Protein Crystallography

BBMB 334 Lab 4

Whitman College
Program in Biochemistry, Biophysics & Molecular Biology (BBMB)
Biophysics Laboratory
Prof. Douglas Juers

Part III. Structure Determination/Refinement and Analysis.

Goals: 1. To learn how structure determination and analysis via crystallography works.
2. Learn a little more about protein structure.
3. Learn about the overall architecture of one particular enzyme in relation to its catalytic activity.

In this last week of the Protein crystallography lab, your goal is to (A) produce a model of your protein based on the diffraction data you collected last week, (B) analyze this model for quality and (C) interpret the model.

I frequently give targets you should be shooting for. If your results don't seem to meet these targets, stop and ask me to look over your data.

Part A. Producing the model.

Step 1. Data averaging and assessment.

1. **Setting up the computing environment.**
    a. On an iMac in the BBMB lab, put a copy of the .mtz file produced by the data collection program (here called "protein.mtz") into a folder on the Desktop or on your thumb drive. The folder name should have no spaces, nor should any subfolders. Open at terminal window (using iTerm), and type the command: ccp4i. A window should open.
    b. Once in ccp4i, choose "Directories & Projectdir". Add project; give your project a name (hereafter called YOURPROJ in this writeup) and Browse to the folder you created above (Don't enter a filename). Choose YOURPROJ from "Project for this session…". Click "Apply&Exit".

2. **Averaging the data.** Click on the button in the upper left, find the "Data Reduction" tab, and choose "Scale and Merge Intensities". Give the job the title of Scale. Check "Ensure unique data…" For "MTZ in" browse to protein.mtz. MTZ out should automatically switch to YOURPROJ with protein_scala1.mtz. Under "Define Output Datasets" give some appropriate names (just a few characters, no spaces). Goto "Run" and choose "Run Now". In the main window you should see that your job Scale is RUNNING.

3. **Initial assessment of data quality and adjusting the resolution.** After the job is finished, goto "View Files from Job" and pick "View Log Graphs". Look at the following graphs under "Completeness, multiplicity, Rmeas v. resolution,…"
    a. Completeness vs Resolution. The %poss graph should be about 100%. Using the cursor, note at what resolution (Å) the %poss drops below 90%.
    b. Multiplicity vs Resolution, which shows the number of times each spot was measured as a function of resolution.
    c. Rmeas, Rsym… vs Resolution. This shows the agreement between multiple observations as a function of resolution. The vertical axis is essentially like a relative error between multiple observations of the same spot. Note at what resolution the Rsym gets above 0.25.

4. **Re-averaging the data.** Close the graphing window. Rerun step 2 (ReRun Job on the right) with some adjustments: Check the box marked "Exclude data resolution less than…". Leave the number showing in the first box. In the second box, put the larger of the two numbers from Parts 3 a & c. This will be the resolution limit of your data set. Run the job, overwriting files from the previous time.

5. **Final assessment of data quality.** When the job is finished, choose "View Log File" from "View Files from Job" and click "Show Summary" on the bottom. The following parameters will be important to consider when assessing the quality of your data. Record these numbers both for overall and the outershell.
    a. Rmerge. This is the agreement in intensities of the same spot measured multiple times. Generally we want this to be less than 10% overall and ~25% in the outershell.
    b. Mean (I/sd(I)). This is ratio of signal/noise. Higher is better. Should be greater than 2.0 for the outershell
    c. Completeness. Gives the % of the possible spots you could measure that you actually did. Should be greater than 90% overall and in the outershell.
    d. Multiplicity. The average number of times you measured each spot.

6. **What is this data?** The data you measured provide parameters for sine waves to be added together to create a picture of the protein. This is analogous to adding together the sine waves to create the whale and torpedo models in the Excel exercise from last week. Here each spot corresponds to one sine wave. To see this use "View Any File" to open the output file from the scale job (protein_scala1.mtz).
    a. Number of reflections. Scroll down until you see "Number of Reflections". This is the number of sine waves that get added together. Record this number.
    b. Parameters associated with each reflection. Scroll further until you get to the list of reflections. Each reflection has about 19 numbers associated with it, but the important ones for us are columns 1-3 and 5&6.
        i. Wavelength of sine wave. Columns 1-3 (usually called *h,k,l*) code for the wavelength of the sine wave. Three numbers are needed b/c this sine wave is in three dimensions. Larger values of *h,k,l*

correspond to sine waves with smaller wavelengths (see the equation in Part II of the Lab Handout).

ii. **Amplitude of sine wave.** Column 5 gives the amplitude, F, of the sine wave (which is the square root of the intensity of the spot). Column 6 gives the uncertainty in the amplitude, σ_F. All of the sine waves will get added together to create a 3D model of the protein.

