LAB 9: UNDERSTANDING THE MOLECULAR EVOLUTION OF SARS-COV-2 BY THE USE OF PHYLOGENETIC TREES

DEVELOPED AND WRITTEN BY
Nikolaos Tsotakos
Penn State Harrisburg
nxt12@psu.edu

Purpose

In this exercise, the molecular origins of SARS-CoV-2 will be investigated through the creation of a phylogenetic tree that contains SARS-CoV-2 isolates, as well as other coronaviruses.

Overview

Phylogenetic trees are tools that biologists commonly use to infer evolutionary relationships among different taxa or organisms, based on partial or complete similarity of DNA or protein sequences. In this exercise, students will use a database to retrieve the complete genomic sequences of several coronaviruses, including different isolates of SARS-CoV-2. They will then use the software MEGA (Molecular Evolutionary Genetics Analysis) to create a phylogenetic tree. MEGA is a free package that allows users to build evolutionary trees in an easy user interface. It is very versatile and powerful, allowing several options for tree construction, including the commonly used Neighbor-Joining and Maximum Likelihood methods, which give good estimates of the relationship between different molecular sequences. For this exercise, students will use data from GenBank, NCBI's genetic sequence database, and other sources, insert them into a text editor, import those data into MEGA, and create phylogenetic trees from them.

Student Learning Objectives

Following the completion of the module, the students should be able to

- retrieve protein and nucleic acid sequences from public repositories;
- align sequences using open license software;
- explain similarities and differences between aligned sequences;
- construct phylogenetic trees and explain principles of molecular evolution.

This exercise aligns with the recommended student learning outcome from ASBMB that students "should be able to use databases and bioinformatics tools" (https://www.asbmb.org/education/core-concept-teaching-strategies/foundational-concepts/skills) and the competency described by the NIBLSE community to "use bioinformatics tools to examine complex biological problems in evolution, information flow, and other important areas of biology" (https://qubeshub.org/community/groups/niblse/core_competencies).

Safety Precautions

While all activities have associated risks, there are no specific safety concerns identified for this particular lab exercise.

Introduction

MEGA Software

Molecular Evolutionary Genetics Analysis (MEGA) software (www.megasoftware.net) is a free package that lets anyone build evolutionary trees in a user-friendly setup. There are several different options to choose from when building trees from molecular data in MEGA, but the most commonly used are Neighbor-Joining (NJ) and Maximum-Likelihood (ML) (Newman, Duffus, & Lee, 2016), both of which give good estimates on the relationship between different molecular sequences (Tateno, Takezaki, & Nei, 1994). The NJ and ML methods for building evolutionary trees rely on different statistical principles to determine connections between individuals within the trees. In NJ, the least-squares method is used, along with pairwise evolutionary distances. In ML, the maximum likelihood is optimized so that the inferred tree is the most likely tree. Generally, they will produce very similar results, but NJ is much faster, as ML considers all probable trees, and is therefore more computationally expensive. Despite slight differences in branching patterns between NJ and ML trees, both are robust methods for building evolutionary trees. The reliability of the robustness (i.e., accuracy) for each method will be tested by bootstrapping, the most common reliability test. In bootstrapping, resampling of the sites in the alignment is used to build new trees.

No matter what the statistical method used, a specific evolutionary model needs to be used. Consider the following change in a DNA molecule: A -> C. How often does an adenine change to cytosine relative to a change of A -> G or A -> T? There are several such models, such as the Jukes–Cantor model, which is simply a mathematical model that describes the change of one of the nucleotides in the DNA sequence to another one, over time. This model assumes that there is an equal probability of any base changing into any other base, and is thus the simplest substitution model. Other evolutionary models, such as the Kimura and Tajima–Nei ones, make different assumptions for the rates or the frequency of bases, which often deviate from ¼, but for our purposes, the Jukes–Cantor model will be sufficient. All of its parameters are automatically estimated by MEGA.

Both NJ and ML produce trees that are unrooted, even though they are frequently drawn from left to right. Unrooted trees are acyclic graphs, meaning that there is no path that can be followed from any taxon that will return to the same taxon. Thus, unrooted trees do not imply a known ancestral root. While they specify a number of routes to follow from one taxon to another, the number of routes taken is limited (restricted because of the tree topology). Rooted trees, on the other hand, reflect the most basal ancestor of the tree. For example, with three terminal taxa, there are six possible trees (Fig. 1b). If we take a particular unrooted tree (Fig. 1a), only two dichotomous solutions are compatible (the top left and bottom right panels of Fig. 1b). This is because the unrooted tree specifies the path; to get from A to C, you encounter a node that branches to B, and you cannot move from A to C by bypassing this node. Thus, A and B will be adjacent in any rooted solution to the unrooted tree (Fig.1b). However, being adjacent (next to each other) in the unrooted tree does not mean that A and B will form a monophyletic group (sharing a most common

recent ancestor) in the rooted tree: note that of the four rooted trees on which A and B are adjacent (next to each other, without C between them), only two (the ones at the bottom of Fig. 1b) contain taxa A and B as part of a monophyletic group. There are many techniques (sometimes competing) for rooting a tree; one of the most common methods is through the use of an *outgroup*. An outgroup is a lineage that falls outside of the clade under study, but is closely related to it. In this case, if one knows the outgroup, then it can be used to properly root the tree. Choosing a proper outgroup can be a difficult task and may require some trial and error. A good outgroup should be similar to the sequences in question, but different enough so that the computer program can see the differences.

