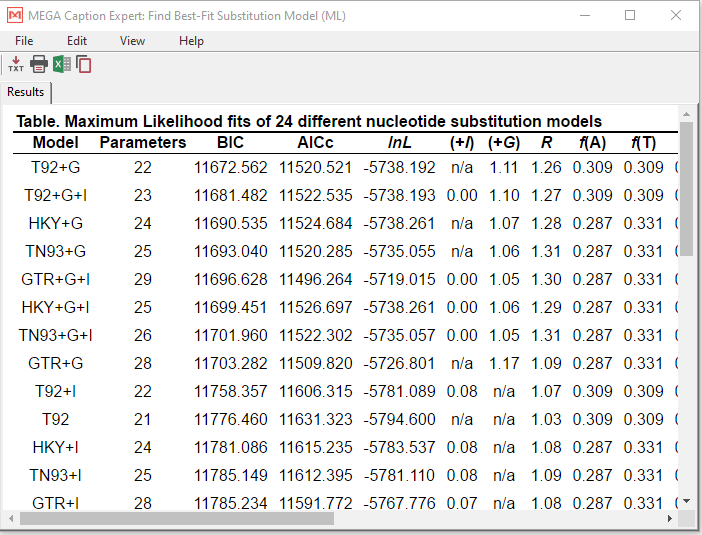


Figure 4. Comparison of the models MEGA can use for Maximum Likelihood analysis. Info from this table is red in the parameters below

* + 1. Close the window when you’ve recorded the information from the second step and then navigate to “Analysis” and select “Construct/Test Maximum Likelihood tree” from the “Phylogeny” menu. For the options in the following menu we recommend:

**Test of phylogeny** = Bootstrap method

**No. of Bootstrap Replications** = 100

**Model/Method =** This is where you put the model from the preceding step using the initials from the model fitting step to choose the correct model (based on the above its T92 or Tamura 3-parameter model)

**Rate among sites: G** if +G, **I** if +I from the model fit above

**No. Discrete gamma categories =** if G+ set to 5

**Gaps/missing data:** Complete deletion

**Branch swap filter:** None

* + 1. Root the tree as before – using the “Botryotinia fuckeliana totivirus” as the outgroup. Observe this and compare it to the tree produced using maximum parsimony earlier. Save this tree as before.
    2. Bootstrap values offer a convenient way of assessing how much trust we can put in the branching we see in our tree. MEGA offers us the ability to filter or “condense” the tree to only see relationships that most models predict to be true. To see this tree select “Compute” from the menu and then “Condensed tree”. The cutoff will hide branches that the model predicts less than a certain percentage of the time – we recommend you set this to 70%. Click OK.
    3. Observe changes to the tree and compare it with your uncondensed tree. Save it and then save any other files before closing MEGA.

**Question 10.** Save this tree for future comparison using the Copy button in the tool bar to save it in a word document. Write a caption for this tree.

**Question 11**. How did you ML and MP trees compare to each other in terms of relationships of coronaviruses? What about in terms of support from bootstrapping?

**Question 12.** Which method seems more useful to you and why? Can you imagine a reason to use the method you think is less useful?

**Question 13.** Based on the support: how confident are you that the use of RdRP is useful in comparing strains of viruses in the *Coronaviridae*?

# LAB 2

# Introduction to PyMol molecular modeling using SARS-CoV-2 S-protein with ACE2 receptor

## Background

Coronavirus entry into host cells is mediated by the transmembrane spike (S) glycoprotein that forms a homotrimer. The CoV S-protein is a key target for vaccines, therapeutic antibodies, and diagnostics. S-protein comprises of two subunits. One is responsible for binding to the host cell receptor (S1 subunit), and the other is responsible for fusion of the viral and cellular membranes (S2 subunit). The S1 subunit comprises the receptor-binding domain that binds to ACE2 receptor located in the membrane of the host cells and stabilize the pre-fusion state of S2 subunit. S2 subunit is membrane-anchored and contains the fusion machinery. Upon binding to the cell membrane, S-protein is cleaved by proteases. This cleavage has been proposed to activate the protein for membrane fusion via irreversible conformational changes. As a result, coronavirus entry into cells is a complex process that requires binding to the ACE2 receptor followed by proteolytic processing to promote virus-cell fusion (Walls et al. 2020).

