	SKILLS CENTER **STANDARD OPERATING PROCEDURE**	**A BIOFIZZ** **PRODUCTON**
Navigating the Protein Data Bank 1 **2 Module Hours**	**Effective Date: 03/15/2021**	**Revision # 1.0** **A. Siclair** **Checked: M. Guzie**

BACKGROUND

One of the most important concepts to have surfaced in biology is the connection between protein structure and function. It has become increasingly notable that this relationship is imperative to bodily systems and that disruption of protein structure can lead to the establishment of numerous different diseases and phenotypes, such as Alzheimer's and prion diseases. Luckily, with the advent of protein data banks like the RCSB PDB and the worldwide PBD, researchers have been able to share a conglomerate of data surrounding the important topic of protein structure.

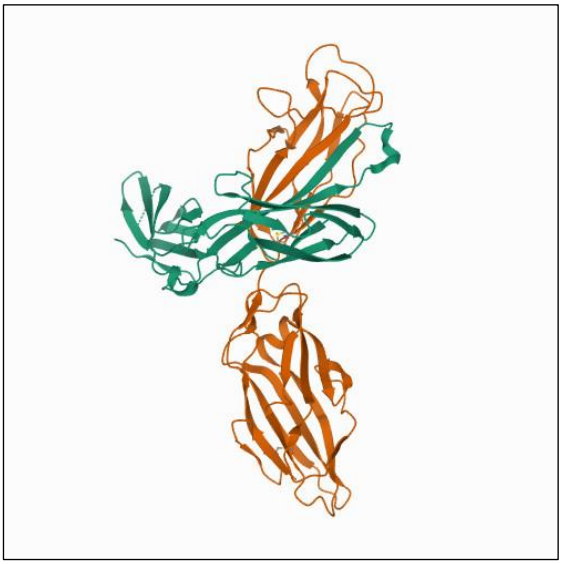

Figure 1: 1QUN, X-ray structure of the FIMC-FIMH chaperone adhesin complex from uropathogenic E. coli

The Research Collaboratory for Structural Bioinformatics (RCSB) protein data bank, established by Walter Hamilton in 1971, began its existence with a mere 7 protein structures to its name (RCSB PDB). As of 2015, over 100,000 structures have been documented in the archive in numerous different organisms, one of which is seen in Figure 1 (RCSB PDB). This program is advantageous because it allows for collaboration between researchers on an international scale, something that is invaluable when considering the big picture in research. The protein data bank information has contributed to fields including pharmacological advancement, such as the establishment of Gleevec for treatment of chronic myeloid leukemia, and will continue to aid in treatment progressions in the future at a macromolecular level (Sullivan et al., 2017).

The RCSB PDB is one of the first stops when analyzing a protein, whether focusing on structures or interactions, because the PDB file can be used within other applications such as UCSF Chimera for visualization. By developing an understanding of the fundamental tools in PDB, one can establish a solid foundation for further analysis down the line.

	SKILLS CENTER **STANDARD OPERATING PROCEDURE**	**A BIOFIZZ** **PRODUCTON**
Navigating the Protein Data Bank 1 **2 Module Hours**	**Effective Date: 03/15/2021**	**Revision # 1.0** **A. Siclair** **Checked: M. Guzie**

1. PURPOSE

The purpose of this procedure is to become comfortable with navigating the RCSB protein data bank (PDB) and understand how this tool can be utilized for analysis in other applications.

2. SCOPE

This procedure applies to qualified skills center users.

3. RESPONSIBILITY

3.1. It is the responsibility of the user to understand and perform the procedure described in this document.
3.2. It is the responsibility of the user performing the procedure to fully document any deviations from the written procedure.
3.3. It is the responsibility of the user to become trained on the procedure.

