

Session 1 Aims

The main aims of these exercises are as follows:

- To become familiar with the **Protein Data Bank**, the online repository for macromolecular structures. At the end of this exercise you should be able to search for structures in the Protein Data Bank, view images of the structures online, and download the atomic coordinates of a structure.
 - To become familiar with the molecular graphics program **MOLMOL**. At the end of this exercise you should be able to read a set of atomic coordinates into MOLMOL, and then display and manipulate the structure. In the case of NMR-derived structures, you will learn on Day 2 how to manipulate an *ensemble* of structures.
 - To become familiar with various tools within MOLMOL for manipulating and analyzing structures. At the end of this exercise you should be able to: (i) produce a schematic of a protein structure showing the location of α -helices and β -strands; (ii) calculate and display the molecular surface of a protein; (iii) map the electrostatic potential onto this molecular surface; (iv) produce and analyze a Ramachandran plot of the structure.
-

Exercises

1. MAKING WITHDRAWALS FROM THE PROTEIN DATA BANK

The online repository of macromolecular structures is known as the Protein Data Bank or PDB. It is located at <http://www.rcsb.org/pdb/>. As of 5/7/02 there were over 18,000 structures deposited in the PDB. Almost all journals require that the atomic coordinates of a structure be submitted to the PDB prior to publication. Most structural biologists make the coordinates publicly available from the PDB as soon as the manuscript describing the structure is published. In this exercise, you will learn how to find a structure in the PDB, view images of the structure online, and then download the atomic coordinates for more detailed analysis on a local computer.

Commands to be typed are indicated by **bold green Helvetica** type. Right-facing arrows are used to indicate menu relationships and are shown in **bold red Helvetica** type; for example, “**choose X→Y→Z**” means go to pulldown menu X, choose command Y, and then choose Z from the submenu that is then displayed. Questions that you should try to answer are indicated by **bold magenta Helvetica** type.

A. FINDING THE STRUCTURE

- All of the computers in the computational facility use the Linux operating system (a version of Unix). Some of you may be unfamiliar with this computer environment, so ask for help if you run into trouble. Dr Maciejewski, Manager of the NMR Structural Biology Facility, will be available to provide help during the practical class. First open up a **shell** (a window where you can type commands) by clicking the shell icon in the lower left hand corner of the screen. Then type **netscape &** to run the Netscape browser.
- When Netscape starts, point the browser to the PDB at <http://www.rcsb.org/pdb/>. You will see that there is a search engine available on the first page. This simple search engine is useful if you know exactly what you are looking for. Let’s say, for example, you have just read a paper describing an interesting structure that you would like to investigate in more detail. The paper should list the PDB coordinate filename for the structure, which usually consists of the number 1 or 2 followed by three alphanumeric characters. Let’s use the example **1vtx**. Enter **1vtx** into the search area, tick the box “**query by PDB id only**”, then click the “**Find a structure**” button. This will take you directly to the 1vtx entry. The menu on the left-hand side of the entry lists various options. We will discuss these options in more detail later.
- Go back to the search engine on the front page. Let’s say you only know the type of compound you are looking for, but not its name or its PDB coordinate filename. 1vtx is the coordinate filename for an atracotoxin from Australian funnel-web spiders. To get an idea of how many atracotoxin structures are available, type **atracotoxin** in the search area, tick the “**match exact word**” box, and then click the “**Find a structure**” button. This brings up a list of five atracotoxin structures, one of which is 1vtx. To get to the 1vtx entry, click on the red **EXPLORE** link. This takes you to the same place we reached by typing **1vtx** directly into the search menu.

- Let's try a more advanced search. Go back to the front page and click on the **SearchLite** link. This slightly more advanced search engine provides additional ways of locating the protein you are interested in. Let's say you know the name of the group that solved the structure (**king**) but you can't remember the name of the compound, just the broad category that it falls into (**toxin**). Try typing **author:king and compound:toxin** in the search area. Click on the **Search** box and you will get a list of five toxins, including **1vtx** again.
- It is possible to gradually refine your search as you go along. For example, go to the SearchLite page and type **DNA repair** in the search area. This brings up 66 entries. Let's say you know that the structure you are interested in was solved by the **Mullen** lab. Choose "**Refine Your Query**" from the pulldown menu, which takes you to a new page. Make sure the **AND** button is highlighted in the upper logic menu and then type **author:mullen** in the search area and click the **Search** button. This limits the search to three structures of DNA repair proteins solved by the Mullen lab.
- It should now be clear to you that the PDB search engines provide various ways to locate the protein you are interested in. You can do more specific searches using the advanced **SearchFields** search engine, but most of the time the simpler search engines will suffice.