In this lab, we will build a 3D model of the S1 subunit of the S-protein (a monomer in our model) from SARS-CoV-2, complexed with its receptor ACE2 and label some of its parts. We will highlight several amino acids important for binding with ACE2 and explore the position of some amino acids that could be recognized by antibodies.

We will use PyMol ("The PyMOL Molecular Graphics System" 2010) program, which is free for educational use; make sure you download the proper version.

### Install PyMol

* 1. Go to<https://pymol.org/edu/?q=educational/>.
  2. The system will ask for your name and email.
  3. Download and install **PyMOL 1.3r1 edu** **(Sept 2010)**.

**Note:** You need to install the educational version, NOT the paid version.

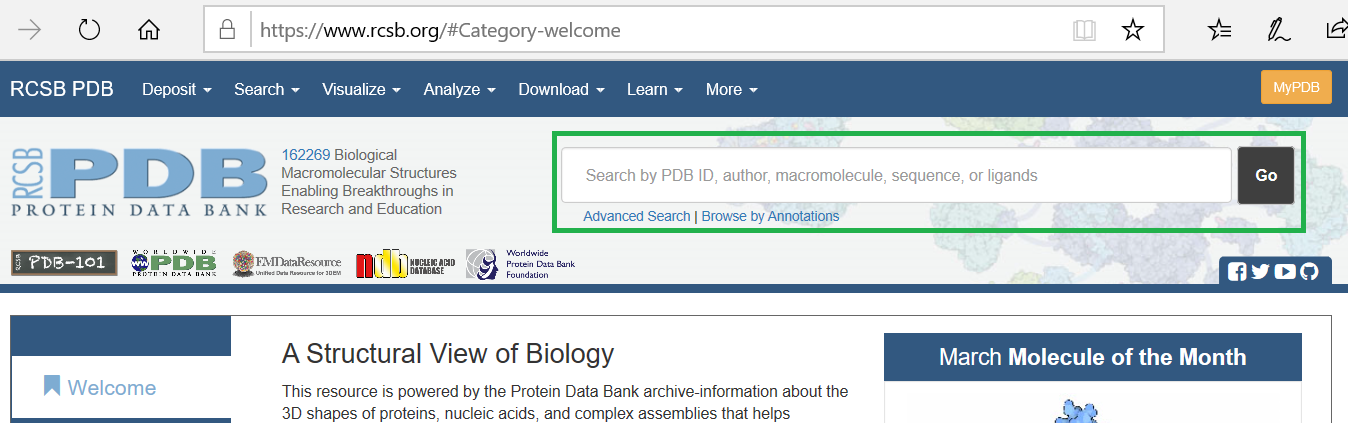
* 1. Once installed, select from the Start menu PyMol + GUI program.

### Accessing the Protein Data Bank (PDB)

The Protein Data Bank website contains thousands of downloadable structures that were created from X-ray crystallography or NMR experiments. These experiments allow scientists to determine the 3D coordinates of every non-hydrogen atom in the molecule. Computer software then adds bonds and secondary structural information based on the protein's known sequence and measured dihedral angles. The resulting structures are then assembled into a pdb file and are deposited into the database for the world's scientific community to access.

