Structural Bioinformatics Exercise – PyMOL

In this Exercise we will use the visualization program PyMOL and also have a brief look at the PDB, the home of all experimental protein structures.

If you have used PyMOL before or even done this exercise previously (and you did it **very** thoroughly), you can go quickly through this exercise or even jump directly to for example item 31. Or ask Jon what to do…

1. Open the internet browser and go to http://www.uniprot.org/uniprot/O15527. Here you can find information on the protein "8-oxoguanine DNA glycosylase", human OGG1. In the isoform α-OGG1 (Isoform 1A) there are 345 residues. The sequence is given by:

```
>sp|O15527|OGG1_HUMAN N-glycosylase/DNA lyase OS=Homo sapiens
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLA
DQVWTLTQTEEQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHF
QEVAQKFQGVRLLRQDPIECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHG
FPSLQALAGPEVEAHLRKLGLGYRARYVSASARAILEEQGGLAWLQQLRESSYEEAHKAL
CILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAKGPSPQTNKELG
NFFRSLWGPYAGWAQAVLFSADLRQSRHAQEPPAKRRKGSKGPEG
```

   In the "Structure" section on the page you will find links to 3D structure databases (in the "Links" column). What is the PDB identifier for the structure with the best resolution? What is the resolution for structure 1EBM? Are any of the structures for the full-length protein? Are there any visible hydrogen atoms in any of the structures? Why not?

Best resolution is for 2XHI. It is 1.55 Å. 1EBM has 2.10 Å resolution. The only full-length structures are 1KO9 and 2XHI. With a resolution ~1.6 Å or worse one cannot reliably predict the positions of the H atoms from the electron density.

2. Open a browser window and go to the PDB at http://www.rcsb.org.

Here you can search for protein structures determined by either X-ray crystallography or NMR. There are various ways to search the database, the most common being by protein name, acronym, organism, protein class or author name.

3. Type in **1EBM** in the search field and press enter to search for this structure. You get only one hit, so you are sent directly to the "Structure Summary" for this structure. Which macromolecules does this structure contain? How many chains are there? What is chain A? Is it mutated? (Now, in the fall of 2018, the database claims that the protein in not mutated. This is a change since the spring, and the last 15 years, and is wrong! It is mutated, Lys249 (K on blue background on previous page) to Gln. This is clearly stated in the original article. This is good example of an error in partially automatically generated databases). Do you have any suggestions for why the researchers behind this study have mutated a single residue in the middle of the sequence? If not, you will have the chance to guess later in this exercise. Is it a full-length protein? What is missing? Do you have any suggestions for why it is not full-length? (Hint: Structural disordered regions are hard to crystallize)

1EBM contains a DNA duplex and a protein, OGG1. We usually consider a DNA duplex to be one macromolecule even if the two strands are not covalently connected, thus 2 macromolecules in total. There are 3 chains, A (OGG1 protein), C and D (two strands of DNA). Chain A is the protein OGG1. It is a K249Q mutant which means Lys249 (K on blue background on previous page) has been mutated to Gln. It is not a full-length protein. Residues 1 to 11 and 326 to 345 have been removed (See the "Sequence" tab, and compare with the bold face residues in the sequence on the previous page). One likely reason why the researchers behind this study removed these two segments is that they are structurally disordered. Possibly they were not able to crystallize the full-length protein. The researchers expressed, in *E. coli*, a DNA construct comprising residues 12 to 325 of human K249Q OGG1, but put a His-tag at the N-terminus. They produced a lot of protein and used the His-tag to purify the protein. Note that residues 9 to 11 are RRM in the wild-type OGG1 (previous page), but GSE in the 1EBM structure. The residues 9 to 11 listed in the PDB file and on the PDB "Sequence" web page are parts of the linker between the His tag and the rest of the construct, and thus not really mutations of OGG1.

4. Click on "Download Files" (up at the right hand side) and download the "PDB Format" file. Save it, for example on the Desktop, as 1EBM.pdb. Take a look at the pdb-file, for example in WordPad or nano. It shows among other things the experimentally determined positions (atomic coordinates) of all the atoms in the

model (including many water molecules). Who are the authors of this work? Is this an X-ray crystallography or NMR study? What is the resolution of the structure? What do you find in the fields "REMARK 200"? Did they use a synchrotron for this work? What do you find in the fields "REMARK 465" and "REMARK 470"? Can you explain the missing residues/atoms? What do you find on the "SEQRES" lines?

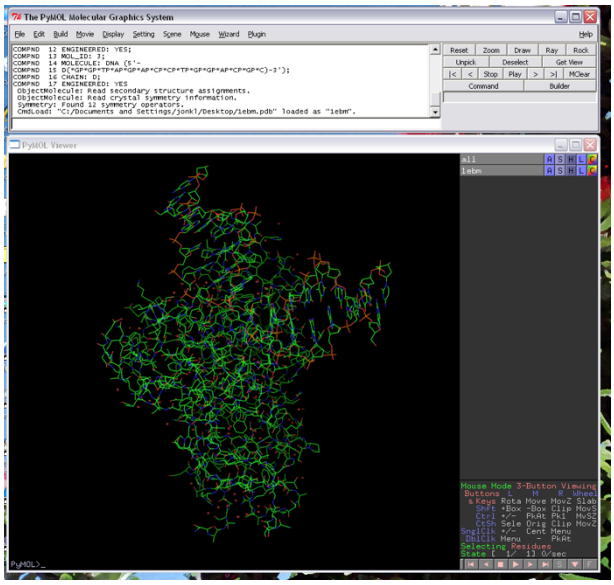
The authors are "S.D.BRUNER,D.P.NORMAN, and G.L.VERDINE". This is an X-ray crystallography study. The resolution is 2.10 Å. In "REMARK 200" we find experimental details. The structure was solved using a synchrotron. "REMARK 465" and "REMARK 470" contains missing residues and missing atoms (in visible residues), respectively, in the crystal structure. These are missing since there was no visible electron density for these residues/atoms. The reason for this is that they are structurally disordered/flexible/floppy. "SEQRES" contains the sequence.