B. VIEWING THE STRUCTURE ONLINE

- Go to the PDB front page, type **cofilin** in the search area and tick the "**match exact word**" box. This should bring up a list of seven structures. Click on the **EXPLORE** link for entry **1COF**, the 2.3 Å resolution crystal structure of yeast cofilin. This should take you to the entry for **1COF**.
- Cofilin is a member of the widely conserved family of actin depolymerizing factors that modulate actin polymerization into filaments. Click on the **View Structure** link in the left-hand menu to take you to a page with various options for viewing the structure online. First, there are several pre-prepared static JPEG images of the cofilin structure. Click on **Ribbons (250x250)** to view a 250x250 JPEG image of **1COF** in which β -strands are shown as arrows and helices as ribbons. Hit the back button on your browser and click on **Cylinders**

(250x250) to view a 250x250 JPEG image of **1COF** in which β -strands are shown as arrows and helices as solid cylinders.

- The next simplest option for viewing the structure online is using Java. Just click on the **QuickPDB** button to get an interactive display of the protein backbone. You can rotate and translate the structure with the left and right mouse buttons, respectively, and you can zoom with the middle mouse button. Choose **Secondary Structure** in the left-hand pulldown menu to view β -strands (blue), helices (red), and non-regular structure (yellow). Choose **Exposure** to view solvent exposure of each residue with the color ranging from red (highly solvent-exposed) to blue (mostly buried). If you click on a residue it will be highlighted (both on the structure and in the amino acid sequence shown at the top of the window) with whatever color you choose in the **Color** menu. You can also use the second pulldown menu to view the location of residues according to their properties (e.g., take a look at the distribution of aromatic versus charged residues in **1COF**).

Click the **Stereo** button to get a stereoview of the structure. Although it takes some practice to view stereo images properly, it is worth the effort as it is the best way to get a real “feel” for the three-dimensional shape of the molecule. Stare at the two images until you see a third image appear between the two stereo pairs. When this third molecule comes into focus it will be in 3D! Don’t waste too much time trying this with **QuickPDB** as the thin lines used to represent the protein backbone makes stereo viewing somewhat difficult. We will return to this later in the **MOLMOL** demonstration.

- The nicest way to view the structure online is using one of the **VRML** (Virtual Reality Modeling Language) options. However, this requires installation of a browser plug-in which is not yet available on the Linux workstations. However, instructions are available in the **Download Help** menu if you want to install the plug-in for your PC or Mac (which I strongly recommend).

C. DOWNLOADING THE COORDINATE FILE

- Click on **Download/Display File** from the menu on the left of the screen. To view the PDB coordinate file, click on **HTML** in the **PDB file-format** column in the first section entitled **Display the Structure File**. This shows the full PDB coordinate file for **1COF**. Note that it is

just a text file. The header gives information about the protein, its source, the authors of the deposition, the primary reference for the structure, and information about the technique and instrumentation used to solve the structure. If you scroll down you will see the protein sequence on the lines marked **SEQRES**. Listed below the sequence are the atomic coordinates that define the structure. Note that there are no coordinates given for residues 1-5 and 141-143 at the extreme N- and C-termini, respectively. This is presumably because these residues are highly flexible, and hence do not produce a visible diffraction pattern in X-ray crystallographic studies.

- For each residue you will see the individual atoms listed (e.g., N, CA, C, O, CB, CG1, and CG2 for Valine 6). The first three columns after the atom name, residue name, and residue number are the x,y,z coordinates for that atom in an arbitrary Cartesian coordinate frame. **Note that there are no hydrogen atom coordinates listed for 1COF – why is this the case?** The second last column is the crystallographic B factor, which is a rough measure of protein dynamics (a higher B factor indicates more flexibility). The B factors tend to increase at the N- and C-termini as we might expect. **But why are there some internal residues with high B factors (e.g., residues 31–38)? Where are these residues located (go back to QuickPDB and display the B factor distribution if necessary)?**
- Go back to the **Download/Display File** page. Go to the **PDB format** column in the table in the **Download the Structure File** section and click on the cross in the first row (no compression). Make sure you download the file into the MEDZ325 directory. NMR-derived coordinate files usually contain an ensemble of structures (typically 20) and hence are much larger. When downloading these files you might want to choose a compressed file format such as **GNU zipped**. Just download the gzipped file (it will have a **.gz** extension) then type **gunzip filename.gz** at the command line to decompress the file once it has downloaded.

