

Determining, Analyzing, and Understanding Protein Structures

The goal of this recitation is to get you comfortable with looking at protein structures. Being able to analyze and understand protein structures is extremely beneficial for understanding function and designing experiments to study mechanism.

Specific topics in this recitation:

- Protein structure determination by X-ray crystallography
- The Protein Data Bank as a protein structure repository
- Analyzing the structure of ubiquitin in PyMOL

Introduction

TA: Shiva Mandala

This course focuses on advanced problems in biological chemistry such as protein synthesis and degradation, assembly line syntheses, cholesterol metabolism, and nucleotide metabolism. In recitations, we will discuss case studies from the literature and review important experimental methods. There will be a reading assignment and a recitation handout posted on Stellar every week. Please go over these materials before recitation.

Determining protein structures

The most widely-used technique to determine protein structures is X-ray crystallography. Since the report of the first atomic resolution protein structure in 1958, crystal structures of proteins have fundamentally changed our understanding of protein function and mechanism. There are now over 100,000 crystal structures deposited in the Protein Data Bank (PDB). 2014 was the “International Year of Crystallography” and many journals including Nature published special features on crystallography for a broad audience (see, e.g., <http://www.iycr2014.org/> and <http://www.nature.com/news/specials/crystallography-1.14540>). More recently, developments in electron microscopy and NMR spectroscopy have led to additional reported protein structures, in particular of flexible proteins or large protein complexes. Electron microscopy will be discussed later in the class.

Overview. The basis of protein structure determination by X-ray crystallography is the elastic scattering of X-rays off electrons in a single crystal. Scattering off the crystal lattice leads to interference of the X-rays, resulting in constructive interference only for a special set of conditions. From the diffraction pattern, the so-called “electron density” can be back-calculated. The three-dimensional protein structure is then modeled into this electron density (**Figure 1**).

Why X-rays? X-rays are high-energy radiation, usually with wavelengths of 0.01 to 10 nm (frequencies in the range of 10^{16} - 10^{19} Hz). Crystallographers often talk in terms of “Ångström”, abbreviated Å, where $1 \text{ Å} = 0.1 \text{ nm} = 100 \text{ pm} = 10^{-10} \text{ m}$. Atomic bonds are on the length scale of 1 Å, e.g. C–C single bond lengths are usually around 1.5 Å (150 pm). Resolution in all types of scattering experiments (including optical microscopy) is physically limited by the wavelength of the electromagnetic radiation used (approximately $\lambda/2$). Thus, X-rays are well suited to study chemical structures because their wavelength matches that of chemical bonds.

Crystallization. For X-ray crystallography, the protein of interest first has to be crystallized, often in a specific state (substrate- or inhibitor-bound, post-translationally modified, etc.). Just like salts, many proteins are capable of forming regular crystal lattices in which the molecules are arranged over and over again in a regular fashion, this is also known as translational symmetry. Crystallization out of a supersaturated protein solution occurs under specific conditions, determined by a buffer, the presence of a precipitant (commonly salt or polymer), and a number of other parameters, including temperature, pH, time, protein concentration, etc. This step is usually the bottleneck in protein structure determination and requires testing a large variety of conditions because the crystallization behavior of a protein cannot be determined *a priori*, at least not with current technology. Some proteins are indeed impossible to crystallize.

Data collection. Once crystals are obtained, one exposes them to X-rays to collect the diffraction data. X-ray data collection is usually carried out at low temperatures of about 100 K to minimize radiation-induced damage in the crystals and resulting structural changes. As protein crystals are usually >50% water and water expands when freezing, which can disrupt the crystal lattice, crystals need to be supplemented with a cryoprotectant such as glycerol prior to freezing. The cryoprotectant prevents ice formation.

For X-ray diffraction data collection, a collimated X-ray beam is directed through the crystal while the crystal is kept at 100 K in a cold nitrogen stream. Constructive interference of the scattered X-rays occurs only under a special set of conditions, leading to a distinct diffraction pattern of regularly spaced spots. This pattern is recorded by specialized detectors. To obtain a full diffraction data set, the crystal is rotated about an axis perpendicular to the X-ray beam and individual diffraction patterns are recorded throughout. This rotation allows access to (almost) all possible conditions for constructive interference.