5. On the "Structure Summary" page for 1EBM, left hand side, there is a picture of 1EBM and links to various "viewers". Click on the links at "3D View: Structure" to open the viewers. Default viewer is NGL. Play a little with these tools and try at least NGL and JSmol. What is the green ball you see with the JSmol and NGL viewers? You might need to go back to the "Structure Summary" page and look at "Small Molecules/Ligands" to get an idea. In JSmol, rotate the structure (by left-clicking and dragging) and zoom in and out (using the mouse wheel). Right click to get a menu and for example choose "Style", "Scheme" and "CPK Spacefill" to get the protein visualized in a different way.

The green ball is a Ca^{2+} ion.

6. Open PyMOL (you will find the shortcut in the Start Menu or in the All Programs listing, under Windows, at least...). You can make the "Viewer" window bigger by pulling at the corner.

7. Open the 1ebm.pdb file in PyMOL (**File → Open →** select file on desktop).



8. Turn on sequence display (**Display → Sequence**).

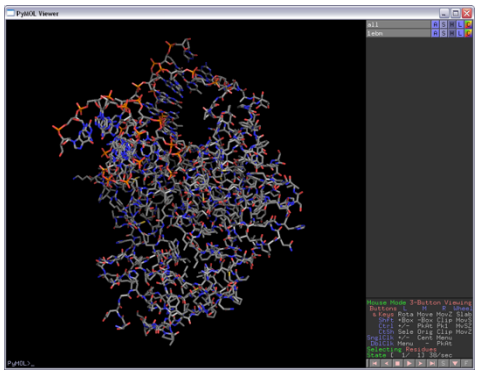
9. Play around with the protein structure using the mouse buttons (**left** = rotate, **middle press** = move, **middle scroll** = clipping, **middle click** = centring and **right** = zoom). If you manage to "get lost in protein space" you can always centre the molecule again by "middle-clicking" with the mouse at any given amino acid in the sequence display on the top. You can also "right-click" on the background and choose "zoom (vis)".

10. The isolated red crosses/stars are single water molecules that are fixed in the crystal near the protein surface. We don't need them right now, so press "**H**" for **hide** and go down to the **waters** item on the pop-up list (this is just next to "1ebm" in the upper right corner of the viewer).

11. Actually, let's turn off the whole protein. Select "**H**" → **everything**.

12. Select "**S**" for show, then select **cartoon.** Now you only see the trace of the main chain of the protein. You can also see the DNA duplex bound to OGG1. Select a nice colour by pressing "**C**" and then pick one from the list according to your liking. Try one of the "**spectrum**" → "**rainbow**" colouring schemes as well.

13. Use "**H**" → **everything** again and then "**S**" and choose (hiding everything again when you want to)
    i. "**lines**". This is "wireframes". Colour "**gray 60**" and then "**by element**", first option. Now you see carbon atoms as gray, oxygens as red and nitrogens as blue.
    ii. "**sticks**".
    iii. "**ribbon**". This is often called a "Cα-trace" and is a line connecting all the Cα atoms.
    iv. "**spheres"**. This is CPK space-filling spheres. Choose a nice colouring scheme.
    v. "**surface**". You get the full surface of the protein complex. Choose a nice colouring scheme.

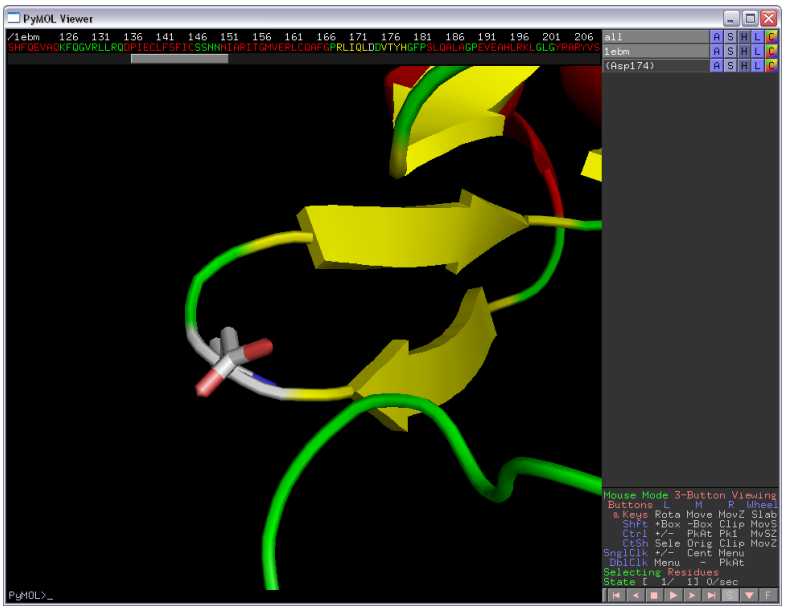


14. Hide everything and show as cartoon. Colour according to secondary structure "**by ss**". How many β-sheets do you find? How many β-strands does each sheet contain? Are there more α-helices or β-sheets in OGG1? Scroll the sequence in the PyMOL Viewer window (at the top) and locate the residue Asp174. Click on this residue in the sequence (*i.e.* click on the "D"). The "D" is now highlighted, it is "selected". Numbering in the sequence is a bit awkward... Residue 171 is Ile and residue 176 is Val, got it? You will also find little pink squares on the structure corresponding to this selection. Finally, you find the selection as "(sele)" in a separate row on the right hand side of the Viewer window. For this selection choose "**A**" → "**rename selection**" and type the name "**Asp174**" before you press enter. If you click on "(Asp174)" the little pink squares disappear, but you can still use this selection later.