2. VIEWING MOLECULAR STRUCTURES USING MOLMOL

MOLMOL is a molecular graphics program for displaying, analyzing, and manipulating the three-dimensional structure of biological macromolecules. You can read more about the program at <http://www.mol.biol.ethz.ch/wuthrich/software/molmol/>. The program is freely available and it runs on PCs, Unix/Linux workstations, and Macs running OS X. In this exercise, you will learn how to read PDB coordinate files using MOLMOL and how to do basic manipulations, including

rotation/translation of the structure and creation of stereo images.

A. READING THE COORDINATE FILE INTO MOLMOL

- Start the program by typing **molmol &** within a shell. After a few seconds you will see the main (large) molecular display window as well as two smaller windows labeled **Log Window** and **MOLMOL**. Dismiss the **Log Window** by clicking OK and dismiss the **MOLMOL** window with the sliders by clicking twice in quick succession in the upper left corner. If there are any molecules displayed then go to the **Edit** pulldown menu and chose **Init All**. Then choose **yes** in response to the **Delete Everything?** question in the pop-up window. You are now ready to load the molecule you wish to view.
- To load a PDB coordinate file, choose **ReadMol→PDB** from the pulldown menu. This will bring up a file list. Choose a filename and click OK or double click on the filename to read the file.
- If at any point you wish to save the current session, choose **File→Write Dump** from the pulldown menu and save the session as a filename with a **.mol** extension. Make sure you save the files into the MEDZ325 directory. To restore a previously saved session (which will have a **.mol** suffix) choose **File→Read Dump** from the pulldown menu. This will bring up a file list. Choose a filename and click OK or double click on the filename to restore the session.

B. MOVING AND RESIZING MOLECULES

- To **rotate** molecule(s), click and hold down left mouse button and then move mouse.
- To **translate** molecule(s), click and hold down middle mouse button and then move mouse.
- To **resize** molecule(s), click and hold down left and middle mouse buttons then move mouse upwards (to increase size) or downwards (to decrease size).
- For quick zooming of the molecule(s), click on the right mouse button until a dialog menu appears then choose either **Zoom in** or **Zoom out**.

C. SELECTING MOLECULES AND ATOMS

MAKING SELECTIONS

- MOLMOL has a complicated syntax for selecting atoms, residues, and molecules. It is very powerful but also very unforgiving. One misplaced punctuation mark will lead to gibberish as far as the program is concerned. The general syntax is as follows:

`#molecule number:residue range@atom names`

For example, the selection command `#3:25-35@CA` selects the CA atoms of residues 25-35 in molecule #3. Note that there are no spaces and the placement of the #, :, and @ symbols are critical.

- *Selecting molecules.* Individual molecules can be selected in several ways:
 - (i) Click on the **Mol** button. This brings up a menu from which individual molecules can be selected for manipulation (**select** button), display (**display** button), or on-screen movement (**move** button).
 - (ii) Click on the **Selection** button. This brings up a Selection Dialog box. In the top row labeled **Mol**, type `#n` where n is the number of the molecule you wish to select. For example, typing `#3,5` will select molecules number 3 and 5.

Since you read in only one coordinate file (1COF), you only have one molecule to display. Thus, you do not need to worry about molecule selection for the moment.

- *Selecting residues.* To select specific residues, click on the **Selection** button to bring up the Selection Dialog box. In the row labeled **Res** type `#A:X-Y` where A denotes the molecule you want to select and X and Y denote the residue range you desire. For example, typing `#1-3:11-15` would select residues 11-15 in molecules 1-3. Typing `:11-15` would select residues 11-15 in *all* molecules.
- *Selecting atoms.* To select individual atoms, click on the **Selection** button to bring up the Selection Dialog box. In the row labelled **Atoms**, select the desired atoms using the syntax outlined above. For example, typing `#3,5:15-19:CA` will select only the CA atoms of

residues 15-19 in molecules 3 and 5.

- *Selection shortcuts.* Sometimes you want to make simple selections such as all of the protein backbone atoms. There are shortcuts for these: you can select **all** atoms, just the **bb** (backbone) atoms (C, N, CA), just the **heavy** (i.e., non-hydrogen) atoms, or just the **sidechain** atoms by clicking the corresponding boxes on the right hand menu of the main window.

