Structural Bioinformatics Workshop - Pymol version

Updated November 2021

Find up to date version here:

https://github.com/elliot-drew/structural-bioinf-workshop/blob/main/workshop_pymol.md

Overview

In this workshop you will learn how to do the following:

- Identify a protein from its sequence
- Select an appropriate template for said protein
- Build a homology model using "traditional" template based methods
- Examine drug/ligand binding through homology modelling and structural alignment

This is the preferred version of the workshop instructions, as Pymol is much more powerful and useful than NGL - and its much more likely that you'd be using Pymol or something similar in a proper research setting.

However, if you have issues with Pymol, feel free to use the NGL version (https://raw.githubusercontent.com/elliot-drew/structural-bioinf-workshop/main/workshop_ngl.md) of the instructions. Until Step 5, both protocols are the same.

Modelling Protein Structures

Step 1

Imagine you are a researcher interested in differences in the clinical outcomes observed for COVID-19 patients being administered the steroid dexamethasone. Through proteomic studies of patients who don’t respond well to treatment, you have identified the following protein variant as significant (sequence shown):

```
>unknown protein
MKWVTFISLLFLFSSAYSRGVFRRDAHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTC
VADESAENCDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPEVDVMCTAFH
DNEETFLKKYLYEIARRHPYFYAPELLFFAKRYKAAFTECCQAADKAACLLPKLDELRDEGKASSAKQRLKCASLQKFG
ERAFKWWAVARLSQRFPKAEFAEVSKLVTDLTKVHTECCHGDLLECADDRADLAKYICENQDSISSKLKECCEKPLL
EKSHCIAEVENDEMPADLPSLAADFVESKDVCKNYAEAKDVFLGMFLYEYARRHPDYSVVLLLRLAKTYETTLEKCC
```

```
AAADPHECYAKVFDEFKPLVEEPQNLIKQNCELFEQLGEYKFQNALLVRYTKKVPQVSTPTLVEVSRNLGKVGSKCC
KHPEAKRMPCAEDYLSVVLNQLCVLHEKTPVSDRVTKCCTESLVNRRPCFSALEVDETYVPKEFNAETFTFHADICTL
SEKERQIKKQTALVELVKHKPKATKEQLKAVMDDFAAFVEKCCKADDKETCFAEEGKKLVAASQAALGL
```

a. Identify what this protein might be by running the protein BLAST search using the above sequence. Choose the “UniProtKB/Swiss-Prot” option for database but leave all other options as default. You can find the webserver here: (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&PAGE_TYPE=BlastSearch&BLAST_SPEC=).

b. Look at the top alignment in your results. What is the sequence identity? Are there any differences? If so, what is the residue number and amino acid substitution? Make a note of this for later.

c. Search for the sequence ID of the top match using UniprotKB (https://www.uniprot.org/). Look at the entry for this protein – What is its name? What is its function? Does it bind dexamethasone?

d. On the UniProt entry for this protein, are there any experimental structures available?

Step 2

You have decided you want to create a homology model of this sequence to see if there is a structural basis for poor response to dexamethasone treatment. This means you need to find a suitable template. Since you are interested in binding to a specific drug, it would begood if you could identify a crystal structure of an albumin protein with dexamethasone bound to it.

a. Go to the RCSB PDB website (http://www.rcsb.org/)

b. Search using the name of the protein and the name of the drug. Do any structures match?

c. Click on the entry for the top structure, and take a note of the 4 letter PDB ID. What organism is the structure from? Is the drug molecule present?

d. Download the FASTA sequence for the protein, and the structure file in PDB format (the download button is blue, on the right of the page).

Step 3

You need to check whether this structure will be a good template. You can do this by looking at the alignment of the template sequence (from the FASTA file you just downloaded) for your target protein.

a. You can do a pairwise alignment with BLASTp. Compare your two sequences, with your target protein sequence as the query and template sequence as the using this link: (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&BLAST_SPEC=blast2seq&LINK_LOC=blasttab).

b. Look at the alignment – what is the sequence identity %? What is the query coverage %? Is this good enough to obtain a good homology model of your protein?

