BioMolecular Model Building 2019/20

Jeremy Harvey & Quico Sabanes

Practical Session 1:
Displaying and Analyzing Biomolecule Structures

In this workshop, we will use graphical display programs, especially PyMol, to investigate a number of structures of biomolecules, revisiting some of the examples taken from the lectures. Important note: This instruction script is designed to be read online rather than printed.

1. PDB File Structure

In this practical and the others, we will often use files in pdb (protein data bank) format. These can be located & downloaded from www.rcsb.org (or its various mirror sites). It can also be useful to use sites such as SCOP to find particular types of structure.

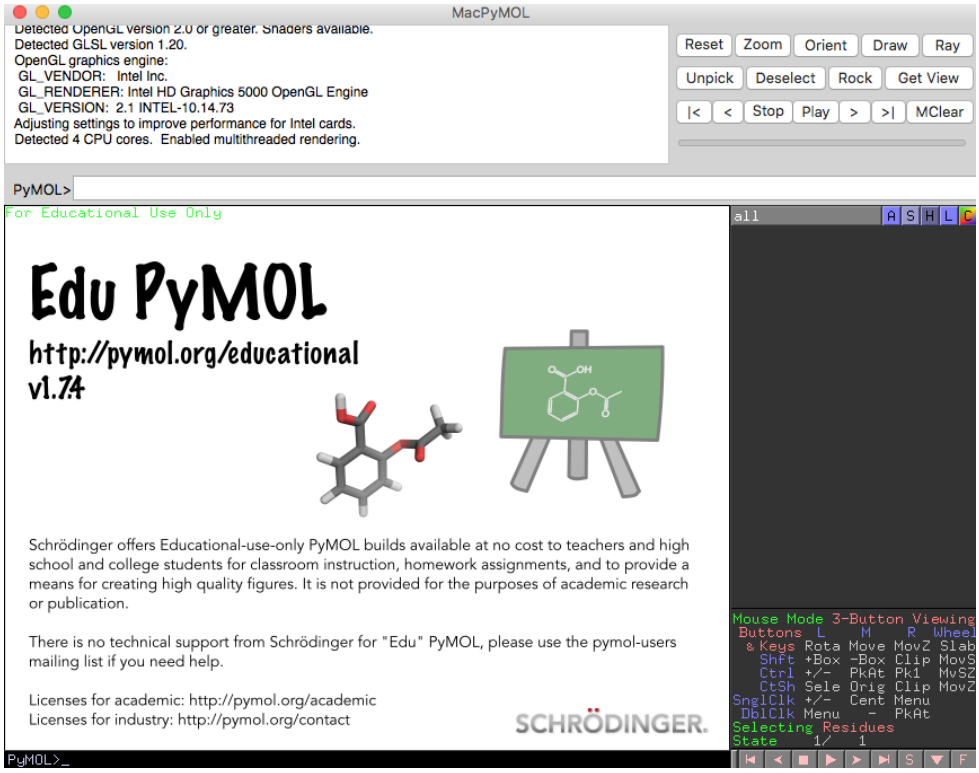
The Protein Data Bank is a website with a databank of structures – but it also defines a commonly used *file structure*. Files in 'pdb format' follow some strict rules, exhaustively described on www.rcsb.org. A short overview can be found on Wikipedia. PDB files are in text, with each line (or 'record') having a particular function, defined by the keyword at the beginning of the line. The most important type of record is that encoding the actual atomic coordinates, with keyword 'ATOM', but other records are useful also, e.g. those that define the biomolecular primary structure *sequence* corresponding to the structure.

First, we will open a crystallographic structure for ubiquitin in Pymol. Find the structure with 4-symbol code 1UBQ on www.rcsb.org, and download the pdb structure file. Open this file with a text viewer (e.g. vim in linux, or WordPad on Windows, or …) and identify HEADER, TITLE, SEQRES, ATOM and HETATM records (see Wikipedia: https://en.wikipedia.org/wiki/Protein_Data_Bank_(file_format)). Can you find the place where it says how the structure was determined? Can you see the resolution to which the structure was measured? Can you see how many amino-acid residues there are, and find the sequence? It can be helpful to compare these to entries in databases such as UniProt (where Ubiquitin has the entry code P0CG48). Which HETATM atoms are present? Can you suggest why they might be present?

