UNIVERSITY OF PENNSYLVANIA

Department of Chemical and Biomolecular Engineering

CBE 5520 Protein Engineering & it’s Applications Spring 2025

Hands-On Module**: Structural Alignment With SPRITE and ChimeraX**

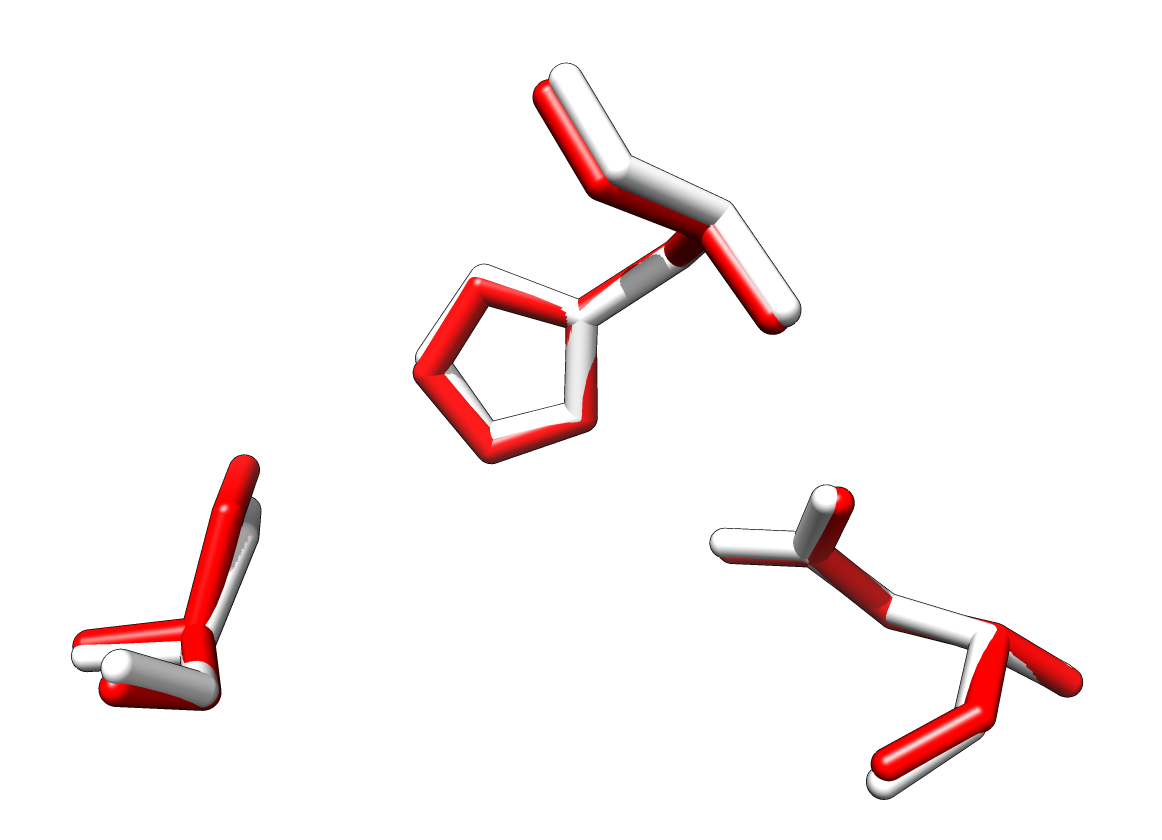
This tutorial engages students with the structure of a protein whose enzymatic function is unknown. This hands-on tutorial uses the program SPRITE (1) to search the structure of a protein for configurations of amino acid side chains that have a similar structure to those of known enzyme active sites. We will also make use of ChimeraX (2) - a structure viewing program, to visualize the results from SPRITE. Using alignments based on presumed active sites allows researchers to form hypotheses about the possible function of an unknown enzyme. This alignment can be a stand-alone activity or can be used as a platform to further explore the structure and function of the protein of interest through various computational and *in vitro* approaches.