There are some shortcut scripts that can be used in the Selection Dialog box to select certain types of atoms. The terms **bb**, **sc**, and **heavysc** can be used to select backbone, sidechain, and heavy sidechain atoms, respectively. The syntax used is **#molecule number:residue number & shortcut**. For example, typing **#3,5:15-19 & heavysc** will select only the heavy sidechain atoms of residues 15-19 in molecules 3 and 5. You can also use the three-letter amino acid code to specific particular types of residues. For example, typing **#3,5:PHE & heavysc** will select only the heavy sidechain atoms of phenylalanine residues in molecules 3 and 5. Note that the amino acid name is uppercase.

MAKING USE OF THE SELECTION

- You normally make a selection because you either want to make something invisible or because you want to display it in a particular way. If you want to make the selected atoms disappear, simply make the selection and then click the **invisible** box in the right hand menu of the main window. If you wish to draw bonds between the selected atoms, click on the **Line** button to connect these atoms by lines, the **Neon** button to connect the atoms by rendered tubes, or **Ball/stick** to connect the atoms in ball-and-stick mode.
- You can also highlight your selection in a user-specified color. After making your selection and choosing how you want it displayed (line, neon, or ball/stick), click on the color box in right hand menu of the main window. This causes a Color Dialog box to appear. Choose a color from the list of pre-defined colors or make your own color by specifying the RGB values (each number must be in the range 0-1). Then click on **bond** or **atom** in the color dialog box to display the selection in your chosen color.

LEARNING BY EXAMPLE

- The best way to learn the MOLMOL syntax is with an example. Currently, MOLMOL should be displaying all atoms found in the coordinate file **1COF**. Let's turn everything off to begin with. Click on **all** to select all atoms and then **invisible** to make everything invisible. Your molecular graphics window should now be blank.

Now click on **backbone** to select just the backbone atoms and then click on **line** to draw the bonds connecting the atoms. It is much easier to see the overall fold of the protein now that the sidechains are not displayed. By rotating the molecule you should be able to make out where the α -helices and β -sheets are located.

- Let's take a look at where certain types of residues are located in the structure. Click **Selection** and then type **:PHE & heavysc** in the row labeled **Bond** in the Selection Dialog box. Then click on **color** to bring up the Color Dialog box, chose a distinctive color like red, and then click on **bond**. Now click on **neon** and the sidechains of all of the Phe residues should be displayed as red tubes. **What is the topological distribution of the Phe sidechains—are they generally buried or are they found on the surface of the protein?** Now let's look at the topological distribution of Lys residues. Type **:LYS+ & heavysc** in the row labeled **Bond** in the Selection Dialog box. Then click on **color** to bring up the Color Dialog box, chose a different color to that you used to display the Phe sidechains (e.g., green), and then click on **bond**. Now click on **neon** and the sidechains of all of the Lys residues should be displayed as green tubes. **What is the topological distribution of the Lys sidechains—are they generally buried or are they found on the surface of the protein? Why do Lys and Phe have such different distributions? How does this relate to protein folding?**

3. ANALYZING STRUCTURES USING MOLMOL

- We saw in the previous exercise that it was much easier to make out the 3D fold of the protein when the sidechains were made invisible. Comparing protein folds is made even easier when the proteins are displayed as cartoons in which the helices are represented by ribbons or cylinders and the β -strands are represented by arrows. This type of schematic was pioneered by Dr Jane Richardson who originally drew these schematics by hand. Today these schematics can be generated automatically by most molecular graphics programs,

including MOLMOL.

Let's make a schematic of the yeast cofilin structure. Click **all** and then **invisible** to turn off any displayed atoms/bonds. Now choose **File**→**Macro**→**Execute Standard** and choose **ribbon.mac** from the pop-up menu. This is a macro that automatically calculates the protein secondary structure and then displays a schematic with β -strands drawn as cyan arrows and helices as red/yellow ribbons. It is now easy to see that the yeast cofilin structure consists of a central β -sheet (composed of six β -strands) surrounded by four α -helices. Note how some of the β -strands are highly twisted. This type of fold is referred to as an **open twisted α/β structure**. It is called "open" because the β -sheet has two open ends, and is surrounded on both sides by α -helices, as opposed to **α/β barrel structures** in which the β -strands form a closed barrel that is surrounded on one side only by α -helices.

