**Day 8 & 9: Protein structure visualization & analysis, modeling, and prediction**

*Everyone must complete these activities individually, but you are welcome to discuss with us, your group, and to help each other out. You will use PyMOL – a protein structure viewer, and a couple of online webservers to predict protein secondary structure and protein 3D structure. It is important to follow the stepwise instructions to not get lost. Further, some theory, mostly short videos, is mixed in with the stepwise instructions, and it is important to not skip these. There are questions for you to answer while completing this 2-day activity. Submit what you have completed by Day 8 by 1pm on Wednesday and the final complete activity by 1pm on Friday.*

**Protein structure visualization and PDB**

1. Open PyMOL.
	1. While PyMOL is opening up, go to your Bioinformatics folder on your computer and make a new folder called structure\_activity.
	2. Once PyMOL is open we can start by adjusting a few of the settings using the menu at the top. First, set your working directory to the structure\_activity folder
	3. Display> Color Space>PyMOL
	4. Display> Quality> Maximum Quality
	5. Settings> Cartoon> Round Helices
	6. Settings> Cartoon> Fancy Helices
	7. Settings> Cartoon> Flat Sheets
	8. Settings> Cartoon> Fancy Sheets

(For options that are not on this list, leave as is, but remember that you can make additional changes here)

* 1. In the command line window (where it says PyMOL>\_) type: fetch 6hyc
	2. Did it load a protein structure into the main PyMOL viewer window?
	3. The “fetch” command will call up the PDB site and load the structure directly into PyMOL. This command will also download the requested PDB file directly into your working directory.
	4. Is the file in your working directory?
	5. Use your mouse buttons to navigate the protein in PyMOL. There can be some slight differences depending on your computer’s settings, but try to be able to perform the following actions:
		1. (Hold down the left mouse button and move the mouse to) **spin the protein around**.
		2. (Hold down the right mouse button and move the mouse to) **zoom in and out**.
		3. (Ctrl/Cmd + Left mouse button allows you to) **move the protein across the screen**.
	6. Object control panel

On the right side of your PyMOL screen, you should see the Object Control Panel (Fig. 1).

This is the object control panel and should include everything in your structure. Each row is a different object. When you select the molecular or parts of the molecule, an additional row for the selected objects will appear. It will be called (sele). You can choose to perform the following actions on all, a pdb structure that you have opened (in this case, there is only one), or on a selection.



Fig. 1. The Object Control Panel: A(action) S(show) H(hide) L(label) C(color)

Try the following:

For all, hide everything

For 6hyc, Show ribbon, what does it look like?

 Hide ribbon

 Show cartoon. This is a common way to display protein structure.

 Color the cartoon by secondary structure (ss).

Export an image, under File.

Insert the image here:

What color are beta strands?

What color are alpha helices?

What color are loops?

Color by Chain. How many chains are in this structure?

(1 = monomer, 2 = dimer, 3 = trimer, 4 = tetramer, etc. This is called the oligomeric state)

Leave PyMOL open while you go to the PDB. You will come back to PyMOL soon.

1. Go to the PDB and learn more about 6hyc. <https://www.rcsb.org/structure/6HYC>

This link will take you to the tab for Structure Summary for this protein structure:



**Fig. 2.** Tabs

Stay in Structure Summary:

a) What is the protein? (The title will give it away, but if it hadn’t you can also check under Macromolecules further down on the same page.)

b) What organism is the protein from?

c) What Method was used to experimentally determine this structure?

d) What is the resolution?

1. Just under the Tabs, on the right side, click to Display Files, select PDB format.

 This will open the PDB file as a text file. Scroll down to where the left column says ATOM and the following lines are shown:

ATOM 1 N THR D 22 14.226 -55.364 81.395 1.00 74.63 N

ATOM 2 CA THR D 22 13.648 -54.078 81.767 1.00 73.48 C

ATOM 3 C THR D 22 13.423 -53.227 80.521 1.00 60.11 C

**This is the format for protein structure used to visualize proteins in e.g. PyMOL. The different columns mean the following:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Atom** | **Atom #** | **Atom type** | **Amino acid** | **Chain** | **Residue number** | **x** | **y** | **z** | **Misc.** | **B-factor** |
| ATOM | 1 | N | THR | D | 22 | 14.226 | -55.364  | 81.395  | 1.0 | 74.63  |
| ATOM | 2 | CA | THR | D | 22 | 13.648 | -54.078  | 81.767  | 1.0 | 73.48 |
| Etc… |  |  |  |  |  |  |  |  |  |  |

**Why do you think there are several lines for each residue?**

1. **Go to the Sequence Tab.** Here the sequence that is in the PDB structure is displayed together with different annotations such as secondary structure and Domains (if not shown you can add it under the Add annotations. It should look like the figure below (Fig. 3).



**Fig. 3.** Protein features

Note that if you mouse over any of the colored shapes, information will appear just above the graphic on the right side.

1. For the one domain that is shown, what is the domain and which sequence range does it cover?
2. Return to PyMOL.
	1. Color the entire protein (shown as cartoon) green.
	2. In Sequence, select all residues that correspond to the domain from the step above. Color the domain from chain A blue.