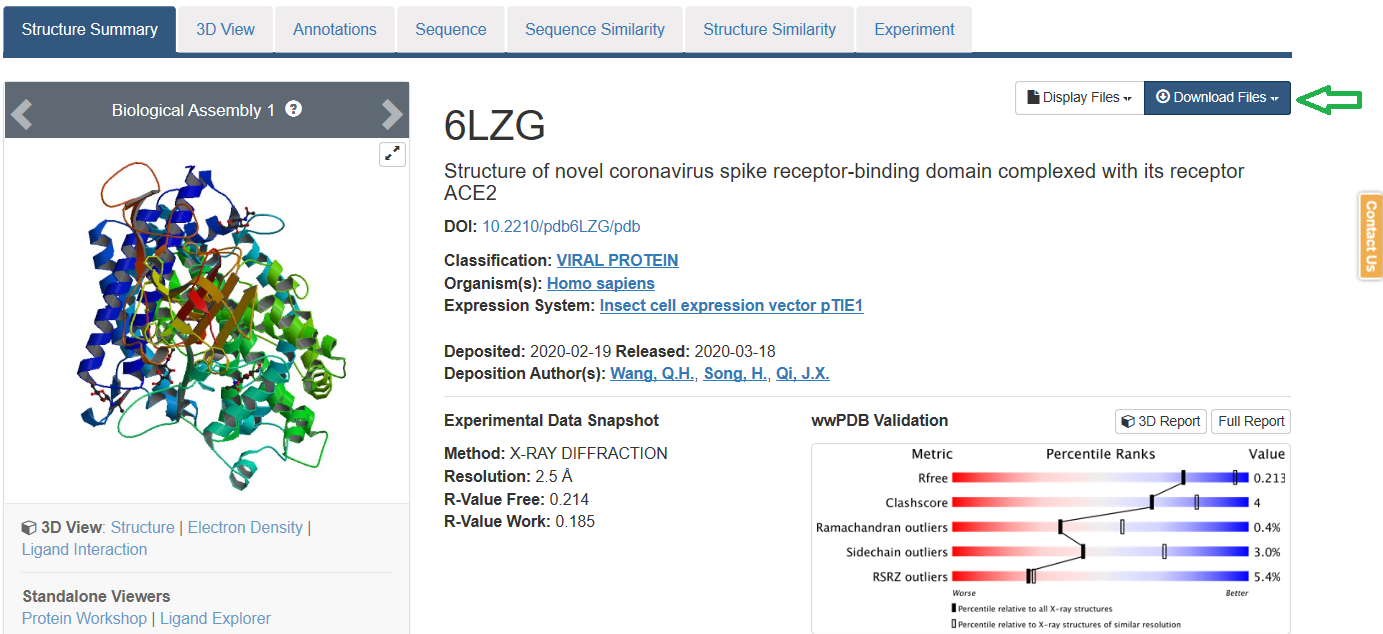
We will practice using the file with the monomer of SARS-CoV-2 spike glycoprotein bound to its receptor ACE2. To view the structure of a protein, you must download a .pdb file into the computer.

* 1. To download a .pdb file, first access the Protein Data Bank<https://www.rcsb.org/> or by googling "protein data bank".
  2. To find a protein's pdb file, use the search bar on the PDB website.

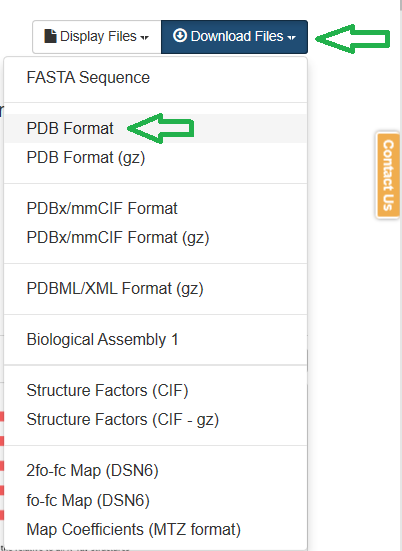


* 1. Enter a pdb code 6LZG.

* 1. Once you access the page for the structure 6LZG, click on "Download Files" on the right side of the page.



* 1. From the popup menu, click PDB Format. **Note:** The file 6LZG.pdb will download onto the computer.



* 1. To open it, go to the computer's downloads folder and open 6LZG.pdb. The structure should automatically open into PyMol.