Processing and modeling. The resulting diffraction patterns are then processed with specialized software. Processing entails assignment of crystal lattice symmetry, integration of spot intensities, and then merging of reflections that were observed multiple times in the process of data collection. To calculate electron density maps, one also needs to obtain the phases for the reflections. Phase information is lost in the X-ray diffraction experiment. Their determination is beyond the scope of this class. Ultimately, one obtains an electron density map into which the protein structure is modeled. The positions of the atoms are refined iteratively until a satisfactory

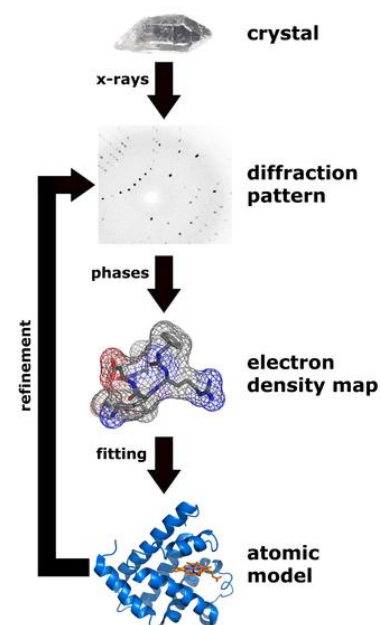


Figure 1: General workflow of protein structure determination by X-ray crystallography.

© [Thomas Spletstoeser](https://commons.wikimedia.org/wiki/Thomas_Spletstoeser), Wikimedia Commons, License BY-SA. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use>.

model is obtained. The final structure is an atomic model of the protein of interest and is, if novel, published and deposited in the Protein Data Bank (see below).

Assessing structures. Several aspects determine the quality and information content of a crystal structure. One of the most important aspects is the “resolution” of a structure. Similar to optical microscopy, resolution is a measure of how far two points have to be apart to appear as separate peaks, although the definitions are slightly different. Even though there is a physical limit to resolution, crystal quality is usually far more limiting in X-ray crystallography. Small irregularities in the crystal, amino acid side chain flexibility, and other sources all contribute to noise in the data that will ultimately limit the resolution.

For a crystal structure, the resolution indicates how many features are discernible in the electron density. The smaller the number on the resolution, the “higher” the resolution. A 1.5 Å resolution crystal structure is higher resolution than a 3 Å resolution crystal structure, and it will reveal far more details about the protein (**Figure 2**). The rule of thumb is that atomic positions are accurate to $\sim 1/10$ of the resolution (e.g. in a 4 Å structure, the average displacement in the model versus the “true” structure is ~ 0.4 Å).

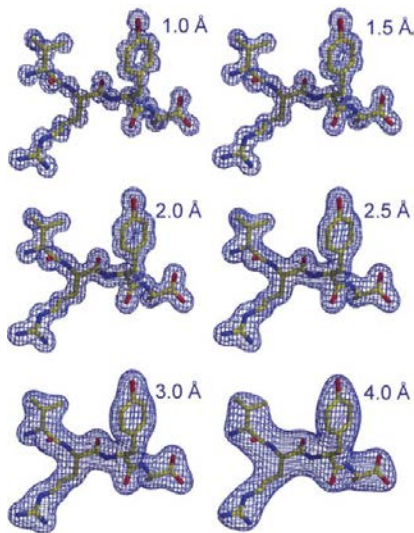


Figure 2: Comparison of electron density maps at different resolutions from 1.0 Å to 4.0 Å.

© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use>.

information on the quality of the data and can serve as an indicator of reliability. The diffraction intensity drops with higher resolution (smaller number) but the noise level stays roughly equal. Therefore, the data quality decays with higher resolution until the signal is indistinguishable from background.

Resolution	Features visible
6 Å	α -helices as tubes, overall protein shape
4 Å	large side chains as tubes, protein backbone traceable, individual β -strands
2.5 Å	most individual side chain orientations, backbone carbonyl orientation, water molecules
2 Å	almost all side chains with correct orientations
1.0 Å	individual atoms, atom type (C, O, N) discernible by electron density

Each published X-ray diffraction data set comes with a table of statistics that includes the resolution, the signal-to-noise ratio, the data completeness, and other statistics (**Table 1**). These statistics provide

Table 1: Examples of basic X-ray data collection statistics

	value	meaning/interpretation
resolution (Å) ¹	100 – 2.15 Å (2.21 – 2.15 Å)	See above. Data are usually binned by resolution (10 to 20 bins with similar number of reflections). Statistics for the “highest resolution bin” are given to justify a given resolution limit.
unique reflections ¹	40511 (2970)	Number of independent observations, essentially the number of data points. Important in refinement.
R_{meas} (%) ¹	6.2 (71.5)	R -factor: Agreement between reflections that should have the same intensity. The lower, the better.
$\langle I \rangle / \sigma(\langle I \rangle)$ ¹	15.2 (2.1)	Signal-to-noise ratio. The higher, the better. Usually cut off at a value of 2 for the highest resolution bin.
completeness (%) ¹	98.5 (99.4)	How many out of the theoretically observable reflections ones were actually measured. The higher, the better. >95% highly desirable.
redundancy ¹	3.6 (3.6)	How many times each individual reflection was measured. The higher, the better. >3 highly desirable.