2. General Overview of Pymol

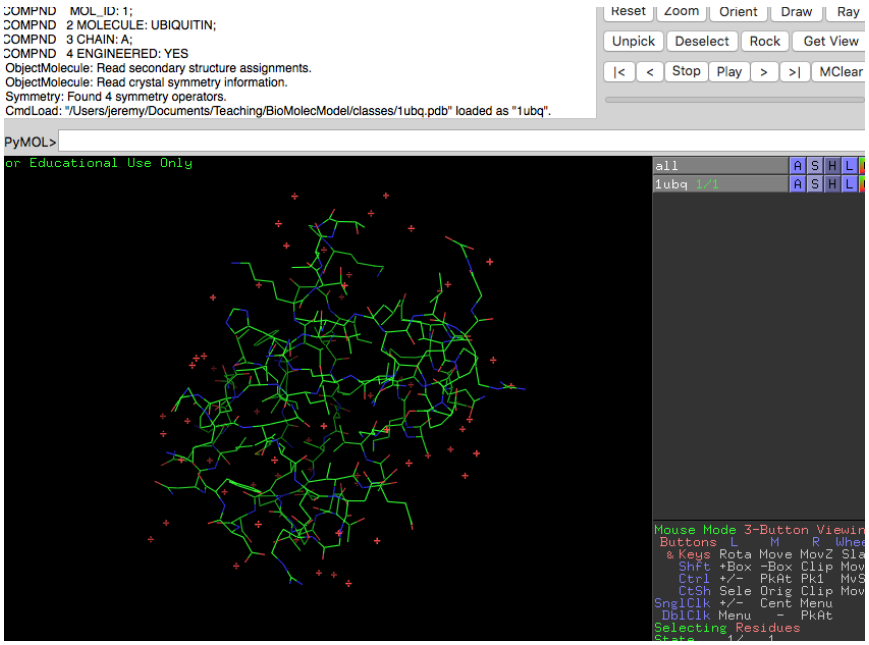
Pymol is available from Toledo for the purposes of these classes. It can also be freely downloaded in its educational version from www.pymol.org (educational version see here). Pymol is a powerful program with many features! These can be discovered through many tutorials available online; we will use only a handful of features here. Pymol is partly a 'point and click' program, but in common with much scientific software, more precise control over the program can be obtained by inputting commands in text form, and you are encouraged to learn some of the associated skills. This allows e.g. '**scripts**' of commands to be saved and executed repeatedly. The top part of the window illustrated below is the place which allows entry of commands in text, at the 'Pymol>' prompt.

When you start Pymol, it will look something like this:



The main region with the logo is the viewer region.

Open the downloaded ubiquitin file in Pymol: choose ‘File/Open’ in the Pymol menu, and open it. You should see something like this:

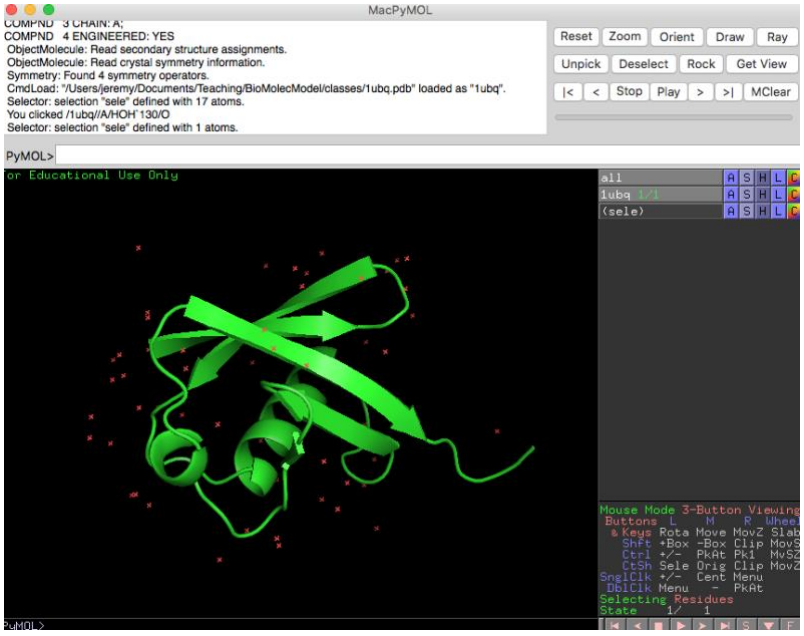


The molecule is shown by default in ‘stick’ representation, with a black background, and with colour-coding by atom.

The mouse allows you to rotate by left-click-and-drag (note if you place the mouse to the right or left of the viewer region, you rotate around the axis perpendicular to the screen), right-clicking allows zooming, middle clicking translating. On a Mac you may have to play around with click+Alt or click+cmd or some such to emulate these effects.

After opening 1UBQ, you see the entry ‘1ubq 1/1’ in the right-hand side, followed by five symbols: A S H L C. This allow certain ‘Actions’, ‘Show’ commands, ‘Hide’ commands, ‘Label’ commands, and ‘Colour’ commands. E.g. ‘S’ then ‘show as/cartoon’ followed by ‘H’ and ‘Hide/lines’ should give something like this, highlighting the helix and sheet secondary

structure. The same effect can be obtained by typing first 'show cartoon' then 'hide lines' at the 'Pymol>' prompt.

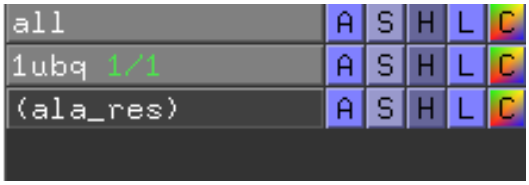


Some preset display modes can also be useful: e.g. click on the 'Action' (A) button, then choose 'preset', and in the subsidiary menu, choose 'publication'

Another useful option for display becomes important if you ever want to **print** images (or display them in PowerPoint): the default *background colour* is black, but this is a poor choice for printing (so that's why I do not recommend printing this script ;-)). From the main menu, choose Display/Background/White to change to white.

3. Selecting Parts of the Structure.

It is often helpful to be able to apply commands only to a specific part of the whole structure. To do this, you need to be able to **select**. This can be done in a variety of ways via point-and-click mouse actions. For example you can use 'Display/Sequence on' from the main menu, which shows the amino-acid sequence with one letter codes, and clicking will select residues from within the sequence. But it is more powerful to do it via text commands. The general format for a select command is: select name=(criterion). 'name' is a name you give to the selection, 'criterion' is a criterion. It can be based on residue name, distance from a residue, residue number in the peptide, or atom name. Type 'select ala_res=(r;ALA)'. This selects all alanine residues. It creates on the right-hand side a new line 'ala_res' for which you can modify the display.



Try these select commands and work out what they do:

'select helix_1=(i;23-34)'
'select backbone=(n;CA,C,O,N)'
'select asp52=(i;52)'
'select round52=(asp52 expand 5)'
'select alanines=(r;ALA)'

To help with selection, it is useful to know that clicking on an atom will tell you which residue it belongs to. Also, typing 'delete name' at the command prompt will delete the selection with the corresponding name.

To practice viewing and selecting, find some of the features listed in the lectures. For this purpose, note that a quick way to open a particular pdb file is to type 'fetch 1UBQ' (or another PDB code) at the prompt – this will automatically download and open the file in one step. From lecture 2: identify the hairpin in thermolysin 3TLN, and the helix-turn-helix Ca-binding site in calmodulin 1EXR, from lecture 3, in the DNA:enzyme complex 5YX2, look at the covalent linkage between cysteine 710 and the DNA, from lecture 6, the bound azole drug in the demethylase, 5FSA, from lecture 7, the active site of trypsin in 2PTC (look for Ser 195).