Step 2. Model Building and Refinement

7. **Refinement.** For this crystallography exercise, we will start with the known structure of your protein. Get the file "lyz.pdb" or "trypsin.pdb" from CLEo. These are pdb files that I downloaded from the Protein Data Bank and adjusted somewhat for this exercise. In each case, I removed all waters and ligands from the file and truncated the catalytic side chains to alanine. Your task is to improve this starting model by (a) building in the catalytic sidechains and (b) adding noncovalently bound molecules such as water (both lysozyme and trypsin) and benzamidine (trypsin only). After you download the files, make sure their names don't include parentheses.
    a. **Step 1. Rigid body refinement.** The first step is to refine the overall location of the protein, keeping it as a rigid body. This is done by calculating the expected diffraction pattern based on the coordinate model, and then moving this model around to minimize the difference between the calculated pattern and the one you measured. To do this, choose "Run Refmac5" under the Refinement Tab. Give the job the title of "Rigid Body". Under "Do" pick "rigid body refinement". For "MTZ in" pick protein_scala1.mtz. For "PDB in" pick trypsin.pdb. Check the Refinement Parameters box. Then check the resolution range box, and change the smaller number to 3.0. Run the job. When the job is finished, View Log Graphs, under tables choose the last line, "Rfactor analysis…" and look at the graph of <Rfactor> vs cycle. The R-factor is like a % error between your observed data and the data calculated from the model. It should drop initially and then level off somewhere between 25 and 45%.
    b. **Step 2. Positional refinement.** The next step is to allow each atom to move independently (the protein is no longer a rigid body). Set up another Refmac job called "Restr". This time, do restrained refinement. For "MTZ in" pick protein_scala1.mtz. For "MTZ out" type protein_refmac2.mtz. For "PDB in", pick protein_refmac1.pdb (the result from the previous step). Under Refinement Parameters, this time use all the data (uncheck the Resolution range box). Run the job and when it's finished look again at the R-factor graph. Depending on your resolution limit, the R-factor may be lower than from the previous step, or it could be somewhat higher. If it's a lot higher (say 10 points higher) something might be wrong with your refinement.

8. **Model Building.**

a. **Display and Manipulate the Coordinates.** The next step is to look at the model and make some adjustments. Start the program called "Coot". Read in the refined coordinates with File -> Open coordinates and then locate protein_refmac2.pdb (Click on Filter, which will make it easier to find protein_refmac2.pdb). Play around with moving the molecule. The first mouse button rotates, the third zooms in and the middle button will let you recenter on an atom you click. You can go to a particular atom with Draw -> Go To Atom…
b. **Display and Manipulate the Electron Density Maps.** Next read in the protein_refmac2.mtz file with File -> Auto Open MTZ…. You'll see some chicken wire, which is a three dimensional contour map of the electron density. The electron density is calculated using the Fourier transform discussed in Part II of the lab – it is the sum of sine waves. This is similar to the whale in that it corresponds to a sum of sines, except there's three dimensions. Your job is to optimize the fit between the model and this electron density. To assist with this, we can use a "difference map". Click on Display Manager. There are two Maps you have at your disposal. The FWT map is the "electron density" map. The DELFWT map displays "difference density", which show differences between your data and the model. Positive difference density shows places that your data predicts atoms that are not in your model, while negative difference density shows atoms in your model that are not present in your data. The scroll wheel on the mouse will change the level of sensitivity in displaying the map. In the Display Manager is a button that chooses which map is controlled with the scroll wheel.
c. **Fixing the catalytic residues.** Use the Go To Atom menu to center on residue 195, which is currently modeled as an alanine, but should be a serine. You should see some difference density corresponding to the part of the side chain that is missing, and also the electron density. You can use these as guides to model in the rest of the side chain.
    i. **Mutate the alanine to glutamate/serine.** Go to Calculate -> Mutate residue range. Pick residue range 195 to 195, enter the single letter code for serine (S) in the textbox, and click Mutate. The alanine should change to a glutamate/serine, but the side chain probably won't match the electron density.
    ii. **Adjust the glutamate/serine position to match the electron density.** Goto Calculate -> Model/Fit/Refine -> Edit Chi Angles (double click). Then double click on an atom of the serine. A new panel will appear that will let you adjust the side chain until it is located into the electron density. Do this by double clicking somewhere in the coot display window (but NOT on an atom of the serine), and then dragging across the screen with button 1 held down. You should see the side chain oxygen move across the screen. To see the situation from another viewpoint, click View Rotation Mode. Once you have it how you like it with the side chain fitting in the electron density, click "Accept."