Step 4

If you are happy that the template you have chosen will be good enough, it is time to make the homology model. This should take around 15-20 minutes – if it takes longer please say. It would be a good idea to try and get this running before the break.

a. You are going to use the “User Template” mode of the SWISS-MODEL homology modelling server to obtain your model. This mode allows you to define a specific template structure (https://swissmodel.expasy.org/interactive#structure).

b. Provide the sequence of your target protein in the box provided, and then upload the PDB file of the template you downloaded using the green “Add Template File...” button.

c. Fill out the Project Title with a relevant name for the job and provide your email address so you get a notification when the job is complete.

d. Run the modelling job. Keep the tab/page open as the results will appear there automatically when they are ready.

e. While the modelling job is running, move on to the next part.

Visualising Protein Structures

You are now going to visualise some protein structures. First we will start by looking at the template structure. You will be using Pymol to visualise and compare structures - this is a program that you install on your computer, and is widely used and quite powerful.

You will need to install it to use it. A good time to do this would be while your model is being made.

There is an open source version, and a paid version (for which you can get a free educational license OR choose to never activate it... it will not stop you using it).

Here are some instructions for linux: https://pymolwiki.org/index.php/Linux_Install

Here are some instructions for OSX: https://pymolwiki.org/index.php/MAC_Install

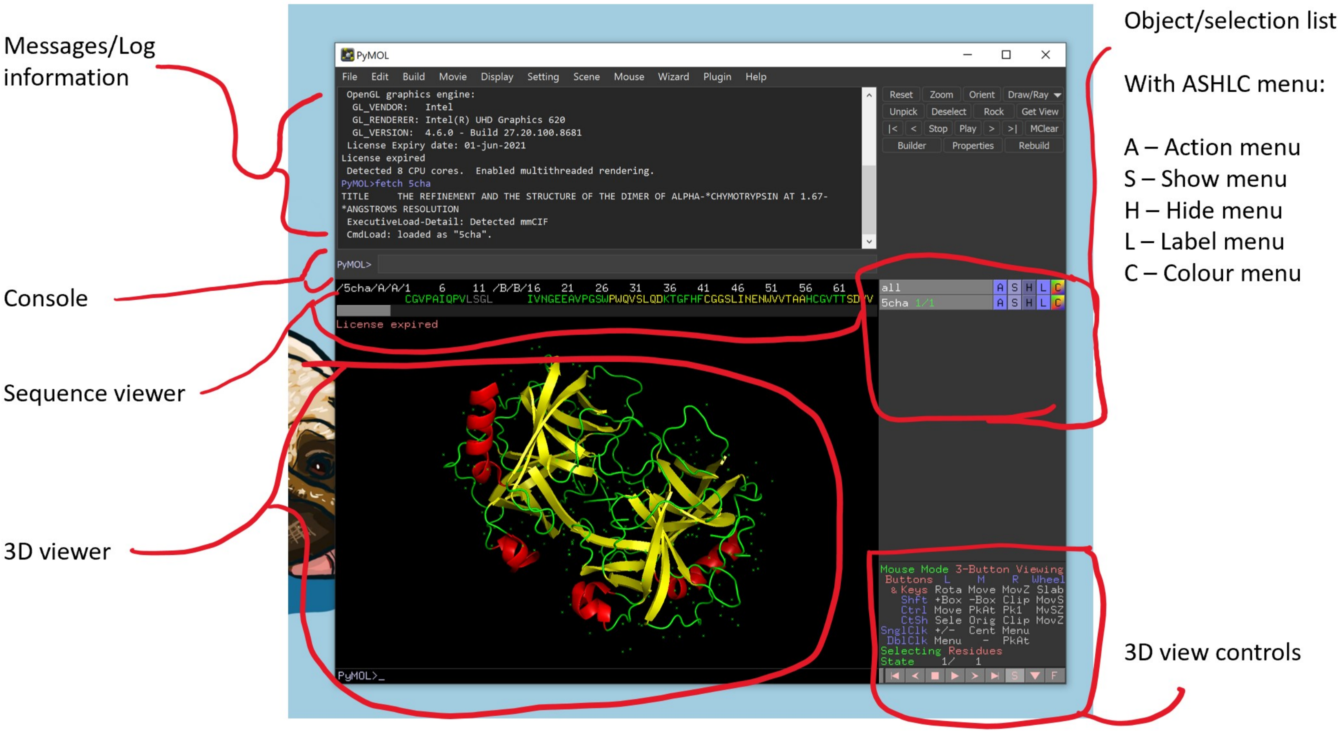
Here are some instructions for Windows: https://pymolwiki.org/index.php/Windows_Install

I'll leave it as an exercise to the reader to get this done.