**Troubleshooting**

If you have other molecular viewing programs on your computer, it is possible that PyMol is not the default-viewing program. If you cannot view the structure simply by clicking on the pdb file download, open the PyMol program itself and load the structure directly from PyMol by clicking **File>Open.**

If the structure does not open in PyMol, make sure you downloaded the PyMol File and not one of the other files accessible from the popup menu.

**Note:** If you did not download PyMol, the structure will not open.

### Viewing PyMol and Basic Commands

#### Once you have opened PyMol and loaded a PDB structure, you will notice two parts of the screen.

* On a PC, there are two separate windows, one thin rectangular window and one large viewing window. You need both windows to be visible to use PyMol.

**Note:** If you cannot see both screens (on the PC), make sure they are not hidden behind other windows. If you lost one or both of the windows, just reopen PyMol, and both windows should appear.

* On a Mac, the two areas are connected into one window.

### Important Menus in PyMol.

Please read through the information below and practice clicking on the menus. This information will be needed in your assignment.

#### Top screen: File

* **Save Session As:** Saves your work as a PyMol file (.pse) for editing later.

Save your structure now so that in case you make any mistakes, you can start again.

**Note:** PyMol does not have an "undo" function, so save your sessions frequently as you work; it may not be easy to correct a major mistake without reloading a saved version of your session.

* **Save Image As:** Saves a PNG image of the viewing screen. This command saves the structure's view as it appears in the main viewing screen as an image, like a jpeg or a tiff file. You must save the image as a PNG picture file if you want to use the structure in a word processing document to hand in. The PyMol session file (pse) can only be used for viewing a structure in PyMol since it contains the 3D coordinates of the structure's atoms. *You cannot insert PyMol session files (pse) files into a word processing document.*

All work done in PyMol must be saved as a PNG image and then inserted into a Word or Pages document for submission.

#### Top screen: Display

* **Sequence On**: Shows the amino acid residue sequence of the structure.

Show the sequence now. After turning on the sequence, you should see numbers and letters along the top of the main window. The numbers are the residue numbers of each amino acid, and the letters are the one-letter codes for each amino acid that the protein is composed of.

Left click and hold the gray slider to scroll to the right to see the entire sequence of the protein. Notice that first there is a sequence more than 600 amino acids long. This is a sequence of ACE2 receptor located in human cellular membranes. After it, notice the NAG – it stands for N-acetyl-D-glucosamine, one of the ligands, and Zn (zinc). After the Zn, notice the string of hundreds of "0". These are oxygen atoms of water molecules which were crystalized along with the protein. At the very end of these oxygen atoms, notice another amino acid sequence which starts with 333; this is the sequence of S1 subunit of S-protein. It also ends with NAG and a series of “0”. Scroll back to the beginning of the sequence and continue to the next steps.

* **Background:** Click on some of the background color options. Using a black background is usually best for viewing and exploring a structure, while white is usually preferred when completing a structure for inserting into assignment. Change the background to white now, then change back to black.
* **Bottom right panel: Mouse Mode:** There are two modes in PyMol: 3-Button Viewing and 3-Button Editing. It is extremely important that you are aware of which mode you are in. Many problems will come from accidentally being in the wrong mode. For simple commands, you must be in the 3-Button Viewing mode.

Practice toggling between Viewing and Editing modes by clicking on 3-Button Viewing and 3-Button Editing. If you do not click precisely on what you want, you could easily enter the wrong viewing mode, and then the program will not behave as expected, so always check your viewing mode if a problem arises. **Selecting State:** At the very bottom right, look at "Selecting State". This is also an extremely important function of PyMol. Selecting State allows you to change what kind of structure you are selecting. When you open a new PDB file, it will be set to "Residues". This means that when you left click on the sequence at the top of the screen, you will select only one residue (amino acid) at a time.

Practice selecting single amino acid residues now. Make sure your sequence is scrolled all the way to the left. Left click on the first amino acid residue, S. Then left click on T, then left click on Q. You have now selected the amino acids serine 1, threonine 3, and glutamine 6. Notice that when you select an amino acid in the sequence, it is highlighted with small pink squares in the viewing window. To unselect the amino acids, simply left click again on the highlighted amino acid residues in the sequence. A way to unselect everything is to left click on the blank space around the structure in the viewing screen. You can also highlight individual amino acid residues directly from the 3D structure itself, but this can be a bit more challenging for new users.