4. DEFINITIONS

4.1. Protein data bank – An international database that houses information about 3D macromolecule structures such as proteins and nucleic acids.
4.2. Coordinate files – The primary information in the databank that list the atoms in each structure, showcases the 3D location of the atoms in space and summarizes information about the structure, sequence, and experiment.
4.3. Biological assembly – The macromolecular assembly that has been presumed or shown to be the functional form of the macromolecule. (Dutta et al., n.d.)
4.4. Asymmetric Unit – The smallest portion of a crystal structure to which symmetry operations can be applied in order to generate the complete unit cell. It can include a portion of the biological assembly, the whole biological assembly or multiple biological assemblies (Dutta et al., n.d.)
4.5. PDB identifier – The 4-digit numerical and alphabetical code that corresponds to a specific structure in the PDB database.
4.6. Taxonomy – The classification of an organism (EX = eukaryote)
4.7. Chaperone – A protein that assists in the conformational folding of other macromolecules, primarily other proteins.
4.8. Resolution – The minimum distance needed to distinguish between 2 atoms.

5. MATERIALS/EQUIPMENT

5.1. RCSB Protein Data Bank: https://www.rcsb.org/

5.2. NCBI Protein Database: : https://www.ncbi.nlm.nih.gov/guide/proteins/

6. PROCEDURE

6.1. Understanding the PDB Search Options

6.1.1. Go to the RCSB PDB: https://www.rcsb.org/

6.1.2. Via direct input of the protein of interest or related search terms into the home page search bar, a basic search can be conducted. (Figure 2)

- Key words can include things like the name of the protein, the organism of interest, gene name, etc.

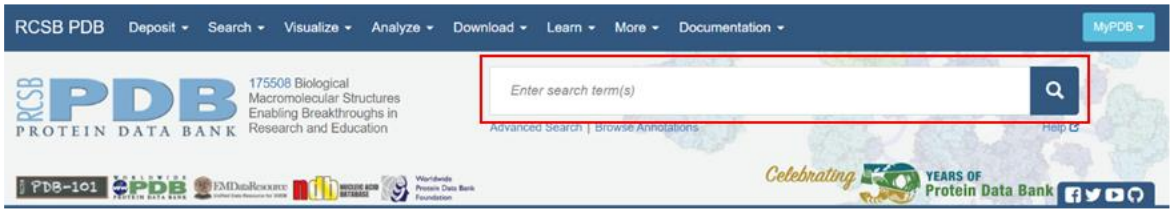

Figure 2: The homepage of the RCSB PDB, which allows for basic search functions.

6.1.3. Searching "chaperone" resulted in the page seen in Figure 3. From here, many related structures show up in the summary section, each with their own separate PDB identifiers.

- Utilizing the Refinements section can help to narrow down the search results by selecting for specific organisms, taxonomy, the experimental techniques used to determine the protein structure, etc. (Figure 3)
- The MyPDB feature allows for structures of interest to be saved to an account for future use. It is possible to sign in using Google, ORCID or Facebook accounts. (Figure 3)