There are 2 β-sheets. They contain 2 and 5 β-strands, respectively. There are more residues in α-helices than in β-sheets.

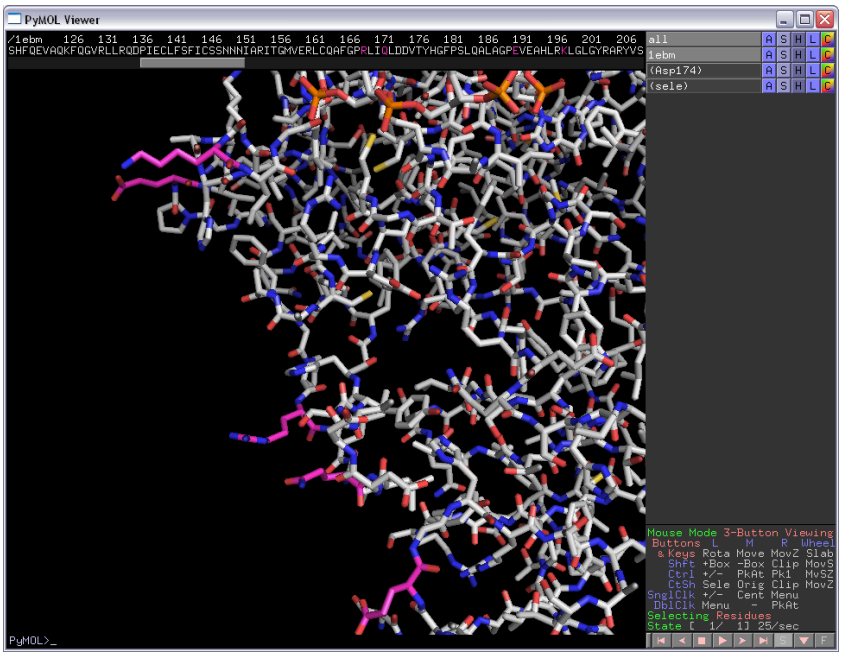
15. Show the whole structure as "cartoon" and Asp174 as "sticks". Zoom in on Asp174 and choose a colouring that makes it easy to see. What kind of secondary structure element is Asp174 located in?

Asp174 is in a loop (connecting two β-strands).



16. Hide "**everything**" (for 1ebm) and then show as "**sticks**". Colour "gray 60" and then "by element" (first option in the list). You see some residues sticking out into the solvent from the surface of the protein. Would you describe the side chains of these residues as hydrophilic or hydrophobic?

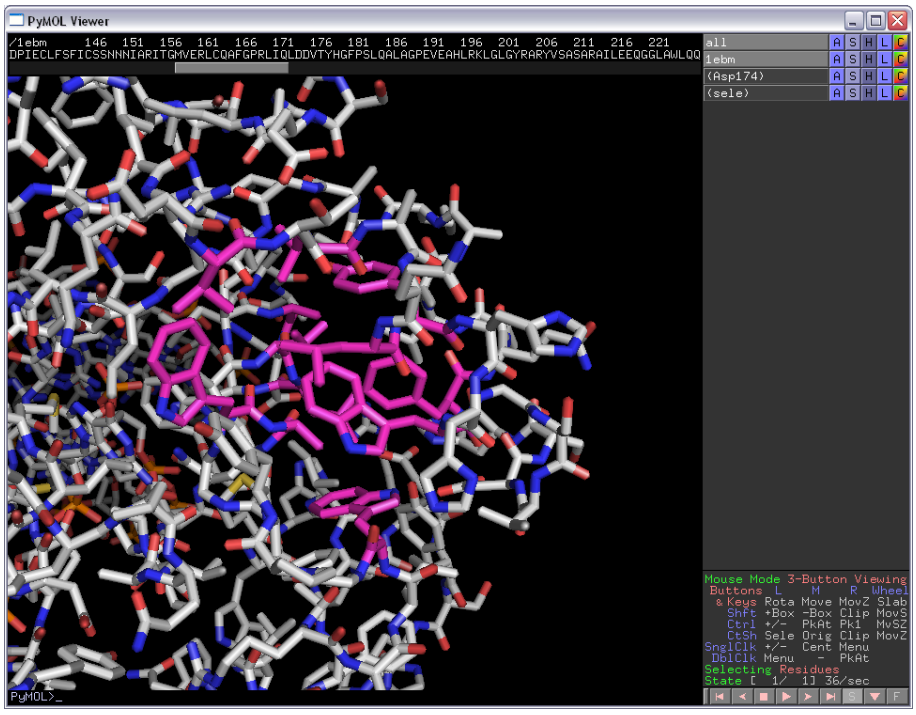
The residues sticking out into the solvent, for example those I have colored pink below, are charged (Lys, Arg & Glu) or polar (Gln). All these are hydrophilic. ***If you want to generate the picture below, you have to color the residues yourself.***



17. Choose one of these residues by clicking on it. You have made a new selection "(sele)". For "(sele)", do "**L**" → "**residues**" to get this residue labelled. Which residue is it? Which amino acid is it?

