Structural Bioinformatics Workshop

Updated November 2021

Find up to date version @ https://github.com/elliot-drew/structural-bioinf-workshop/blob/main/workshop_ngl.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 workshop uses NGL viewer, an online protein visualisation tool. It should be used if its not possible for you to install Pymol on the computer you are currently working on - NGL is not nearly as advanced as Pymol. BUT if you cant get Pymol working you can still do everything, it just takes a few more steps sometimes.

So, if you can try to use the Pymol version. Both sets of instructions are identical up until Step 5, so you can easily change halfway.

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.

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 the NGL viewer to visualise the proteins. NGL viewer is an online, Javascript library that allows you to view and manipulate biological molecules in 3D. It is more limited than desktop apps like PyMol, Chimera or VMD – but it does a good job for simpler tasks. You can find it here.

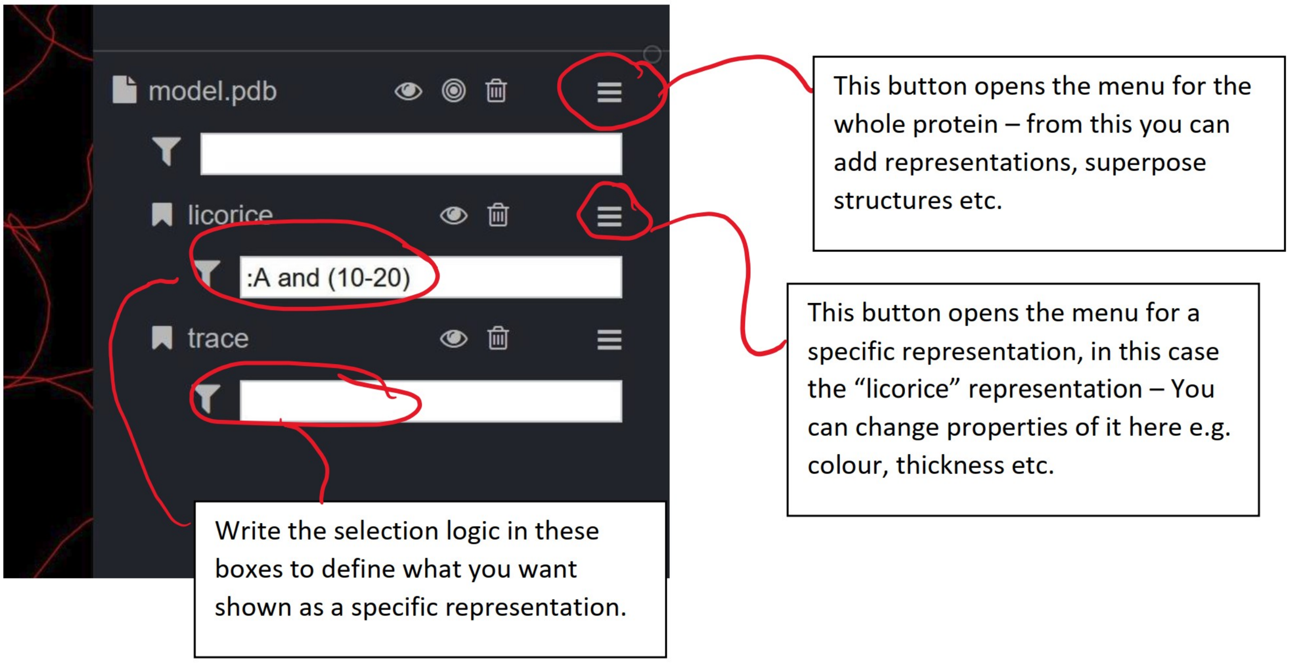
> Attention! Currently NGL viewer is a buggy mess. However, in lieu of installing Pymol or another protein visualisation software, its the next best thing. You will find issues with tasks like structural alignment, fetching structures using PDB codes and probably other aspects of the viewer. I have provided alternate ways to do tasks like alignment that would be usually done in the viewer if you were using Pymol, for example. If you run into an issue, ask a demonstrator for help.