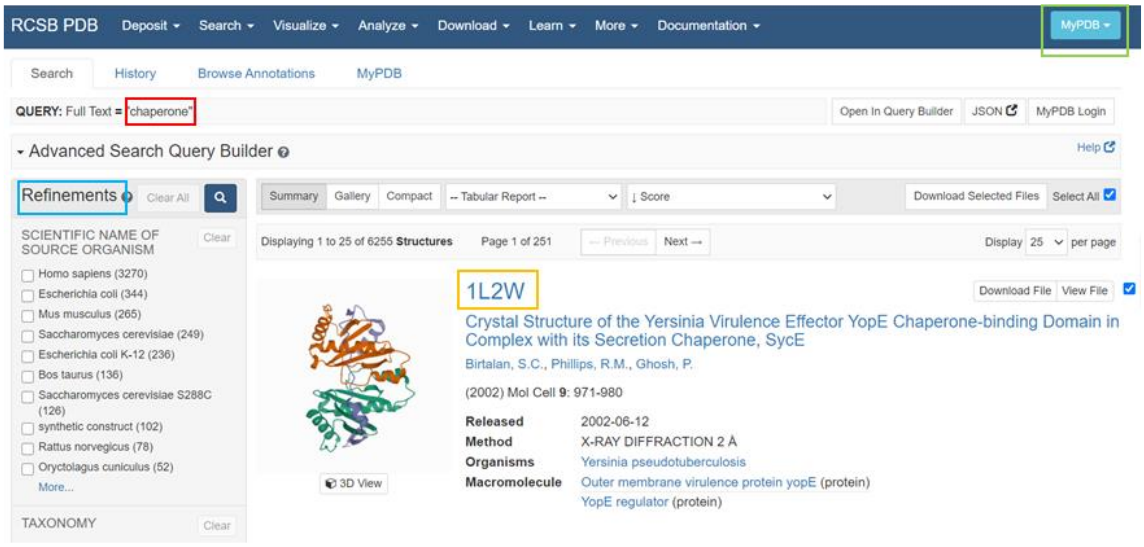

Figure 3: The results for a basic search for chaperone proteins.

6.1.4. It is also possible to further narrow the search using the Advanced Search Query Builder. This search feature is mainly applicable when you are looking for a specific entry in the PDB. The question mark next to each search option tells a bit more about how each specific tool functions. (Figure 4)

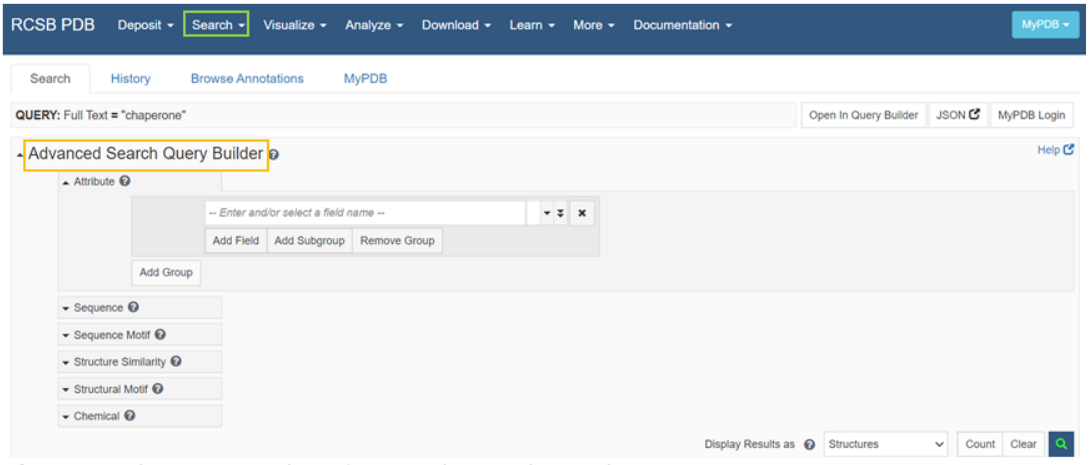

Figure 4: The query options for an advanced search.

- This function can also be accessed from any page by clicking Search and Advanced Search. (Figure 4)
- Attribute: This search tool has many different categories that can be applied either separately or together for further searches. Example search parameters include:
  1. PubMedID: if you are looking for structures specifically mentioned in a paper.
  2. UniProtKB Accession Number: if you are looking for structures of specific annotated proteins.
  3. Protein stoichiometry: A = monomer, AB = heterodimer and A2 = homodimer
  4. Secondary structure content: you can document the percentages for alpha helices versus beta sheets here.
- Sequence: Search using either the nucleic acid or protein sequence of the molecule of interest. Protein sequences can be found using the NCBI protein database: https://www.ncbi.nlm.nih.gov/guide/proteins/. Note that only the letter codes for the amino acids can be used, and that the numbers found on the NCBI site must be taken out of the sequence for the sequence query search to work properly. When including a sequence in the search, make sure to change Display Results as to Polymer Entities. (Figure 5)