In this module, we will use SPRITE to search the structure of a protein to determine if it has an active site with a motif of known catalytic function. The active site of an enzyme is the region of the protein where the substrate binds and the chemical reaction takes place. This is often a groove or pocket on the surface of the enzyme that allows the substrate to bind. This region also contains the amino acids necessary for catalysis to occur. The arrangement of those amino acids in space is called a motif. The Mechanism and Catalytic Site Atlas (3) is a database of motifs with known catalytic function. SPRITE uses this database to identify possible motifs in a protein of interest. Although catalytic site structural homology alone is not sufficient to define the function of a protein, it provides one mechanism which, when combined with other structural and sequence motifs, can suggest candidates for experimental verification.

These tools can be used only to develop a hypothesis for the potential function of a protein. Final confirmation of the function often relies on biochemical techniques, assays, and characterization (wet-lab). It is not sufficient to simply identify structural homologs (proteins with similar structures and identified with tools like Dali and MarkUs) or sequence homologs (proteins with similar sequences that are identified with tools such as BLAST and HMMER).

The goal of this tutorial is to choose a previously uncharacterized protein structure from the PDB and use currently available computational tools (SPRITE, BLAST, Pfam, Dali, etc.) to develop a hypothesis about the putative function.

The first step in our function prediction process is to compare a protein of unknown function against a library of motifs from the Catalytic Site Atlas (<https://www.ebi.ac.uk/thornton-srv/m-csa/>) and other sites that together constitute the motif template library of SPRITE. Each catalytic site motif template typically consists of 2-5 amino acid residues that have a fixed spatial and distance relationship. The example shown in Figure 1 is an alignment for a serine protease.



**Figure 1.** Alignment for a serine protease. Alignment of PDB entry 1AFQ (bovine gamma chymotrypsin; the query in red) with a motif template based on 1A0J (in white) a trypsin structure from Atlantic salmon. Three residues from 1AFQ (His 57, Asp 102 and Ser 195) aligned with the three homologous residues from 1A0J (His 57, Asp 102 and Ser 195).

### 

### The purpose of this tutorial is for you to become familiar with SPRITE so that you can complete a structural alignment of any protein or your engineered protein of unknown function. This will provide an early clue to the function of the protein.

* For this hands-on activity, the only physical equipment you will need is a computer with access to the internet and the following digital data/resources:
  + The four-character PDB code of protein (e.g., “4EZI”)
  + [The SPRITE database website](http://211.25.251.163/sprite/)
  + [ChimeraX](https://www.cgl.ucsf.edu/chimera/download.html) 1.9 software

Navigate to SPRITE

The SPRITE website is located at the URL <http://211.25.251.163/sprite/> (Figure 2)

|  |
| --- |
| The database page that searches an input protein structure for motifs and matches similar proteins |
| **Figure 2.** The opening screen for SPRITE includes general information about the program. The SPRITE interface allows you to upload your own file in a PDB format or to enter a 4-character PDB ID. You can choose to include or exclude 2-residue patterns (exclude is best). |

Finding a Motif in a Query Protein

Start the exercise by entering the PDB ID “2HNT” in the SPRITE query box (Figure 2). Once you have gone through this process with this structure of known function (human thrombin), you will conduct this exercise with a protein of unknown function (search the PDB for “uncharacterized proteins”. After completing your search, you will be asked to choose a set of motif templates for comparison with your query structure.

1. Using the SPRITE interface, submit your structure (either 2HNT or, later, your choice of protein of unknown function from PDB) as a query and exclude 2-residue hits. You can wait for your search to run.

|  |
| --- |
| Generated links to the active site aligned proteins in SPRITE |
| **Figure 3.** Results Pane. Initial SPRITE results are presented as a text screen, with options for how to view the results. |