In the instructions that follow, there may be slight discrepancies between what I say/show and what is shown in your version of Pymol - if that happens ask one of the demonstrators for help.

Step 5

The pymol window:



```
 OpenGL graphics engine:
  GL_VENDOR:   Intel
  GL_RENDERER: Intel(R) UHD Graphics 620
  GL_VERSION:  4.6.0 - Build 27.20.100.8681
 License Expiry date: 01-jun-2021
License expired
 Detected 8 CPU cores.  Enabled multithreaded rendering.
PyMOL>fetch 5cha
TITLE     THE REFINEMENT AND THE STRUCTURE OF THE DIMER OF ALPHA-*CHYMOTRYPSIN AT 1.67-
*ANGSTROMS RESOLUTION
 ExecutiveLoad-Detail: Detected mmCIF
 CmdLoad: loaded as "5cha".
PyMOL>
```

```
/5cha/A/A/1     6     11 /B/B/16   21    26    31    36    41    46    51    56    61
          CGVPAIQPVLSGL      IVNGEEAVPGSWPWQVSLQDKTGFHFCGGSLINENWVVTAAHCGVTTSDVV
License expired
```

```
PyMOL>_
```

```
Mouse Mode 3-Button Viewing
 Buttons  L    M    R  Wheel
  & Keys Rota Move MovZ Slab
    Shft +Box -Box Clip MovS
    Ctrl Move PkAt Pk1  MvSZ
    CtSh Sele Orig Clip MovZ
SnglClk +/-  Cent Menu
 DblClk Menu   -  PkAt
Selecting Residues
State     1/   1
```

While you are waiting for the modelling to finish, use File -> Open... to load in the PDB file you used as a template for the modelling. The mouse controls for the camera are: right click and drag to zoom the camera in and out; left click and drag to rotate the camera around the center of scene; click and hold mouse wheel and move mouse to pan the camera. If you click on an atom, information about it will be shown in the Messages/log info panel.

a. Use the information in the Pymol wiki (https://www.pymolwiki.org) to help you. There is a good Pymol for beginners section here: (https://www.pymolwiki.org/index.php/Practical_Pymol_for_Beginners). There is a good section on selection algebra here: (https://pymolwiki.org/index.php/Selection_Algebra).

- For example, if I wanted to select residues 10-20 in chain A with the name "name_you_choose", I would type the following in the console, then hit enter:

```
sele name_you_choose, (resi 1-20 and chain A)
```

- Selections will appear in the object/selection list, and have the same ASHLC menus as any other object. Any changes you make to the representation e.g. colour or rep style will only apply to the selection you made. If you gave the selection a name, it will be called by that name with parentheses around it, like so: (named selection). If you didn't give it a name - or made the selection by clicking on atoms - it will be called the default: (sele).

b. There will be several ligands present in the structure. Make just the dexamethasone ligand visible.

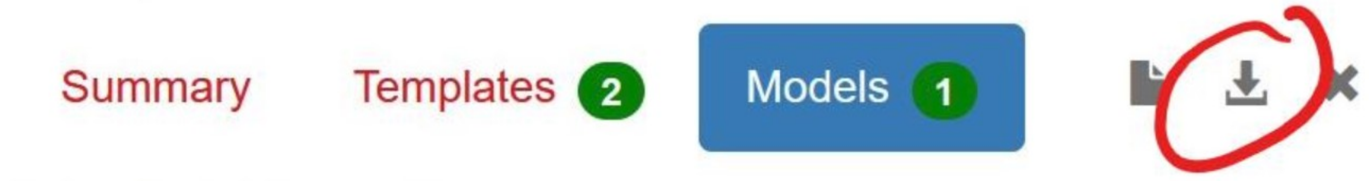
> Tip: Look on the RCSB PDB entry page for the template to find out what the 3 letter name of the dexamethasone residue is. You can either make a selection that includes it (probably using the `resn` selection algebra) or select it in the >sequence viewer, which can be toggled using the "S" in the 3D view controls at the bottom right of the window.

c. Make the sidechains of the protein visible by creating a new licorice representation, and selecting just the sidechain atoms (look at the documentation). If you want, try and identify just those residues close to the drug molecule, and only make those sidechains visible (Tip: click on atoms to get the residue numbers).

d. You can also make the “surface” of the protein visible by adding a surface representation. Make the selection only protein, and change the opacity of the representation so you can see the sidechains and backbone. Opacity of the surface rep can be changed by going to Setting -> Transparency -> Surface and selecting a % option.