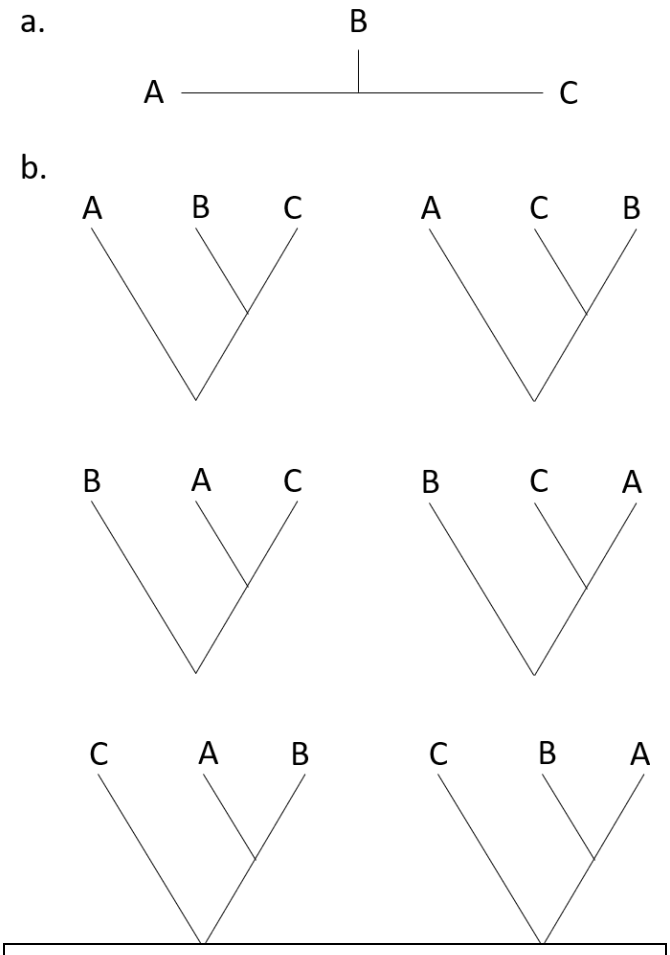

Figure 1: (a) Unrooted tree, (b) rooted trees for three taxa.

SARS-CoV-2 and the COVID-19 Pandemic

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pneumonia was first noted in Wuhan (China) in December 2019 and the disease induced by the virus has been termed coronavirus infectious disease 2019 (COVID-19). The disease spread throughout the world, forcing the World Health Organization (WHO) to declare a pandemic in March 2020. The exact origin of SARS-CoV-2 has been unclear and a cause for disputes (Lee & Morling, 2021). The current consensus is that it originated in bats (Zhou et al., 2020), but some scientists dispute that (Deigin & Segreto, 2021).

Effective vaccines protecting against COVID-19 were developed very quickly; the U.S. Food and Drug Administration (FDA) issued the first emergency use authorization (EUA) for a vaccine in December 2020. Despite the speed of vaccine development, different variants of the virus emerged. Some of these variants harbor mutations in the S gene, the one responsible for the production of the Spike protein, which the virus uses to enter human cells. These variants caused public health experts to have concerns, as divergent strains with an accumulation of mutations in the different S domains are potentially capable of evading infection- or vaccination-induced neutralizing antibodies (Geers et al., 2021). Additionally, the variants have different transmission dynamics, and knowledge of their prevalence may guide public health policies (Boehm et al., 2021; Dellicour et al., 2021).

In this exercise, students will select a few genomic sequences (out of thousands of possible ones) and use them to create a phylogenetic tree of coronaviruses to try to explain the molecular origins

of SARS-CoV-2 and the differences among the variants of concern. It is recommended that each student focus on acquiring sequences with differences on one parameter (e.g., geographical location or time of sample collection) to try to answer a specific research question. An example of such a question could be "How many viral lineages were there in circulation in the United States in September 2020?"

References

Cited Sources

Boehm, E., Kronig, I., Neher, R.A., Eckerle, I., Vetter, P., & Kaiser, L. (2021). Novel SARS-CoV-2 variants: The pandemics within the pandemic. *Clinical Microbiology and Infection*. **https://doi.org/10.1016/j.cmi.2021.05.022**

Deigin, Y., & Segreto, R. (2021). SARS-CoV-2's claimed natural origin is undermined by issues with genome sequences of its relative strains: Coronavirus sequences RaTG13, MP789 and RmYN02 raise multiple questions to be critically addressed by the scientific community. *Bioessays*, e2100015. **https://doi.org/10.1002/bies.202100015**

Dellicour, S., Hong, S.L., Vrancken, B., Chaillon, A., Gill, M.S., Maurano, M.T., . . . Heguy, A. (2021). Dispersal dynamics of SARS-CoV-2 lineages during the first epidemic wave in New York City. *PLoS Pathogens*, 17(5), e1009571. **https://doi.org/10.1371/journal.ppat.1009571**

Geers, D., Shamier, M.C., Bogers, S., den Hartog, G., Gommers, L., Nieuwkoop, N.N., . . . GeurtsvanKessel, C. H. (2021). SARS-CoV-2 variants of concern partially escape humoral but not T-cell responses in COVID-19 convalescent donors and vaccinees. *Science Immunology*, 6(59), eabj1750. **https://doi.org/10.1126/sciimmunol.abj1750**

Lee, A. C. K., & Morling, J. (2021). COVID-19: Viral origins, vaccine fears and risk perceptions. *Public Health*, 196, A1–A2. **https://doi.org/10.1016/j.puhe.2021.04.013**

Newman, L., Duffus, A. L. J., & Lee, C. (2016). Using the free program MEGA to build phylogenetic trees from molecular data. *American Biology Teacher,* 78(7), 608–612. **https://doi.org/10.1525/abt.2016.78.7.608**