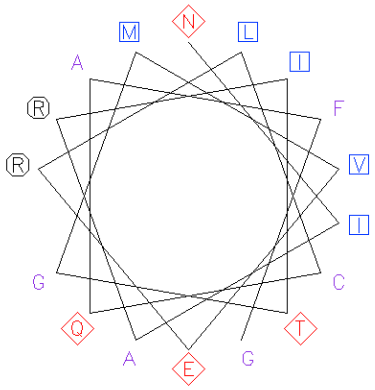
18. Can you find any areas within the structure that are rich in hydrophobic residues? What forces are stabilizing the structure in these regions?

The pink residues (colored manually by me) in the figure below are forming a network of hydrophobic interactions. Hydrophobic interaction forces are stabilizing this region.



19. The segment "NIARITGMVERLCQAFG" is a part of OGG1, residues 151-167. Go to http://emboss.bioinformatics.nl/cgi-bin/emboss/pepwheel and generate a helical wheel plot with this sequence (Use default settings). Does the helical wheel plot suggest/hint that this is an α-helix or a β-sheet in this enzyme? Why? Can you find the sequence in the structure? Is it an α-helix or a β-sheet? Which part of the segment is facing the solvent? What do we call something that is hydrophobic on one side and hydrophilic on the other?

The helical wheel plot gives,

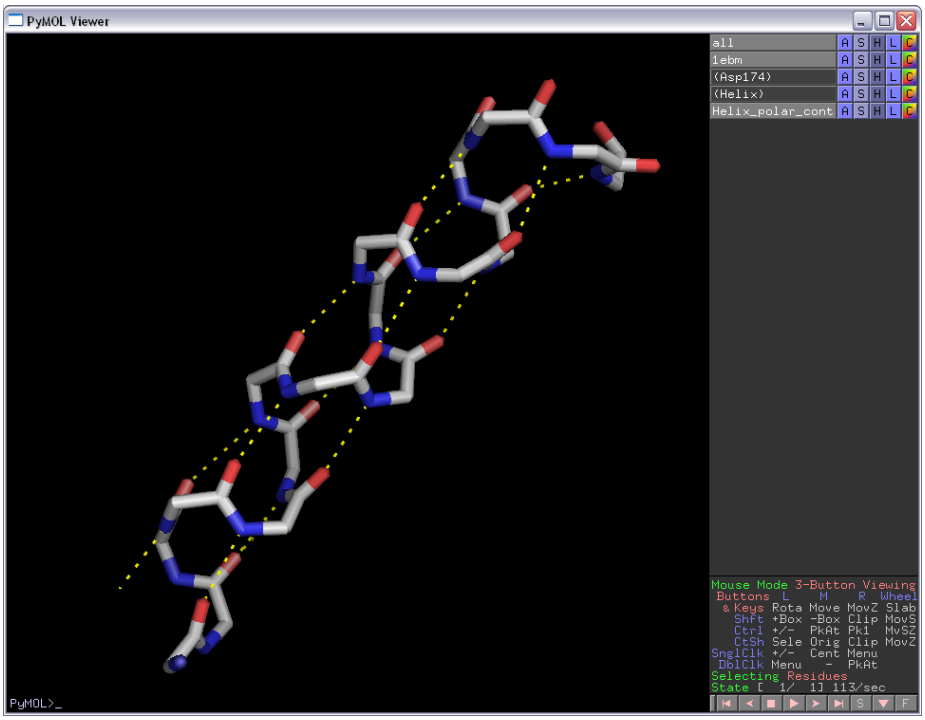


If this is an α-helix, we get a "cylinder" that is hydrophobic on one side (the "blue" half) and hydrophilic on the other ("red" and Arg residues). This is typical for α-helices on the surface of proteins. If this is a β-sheet sitting on the surface of a protein, we might expect a rough pattern of alternating hydrophobic and hydrophilic residues. However, the residue pairs 2 & 3, 8 & 9, and 15 & 16 are hydrophobic. There is thus not a strong support for this being a β-sheet. The *pattern* of hydrophobic/hydrophilic residues suggests that most likely this is an α-helix. Pattern recognition is an important discipline within informatics, and it can be used, for example, to predict secondary structure form the protein primary structure.

Residues 151-167 in OGG1 is indeed an α-helix. This is an amphipathic structure with the hydrophobic side (the "blue" half) facing the core of the protein and the other side facing the solvent.

20. Select the residues 151-167. Display only this alpha helix and hide everything else. Show as "**sticks**" and colour "**by element**". Approximately how many turns are there in this helix? Which atoms are involved in H-bonds stabilising the α-helix? There are two residues involved in each of these H-bonds. How many residues is it between them in the sequence? Can you locate the bonds which define the rotations phi and psi? Where are the Cα atoms? It may be easier to answer the questions if you first do "**Hide**" → "**Side chain**" to see only the main chain atoms. You can let PyMOL make an attempt on localizing H-bonds by doing "**Actions**" → "**find**" → "**polar contacts**" → "**within selection**". As you see, there are too many H-bonds. You may fix this by typing (in the Viewer window at the command line/lower left corner after "PyMOL>" the following: "set h_bond_max_angle, 30". If you repeat the "find polar contacts" procedure you should now get a slightly better result.

The figure below shows the hydrogen bonds stabilizing the α-helix. The H-bonds are between the amide O-atom residue *n* and the backbone N-atom of residue *n*+4. There are approximately 4 turns.