- You can get a better feel for the overall shape and contour of the protein by viewing it in stereo. With the 1COF schematic displayed, choose **Options**→**Stereo**→**Side_by_side** from the pulldown menu. This command produces a so-called **relaxed, parallel, or wall-eyed** stereo image (as opposed to a **cross-eyed** stereo image). This is the format used for Magic Eye stereograms. Check out <http://www.polarimage.fi/stereo/stereo.htm> to practice on some cool (non-scientific) parallel stereo images. To practice on some simple protein structure images, try our lab website at <http://psel.uchc.edu/structures.html>.

Reduce the MOLMOL window size to about 15 cm and make sure the two stereo images are only separated by ~1 cm. Parallel viewing works best when the centers of the two images are separated by about 2.5 inches. Look directly at the center of the two images (do not attempt to go cross-eyed), relax your eyes, and focus at a point apparently beyond the two images until you see a third image appear between the stereo pair. Once this image comes into focus it will appear in glorious 3D. You can now rotate/translate the molecule and get a real feel for the three-dimensional shape of the protein. If you are having trouble viewing the stereo image, borrow the stereo glasses from Dr Maciejewski.

- Not all proteins are ideally behaved for structural analyses and consequently the quality of an experimentally-determined protein structure can vary considerably. For X-ray structures, the resolution in Ångstroms is a rough guide to quality, while for NMR-derived structures the rmsd of the ensemble of structures is usually a good indicator of the quality (even

though it measures precision rather than accuracy). A good independent measure of the quality of a structure, regardless of the technique used for structure determination, is the **Ramachandran plot**. In the 1960s, the Indian biochemist G.N. Ramachandran determined the most energetically favorable combinations of backbone ϕ, ψ angles; the ϕ, ψ map showing the favored regions is now known as the Ramachandran plot. As the quality of a protein structure increases, so does the percentage of residues found in the favored regions of the Ramachandran plot; the highest-resolution crystal structures (<2.0 Å resolution) usually have >90% of residues in these regions. Good structures should have more than 80% of residues in the favored regions of the Ramachandran plot.

Let's take a look at the Ramachandran plot for 1COF. Click **all** to select all residues and then select **Fig→Ramachandran**. Select **normal** for background and then click **OK**. You should now see the Ramachandran plot with the most favored regions in green, additionally allowed regions in yellow, and generously allowed regions in pink. The uncolored regions are highly disfavored; 1COF has no non-glycine residues in these regions. The four crosses correspond to glycine residues which have a less restricted Ramachandran space because of their small sidechain. Most residues (>90%) in 1COF fall into the highly favored green regions of the Ramachandran plot; this indicates that it is a high-quality structure.

- The molecular surface of a protein sometimes provides an insight into function. DNA-binding proteins often have surface patches with a high density of positive charges that direct interactions with the negatively charged DNA. Solvent-exposed hydrophobic patches often indicate a possible protein-protein interaction surface. Let's draw the molecular surface of yeast cofilin and examine the surface charge (the so-called electrostatic potential).

First we will calculate the electrostatic potentials and save these in a file. Select **Fig→Off** to get out of Ramachandran plot mode. Click **all** and then **invisible** to turn off any displayed atoms/bonds. Execute the **pdb_charge.mac** macro from the **File→Macro→Execute Standard** pulldown menu. Click **all** to make sure all atoms are still selected. Calculate the electrostatic surface potential by selecting **Potential** from the **Calc** pulldown menu. When the dialog box appears set **Atom charge = simplecharge** and **Mol. Dielectr. = 30**. Choose a filename (or just use the default filename = tt.pot) then press **OK**. The program will write a file containing the potentials. The calculation will take a little while.

Now we will draw the molecular surface. Click **all** to make sure all atoms are still selected.

Draw the molecular surface by selecting the **Prim→Surface→Add** pulldown menu. When the dialog box appears, choose **contact surface** then press **OK**. The calculation will take a few seconds then the surface should automatically appear on screen.

Now we will map the electrostatic potential onto the molecular surface. Read in the electrostatic potential information by choosing the **File→Potential** pulldown menu and selecting the file containing your data when the dialog box appears (default filename = tt.pot). Now you can paint the electrostatic potential onto your molecular surface by selecting **Prim→Surface→Paint** from the pulldown menu and choosing **pot** when the dialog box appears. Regions of significant positive and negative charge are indicated by blue and red, respectively. **Is there anything unusual about the surface charge distribution for 1COF?**

- At the end of the lesson, you can store your current MOLMOL session by choosing **File→Write Dump** from the pulldown menu and saving the session as a filename with a **.mol** extension in the MEDZ325 directory. Otherwise, quit MOLMOL by choosing **File→Quit** and selecting **no** in response to the question **Save State?** in the pop-up Dialog Box.