Step 6

Once your model is finished, download it from the results page on the SWISS-MODEL website. You can download everything in a single .zip file:



Step 7

Extract all the files, and open the model in the same NGL window as your template structure. You will find the model PDB file inside the extracted folders.

a. Are the two structures aligned already? If so, why is that?

b. Make the sidechains of your model visible as you did before.

c. Make a selection that only includes the substitution you identified in part 1b. You may want to also select the original residue in the template as well.

d. Add a label to this selection using the "L" menu for the selection. You will want to show the residue number and amino acid type. You may have chosen to select the original amino acid and the substituted one. Make sure both are labelled if that is the case.

e. What do you think the effect of this mutation would be upon dexamethasone binding? Why?

f. You can take a screen shot showing the ligand and the substitution on your model using the File... menu. Play around with different representations/options to make the image clear. For example, you could change the representations of the ligands from licorice to spheres.

Step 8

During the course of your research, you also noticed that male patients responded worse to treatment than female patients. You suspect it might be due to hormones.

a. Look back at the function of albumin, and think of a reason why this might be the case.

b. Are there any structures of albumin proteins bound to hormone molecules? Search on the RCSB website to find them. Look at the organism.

c. If you find one, download the structure file and load it into Pymol.

d. You will need to align this new structure to your template and/or model to compare the structures properly.

e. There are a few ways to do this - you can use the align option in the "A" action menu or you can use a console command.

> here (https://pymolwiki.org/index.php/Align) is a link to a Pymol wiki page that describes one of the methods for performing an alignment... there are several.

f. Look in the Message/Log info panel after you perform the alignment. Is there any information there? Do you see something called RMSD? What is that? What value does it say? How would you intepret it?

f. Look at the aligned structures, specifically the binding site for dexamethasone. How similar are the poses of the ligands between the template and this new structure? Why do you think the drug is less effective for male patients?

Step 9

Alphafold2 is the exciting new prediction algorithm - so lets have a look at one of their models. Unfortunately, I've found the method takes a bit longer to run than is realistic given the length of the workshop, so you won't be making the model yourself - instead we will use the new Alphafold protein structure database.

a. Go to the AF database (https://alphafold.ebi.ac.uk/) and find the entry for albumin in humans - you could search using the protein name or the Uniprot accession number.

b. Have a look at the info on the page - the confidence values for the prediction are displayed on a 3D model of the structure. What do you see when you look at this?

c. Download the file in PDB format. You will want to align it to the template structure as you did before so you can compare it. How similar is it? Are there differences? Why do you think there are those differences?

Step 10

If you have time, you can try and predict the different binding affinities of the two molecules for albumin. For this, you can use CSM-LIG (http://biosig.unimelb.edu.au/csm_lig/prediction).

a. You will need to supply the PDB file of the complex, the residue name of the ligand and the structure of the ligand in SMILES format.

b. For example, for dexamethasone the residue name is “DEX” in the PDB file.

c. The SMILES string for dexamethasone is:
“CC1CC2C3CCC4=CC(=O)C=CC4(C3(C(CC2(C1(C(=O)CO)O)C)O)F)C”

d. You can find SMILES strings for chemicals using a website like PubChem (https://pubchem.ncbi.nlm.nih.gov/).

Step 11

Run the analysis for both complexes separately. On the results page you will be given a value for the affinity. Remember that a lower Kd, the dissociation constant, indicates a higher affinity of the ligand for the protein. However, this website gives the result as -log10(Kd), which means higher affinity is indicated by a higher score.

Step 12

Compare the values you get for both complexes. Does this support the idea hormone binding of albumin is a potential factor in the decreased effectiveness of dexamethasone in male patients?