Tateno, Y., Takezaki, N., & Nei, M. (1994). Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Molecular Biology and Evolution,* 11(2), 261–277. **https://doi.org/10.1093/oxfordjournals.molbev.a040108**

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature,* 579(7798), 270–273. **https://doi.org/10.1038/s41586-020-2012-7**

Additional Sources for Information

LaTourrette, K., Holste, N.M., Rodriguez-Peña, R., Leme, R.A., & Garcia-Ruiz, H. (2021). Genome-wide variation in betacoronaviruses. *Journal of Virology*, JVI0049621. Advance online publication. **https://doi.org/10.1128/JVI.00496-21**

Das, J.K., & Roy, S. (2021). Comparative analysis of human coronaviruses focusing on nucleotide variability and synonymous codon usage patterns. *Genomics*, 113(4), 2177–2188. Advance online publication. **https://doi.org/10.1016/j.ygeno.2021.05.008**

Methods

Retrieval and Downloading of Viral Sequences

In this section, you will visit the Virus Pathogen Database and Analysis Resource (ViPR) and find the accession numbers for different viral genomes. You will then visit the Nucleotide Database of NCBI and download the sequences to your local workstation, to be analyzed later with MEGA.

1. Visit the BV-BRC at www.bv-brc.org.
2. Click on the VIRUSES link and click on the *Coronaviridae* icon under the Virus Families section of the webpage.
3. Under the Taxonomy tab, click on the arrow next to the *Orthocoronaviridae* subfamily to expand it, and check the box next to the *Betacoronavirus* genus.
4. Click on the Genomes tab. A list of the available genomes from the specific genus of viruses will come up.
5. Filter the results by clicking on the "Filters" button close to the top right of the table.
6. Make sure you select for complete genome only under the "Genome Status" filter.
7. Note that the total number of genomes that match your search criteria is in the millions. You will need to decrease the overall number of sequences to something more manageable. You can filter the results by limiting the time of collection, the geographic location or country, and the host species. It is recommended that you select at least two bat viruses, and three human isolates of SARS-CoV-2. While selecting your sequences, consider what your biological question is.
8. You can click the “Apply” button to generate a list of sequences that match your search parameters. Take notes of the NCBI Accession Numbers for the virus of your choice. *Alternatively, you can click the check box to the left of the sequences you want to save and then click download at the bottom of the page to save the list as an Excel file. You can then peruse the Excel file to select the sequences of choice and take notes of the respective Accession Numbers.*
9. Visit the Nucleotide Database of NCBI at https://www.ncbi.nlm.nih.gov/nucleotide/.
10. In the search box, write the Accession Numbers that you saved in step 7, separated by the Boolean operator OR[1], so that you can retrieve all the results in the same page. Include the following accession numbers as well and hit the *Search* button.
    - *GU553364 (SARS, China, 2003)*
    - *KM015348 (MERS, UK, 2013)*
    - *KY967354 (Human respirovirus 1, USA, 2015), to be used as an outgroup.*
11. After you receive the results, select the relevant strains by checking the box to the left of the description line of each result.
12. At the top right, click on the *arrow next to the Send to:* option. Select *File* under the *Choose Destination* option. Make sure you select *FASTA* as the *format*, and click on the *Create File* button.
13. Download the multiple .fasta file to a location of your choice in your computer.

[1] A Boolean operator is a simple word or phrase, such as AND, OR, NOT, AND NOT, that is used in searches to combine or exclude keywords.

Alignment of Multiple Sequences

In this section, you will perform a multiple alignment of the downloaded .fasta sequences, and then you will generate a phylogenetic tree based on this alignment.

1. Open the MEGA-X software.
2. Click on the circular *Align* menu, and then select *Edit/Build Alignment*.
3. Create a new alignment. A secondary menu will appear that requires you to select an option. Choose *Create a new alignment* option and click *OK*.
4. A second submenu will appear asking you to select the type of sequence data that will be used to build the alignment. Select the *DNA* option. This will open the MEGA Alignment Explorer in a new window.
5. At the top of the MEGA Alignment Explorer Window, select the *Edit* menu by clicking on it. From this menu, select the *Insert Sequence From File* option. This will open a new window.
6. Select the .fasta data file(s) you saved earlier.
7. Once you have selected all of the files that you wish to upload, select the *Open* button by clicking on it.
8. Once the sequences are loaded into MEGA, we want to align them. However, MEGA normally introduces an empty sequence into the alignment. This is going to interfere with your downstream analysis, and may lead to errors, so it needs to be removed. This is done by clicking on the sequence name (usually "*Sequence 1*") to select it, followed by *right-click* and *delete*.
9. To select all the sequences of choice, click on the *Edit* button and then *Select All*. Their background color will change.
10. Go to the *Alignment* menu at the top of the Alignment Explorer window. Click to open the dropdown menu and select *Align by ClustalW* by clicking on it.
11. This will open another window that is filled with ClustalW parameters. For our purposes, the default settings are adequate. Select the *OK* option at the bottom of this menu to proceed. This will set the alignment algorithm in motion. Depending on the size and number of the sequences being examined, it is common for this step to take 30 minutes or longer, so do not close the program.
12. Now, we need to save the Alignment Session so that the data are saved in a format that MEGA can use to build the phylogenetic trees. Save the Alignment Session by selecting the *Data* menu at the top right of the Alignment Explorer window. Click on the *Save Session* option. After this has been completed, save the session to a known location in your computer as a .mas file.
13. Take a minute to scroll through the alignment. Are there any differences among the sequences you selected? Do you think that you made a proper selection of an outgroup?
14. Close the Alignment Explorer window.