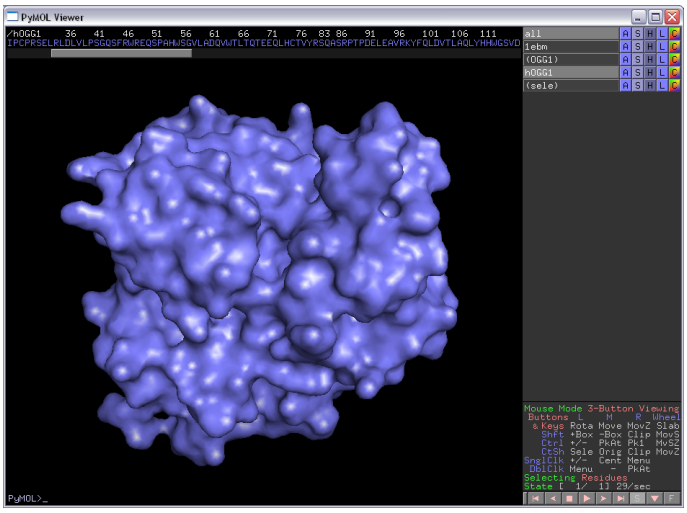
21. Show the whole protein again. Far down on the right-hand side you see "selecting residues". That means that if you click on the structure, you can choose single residues. Click on the word "selecting" until you get "selecting chains". Click on a part of the protein in the main window and confirm by looking at the sequence bar that you have chosen the full "chain A" and nothing more.

22. Create a new object by typing "create OGG1, sele". Click on "1ebm" to make this object inactive. You now only see your new object "OGG1" in the main window. Locate the N-terminus and the C-terminus. Are their any gaps in the protein backbone? Which residues are missing? Why are they missing?

There is one gap, between residues 79 and 83, *i.e.* residues 80-82 are missing. Very likely, they were not visible in the electron density due to structural disorder. In the most recent version of PyMOL, gaps like this are no longer "missing" completely in cartoon rendering. Instead, a dotted loop is shown.

23. For the object OGG1, select "**H**" for hide, and then select **everything**. The molecule disappears.

24. For the same object, select "**S**" for show, and then select **surface.** A nice protein surface is calculated and displayed.

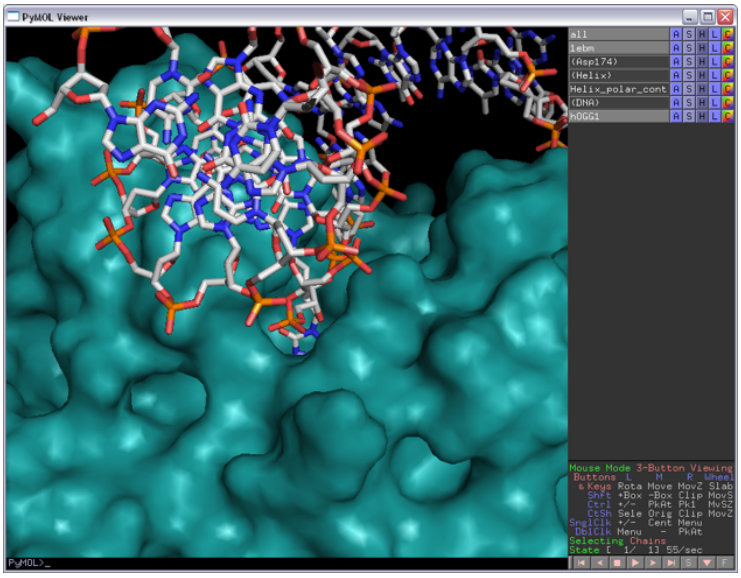


25. Active sites are often located in grooves or holes in the surface. Where do you think the active site is located?

26. This protein is a DNA glycosylase that scans DNA and recognises and removes 8-oxoguanine (8-oxoG), which is guanine (G) with an oxygen atom bound at position 8 on the DNA base. The protein belongs to the helix-hairpin-helix superfamily of DNA glycosylases. Make 1EBM active again and hide everything in this object. Create a new selection that contains the DNA chains. For this selection choose "**A**" → "**rename selection**" and type the new name "**DNA**" before you press enter. Use "**S**" for show, and then select "**sticks**" for "DNA".

27. DNA containing 8-oxoG is the substrate for this enzyme. Can you now locate the active site?

The active site is located in the region where the DNA base is flipped out of the DNA-duplex and inserted into a cavity on the OGG1 surface. See figure below.