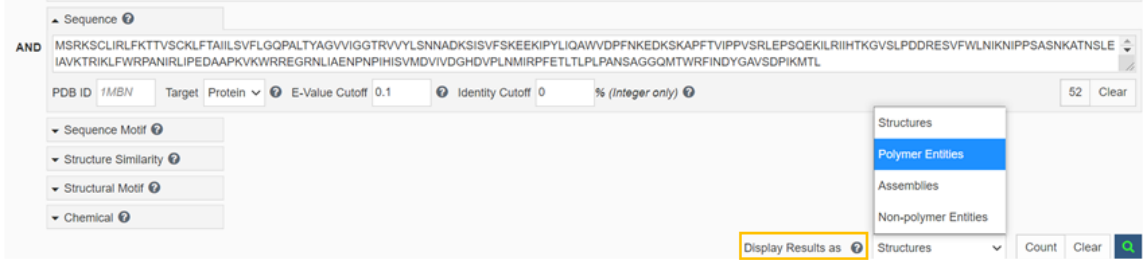

Figure 5: The sequence search function.

- Chemical: The chemical formula can be used to identify interactions between ligands and their corresponding binding sites/macromolecules.
- See this link for more information on advanced search options: https://www.youtube.com/watch?v=zVhfJbplAXY

6.1.5. Once on the page of a specific PDB structure, it is also possible to search for proteins with a similar sequence to that singular structure. By scrolling down on the page, clicking sequence and choosing the percent identity cutoff desired, this can be achieved. (Figure 6)

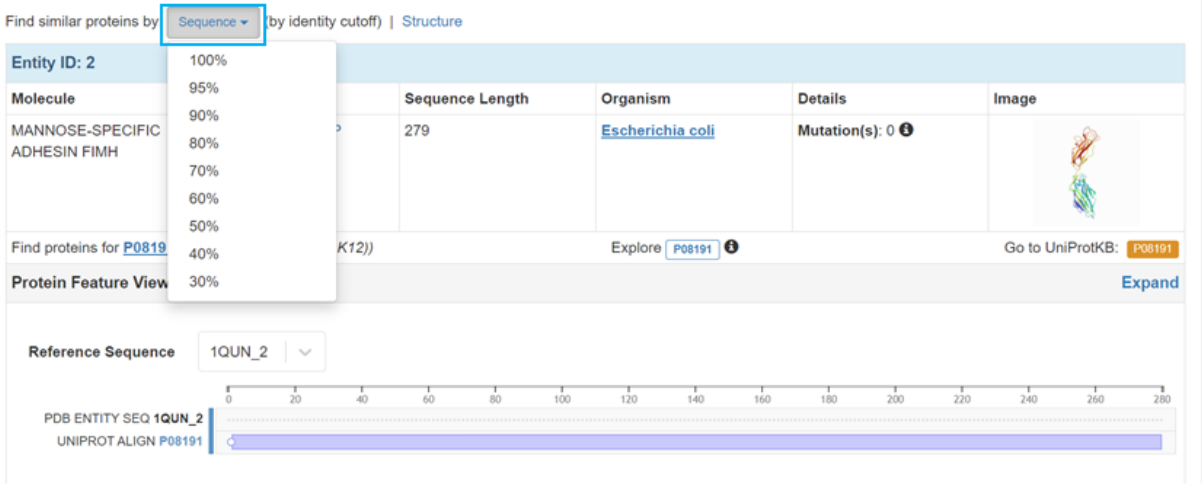

Figure 6: Searching for proteins with a similar sequence to a specific PDB entry.

6.2. Understanding and Analyzing Biological Assemblies and Asymmetric Units

6.2.1. Go to the RCSB PDB: https://www.rcsb.org/

6.2.2. Search the macromolecule of interest.

- The example search for this procedure was E. coli chaperone protein.

6.2.3. Click on one of the resulting PDB files to open a new page specific to that molecule and its experiments.

- The example used for this procedure was 1QUN.

6.2.4. A 3D structure will pop up in the left-hand portion of the screen, showcasing either an asymmetric unit or a biological assembly (see definitions for a more detailed description). (Figure 7)

- There will likely be multiple different biological assemblies and/or asymmetric units for the given protein.
- Utilize the arrows along the top of the image to scroll through the different available assemblies.
- Each structure will have its global stoichiometry and symmetry listed.
- It will also be noted if the structure was provided by the author or created by a specific software.

- The asymmetric unit and biological assembly can be the same but are oftentimes different from each other.