OR

Type in “sele i. X-Y in c. A” in the command prompt (with X equal to the number of the first residue in the domain and Y equal to the number of the last residue in the domain)

* 1. Under File, save session as 6hyc\_domain.pse

**Homology modeling**

1. You will build a model for the HN\_DM sequence that was in your phylogenetic tree (P17276 is the accession number for HN\_DM (Drosophila melanogaster (fruit fly) phenylalanine hydroxylase)).
2. Are there any protein structures for HN\_DM? Use BLAST at NCBI to find out (you can paste the accession number P17276 in the sequence box in BLAST). Specify the Protein Data Bank (PDB) as the database. Add the information for the best (top) hit and for the hit for pdb id 5jk5 (from PAH\_DD).

|  |  |  |  |
| --- | --- | --- | --- |
|  | From NCBI BLAST |  | From PDB |
|  | Query cover | E-value | Per Ident | PDB ID | Resolution | Method |  |
| Top hit |  |  |  |  |  |  |  |
| First hit from human |  |  |  |  |  |  |  |
| First hit from Dictyostelium |  |  |  | 5jk5 |  |  |  |

1. For the first hit in human, click on the description to show the alignment. Note that you have additional hits from identical sequences:



Note that 6hyc is listed here.

1. Since there is no structure of HN\_DM in the PDB, we are going to build two 3D models. One based on the top hit and one based on 5jk5. Go to SwissModel <https://swissmodel.expasy.org/> make an account and login. Then click on Start Modeling.
2. Your ***target*** sequence is HN\_DM (this is the sequence you are building a model for - sequence accession: P17276). Paste the FASTA sequence for HN\_DM or the accession number in the box for the target sequence. Click Validate.
3. To build a model, we need a ***template*** (a protein structure of a homolog) to base our model upon. To search for templates, click on Search For Templates.
4. While the search for templates is running, watch the following short tutorials

<https://www.youtube.com/watch?v=W2Sy3ZUjE88>

<https://www.youtube.com/watch?v=XDbv1OQ9GWI>

<https://www.youtube.com/watch?v=lfZRUZn2UQc>

1. There are many ways to select a template from the template search. For this activity, you will choose a human PAH tetramer (chain A from pdb id 6hyc: it should say 6hyc.1.A) and a PAH dimer from *Dictyostelium discoideum* (chain A from pdb id 5jk5: it should say 5jk5.1.A). Select only these two and verify that it says Build Models ❷ on the big blue button on the right side. Click on the button.
2. While the models are being built, check out this tutorial (from beginning to 4:40)

<https://www.youtube.com/watch?v=WdzHllIay6U>

What do the Global Quality Estimation, Local Quality Estimation, and Comparison show? (These are mentioned in the YouTube tutorial, but you can also read about them by clicking on the ? next to the Model results header.

Global Quality Estimation:

Local Quality Estimation:

Comparison:

1. When the models are done, fill out the table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model with Template | GMQE | QMEAN | Sequence identity | Residues within 4 $Å $of FE1 (for 5jk5) |
| 5jk5 |  |  |  |  |  |  |
| 6hyc |  |  |  |  |  |  |

1. Use the graph of the local quality estimation and the QMEAN coloring of the alignment (and structure) to determine which Pfam domain (ACT or Biopterin, see below) in the two models has lower quality and which have greater quality (for the model with template 6hyc).

For the target HN\_DM <http://pfam.xfam.org/protein/P17276>



For the template PAH\_HS <http://pfam.xfam.org/protein/P00439>



1. You can download your model and visualize it in PyMOL, like a structure from PDB. But, we are not going to do that for this activity.

**Secondary Structure Prediction**

Sometimes we work with proteins that do not have a structural representative. When that happens, it can be useful to predict the secondary structure elements from the amino acid sequence of the protein.

1. Go to the PSIPRED webserver <http://bioinf.cs.ucl.ac.uk/psipred/>
2. Make sure only the box for PSIPRED is checked:



1. Paste in the sequence that corresponds to the domain region of PAH\_HS from PDB 6hvc from the PyMOL activity. It should be:

>sp|P00439.1|PH4H\_HUMAN:36-114 RecName: Full=Phenylalanine-4-hydroxylase; Short=PAH; AltName: Full=Phe-4-monooxygenase

SLIFSLKEEVGALAKVLRLFEENDVNLTHIESRPSRLKKDEYEFFTHLDKRSLPALTNIIKILRHDIGAT

VHELSRDKK

1. While PSIPRED is running, return to PyMOL and open your saved PyMOL session (or fetch 6hyc).

Color by secondary structure. Show sequence. Screen shot the part that corresponds to the sequence above. Paste it into a Powerpoint slide.

1. When you PSIPRED prediction is completed, save the png for the Sequence Plot. Open the png in the same Powerpoint slide as for the previous step.
2. Compare the prediction from PSIPRED with the experimental structure from 6hyc. How many residues are predicted correctly (helix in both, or coil/loop in both, or strand/sheet in both) and how many are incorrect? Accuracy is sometimes reported as the percentage of all residues that were correctly predicted. What accuracy does this prediction correspond to?
3. Focus on helix.

How many residues from 6hyc that are in a helix are predicted to be in a helix (these are the True Positives (TP)?

How many residues from 6hyc that are in a helix are not predicted to be in a helix (these are the False Negatives (FN)?

How many residues from 6hyc that are not in a helix are predicted to be in a helix (these are the False Positives (FP)?

How many residues from 6hyc that are not in a helix are not predicted to be in a helix (these are the True Negatives (TN)?

Plug your numbers into the equations below to calculate Specificity and Sensitivity for predicting a helix.

****

1. Add a picture of your Powerpoint slide with the PSIPRED prediction and your PyMOL secondary structure here:
2. Submit to Canvas by 1PM on Friday Feb 12.