Troubleshooting: If you are having trouble selecting residues, check that you are clicking on the sequence letters, and not the numbers. PyMol is very sensitive to where you click, and it will not select anything unless you click directly on the amino acid letters. Also always check that you are using the left mouse button and that you are in the correct selecting state and viewing modes.

Change the Selecting State by clicking on Residues to toggle through the options. (To toggle, left click repeatedly on the word next to "Selecting State" so that new options appear. Toggling through the selecting residues shows that you can select Chains, Segments, Objects, Molecules, C-alphas (alpha carbons), and atoms. Toggle until you get to "Objects". \*Be sure not to accidentally switch the **mouse mode**; if you click the middle of the bottom right box, you could easily change it and then you will not be able to toggle Selecting State.

Now that you have selected "objects" click anywhere on the letters of the sequence at the top of the screen. Notice that the entire sequence will become highlighted, and that the structure will become highlighted in small pink squares. Practice highlighting and un-highlighting the whole structure. *Notice that you can also un-highlight by clicking on the blank space around the structure.*

### 3D modeling of SARS-CoV-2 and ACE2 protein complex

**A.** With the whole object selected, find the gray bar labeled "(sele)" on the right panel. The five boxes A, S, H, L, and C allow you to control many attributes of what is selected. Left click on "H" for the "Hide" popup menu. From the Hide popup menu, click "waters" to hide all the water molecules. Once you have done this, unselect the object by clicking on the amino acid sequence or the blank space around the structure. Notice that all the red stars around the structure have been erased. It is usually a good idea to remove the water molecules from a crystal structure before rendering and viewing the structure. Showing the position of water molecules in the crystal structure is *usually* not very useful.

\* Altering a molecule's parameters can be accomplished by *right clicking* on the selection, or by *left clicking* on the small boxes in the "(sele)" selection on PyMol's right panel, as directed in the written directions. Either selection method can perform the same commands.

**B.** Our structure is still very difficult to understand. We will now begin to use PyMol to render a structure that displays useful information. Select the entire object again and hide everything. This will give a blank screen once we unselect by clicking the blank space around the structure.

**C.** Once you see a blank screen, toggle the Selecting State from “Objects” to "Molecules" as described above. Click anywhere on the letters of the amino acid sequence to select the ACE2 molecule. Selecting "Molecules" will only select the protein, not the waters or the oxygen, because these are considered to be separate molecules. With the ACE2 sequence highlighted find the "S" box on the selection "(sele)" bar on the right panel. Left click on the "S" for the "Show" popup menu. Select Cartoon. You should now see the alpha-helical structure of ACE2. Unselect the molecule by clicking on the sequence or in the blank space around the structure.

**D.** Do the same for the S1 subunit of the spike protein.

*If you are adept at 3D free-viewing you can also try to do so Display > Stereo Mode > Cross-eyed Stereo*

**E.** Now let's learn how to use the three buttons of the mouse to explore the structure.

* Left-click and hold the center of the molecule and move the mouse. This **tumbles** the structure, allowing you to see it from any angle.
* Left-click and hold near the edge of the screen and move the mouse up and down. This **rotates** the structure, allowing you to turn it upside down.
* Click the scroll wheel and hold in the center of the screen and move the mouse in any direction to pan the structure. (Do not scroll the scroll wheel; just click and hold.) This allows you to **center the scene** on different parts of the molecule.
* Click and scroll the scroll wheel changes the depth of view. This function can be used to **clip parts** of the molecule away so that structures behind one another can be more easily seen.
* Click and release the scroll wheel on the molecule to **center the tumbling pivot point**. This function is useful because you will sometimes find that it is difficult to tumble the molecule easily; re-centering the pivot point fixes the problem.
* Right-click and hold the molecule and move it up and down to **zoom in and out**.