Construction of the Phylogenetic Tree

15. Open the saved alignment session by selecting the *File* menu in the general MEGA window and clicking on the *Open a File/Session* option.

16. This will bring up a second window where you have the option to open the .mas file to analyze it or align it. Select *Analyze* by clicking on it.
17. To construct a phylogenetic tree, select the *Phylogeny* menu midway through the menu bar at the top of the MEGA window. Here we will continue our example with a Neighbor-Joining tree, but the process is the same for other types of phylogenetic trees. Select *Construct/Test Neighbor-Joining Tree*.
18. A menu will appear that asks if you want to use the currently active data sheet. Select *Yes*.

Estimation of the Reliability of the Tree

19. This will open a dialogue box called Analysis Preferences. For the *Test of Phylogeny* select *Bootstrap Method*. In the field entitled *No. of Bootstrap Replications*, select *1000* to obtain stable estimates of reliability of the tree. For the *Substitution Type* start by selecting *Nucleotide*, followed by selecting the *Jukes–Cantor Model*. Here all other fields are left at their default values. To generate the tree, click on *OK*.
20. After a few minutes, a tree will be generated. The length of time that this takes will depend in part on the length and number of sequences that are being used to create the tree.
21. The numbers on the branches of the tree represents the Bootstrap value, which is the statistical support that each branch receives by the Bootstrap analysis. Larger numbers mean that the branch has higher support and is more likely to be a real branch.

Rooting and Presentation of the Tree

22. To simplify the tree, we now want to condense or cut out the branches that have less support and are less likely to be true branches. To do this, go to the *Compute* menu on the TreeExplorer menu and select *Condensed Tree*.
23. This opens a new menu, Tree Options. Select the *Cutoff* submenu and input *50* for the *Cut-off Value for Condensed Tree*, then click *OK*. Leave all other values at their defaults.
24. Now, the tree in the TreeExplorer will reflect the changes. All branches that have less than 50% support will have been removed.
25. The last thing that we need to do is to set our outgroup. In our example, it is the human respirovirus 1. To do this, right click on the branch that has the human respirovirus. This brings up a submenu. In this submenu, select the *Root* option. This provides us with a rooted phylogenetic tree.
26. The tree can be saved as a PDF or printed out. To save the tree as a PDF, go to the Image menu and select Export as PDF. A window will pop up and you can save the file there. To print the tree, there is a Printer Icon that you can use just below the upper menu.
27. (optional) Repeat the process (steps 17–26) by creating a phylogeny with a Maximum-Likelihood Tree.

Student Assessment

1. After constructing a phylogenetic tree with both the Neighbor-Joining and the Maximum Likelihood methods, do you observe any differences between the trees? What are they?
2. Do you observe differences while condensing the tree if you change the cutoff value (step 22 in module 2)?
3. What criteria did you use to select your original sequences? Do the sequences look similar to or different from each other? Spot at least three sites of the viral genomes you aligned that do not contain a consensus nucleotide (i.e., at least one genome contains a different nucleotide than the others). Using the resources found in NCBI, do you think that these nucleotide differences may play a functional role?
4. What was the biological research question that you set out to study? Do your phylogenetic trees help you to begin to answer that question? What additional research would you need to do to better address your research question?
5. As discussed in the *Introduction* in Module 1, changes in the S gene can be of concern, as the spike (S) protein binds to ACE2 (angiotensin-converting enzyme 2) on the surfaces of human cells, which is the first step toward viral entry. Based on your knowledge of the resources found in ViPR and NCBI, perform a sequence alignment of the different versions of the S proteins of variants of concern. Prepare the phylogenetic tree using the NJ method. What conclusions can you make with regard to the evolution of the protein based on the geographic location and time point at which the samples were collected?

Instructor's Notes

Laboratory Preparation

This is an in-class activity that guides students through the retrieval of nucleotide sequences from NCBI and the step-by-step construction of an evolutionary tree using MEGA software features, including the sequence import, alignment, and phylogeny settings. The lesson centers on using free software (MEGA, www.megasoftware.net) to construct evolutionary trees of several members of the Coronaviridae. The activity can take place in a computer lab, or students can download MEGA to their own devices, which they can bring to the lab.

It is highly recommended that students have a basic understanding of DNA structure and the central dogma of molecular biology, as well as knowledge of mutations and their role in evolution. There is no need for any prior knowledge of virology. It is advised that phylogenetic trees and sequence file formats, such as .fasta, be introduced prior to the beginning of the activity.

Because of the volume of sequences submitted to GenBank and the necessity of timely submissions, some sequences may be removed, while others may be available. In NCBI, it is difficult to tell why a sequence has been removed, but more often than not, the sequences or annotations are corrected, amended, or updated, and they receive a different accession number. There is a summary of SARS-CoV-2 lineages of concern or of interest in the ViPR (https://bv-brc.org/view/VariantLineage/#view_tab=overview). Some that can be used in the activity are

- SARS-CoV-2/human/USA/FL-CDC-STM-P012/2020 (B.1.1.7, WHO "variant alpha"),
- SARS-CoV-2/human/GHA/WACCBIP_nCoV_GS73/2021 (B.1.351, WHO "variant beta"),
- SARS-CoV-2/human/USA/TX-CDC-9N4G-9005/2021 (B.1.427),
- SARS-CoV-2/human/USA/CA-LACPHL-AF00141/2021 (B.1.429, WHO "variant epsilon"),
- SARS-CoV-2/human/ITA/ABR-IZSGC-TE30968/2021|gb|QRX39425 (P.1 or B.1.1.28.1, WHO "variant gamma").
- SARS-CoV-2/human/USA/NJ-CDC-LC0038223/2021 (B.1.617.2, WHO "variant delta").