Step 5

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: mouse wheel for zoom; right click and drag to pan the camera; left click and drag to rotate thecamera. If you click on an atom, information about it will be shown in the bottom left corner of the screen (this is useful for identifying things for selection syntax logic).

a. Use the information in the documentation to help you. The key section is the “Selection Language” part. There are lots of examples on this page which will help you when doing your own.

- For example, if I wanted to select residues 10-20 in chain A, I would use “:A and 10-20”
- Selections go in the boxes for each representation, as below:



:A and (10-20)

This button opens the menu for the whole protein – from this you can add representations, superpose structures etc.

This button opens the menu for a specific representation, in this case the “licorice” representation – You can change properties of it here e.g. colour, thickness etc.

Write the selection logic in these boxes to define what you want shown as a specific representation.

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 residue name of the dexamethasone residue is. You can then use this in the selection syntax box to select just the drug.

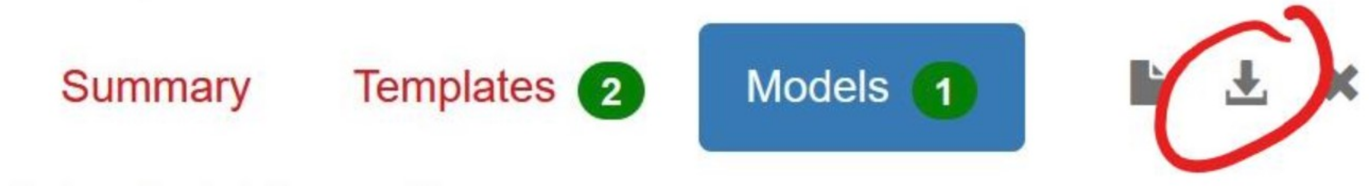
c. Make the sidechains of the protein visible by creating a new licorice representation, and selecting just the sidechain atoms (look at the documentation). If you want, try and identify just those residues close to the drug molecule, and only make those sidechains visible.

> Tip: click on atoms to get the residue numbers - they will appear in the bottom left of the screen, and also appear when you mouse over atoms.

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.

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:



If there are any issues with this - its taking forever, or you get errors, you will find a model I made earlier in the data folder.

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 in the 6c.

c. Add a representation of the type “Label” to the model using the menu (see image above) and use the selection syntax to only label the substitution you identified in part 1b. Tip: if you only want one label to show up on the residue, add “and .CA” to your selection syntax – this will mean only the C-alpha atom of the residue is labelled.

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

e. 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.

Step 9

You will need to align this new structure to your template and/or model to compare the structures properly. Unfortunately the superpose function in NGL has been buggy recently, so we are going to use another webserver to do the structural alignment - TMalign.

a. You should upload two PDB files to the server. First file should be the new structure, the second should be the template - this should mean the new structure is translated in space to align the template coords.

b. Submit the job - don't worry about putting in an email if you don't want, the server is very fast usually.

c. On the results page will be a bunch of information. Near the top will be information about the alignment, including a score called the TM-score and RMSD. These are both ways of describing the quality of the alignment. Feel free to look up what they mean, as they are very useful in structural bioinformatics and appear a lot.

d. Below that there will be an interactive 3D view of the alignment. Below *that* will be the output files. Download the one that says "...full-atom structure of the entire chain with ligands/solvents...". We choose this one because we want the ligands to be included in the alignment. It describes the files as Rasmol scripts - Rasmol is an ancient protein viewer that is still fairly popular... luckily a rasmol script is basically a PDB file with some extra information/gunk.

e. Open up the aligned structure file in NGL viewer. The structure consists of two chains - A will be the first protein you input, B the second. Feel free to colour/change the reps of these as you wish.

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 10

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 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 11

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.

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.

Step 12

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 13

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?