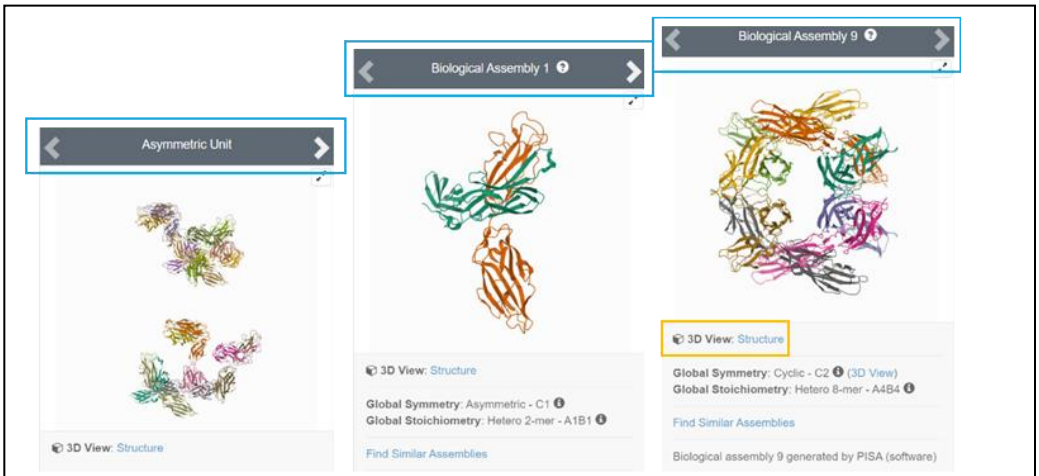

Figure 7: The asymmetric unit and 2 of the biological assemblies for 1QUN.

6.2.5. Clicking on the Structure tab in 3D view opens a new window that gives more details on the specific biological assembly/asymmetric unit chosen. (Figure 7)
- Ensure that the Mol* Javascript tool is being used for this procedure by looking at the lower right-hand corner of the new window near the Select a Different Viewer tab.

6.2.6. The Measurements tab can be used to measure things like distances and angles and to label specific points in the macromolecule. (Figure 8)
- 1 point must be selected to label something, 2 points must be selected for measuring distance and 3 must be selected for measuring angles.
- To enable selection mode to begin selecting points for measurement, select the cursor icon. (Figure 8)
- Clicking the cursor will bring up a new menu at the top of the 3D view of the protein. Each point will allow for a different form of selection when analyzing the molecule, as described in Figure 7. (Figure 8)
- By clicking on the Add/Union, choosing a single residue on the protein, clicking the Add button under Measurements and selecting Label, the Asp192 label was created. A similar process to select a second residue will allow for distance to be measured. (Figure 9)