The resource contains the amino acid substitutions in the spike protein, among other valuable information, so the instructor can decide to discuss it as deemed appropriate. Caution is suggested in the association of variants to certain parts of the world, as this is more of a media soundbite than true scientific information. The WHO label uses letters of the Greek alphabet to avoid stigmatizing these names for variants of concern or variants of interest (https://www.who.int/en/activities/tracking-SARS-CoV-2**Error! Bookmark not defined.**-variants/). The detection of a specific variant in a certain geographic location does not indicate that the variant is limited to that location, or even that it first originated there.

The ML method will require more time for tree construction than the NJ method, and may lead to software crashes, depending on the hardware capabilities. It is not recommended that students try to align large genomes, as the capabilities of the students' or the classrooms' computers may be limited.

Connections to Other Lab Exercises

Depending on the main focus of the course, the activity can be paired well with the NCBI activity. It can also be modified to accommodate the phylogenetic tree creation and molecular evolution of any given gene or protein.

Recommended Schedule for the Lab Exercise

The activity can be demonstrated within a 1-h time period (lecture or lab) if the steps have been previously recreated. Students can follow along, using the instructions seen in the methods section. However, the alignment and tree creation steps (steps 11 and 17 of the methods section, respectively) can take a long time (in the case of alignment, several hours). If the instructor has previously recreated the activity, they can provide the previously prepared alignment to the students (like a cooking show), so that the class can move on with the tree construction. Alternatively, the activity can be interrupted; the instructor may assign step 11 as homework and the class can continue work in the following meeting. A 3-h lab would be sufficient for the full activity. However, it is still recommended that the instructor generate the intermediate files previously, particularly so that unforeseeable events (such as computer crashes) can have a minimal effect on the class. Such a “cookbook” approach is not recommended, as it takes away the originality of the research question the students can pose and try to answer.

LAB 10: STRUCTURAL ANALYSIS OF THE ATTACHMENT OF SARS-COV-2 TO HUMAN CELLS

DEVELOPED AND WRITTEN BY
Nikolaos Tsotakos
Penn State Harrisburg
nxt12@psu.edu

Purpose

In this exercise, students will explore the structural elements that contribute to the attachment of SARS-CoV-2.

Overview

Coronaviruses bind to their host cells via the Spike (S) protein on their surfaces. In humans, SARS-CoV and SARS-CoV-2 use the cell surface protein Angiotensin Converting Enzyme 2 (ACE2) to attach to cells that express this viral receptor on their surfaces. Since the declaration of the COVID-19 pandemic by the World Health Organization (WHO) on March 11, 2020, different variants of concern have been identified, several of which contain mutations in the gene encoding the S protein, thus challenging the efficacy of the vaccines that have been approved or authorized for emergency use by health authorities around the world. In this exercise, students will visually explore the interaction between S and ACE2 and get acquainted with the molecular visualization tool iCn3D and the Protein Data Bank (PDB).

Student Learning Objectives

Following the completion of the module, students should be able to

- Describe the chemical interactions that drive the attachment of SARS-CoV-2 to human cells.
- Predict the effects of mutations on the structure of a given protein.
- Use visualization software to explore molecular structures and interactions.

This exercise aligns with the recommended student learning outcome from ASBMB that students should be able to "compare and contrast the effects of chemical modification of specific amino acids on a three dimensional structure of a protein" and "discuss the interactions between a variety of biological molecules (including proteins, nucleic acids, lipids, carbohydrates and small organics, etc.) and describe how these interactions impact specificity or affinity leading to changes in biological function" (https://www.asbmb.org/education/core-concept-teaching-strategies/foundational-concepts/structure-function).

Safety Precautions

While all activities have associated risks, there are no specific safety concerns for this particular lab exercise.

Introduction

The COVID-19 pandemic

In December 2019, China reported a cluster of pneumonia cases in Wuhan, a city in Hubei province. It was determined that the infectious agent causing the disease was a novel coronavirus, initially named nCoV-2019, but later formally named SARS-CoV-2 given the genetic relationship and similar symptomatology to another virus, SARS, which was behind an outbreak in Hong Kong in 2002. Following reports of infections and associated mortality in countries around the world, WHO declared the spread of COVID-19 (coronavirus-induced disease 2019) a pandemic on March 11, 2020. As of early November 2021, the pandemic is ongoing, with over 5 million people dead and over 245 million infections worldwide.

Coronaviridae, or coronaviruses, are a family of viruses that use single-stranded (+) RNA as their genetic material. Their characteristic morphology under the electron microscope is due to the presence of the Spike (S) protein on their surface. The S protein is a glycoprotein that forms homotrimers protruding from the viral surface (in what is called the "prefusion conformation," Xia 2021). This trimer is responsible for the first step in the life cycle of the virus, as it binds to a protein on the surfaces of the host cells, thus facilitating the attachment of the virus to the cell. In the case of SARS-CoV-2, the receptor protein on human cells is Angiotensin Converting Enzyme 2 (ACE2). The three molecules composing the trimer form two functional subunits; the S_1 subunit is responsible for binding to the receptor, while the S_2 is responsible for the fusion of the viral and host cell membranes (Xia 2021). To engage a host cell receptor, the receptor-binding domain (RBD) of S_1 undergoes hingelike conformational movements that transiently hide or expose the determinants of receptor binding. These two states are referred to as the "closed" conformation and the "open" conformation, where "closed" corresponds to the receptor-inaccessible state and "open" corresponds to the receptor-accessible state, which is thought to be less stable (Wrapp et al. 2020). Following this initial binding event, the S protein gets cleaved between the S_1 and S_2 subunits by another host protein, a protease named Transmembrane Serine Protease 2 (TMPRSS2; Hoffman et al. 2020). This priming cleavage allows the S_2 subunit to transition to a different conformation, called the "postfusion conformation." Another cleavage allows the virus to enter the host cell, thus infecting it.