iii. **Repeat** this for residues 102 (to aspartate) and 57 (to histidine) for trypsin.

d. Adding ligands.

i. While centered on histidine 57, zoom out so you can see the whole molecule. Turn off the "electron density" map (FWT) with the display manager, leaving on only the difference map (DELFWT). On the display manager, click on the "scroll" radio button. Now you can change the level of the map with the scroll wheel on the mouse. Increase the contour level to see the strongest features in the map. You will see some water molecules, but the strongest feature will probably be for a benzamidine molecule. This was part of your crystallization buffer, and is an inhibitor of Trypsin. Try to locate the electron density for the benzamidine in your difference map.

ii. To model the benzamidine, download "benzamidine.pdb" from CLEo and read it into coot. Next choose Calculate -> Other modeling tools -> Fit ligands. This brings up a dialog box.

iii. Under Select Map pick the difference map. Under Select Protein pick your protein model. Under Select ligands, pick the benzamidine you just loaded. For the contour level, choose the contour level at which you could clearly see the whole benzamidine in the difference density, but not much else (probably about 3). Click "Find 'em", and the program should locate the benzamidine in the difference density.

iv. Check to make sure it looks ok, then merge this fitted benzamidine into your protein model with Calculate -> Merge molecules. Here you should Append/Insert the fitted ligand into protein_refmac2.pdb.

e. Adding water molecules. There will be other features in the difference map – spherical objects that are water molecules. Water plays a structural role in nearly all protein molecules. In many enzymes it is a key catalytic player in the mechanism (this is true of both lysozyme and trypsin).

i. **Add waters.** Choose Calculate -> Other Modelling Tools -> Find waters. Select the difference map, lyz_refmac2 as the mask, find peaks above 5.0 sigma, and add waters to "Molecule that masks the map". Then click "Find waters".

ii. **Check the waters.** We want to check whether we agree with the programs guess as to the locations of water molecules. They should be centered in some electron density and make polar contacts to the protein or to another water moleceul. Goto the first new water molecule with Go To Atom. The water will be in a different chain from the protein. To see the contacts the water makes, click Measures -> Environment Distances and click the "Show Residue Environment" box. Step through the waters ("Next Residue") on the Go To Atom dialog box, checking each water for

good polar contacts (between 2.4 and 3.2 Å to red (oxygen) or blue (nitrogen) atoms) and electron density. If there are any waters you disagree with, make note of which numbers they are.

f. Refining the new model.

i. Write out your improved model (with active site side chains and water molecules) with File -> Save coordinates. Select the molecule to write out, and then Click OK on the window that appears (it might appear behind the main coot window).

ii. Do another restrained refinement run, similar to the second run you did above, but use these new coordinates as your starting model. When the job is finished, look at the log graph and note how much the R-factor went down from the previous refinement run. The next thing to do would be to look through the model for errors, add more waters, do more refinement etc…

Part B. Model Assessment We now would like to assess the quality of the model. There are several ways to do this.

1. **R-factor/R-free.** You already know the R-factor from the previous step. Look at the log file and write down the R-factor and the R-free. These should typically be around 20% for R-factor and 25-30% for R-free. These tell how well your model predicts the observed data. The R-free is for data that was not used in refinement.

2. **Model quality - ramachandran plot.** In Coot, read in the newly refined coordinates. do Validate -> Ramachandran plot. Pick your final model. This plots each residue on a Psi vs Phi (the backbone angles) plane. Only certain combinations of Psi and Phi are permitted energetically. Most of your residues should be located in the pink regions. Coot will give statistics about the Ramachandran plot. Record these statistics. You can pick various points on the Ramachandran plot to check out those residues in your model. Pick some that appear to be out of the optimal range and investigate why this is so. Pick a glycine (triangle). Why do you think the Ramachandran plot changes when you pick a glycine?