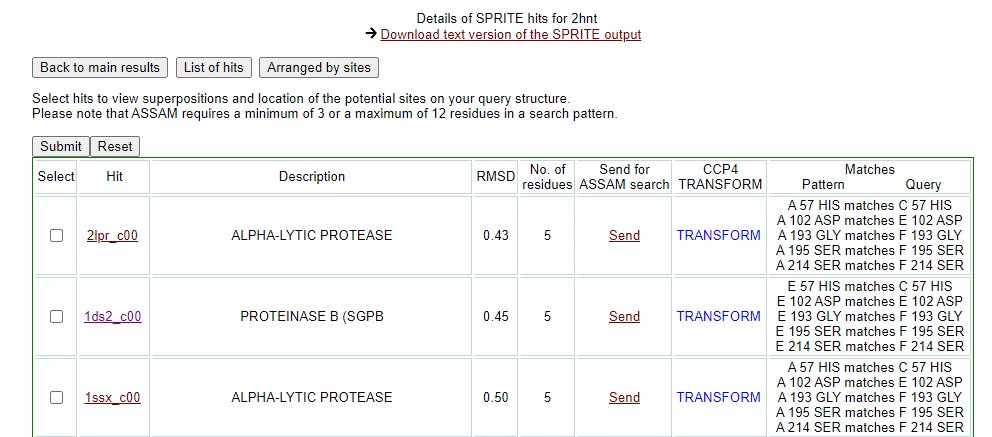
1. The Results Screen. Once a search is complete, the results will appear in the Results window. You can choose to view a list with limited details (“List of hits”), a list with full details (“Full details”) or arranged by all the matches for a given site in your search protein (“Arranged by sites”). The “List of Hits” results shown below were obtained by searching 2HNT. At this point, you should browse the “List of hits”.

|  |
| --- |
| The protein hits generated by SPRITE with their description and RMSD |
| **Figure 4.** List of Hits for 2HNT Search. List of hits for 2HNT are presented as a table, with limited information. |

The information shown:

* 1. *Hit*: the PDB ID that matched the search protein to generate the hit
  2. *Source PDB ID*: a clickable link to the matched PDB file at the Protein Data Bank.
  3. *Description:* a brief description of the function of the protein matching your protein
  4. *RMSD*: the root mean square deviation. A perfect match would have an RMSD of zero. A high-quality alignment will have an RMSD value of 2.0 Angstroms or less.

1. Dig deeper by viewing the “Full details” results alignments that also report the number of matched amino acid residues, along with a list of the residues that aligned. Remember that a hit with only 2 amino acid residues aligned is a poor hit, and you likely excluded those using the settings prior to doing the search.



**Figure 5.** Full Details for 2HNT search.

1. To view the alignment between your protein and the matched protein, check the box in the “Select” column (you can select more than one hit) and click “Submit”. This can be done from the “List of Hits” or the “Full details” screen. Each will display a different version of the overlay (Figure 6a and 6b). Explore them both and decide which is most helpful for your specific protein. The view from the “Full details” screen also allows you to turn on the protein backbone, and to mouse over parts of the backbone to get the identity of the amino acid at a given position.

|  |
| --- |
|  |
|  |

|  |  |  |
| --- | --- | --- |
| Two protein active site residue ball and stick structures superimposed |  | Two protein active site residue stick structures superimposed |
| **Figure 6a.** Alignment of 2HNT with 2LPR (green). Alignment was generated from the “List of hits” screen. The “Show pattern match” button was used to show the overlay. |  | **Figure 6b.** Alignment of 2HNT (yellow) with 2LPR (green). The “Superposed motifs” function was utilized such that both proteins are visible. Alignment was generated from the “Full details” screen. |

1. It is also possible to view hits by each site in your protein using the “Arranged by sites” function (Figure 7). This view shows all the hits found for a specific set of amino acids. There may be anywhere from one to several hits for each specific set of amino acids. Multiple alignments for one site, if they all have a similar function, means higher confidence in the results.

|  |
| --- |
| A compiled list of proteins that matched with a specific set of active sites found in the target protein |
| **Figure 7.** Results for the alignment of 2HNT arranged by sites. Less information is provided for each alignment, but the confidence is better if there are multiple proteins of similar function that align with a specific site. |

Review the alignments with an RMSD below 2.0 Angstroms.

1. Now repeat this exercise using your chosen protein structure of unknown function. Be sure to record the same data for your alignments.