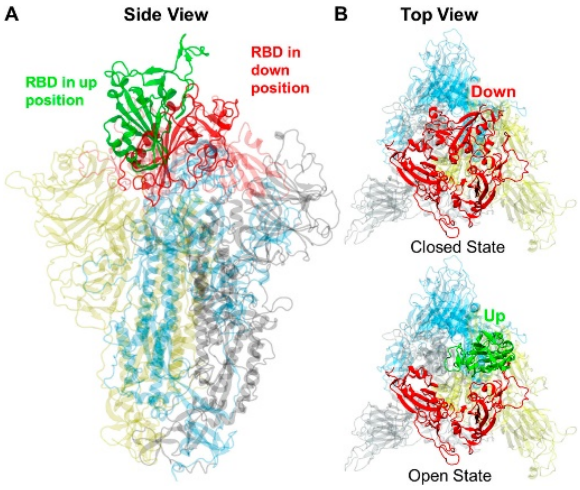

Figure 1: Down and up positions of RBD of S protein closed and open states. (A) The trimeric S protein conformation is shown from a side view. The closed and open state conformations sampled from molecular dynamics (MD) simulations structures are aligned based on their secondary structures in the S2 subunit. RBDs in the down and up positions are shown in red and green, respectively. Since the remaining structure is almost identical between down and up positions, for the up position only the RBD is shown. Protomers A, B, and C are shown in gray, yellow, and light blue, respectively. (B) The closed and open states of the S protein trimer are shown from a top view. Taken from Gur et al. (2020).

Effective vaccines protecting against COVID-19 were developed very quickly; the U.S. Food and Drug Administration (FDA) issued the first emergency use authorization (EUA) for a vaccine in December 2020. Most vaccines are designed to induce the production of antibodies that target the S protein from the immune system, thereby neutralizing the virus by blocking its ability to bind to ACE2 and enter human cells. Despite the speed of vaccine development, different variants of the virus emerged. Some of these variants harbor mutations in the S gene, generating slightly different S proteins with various efficiencies in ACE2 binding. Since the S:ACE2 binding is critical to the viral infectivity, a lot of effort has gone into understanding the molecular details of the interaction between the two proteins.

In this exercise, students will use the Protein Data Bank (PDB) and the NCBI visualization tool iCn3D (pronounced "I see in 3D") to visualize the S protein and investigate the molecular bonds that contribute to the interaction between S and ACE2.

References

1. Xia, X. (2021). Domains and functions of spike protein in Sars-Cov-2 in the context of vaccine design. *Viruses*, 13(1), article 109.
2. Wrapp, D., Wang, N., Corbett, K.S., Goldsmith, J.A., Hsieh, C.-L., Abiona, O., Graham, B.S., McLellan, J.S. (2020). Cryo-EM structure of the 2019-nCoV Spike in the prefusion conformation. *Science*, 367(6483), 1260–1263.
3. Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T.S., Herrler, G., Wu, N.-H., Nitsche, A., Müller, M.A., Drosten, C., Pöhlmann, S.(2020). SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell*, 181(2), 271–280.
4. Li, B., Deng, A., Li, K., Hu, Y., Li, Z., Shi, Z., Xiong, Q., Liu, Z., Guo, Q., Zou, L., Zhang, H., Zhang, M., Ouyang, F., Su, J., Su, W., Xu, J., Lin, H., Sun, J., Peng, J., Jiang, h., Zhou, P., Hu, T., Luo, M. ,Zhang, Y., Zheng, H., Xiao, J., Liu, T., Tan, M., Che, R., Zeng, H, Zheng, Z., Huang, Y., Yu, J., Yi, L., Wu, J., Chen, J., Zhong, H., Deng, X., Kang, M., Pybus, O.G., Hall, M., Lythgoe, K.A., Li, Y., Yuan, J., He, J., Lu, J.(2021). Viral infection and transmission in a large, well-traced outbreak caused by the SARS-CoV-2 Delta variant. Preprint 2021.07.07.21260122.
5. Gur, M., Taka, E., Yilmaz, S.Z., Kilinc, C., Aktas, U., Golcuk, M. (2020). Conformational transition of SRAS-CoV-2 Spike glycoprotein between its closed and open states. *Journal of Chemical Physics*, 153, 075101.

Methods

The PDB and iDn3D

The Protein Data Bank (PDB) is a repository of solved structures and can be accessed at https://www.rcsb.org/. It provides access to 3D structural data on biological macromolecules. Additional information contained in the PDB includes information about the experiment used to derive the data, details about the molecules included in the experiment, and links to bioinformatics resources that can provide additional information about the molecule of interest. Each structure in the PDB has a unique identifier called its PDB ID. Atomic coordinates from the PDB can be explored using various visualization tools.

iCn3D is a WebGL-based viewer for interactive viewing of three-dimensional macromolecular structures. It is open source and there is no need to install any application locally. It can be accessed at https://www.ncbi.nlm.nih.gov/Structure/icn3d/full.html. Users can interactively rotate the molecule/complex, compare structures, and analyze interactions.

Structure of the SARS-CoV-2 Spike (S) Protein

In this section, you will explore the structure of the S protein. The structure was solved by different groups independently by cryo-electron microscopy (cryo-EM). Cryo-EM is a version of transmission electron microscopy where the sample is examined at extremely low temperatures, so that the native state of the specimen is captured, without the need for artificial treatments, such as fixation or dehydration.