4. Measuring Properties.

To *measure* a particular distance or angle, you need to enter the 'Measurement' tool or wizard. Choose 'Wizard/Measurement' on the main menu, then e.g. click on two atoms, then 'done' on the right-hand menu to leave 'measurement' mode. The distance will appear next to a dotted line separating the atoms. Note that this line may be hard to see depending on the display type active at that time. You may also first need to zoom in on a particular zone, or even 'select' then modify the display more, in order to make it easy to pick the right atoms. Changing measurement mode on the right-hand 'Measurement' box will allow you to measure angles, or dihedrals.

Identify a *hydrogen bond* in the alpha-helix part of ubiquitin, and measure the N-O distance (remember, the hydrogen atoms are not typically present in XRay structures such as 1UBQ). Check that what you find is consistent with e.g. what you read on this Wiki site: http://proteopedia.org/wiki/index.php/Hydrogen_bonds.

Measuring dihedrals allows you to work out where a particular residue sits on a Ramachandran plot. For that residue, do a dihedral measurement for (in this order), the C(O) of the neighbouring residue, then the N, Cα and C(O) of the residue. This yields φ. For ψ, click on N, Cα, C(O) then N of the neighbouring residue. Do this for one residue in the helix, then for a residue in one of the β-sheet parts of ubiquitin, and compare with a Ramachandran plot.

Another nice example here is to look at the dihydrouracil non-standard base in the tRNA structure 1VTQ. What is the C-C-C-N dihedral angle in the dihydrouracil, and in a normal uracil in the same structure? Note: you may need to look at the pdb file in a text viewer first to work out where the dihydrouracil is.

Another example: the tryptophan synthase repressor complex 1TRO. Load this structure, and identify some key hydrogen bonds between the protein and the DNA, as well as between the bound tryptophan and the repressor.

5. Generating a Ramachandran Plot for a Whole Structure

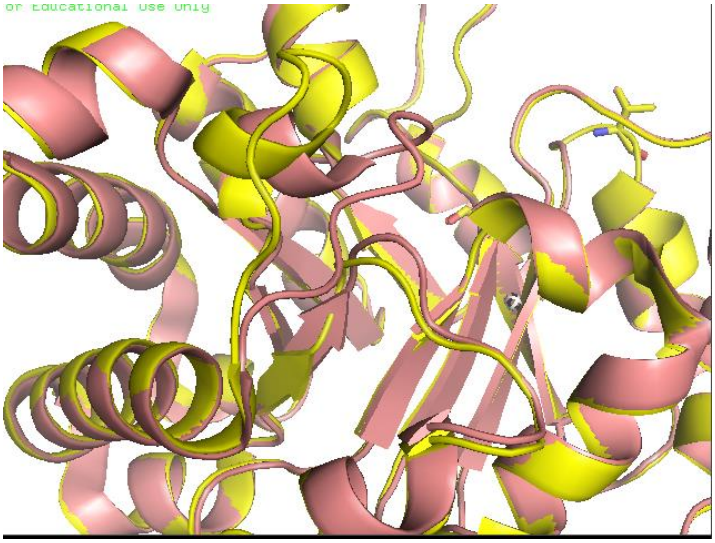
Pymol is a molecular viewer with many useful features, but it is not the only one. Each program has its own quirks, unique features, missing features, and strengths. In this course, we will also use two other viewers: **Chimera** and **VMD** (Visual Molecular Dynamics). Chimera can be freely downloaded from the Chimera website, through this link: https://www.cgl.ucsf.edu/chimera/download.html. There is an introductory tutorial somewhat similar to the present one that can be found at https://www.cgl.ucsf.edu/Outreach/Tutorials/GettingStarted.html. Carry out the first few basic steps in this tutorial so as to have a rough understanding of Chimera.

Then use Chimera to open the 1UBQ structure (To clear any previous work:"File/Close Session", then "File/Fetch by ID" and enter 1ubq in the box then hit enter. Like Pymol, Chimera has *many* features, and opens several windows by default. Here we will use it to view a Ramachandran plot for 1UBQ. Choose "Tools/General Controls/Model Panel". Then select 'Ramachandran Plot'. The plot shows each residue at the place it occupies on the plot, with

little blue dots for each residue. You can see that there are residues in three main areas, and one outlier. Identify the nature of each of the three main areas. Clicking on a blue dot selects the corresponding residue on the main viewer window. Hovering over that residue then gives you its identity.