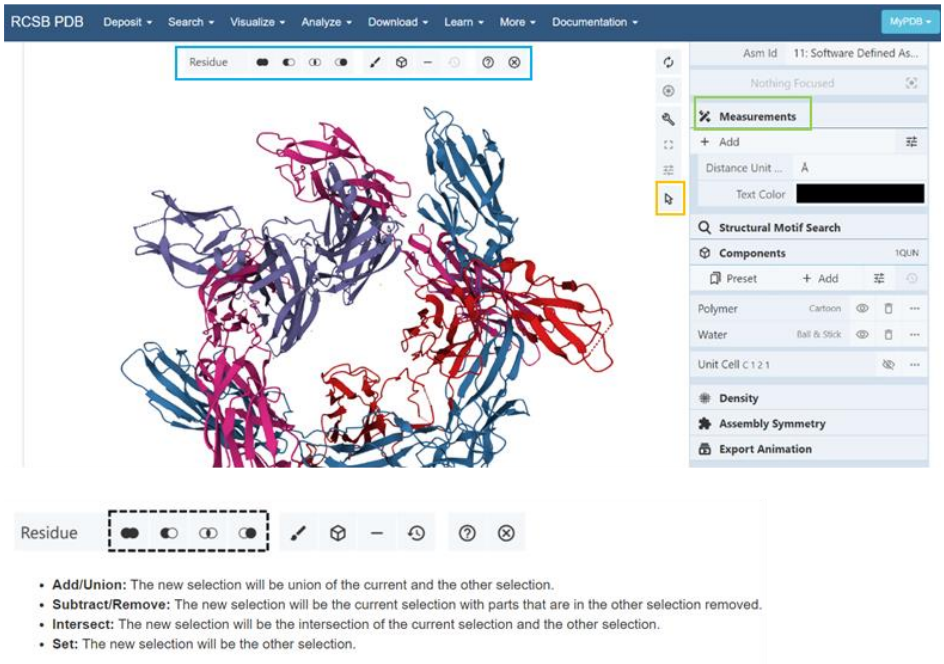

Figure 8: The 3D viewer of the molecule of interest in Mol Javascript, along with descriptions of functions within the selection viewer.

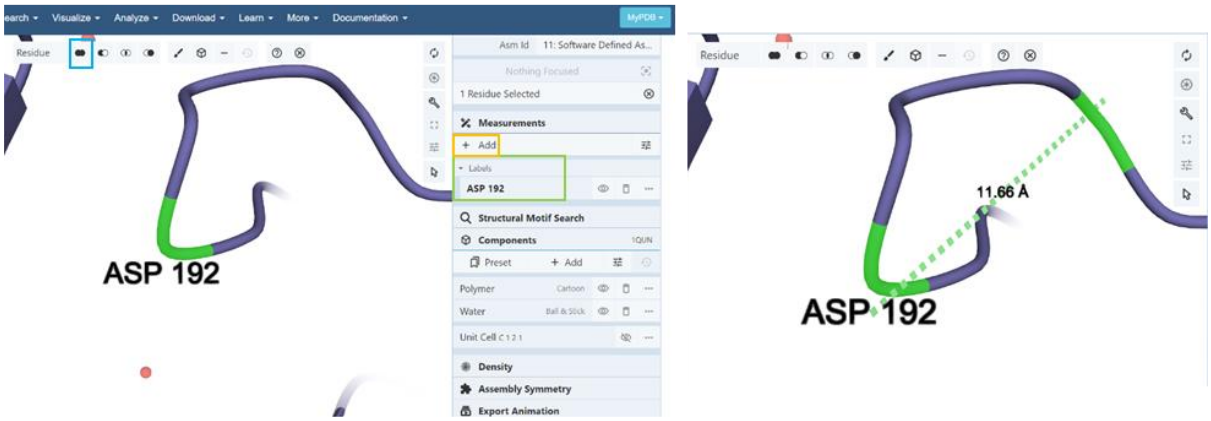

Figure 9: Labeling of the Asp192 residue in the protein (left) and labeling of the distance between the Asp182 residue and another residue (right).

6.2.7. After determining which structure is most applicable to your needs, it is possible to download the desired file from the original PDB page by clicking Download Files and then choosing the desired biological assembly or asymmetric unit. (Figure 10)

- This download can be used in other applications, like Chimera.