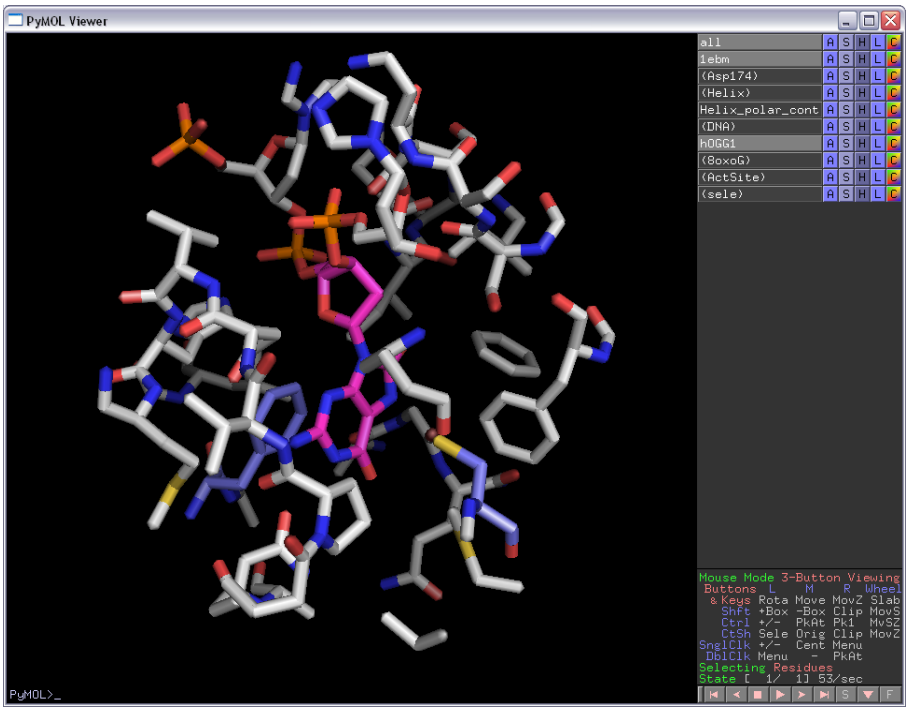


28. DNA glycosylases flip the damaged base into the binding pocket. This leaves a vacant space in the DNA duplex. OGG1 inserts a residue into this vacant slot. Which residue is that? You might want to change the rendering of the protein, show as "**sticks**" for example, in order to find this residue.

The intercalating residue is Asn149.

29. Make a new selection containing only the single residue 8-oxoG (*i.e.* the "8OG" in the sequence!). Rename the selection "8oxoG". The command "select ActSite, 8oxoG around 8" will make a new selection called "ActSite". It contains all atoms within 8 Å of "8oxoG". Show both 8oxoG and ActSite as "**sticks**" and hide everything else. One Phe and one Cys residue are "stacking" against 8-oxoG in the active site pocket. Which ones?

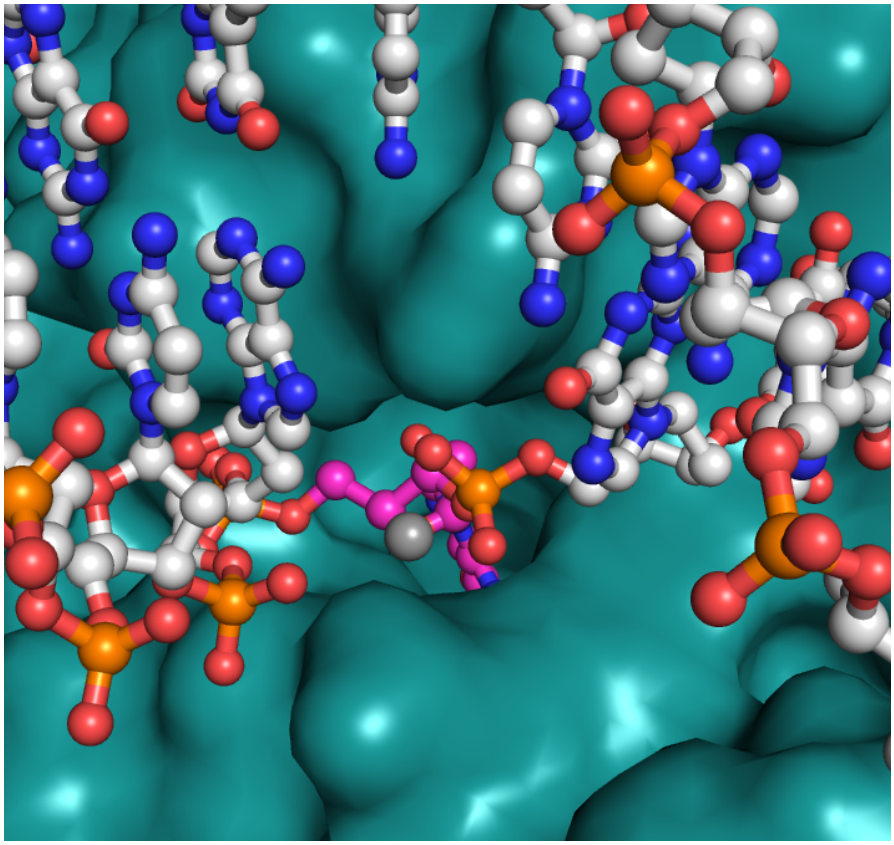
The two residues stacking against 8-oxoG (pink) are Cys253 and Phe319 (blue) show in the figure below.



30. This structure contains the mutant OGG1 K249Q. Can you find residue 249 in the structure? Do you have any suggestions for why the researchers behind this study have mutated this residue?

Wild-type OGG1 (*i.e.* non-mutated OGG1) is an enzyme that grabs DNA, flips 8-oxoG out of the duplex, excises 8-oxoG using Lys249 as a nucleophile, lets go of the DNA and the excised base, and finally starts over again on the catalytic cycle. Wild-type OGG1 would have "eaten up" all the DNA substrates and would not have created any stable complexes. Lys249 was mutated to make the enzyme inactive. The K249Q mutant grabs DNA, flips the base, but then forms a stable complex that can be crystallized and studied.

31. Make a nice illustration of 8-oxoG containing DNA bound to OGG1 by generating an appropriate "scene" and do ray-tracing by typing "ray". You might try
    i. Type "bg_color white"
    ii. Type "set ray_trace_fog, 0" and "set ray_shadow, 0" to remove shadows and a "foggy" look of the ray-traced image
    iii. Typing "ray 2000" gives a picture that is 2000 pixels wide. You may save it by typing for example "save *directory*/OGG1.png".



32. Take a look at the PyMOL wiki page: http://www.pymolwiki.org. Here you find, in addition to a lot of other useful stuff, small scripts and plugins that people have written in order to make PyMOL even more useful. Click on the "Script Library" link. In the "Structural Biology" category, click on the "AngleBetweenHelices" link. What will this script do?

This script adds several commands to PyMOL, making it easy to calculate the angles between alpha-helices and between beta-strands.