6. Alignment of structures

Pymol includes a very powerful *structure alignment* tool that is also quite easy to use (there are also powerful alignment and superposition tools in Chimera). Here we can use it to reproduce the picture shown on the top of page 7 in the pdf version of lecture 4 (titled 'Lid in TIM': the open and shut forms of the 'lid' in TIM. To do this, we need to first download the two crystal structures, 1TRD and 1TPD. Open both in Pymol. Both structures contain two copies of TIM, and it is helpful to display only one of them. To do this *select* chain B (by typing "select chain_b=(chain B)" at the Pymol prompt), and then *hide* it, in both structures. Then type align 1trd,1tpd at the Pymol> prompt. Using 'Show/Cartoon' will generate something like this:



The align function is useful because it allows alignment not only of structures with the same sequence, but also of homologous sequences (the program automatically chooses the alignment of sequences to use as a basis for structural alignment). Take for example the related globin proteins cytoglobin and myoglobin. These have fair sequence overlap (29% identical residues), but are rather different in detail. Load 1A6G (carbonmonoxymyoglobin) and 2DC3 (cytoglobin). You'll notice that the pdb structure 2DC3 contains two molecules – use select to show just chain A. Then use 'align' to superimpose. Note the similarities and differences. View the heme group in both cases, and see the difference in iron coordination.

7. Viewing the Electron Density

As mentioned in section 5, Pymol is not the only viewer… In fact, the PDB website has its own viewers built in to the site, and indeed some of the figures in the course were generated from that. When viewing the page for a particular PDB entry, the left-hand side of the page shows various 3-D viewers that can be used. Here we will use a useful feature of the *European* PDB site, http://www.ebi.ac.uk/pdbe/ or http://pdbe.org. This website contains essentially the same *data* as the standard rcsb webpage, but includes different visualization tools. One of these provides a very convenient mechanism to look at an important intermediate step between the actual experimental data on which the structure is based and the structure itself: the electron density. The experimental data itself is a set of observed diffraction spots from X-Ray diffraction, in terms of angles and intensities. This gets converted into the electron density distribution within the unit cell of the crystal, and this density is well worth looking at.

Here we will consider the case of an enzyme from the HIV virus, in complex with an anti-AIDS drug, nevirapine. The PDB entry code is 3QIP. Search for this in the pdbe (not RCSB!)

page, then click on the NVP ligand in 'ligands and environment'. This will show one view of the electron density in the vicinity of the NVP ligand. A better view is obtained from the 3QIP entry page by clicking on '3D visualization'. After finding the NVP ligand and clicking on one of its atoms, you can view the density as an isosurface, and superimposed on it, a *difference* plot – this shows the difference between the actual (derived from the diffraction pattern) and modeled (derived from the proposed molecular structure) density. Negative regions are shown in red, and denote less actually experimentally present density than implied by the modeled structure, suggesting that some atoms may be absent in the real crystal compared to the model. Positive regions (extra density not accounted for by the modeled structure) are shown in green. In the present case, there is relatively little red or green: the modeled structure fits the observed X-Ray diffraction pattern very well.

This sort of difference plot is very convenient for checking the accuracy of a structure. The 3D visualizer implemented in the pdbe webpage can be used for published structures, while Pymol can generate the same kind of plot provided one has a density map as well as the standard coordinate file – as would be the case in experimental groups generating in-house data. While a structure is being 'fit', there may of course be quite big differences between model and actual densities. For published structures, though, there can also be differences. These are due to a number of factors. One is simply experimental noise and resolution: the actual density never *exactly* matches the modeled one, and the differences are in fact fairly big. Another source of error is misinterpreted experimental data, hence incorrect structures. Many such cases are known, with small or indeed quite large errors in the published structures, leading in some cases to structures being retracted from the pdb some time after publication.