1. Visit the PDB home page and enter the PDB ID 6VYB in the Search box.

You are now on the "Structure Summary" page of the PDB entry for the molecule of interest. The Structure Summary page contains a title for the molecular structure, a snapshot of what the molecule/complex looks like, and the names of the authors who solved the structure, along with literature information for the article describing the structure, the macromolecules and the small molecules present in the structure, and several experimental details.

2. Explore the Structure Summary page for the structure with PDB ID 6VYB. What is the molecule/complex explored? Fill in the following table.

PDB ID	6VYB
Author(s) of entry	
Year when the structure was released	
Method of determining the structure	
# of entities	
# of protein chains	
# of carbohydrate chains	
# of small molecules	

3. Why does the protein subunit contain three chains, named A, B, C?

4. What other structures did the authors publish along with the one you are studying? Name the structure(s) and give the PDB ID(s).

5. In a new browser tab/window, visit the home page for iCn3D at https://www.ncbi.nlm.nih.gov/Structure/icn3d/full.html.
6. Click on the button called File >> Retrieve by ID >> PDB ID. A new window will open up.
7. Enter the PDB ID for the open state SARS-CoV-2 Spike ectodomain structure and click on the Load button.
8. If you are unfamiliar with iCn3D, click on Help >> Tutorial >> Use iCn3D and spend some time familiarizing yourself with the controls.
   Alternatively (or additionally), you can watch the videos by clicking on Help >> Tutorial >> iCn3D Videos.
9. How many polymer chains do you see?

10. Click on the structure and explore rotating it freely to view the different perspectives on this protein complex in three dimensions. Also try zooming in and out using your mouse or touchpad to see additional structural details.
11. After exploring the three-dimensional structure, go to the pulldown menu under Analysis >> Sequences and Annotations. This will open a window with information about the proteins and chemicals present in the structure.
12. Scroll down in the "Sequences and Annotations" window to see the different subunits that make up the ternary protein complex you are viewing. These subunits are labeled 6VYB_A through 6VYB_C. Since this protein complex is a homotrimer, each chain is a copy of the S protein.
13. Scroll down to the section that lists any chemicals, ions, or water molecules that are included in the structure file. What chemical is included in the structure? (Hint: You may need to go back to the "Structure Summary" page in the PDB.) How many molecules of the chemical are there in the structure?

14. Scroll back up in the "Sequences and Annotations" Window and click on the 6VYB_A sequence name. The name should be highlighted in yellow. This selects the protein in the structure, which is now highlighted in yellow in the Viewer Window as well.
15. From the "Color" pulldown menu, select Unicolor >> Red to turn the selected subunit structure red.
16. Repeat steps 13 and 14 for the other two subunits, marking them with colors of your choice.
17. In the "Sequences and Annotations" Window, click on the plus sign next to the "domain: SARS-CoV-2_Spike_S1_RBD" of chain 6VYB_A (the window may not allow you to see the full name, but this should be the first domain under the name of the chain).
18. Read the information about this domain contained in the menu. Given the information in the menu, do you expect the RBD to be in the S1 or in the S2 subunit of the S protein?

19. Close the menu with the details on the RBD, and click on the name of the "domain: SARS-CoV-2_Spike_S1_RBD" of chain 6VYB_A to select it.
20. With the RBD of chain A selected, select the RBD domains of the other chains by holding the Shift key and selecting each name. This may be challenging, so you may have to start over until you get used to the website controls.
21. Rotate the colored structure so that the RBD domains are at the bottom.
22. Save this image of your structure by going to File > Save File > iCn3D PNG image.
23. In the "Sequences and Annotations" window, select chains B and C. Go to Style >> Proteins >> Hide. Those subunits will no longer be visible in the image.

    In this structure, not all features can be hidden. Some oligosaccharides will remain visible, even when the selection is to hide them. This is because of the way that the structure file was written for this particular PDB ID.
24. To identify the N- and C-termini of the Spike protein, click on 6VYB_A in the "Sequences and Annotations" window. With the protein highlighted in yellow, go to Color > Spectrum > for selection. That Spike protein should now be a rainbow of colors (spectrum). The N-terminus is colored purple, and the C-terminus is colored red.

25. In the file you saved from step 22, is the N-terminus facing up or down? How about the C-terminus?

26. To display the secondary structures in the Spike protein, click on Color >> Secondary >> Sheet in Yellow. This will color all the β-strands in yellow and all the α-helices in red.
27. The structure you have been studying is the open conformation of the trimer. The team that published this structure also resolved the closed conformation, with PDB ID 6vxx. You can align the two structures by clicking File >> Align >> Structure to Structure >> Two PDB structures. Put the two PDB IDs in the two boxes and click on "All Matching Molecules Superimposed." The resulting structure is an alignment of the two structures with the overlapping molecules highlighted in red.
28. You can explore how well the structures have aligned by clicking on Analysis >> Aligned Seq. This will open a window with the sequences of the aligned structures. The overlapping amino acids will be in red lettering, but the ones that do not overlap will be shown in blue letters. Which amino acids are not overlapping? Which chains do they belong to?

29. Click on Analysis >> Sequences and Annotations. Use the Details tab in the Sequences and Annotations Window to determine whether the open conformation comes as a result of relative motion of the S1 or the S2 domain. (Hint: The residues that do not align well between the two structures are the ones that are flexible and differ between the open and closed conformations.)

Structure of the Human Angiotensin-Converting Enzyme 2 (ACE2)

In this section, you will explore the structure of ACE2, the SARS-CoV-2 receptor in human cells. The ACE2 protein is a membrane-bound carboxypeptidase, a protease that cleaves amino acids from the C-terminus of proteins, in the presence of a zinc ion.