3. **Further checking – if time allows.** Back in CCP4i, Under Validation & Deposition run Procheck on your final model. This produces a bunch of *.ps files that you can view with the program GSView (Start -> Program Files -> Ghostgum -> GSView.

   a. Ramachandran quantification. Look at the file ending in ramchand.ps. This is another view of the ramachandran plot, with some statistics. Record the fraction of residues in the most favoured regions.

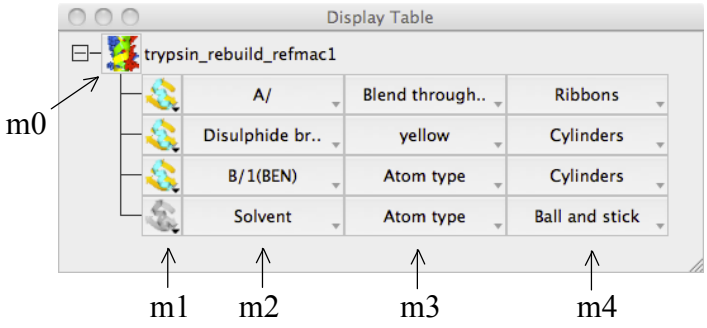
   b. Chi1-Chi2 plots. Look at *_chi1_chi2.ps. This is like a Ramachandran plot, but for the side chain torsion angles (these are the angles you had to change to fit the the active site side chains into electron density).

   c. **Other plots.** Look at other plots as time allows.

d. **Bond angles/bond lengths.** All of the bond angles and bond lengths in the protein are well known bonds and should be within a small range of the expected angle and distance for these types of bonds. The deviations of your bond angles/distances from the expected ones are quantified in the log file for refmac in CCP4i. Open this log file, scroll to the bottom and then scroll up until you see a table in which column 1 is "Restraint type". Record the RMS Delta for bond distances and bond angles (this is the root mean square deviation of your bond distances and angles from the expected distances and angles). These should be < 0.02Å and < 3.0^{o} for distances and angles respectively.

Part C. Model interpretation. In this last part we will use your model to look at some general aspects of protein structure and then a few specific things about your enzyme.

1. **General aspects of protein structure.** Open QtMG and read in your final refined model.
    a. Moving the molecule. In this program you can rotate with the left mouse button, zoom with the scroll button, and recenter around an atom by double clicking on the atom with the left button.
    b. Changing the display mode. The method of displaying the molecule can be adjusted with Windows -> Display Table. This will give something like:



Here, there is one molecule called trypsin_rebuild_refmac1. There are currently 4 objects in the display table:

A/	(the protein chain)
Disulphide bridge	(the disulfide bonds)
B/1(BEN)	(the benzamidine)
Solvent	(the water molecules).

For each object there are 4 menus (m1 - m4) controlling:

m1: Whether the object is visible (here the first three objects are visible)
m2: What atoms are shown.
m3: The coloring scheme.
m4: The type of representation.

Play around with the representation type for the first object (the protein chain) to get a feel for the various representations. In particular, try at least the following:

i. Ball and stick. Shows every atom as a small sphere.
ii. Spheres. Shows every atom as a sphere with radius closer to its actual Van der Waals radius.

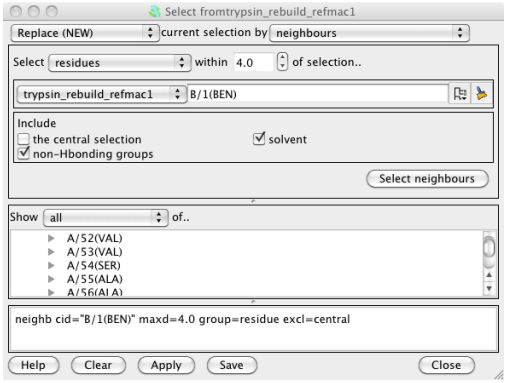
iii. Ribbon. A stylized representation showing just secondary structure.
iii. Surface. Shows the surface of the molecule. Look for how the benzamidine binds in a pocket on the surface. (If you can't see the benzamidine, go to QtMG -> Preferences -> Surfaces -> Surface drawing style and change the surface probe radius to 1.40 Å.)

c. Change the representation of your protein to ball and stick and change the coloring scheme using m3 -> Residue property -> Residue type. Then use m3 -> Edit colour scheme to see what the coloring scheme is, and answer the following questions:
   i. Where are the hydrophobic residues located?
   ii. Where are the polar residues located?
   iii. Are there more basic or acid residues? Based on this, what do you think the charge of your protein is at neutral pH?