**Interpreting Results**

Each time you explore the alignment of a query with a template (e.g., 2NHT with 1A0J), you should record the RMSD values and a screen capture of the visual alignment. It is also important to inspect the visual alignment by rotating the molecules. Notice that you can also turn on the protein backbone in some views, allowing you to see if the amino acids are on the surface, in a pocket, or buried in the core.

At this point, you may find it useful to visualize your whole enzyme using another visualization program. You can now open your enzyme structure in Chimera and find where the amino acids are that make up the predicted active site. Depending on your SPRITE results, you may have found a single active site location that aligned with a few other structures, or you may have found multiple possible active sites from different amino acid residues aligned to different structures. It is best to keep SPRITE open at this point, in case you wish to go back and review your findings. This is especially important for inspecting the visual alignment by rotating the molecule. You can turn on the labels to identify the residues in your alignments if it is not apparent by the structure.

**Opening ChimeraX**

Open ChimeraX software.

Aligning Structures

To load the unknown protein, select the drop down menu “File” to choose “Fetch by ID…” and enter the unknown protein PDB (2HNT) next to “PDB” (Figure 8). Click “fetch” and the structure of the protein will be shown. Note this will be protein #1. Using results from SPRITE, load the protein of known function into ChimeraX by repeating the above step with the known protein PDB (1A0J). Note this will be protein #2.

You can even use the command line to put the syntax as

* $ open 2HNT
* $ open 1A0J

Start this exercise by using thrombin, 2HNT, as protein #1 and trypsin, 1A0J, as protein #2. (note: this is 1A[zero]J, not 1A[letter O]J).

|  |
| --- |
| A panel in Chimera that allows users to search and load specific structures, in this case from the protein data bank |
| **Figure 8.** Fetch Structure by ID table. Here, the protein thrombin, 2HNT, is being fetched by ChimeraX using the manual ID box. |

Since 1A0J is a tetrameric structure, it will be helpful to hide all but the A subunit. There are either way to do this: click the “Select” menu and choose Chains -> trypsin and then Chain A. Click the “Select” menu again and choose “Invert (Selected Models)”. You should now see the blue protein with a green outline, except for the A subunit. Now click the “Actions” menu and “Cartoon” -> hide. The “Actions” again and “Atoms/Bonds” -> hide. At this point, you should see just two protein globs. OR

You can even use the command line to put the syntax as

* $ delete #2 /b-d

## \*Notes:

Selection specific regions of your model is essential for structural analysis and figure making in ChimeraX. In ChimeraX, selected regions will be displayed with a green outline. There are several ways to select residues:

* **Through the Select menu in the toolbar**. This is good for beginners, but offers only corse control over the selection
* **Through the command line (recommended)**. Although it takes a little practice initially, the command line is the most powerful tool in ChimeraX and allows scripting your workflows, which is essential to make figures reproducibly and change them quickly. The following describes the selection syntax

### Hierarchical Specifiers

Hierarchical specifiers are the most common way to select items. They have up to four levels:

1. **Model** (#)
2. **Chain** (/)
3. **Residue** (:)
4. **Atom** (@)

| **Symbol** | **Level** | **Description** | **Example** |
| --- | --- | --- | --- |
| # | Model | Model number in ChimeraX, separated by dots (e.g., #1, #1.3). | #1, #1.3 |
| / | Chain | Chain identifier (e.g., A, B). | /A |
| : | Residue | Residue number or name (e.g., :51, :glu). | :51, :glu |
| @ | Atom | Atom name (e.g., @ca). | @ca |

To align the residues, the motif found via SPRITE needs to be entered in the command line of ChimeraX. Type “align” in the command line, followed by a space. Then enter “#1:” followed by the “/” with chain letter of the unknown protein followed by “:” residue number separate the list of residues with commas. Add a space and repeat for all the chains for protein#1. After inputting all unknown protein residue data, enter a space before repeating the process, this time using the known protein as protein “#2:” followed by residue information.