Troubleshooting: If you clip away too much of the structure using the scroll wheel or you lose the structure some other way, go to the top of the screen and click Display>Zoom>Complete to re-center the molecule. This will cause the structure to reappear.

Understanding how to use the mouse is essential to using PyMol effectively. Practice using all the mouse commands to explore the structure. For the remainder of this exercise, continue to use all the mouse buttons to monitor your progress and observe the structure. Be sure to zoom close enough to see what you are working on.

**F.** Again, select the entire molecule of ACE2. Left click the "C" on the selection (sele) bar on the right panel to show the "Color" popup menu. Change the color to "spectrum" – “rainbow”. Unselect the structure by clicking on the blank space around the structure. The N-terminus of the structure is blue and the C-terminus is red. Again, select the entire molecule and go to the "Color" menu. Select "by element". You will now see several options on the popup menu. Select the 7th option which has a gray "C". This scheme will color carbons gray and other atoms in different color. The molecule cartoon should now be gray. You can try other color schemes as well. It is usually helpful to color molecules "by element" so that when it is time to show atoms, it is easy to identify which element they are. Unselect the molecule to remove the pink squares.

Take this opportunity to save the session. File>Save Session. Save the file on the desktop so that you can find it later.

**G.** Color the S1 subunit of the S-protein in green.

**H.** Change the Selecting state to "Residues" by toggling as described above. Now you will be able to select individual amino acids or other small components of the structure. Use the scroll bar at the top of the screen to scroll through the sequence until you see amino acid 437 in S1 domain. The letter under aa 437 is S, followed by NNLD etc. Select the amino acids until amino acid 505 (Q). Include this amino acid, but do not continue with the rest (PYR etc). Notice that left clicking and dragging selects multiple amino acids at a time. Color this region in red. This is the receptor-binding motif (RBM) of S-protein.

**I.** Analyze Figure 2 (Lan et al. 2020). This is a comparison of the receptor-binding domain of S1 of SARS-CoV virus that caused SARS epidemic in China 2002, and a novel SARS-CoV-2 that you are working on.

Find the specific amino acids that interact with ACE2 in your SARS-CoV-2 3D model, select them as residues, and show them as sticks.



Figure 5. Sequence alignment of the receptor-binding domain of S1 from SARS-CoV-2 and SARS-CoV. Red letters: RBM. Dots: amino acid residues contacting with ACE2. Black dots show the residues in SARS-CoV-2, red dots show the residues in SARS-CoV (Adapted from Lan et al, 2020).

**Troubleshooting:**

If you do not click in the center of the letters on the sequence, selecting will not work. If you left click more than once, or drag the mouse over a selection, you will not select any of the amino acid residues.

**J.** In S1 subunit, select amino acids G503Y504Q505 and Y507. Show them as sticks and color them yellow. These amino acids are the same between two SARS CoV viruses, and are among other amino acids that are recognized by an antibody called *m396* which is able to neutralize SARS-CoV. However, this antibody was not shown to neutralize SARS-CoV-2 and the reasons are currently unknown (Lan et al, 2020). Other antibodies that have been shown to work for 2002 SARS are being tested, but it is anticipated that the working ones will be specific to S2 subunit rather than S1.

**K.** Change the selecting state to "Molecules". Scroll back to the beginning of the amino acid sequence and select the entire ACE2 molecule. Click "S" and show the surface. Unselect the molecule. You should now be able to see the region where S1 is bound. The blue and red blotches are surface exposed nitrogen and oxygen atoms, yellow is sulfur. Use the mouse to explore the structure.

**L.** Change the transparency of the surface by selecting on the top menu: Setting>Transparency>Surface>60%. Now we can see both the inside and outside the surface of the molecule.

**M.** Now, we are going to customize some of the settings in PyMol to make the structure better.

Go Setting - Edit All window and find "surface\_color". Double click on "default" and type the word, white, and then press enter. The structure should now look much cleaner because the distracting colors on the surface have been removed.