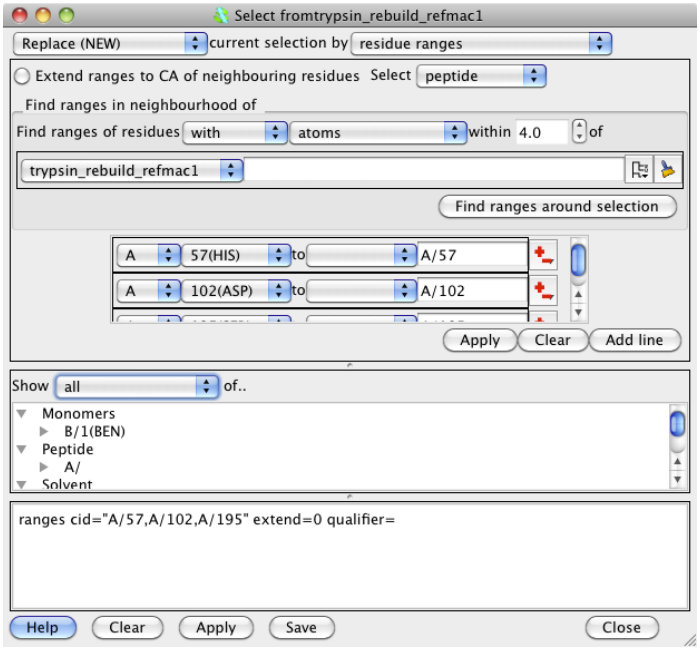
d. Make sure the representation of your protein is ball and stick, and color by atomtype. What color are the polar atoms? Under the main protein menu in the display table (m1), choose Add display object -> Hydrogen bonds. This displays hydrogen bonds with dotted lines. Try to find a polar atom without a hydrogen bond.

e. Change the representation of your protein to "Thermal ellipse". You will see an ellipsoid centered on each atom whose size corresponds to the mobility of the atom (how much it moves around due to contact with the thermal bath). In general, where are the most mobile and least mobile atoms?

2. **Specific things about trypsin.** Hide all objects except the protein chain. Change the color scheme to secondary structure and the display type to ribbons.
   a. Benzamidine binding site. Add a display object (m1 -> Clone). Show just the residues near the benzamidine (m2 -> Selection browser; Replace (NEW) current selection by neighbors; pick the benzamidine, and click applye (see below)). This will create an object of the neighbors of the benzamidine. Examine the interactions benzamidine makes with the protein. Do there appear to be specific hydrogen bonds involved in benzamidine binding? What does the aromatic ring of the benzamidine interact with?



b. Active site and benzamidine inhibition. Add another display object (m1 -> Clone). Show just the catalytic triad (Ser 195, Asp 102 and His 57) (m2 -> > Selection browser; Replace (NEW) -> Atom Selection -> ..of residue ranges; create a range for each residue – i.e. to show residue 57, the range is from 57 to 57; see below). Change the display to ball and stick. Think



about what type of molecule the substrate is. Does it make sense that the active site residues are located where they are on the protein? Make another clone of the molecule and display it as a white surface. You should get a better sense for how the substrate will be interacting with the enzyme. How does the structure help explain why the benzamidine is an inhibitor?

Writeup.

Do a standard writeup (Abstract, Introduction, Materials & Methods, Results & Discussion). It should summarize the overall experiment. Crystallization -> Data collection -> Data averaging -> Refinement and model building -> Interpretation. Include the various descriptors of data and model quality that were described as you went through this procedure. You might also want to include a picture or two of your model.

Software

CCP4i, CCP4mg and Coot are all freely available at http://www.ccp4.ac.uk/download/
The postscript viewer is available at http://pages.cs.wisc.edu/~ghost/gsview/index.htm