30. In a new browser tab, go to www.rcsb.org and enter PDB ID: 1R42 in the top search box. This is the structure of human ACE2 determined on 10/07/2003 via X-ray crystallography.

31. Without closing the PDB browser tab, open a new tab and visit www.uniprot.org. Uniprot is a database that contains information on protein sequences and functions.
32. Type "Q9BYF1" (without the quotation marks) in the search box. This is the Uniprot ID for human ACE2.
33. Explore the information provided on the results page. What amino acids correspond to the active site? (Hint: You can click on the position number to recover the residue.)
34. Scroll down to the "Pathology and Biotech" section. Are there any possible mutations that may abolish the interaction with the Spike protein from SARS-CoV or SARS-CoV-2?
35. Visit the iCn3D site (https://www.ncbi.nlm.nih.gov/Structure/icn3d/full.html) and open the structure with PDB ID: 1R42.
36. Go to the pulldown menu under Analysis >> Sequences and Annotations. This will open a window with information about the proteins and chemicals present in the structure.
37. Click on the Details tab. Scroll right to find the amino acids that you identified as the ones forming the active site. You can select them by clicking and dragging the mouse over the amino acid letter. The amino acid will then get highlighted in yellow. If you selected the wrong amino acid, or you selected more than your targeted one(s), you can unselect by clicking and dragging the mouse over the amino acid letters you want to deselect. You can select multiple sites by holding down the shift key and repeating the above process.
38. Click on View >> Center Selection.
39. Click on View >> Zoom In Selection.
40. Rotate the view, so that you can get a good look of the active center. What metal ion is present in the active center?

41. Explore the molecular interactions that stabilize the active center of ACE2. Click on Analysis >> Interactions. This will open a new window, in which you can select what types of interactions you want to study. Uncheck the boxes for all types of interactions, except for hydrogen bonds and salt bridge/ionic interactions. Leave the distances as set by default. Click on 3D Display Interactions.

    The types of interactions are color-coded, so that hydrogen bonds will show as green dotted lines, and ionic interactions as teal dotted lines.

    How many hydrogen bonds do you observe? How many ionic bonds do you see? Report all interactions in the form of a table.

42. Because it can be difficult to see the interactions in 3D, click on the "Highlight Interactions in Table" button under the 3D Display Interactions button. This will open a new window with a table for each type of interactions detected. Which residue binds to the Zn^{2+}?

Structure of the SARS-CoV-2 S:ACE2 Complex

In this section, you will use the skills you developed before to explore the interactions between the S protein of SARS-CoV-2 and the human ACE2.

43. Close all previous browser windows and visit iCn3D (https://www.ncbi.nlm.nih.gov/Structure/icn3d/full.html).
44. Open the structure with PDB ID 6m0j. What is this structure? (Hint: Visit the PDB site to gather information about it.)

45. Use the Sequences and Annotations menu to identify the components of the complex. Write these down in the space below.

46. In the Details tab of the Sequences and Annotations window, check the box next to the work Interactions in the Annotations box, near the top of the window. This will reveal a new track in the annotations of the 6m0j_A chain, which will contain all the residues that interact with chain 6m0j_E.
47. Following the steps you performed earlier (steps 40–42), report three amino acids from each protein that participate in the interactions between the SARS-CoV2 S protein and the human ACE2 protein.

Student Assessment

The structure under study with PDB ID 6m0j was released on March 18, 2020, when the original strain of SARS-CoV-2 was the dominant one worldwide. As of July 2021, the delta variant, first reported in December 2020, became the most dominant variant in the world, likely due to its increased transmissibility (Li et al. 2021). The delta variant is defined by the following mutations in the Spike protein: 19R, (G142D), 156del, 157del, R158G, L452R, T478K, D614G, P681R, D950N. Of these mutations, positions 452 and 478 are found within the ACE2-binding domain.

Based on the structure that you are studying, what do you think the effect of mutations L452R and T478K will be? To explore that, you can look for interactions of the wild-type amino acids in positions 452 and 478. You can also consider whether the substitution will potentially generate new interactions. One way to think about this is to see what amino acids are found at a certain distance from the ones we study. After having selected L452 in the S protein RBD, click Select >> By Distance and increase the sphere to a radius of 5 Å. Click on the Display button. This will select (and highlight) all the molecules within a 5 Å radius of L452.

Using this method, identify the neighborhoods of residues 452 and 478. Given your knowledge of amino acid structures and bond formations, what do you expect to happen in the respective sites in the presence of the above mutations?

Instructor's Notes

Laboratory Preparation

This is an in-class or homework activity that guides students through the retrieval of protein structures from the PDB and their visualization using NCBI's online visualization tool iCn3D. The activity can take place in a computer lab, or students can work on their own.

It is highly recommended that students have a basic understanding of amino acid and protein structure and an understanding of intra- and intermolecular bonds. There is no need for any prior knowledge of virology.

Instructors who are unfamiliar with iCn3D can read the tutorial (https://www.ncbi.nlm.nih.gov/Structure/icn3d/icn3d.html#useicn3d) and watch the videos (https://www.ncbi.nlm.nih.gov/Structure/icn3d/icn3d.html#videos). A quick YouTube search will find unofficial videos that can be shared with students prior to the beginning of the activity.

Connections to Other Lab Exercises

Depending on the main focus of the course, the activity can be paired with the NCBI and/or the phylogenetic tree creation.

Recommended Schedule for the Lab Exercise

The activity can be demonstrated within a 3-h time period (lecture or lab). Students can follow along, using the instructions seen in module 2.

Chapter 4: in silico Design and Execution of Gene Cloning and Genome Editing