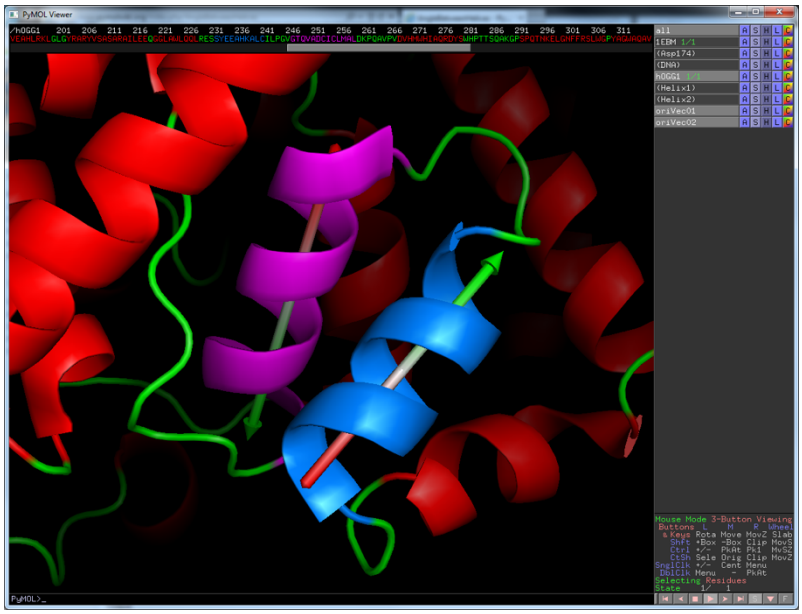
33. Download the script (at the "Download" link in the yellow box) and store it on your computer in a directory somewhere. Make sure you know the full path of that directory. Under Microsoft Windows, for example, store it in "C:\temp\". If you store the file anglebetweenhelices.py somewhere else, use that path name below.

    In PyMOL, type "run C:\temp\anglebetweenhelices.py" to run the script (use the correct path name!). Apparently not much happens, but you now have some more

commands that you may use, for example "angle_between_helices sele1, sele2". Hide the 1EBM object in PyMOL by clicking on 1EBM in the upper right corner. Make sure the OGG1 object is active. Show this object as cartoon and colour by "ss" (secondary structure). Select residues 232 to 241 (this is an alpha-helix) and rename the selection Helix1. Do the same with residues 247 to 259 and name this Helix2. Give Helix 1 and 2 nice new colours.

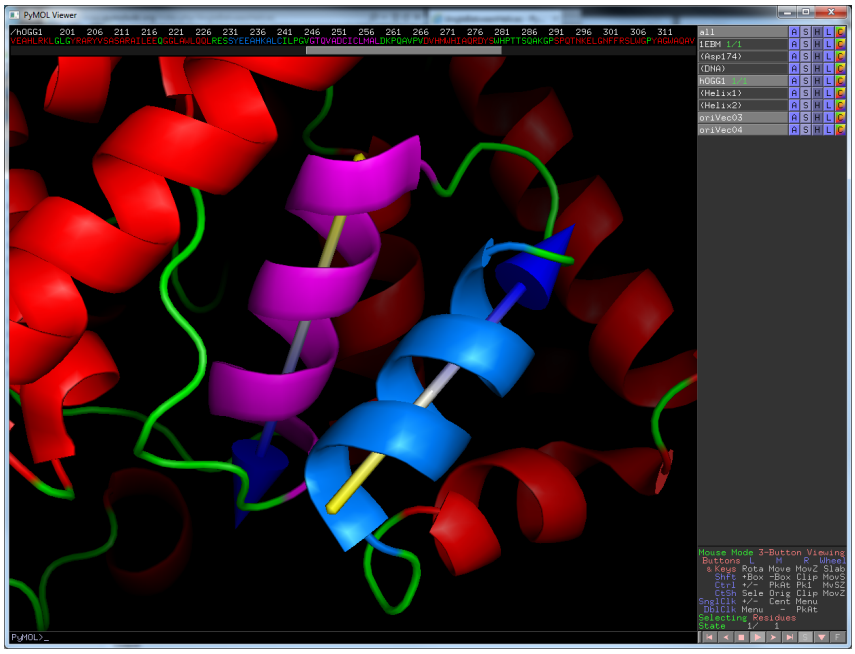
Now type the command "angle_between_helices Helix1, Helix2". What happens?

This creates arrows through the centers of the helices Helix1 and Helix2. They have a red to green color gradient. The angle between the two arrows is calculated to be 129.94°.



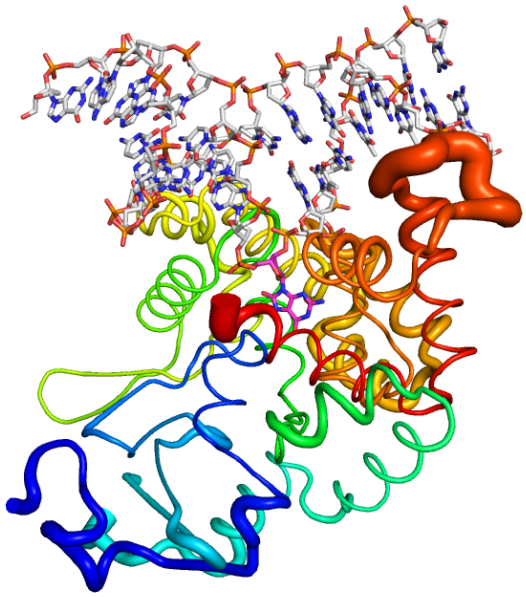
Open the anglebetweenhelices.py file in a text editor. As you might see, this script is written in Python. Actually, PyMOL itself is written in Python. Now let us modify this script a bit. In the script you find the statement "coneend = cpv.add(end, cpv.scale(direction, 4.0*radius))". Change the "4.0" to "15.0". Further down you find "radius * 1.75,". Change "1.75" to "5.0". Finally, a bit further up you find "symmetric=False, color='green', color2='red'):". Change "green" to "blue" and "red" to "yellow". Save the file and in PyMOL run it again by typing "run C:\temp\anglebetweenhelices.py". Now do "angle_between_helices Helix1, Helix2". What happens this time?