Full format to compare 2HNT and 1A0J is as follows (Figure 9):

* $ align #1/C:57 #1/E:102 #1/F:195 to #2/A:57,102,195

A screenshot of a computer

AI-generated content may be incorrect.

|  |
| --- |
| **Figure 9.** Aligned proteins thrombin, 2HNT (gold) and trypsin, 1A0J (blue). The proteins are aligned by active site residues His C 57, Asp E 102, and Ser F 195 from 2HNT and His A 57, Asp A 102, Ser A 195 from 1A0J. The RMSD value, 0.335 angstroms, is found in the log pane of the ChimeraX window. |

The RMSD value will be shown in the log pane of ChimeraX. This is shown in angstroms. Use the RMSD to determine the quality of the alignment; any alignments with an RMSD value below 2.0 angstroms are considered high quality.

**Visualization of Alignment**

To better visualize the alignment, delete the words “align” and “to” in the command line and at the start add “sel” leave the chain identifiers and residues as they were (Figure 10). This selects only residues of interest.

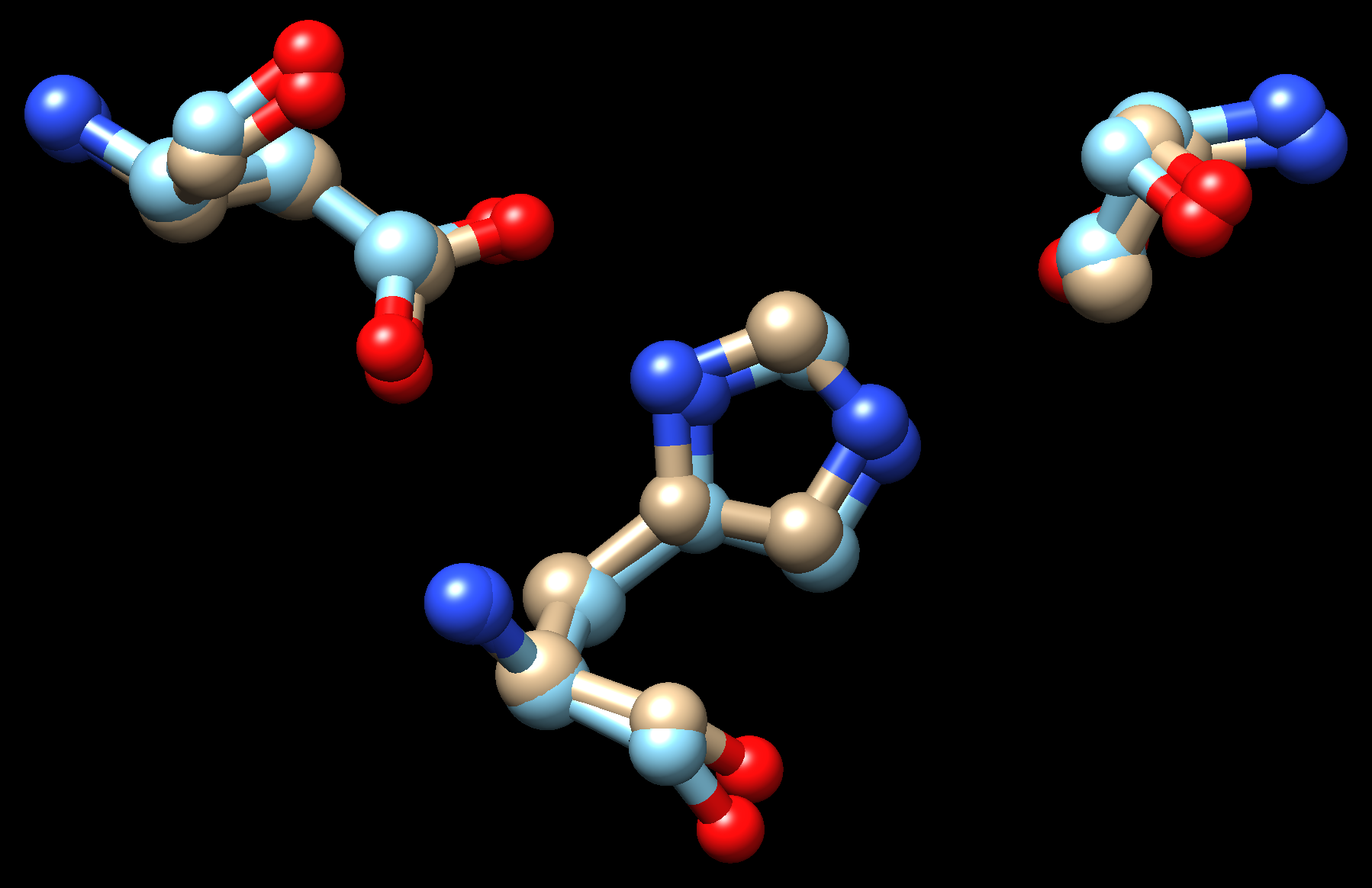
Full format to select active site residues for 2HNT and 1A0J is as follows:

* $ sel #1/C:57 #1/E:102 #1/F:195 #2/A:57,102,195

It will be easier to see the residues you selected and overlapped if they are showing in Ball and Stick format. To do this, with these residues selected from the previous paragraph, pull down the “Actions” menu, choose “Atoms/Bonds” and “show”. You should see the full residues now. Go to the “Actions” menu again and “Atoms/Bonds” and choose “ball & stick”.

To get the same view you saw in SPRITE, with the desired residues still selected, use the drop down menu option “select” and choose “invert (all models)” which will select all parts of the visualization except for the residues of interest. Then, under “action” hover over “atoms/bonds” and select “hide.” Repeat this to hide “ribbon.” Only the ribbons of residues of interest will remain in the visual field. Hold down “ctrl” on the keyboard and drag over the residues or press “enter” on the command line to reselect desired residues. If residues are not shown, use the drop down menu “action” choose “atoms/bonds” and select “show”. Next, hide the ribbon as directed above from the action menu. To enlarge the image, select the residues and go to “action” and select “focus.” Hold and drag the mouse to rotate the image as desired. The active site residue alignment for 2HNT and 1A0J are shown in Figure 12.

To name the selection for future use, use the drop-down menu “select” and choose “Name Selection…” to name as desired. To save the session, use the drop down menu option “File” and select “Save Session As…” to save in an accessible location. To see the rest of the proteins, invert the selection again and show the ribbon. You can now see how the two proteins are related to each other.



**Figure 12.** Visualization of active site residues His C 57, Asp E 102, and Ser F 195 from 2HNT (gray) and His A 57, Asp A 102, Ser A 195 from 1A0J (blue).

Once you have completed this exercise with thrombin (PDB ID 2HNT) and with your own chosen protein structure of unknown function, you will have generated results about the following:

* alignments between your chosen protein (uncharacterized proteins from PDB) and one or more motif templates, including the PDB ids for the motif templates;
* the residues in the alignments (for both the query and the motif template);
* RMSD values for the alignments; and
* images of the alignments.

 Here is the documentation for the ChimeraX align command.

[https://www.cgl.ucsf.edu/chimerax/docs/user/commands/align.html](https://urldefense.com/v3/__https:/www.cgl.ucsf.edu/chimerax/docs/user/commands/align.html__;!!IBzWLUs!UsKzz3rMCalD1pBHJVwgrv4F-ijdcqP9cBIYRv5t4HGqE_upaudk26OgBxpX45OQILgXHjNzNeelHR_RCxM$)

**References**

1. Nadzirin, N., Gardiner, E. J., Willett, P., Artymiuk, P. J., and Firdaus-Raih, M. (2012) SPRITE and ASSAM: web servers for side chain 3D-motif searching in protein structures. *Nucleic Acids Res.* **40**, W380–W386

2. Meng EC, Goddard TD, Pettersen EF, Couch GS, Pearson ZJ, Morris JH, Ferrin TE. Protein Sci. 2023 Nov;32(11):e4792.

3. Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE. *Protein Sci.* 2021 Jan;30(1):70-82.

4. Ribeiro, A. J. M., Holliday, G. L., Furnham, N., Tyzack, J. D., Ferris, K., and Thornton, J. M. (2018) Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res.* **46**, D618–D623