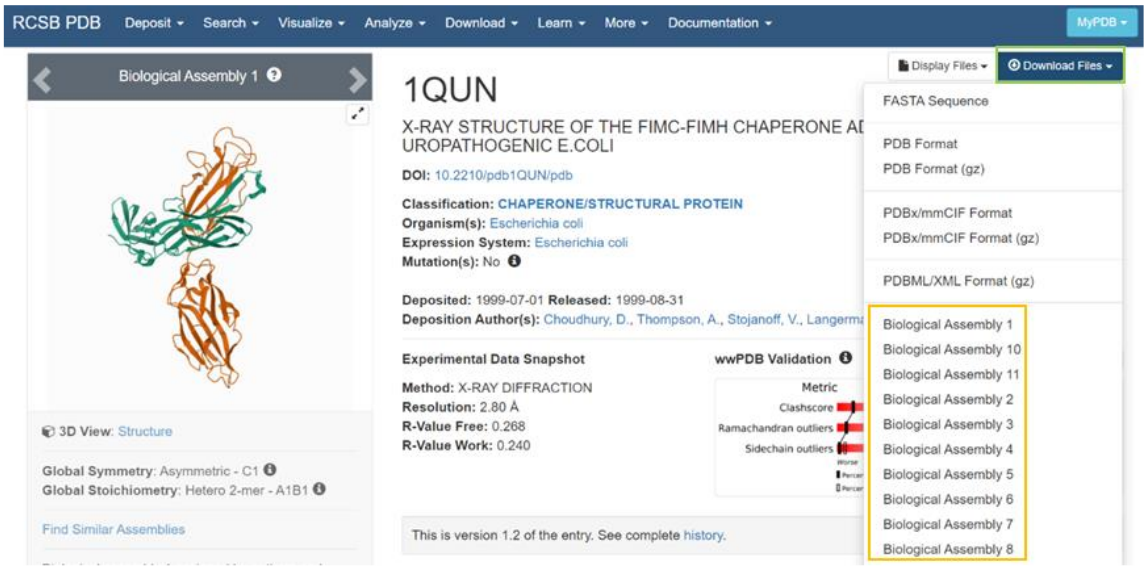

Figure 10: Downloading the 1QUN structure.

7. REFERENCES

[BioPandit]. (2018, Mar 1). *Basic Guide to Protein Data Bank and PDB Files.* Youtube. https://www.youtube.com/watch?v=Fz5t5QPUyi8

[BioPandit]. (2018, Mar 3). *Protein Data Bank Advance Search Guide.* Youtube. https://www.youtube.com/watch?v=zVhfJbplAXY

Dutta, S., Kramer Green, R. & Lawson, C.L. (n.d.) *Introduction to Biological Assemblies and the PDB Archive.* Educational Portal of RCSB PDB. https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/biological-assemblies

[EpicSelenium34]. (2017, Nov 18). *PDB Tutorial: A Basic How-to.* Youtube. https://www.youtube.com/watch?v=DgVfSQUSdlk

RCSB Protein Data Bank. (1999). *1QUN: X-Ray structure of the FIMC-FIMH chaperone adhesin complex from uropathogenic E. coli.*

RCSB Protein Data Bank (2021, Feb 6). *Making Selections.* RCSB Protein Data Bank.

https://www.rcsb.org/docs/3d-viewers/mol*/making-selections#selection-mode

RCSB Protein Data Bank. (2021, Feb 6). *Managing the Display*. RCSB Protein Data Bank. https://www.rcsb.org/docs/3d-viewers/mol*/managing-the-display #structure-panel

Sullivan, K.P., Brennan-Tonetta P. & Marxen, L.J. (2017). Economic Impacts of the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank. *Rutgers Office of Research Analysis.* Retrieved Saturday March 23, 2021 from https://cdn.rcsb.org/rcsbpdb/general_information/about_pdb/Economic%20 Impacts%20of%20the%20PDB.pdf

[5 min Sci Chores]. (2020, Aug 15). *How to Search Protein Data Bank (PDB) Tutorial.* Youtube. https://www.youtube.com/watch?v=DogX_vHezBs

8. MODULE MASTERY TASK

This task will test your knowledge and skillset searching through the RCSB Protein Data Bank.

1. Choose a protein using the basic search, narrow the results to a specific organism of your choosing. Document the PDB code and organism of the protein you chose.
2. What was the method used to establish this structure?
3. Are the asymmetric unit and the biological assemblies the same for your protein?
4. Describe the stoichiometry and symmetry of both the asymmetric unit and biological assembly if available.
5. Scroll down and search for proteins with a similar sequence to your chosen protein. What percent identity did you use? List and describe one of the resulting proteins. Is its function similar to that of your original protein?
6. Using a protein sequence found on NCBI, try an advanced search and describe what you find on the PDB database.
7. Open your protein in the Mol Javascript 3D Viewer and label a residue of your choice. Submit the resulting image.
8. Select another residue of your choice and measure the distance between the 2 residues. Submit the resulting image.
9. Select a third residue of your choice and calculate the angle. Submit the resulting image.