The arrow heads are much bigger and the color gradient goes from yellow to blue. You might have to delete the two objects oriVec01 and oriVec02 to see it really well (See below).

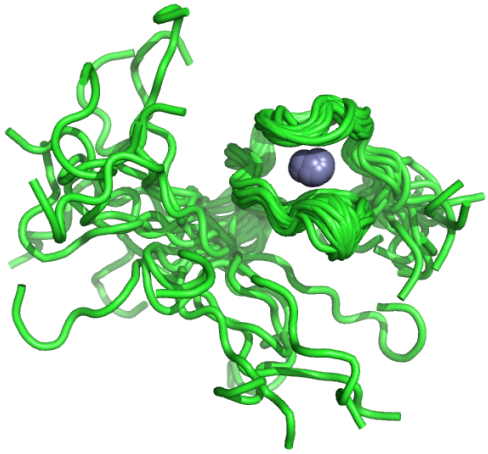


Additional things you might do if you have time:

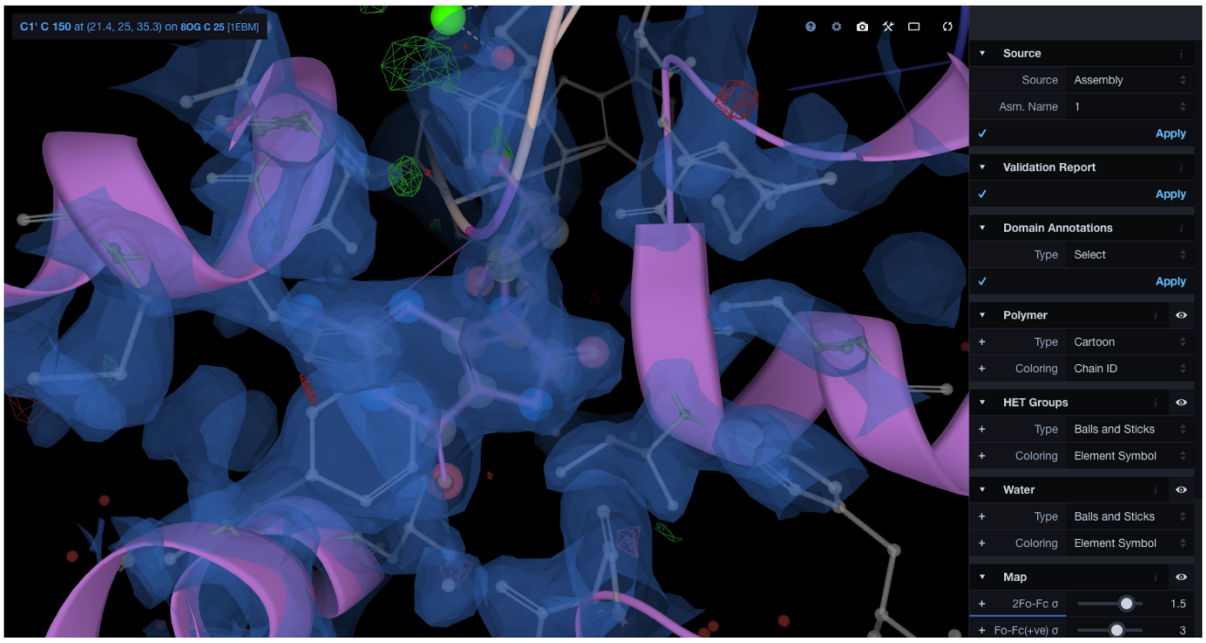
- Experiment with anglebetweenhelices.py and make more changes to the script.
- Try a cartoon rendering in PyMOL, but type "cartoon putty". You will get a tube following the trace of the protein, but with the diameter of the tube showing the B-factor. Some regions are thin (little thermal motion) while others are fat (a lot of thermal motion).



- Have a look at an NMR structure ensemble in PyMOL, for example 1N0Z. Run through the various states. Which parts of this protein segment is structurally disordered and which parts have a regular 3D structure?



- Look at electron density that was used to generate 1EBM at the PDBe (http://www.ebi.ac.uk/pdbe). On the page for 1EBM, click on "3D Visualisation" under "Quick Links" to open the LiteMOL viewer. Check the density for residues 78 to 84. Is anything missing? Also look at the electron density around 8-oxoG.



- PyMOL is open source and user sponsored software. There is a lot of stuff you can do in PyMOL. You might start here:
  - http://pymolwiki.org/images/7/77/PymolRef.pdf
  - http://pymolwiki.org, explore on your own, for example
    - Look at the tutorials
    - Learn selections properly: http://pymolwiki.org/index.php/Selection_Algebra
    - Check out the various plugins or the "gallery" and learn how to make the various illustrations (http://pymolwiki.org/index.php/Gallery)
    - Make a nice illustration of your favourite structure from the PDB?
    - Learn how to make this picture: http://pymolwiki.org/index.php/File:QuteMolLike.png
  - http://www.pymol.org