**N.** We are now finished modifying the structure of ACE2-S1 complex. The structure is much easier to interpret from when we began. We can now clearly see the alpha helices and RBM, and amino acids important for binding of the receptor and S-protein. To complete the assignment, we need to frame our view, create a high resolution image, save it, and paste the image into a word document.

**O.** First use the mouse to create a meaningful view of the molecule and center it in the frame; position the molecule so that the ACE2, S1 and RBM are clearly visible, and so that the molecule is large enough to be seen. When you present your PyMol images, it is very important that they express something meaningful, so careful framing of the picture is essential. Also, you might need to change the Background to white. When you have settled on a good view of the molecule, click the "Ray" button on the top window menu right. This button "Ray Traces" the molecule and creates a high resolution image of your view. Wait; this calculation may take your computer several seconds to complete. Once the ray trace is complete, the image can no longer be manipulated without undoing the ray trace; clicking anywhere on the image will undo the ray trace. When ray tracing is complete, immediately save the image: File>Save Image As>PNG. Name the image and save to the desktop. Note: if you have downloaded the modern PyMol build, ray tracing is disabled. In this case, hand in the low resolution image.

**P.** Open a Word document and insert the png image into the document. Note: you cannot insert the PyMol session file (pse); only the png image file will work.

For a PC: Insert>Photo>Picture from file, and then find the png image file.

For Mac: Insert>Choose, and then find the png image file.

**Q.** Position the image file and type a figure caption. This consists of a sentence or two below the image explaining what the image is. Be sure to reference the pdb entry code 6LZG, the name of the protein, and anything you have highlighted in the structure. Be as descriptive as you can in the figure caption so that the figure can be completely understood by someone unfamiliar with the assignment.

\*See Figure 6 for a sample image created using the above directions.



Figure 6. Crystal structure of sperm whale myoglobin. Amino acid residues H64 and H93 are shown as sticks and the bound oxygen molecule (ball and stick model) is shown coordinated to the heme.

Finally, convert your word processing document to PDF and submit for grading.

**Acknowledgements**

The authors thank Daniel Fried and Christopher Zambell for their original tutorials, and Julia Annuzzi and Christian Meekins for critical reading of the manuscript.

**References**

Andersen K, Rambaut A, Lipkin WI, Holmes EC, Garry RF (2020) The proximal origin of SARS-CoV-2. Nature Medicine. doi:10.1038/s41591-020-0820-9

Domingo E, Perales C (2019) Viral quasispecies. PLOS Genetics 15. doi:10.1371/journal.pgen.1008271

Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular Biology and Evolution 35: 1547-1549. doi:10.1093/molbev/msy096

Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, Zhang Q, Shi X, Wang Q, Zhang L, Wang X (2020) Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature. doi:10.1038/s41586-020-2180-5

Martinez J, Longdon B, Bauer S, Chan Y-S, Miller WJ, Bourtzis K, Teixeira L, Jiggins FM (2014) Symbionts Commonly Provide Broad Spectrum Resistance to Viruses in Insects: A Comparative Analysis of Wolbachia Strains. PLOS Pathogens 10. doi:10.1371/journal.ppat.1004369

Stadler K, Masignani V, Eickmann M, Becker S, Abrignani S, Klenk HD, Rappuoli R (2003) SARS - Beginning to understand a new virus. Nature Reviews Microbiology 1: 209-218. doi:10.1038/nrmicro775

The PyMOL Molecular Graphics System. (2010). 1.3r1 edu ed. Schrödinger, LLC.

Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D (2020) Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell. doi:10.1016/j.cell.2020.02.058

Wang C, Liu Z, Chen Z, Huang X, Xu M, He T, Zhang Z (2020) The establishment of reference sequence for SARS-CoV-2 and variation analysis. Journal of Medical Virology. doi:10.1002/jmv.25762

Zhang T, Wu Q, Zhang Z (2020) Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. Current biology 30: 1-6. doi:10.1016/j.cub.2020.03.022