¹ values in parentheses are for high resolution bin

All protein structures are modeled into the electron density and refined against the collected diffraction data. Refinement is performed by calculating theoretical diffraction data from a given model and then minimizing the difference between the calculated and the actual data, within constraints provided by stereochemistry (bond lengths, bond angles, dihedral angles, etc). The so-called crystallographic R -factor (also R_{cryst} , R_{work}) is a measure of this difference. The smaller the R -factor, the better does a model match the data.

A fundamental problem in refinement of crystal structures is overrefinement: the number of observed reflections is small compared to the number of parameters to be refined (individual positions for each atom, thermal mobility factors). Imagine fitting a third-degree polynomial against four data points: the fit will be great, but it will likely not be a good description of the real behavior in the data. In fact, the relationship might be linear with some error. To prevent overrefinement, in the 1990s crystallographers started setting aside a subset of the observed reflections, usually ~5%, that were not used for refinement. Instead, these reflections are only used to calculate the final difference between the model and the data. Because they were never refined against, they provide a true indicator of how well the model matches the data, without fitting bias. This R -factor is called the free R -factor or R_{free} . A good rule of thumb is that the R_{free} should be about 1/10 of the resolution (e.g. 0.25 for a structure of 2.5 Å resolution) and that the R_{cryst} should be within 0.05 of the R_{free} . Again, the smaller these values, the better. In addition, other refinement statistics such as deviations from ideal bond lengths and bond angles as well as Ramachandran statistics are usually reported.

There are no hard limits on all these statistics and there can be surprises either way: a bad structure can have good statistics, and a good structure can have bad statistics. Nonetheless, being able to critically evaluate both data statistics such as completeness, signal-to-noise and refinement statistics such as the R_{free} can be tremendously helpful in determining which protein structures to trust and which protein structures to approach with caution.

Limitations. X-ray crystallography has revolutionized our understanding of structure and function of biomacromolecules by allowing for visualization at near atomic resolution. One must, however, view structures with a discerning eye. It is important to consider the conditions under which a protein is crystallized. For example, what is the pH and is it physiologically relevant? If a protein is crystallized at unusually high or low pH, the observed structure may not necessarily reflect the native structure of the protein. Furthermore, it is important to remember that crystal structures are static snapshots, yet biomacromolecules are dynamic and conformational change is often key to function. In fact, portions of proteins in crystals that move too much cannot be modeled accurately.

The Protein Data Bank (PDB)

Website: <http://www.rcsb.org>

What is the PDB? The Protein Data Bank (PDB) is a repository of protein and nucleic acid structures determined by X-ray crystallography, NMR spectroscopy, and electron microscopy. Every structure reported in a publication is deposited here, as well as some structures that are not part of publications (such as those determined by structural genomics consortia). Each structure is linked to a unique 4-character ID, called a PDB ID. This unique ID is used to refer to the structure in the literature and can be used to find a certain structure.

Searching for a Protein Structure: The most direct and accurate way to locate a protein structure is by using its PDB ID. You can also use the search function to search for structures of certain proteins (lots of proteins have multiple structures reported) or to search by different categories such as author, sequence, ligands, function, etc. Operator syntax (and, or, not) and exact phrase syntax (using double quotes) are supported for full text search. For example, the search term (ubiquitin and “protein degradation”) returns results that contain both ubiquitin and the exact phrase protein degradation.

For every entry, a range of information is available. The experimental details box contains a few pieces of essential information (**Figure 3**), other statistics are available in a few clicks. The literature reference is also given, if available. Note that this particular structure of ubiquitin (PDB ID 1UBQ) was determined before crystallographers started reporting the R_{free} . Thus, it is hard to assess the true quality of the structure. There are newer reported structures of ubiquitin that are slightly improved.


Experimental Details		Hide
Method: X-RAY DIFFRACTION		
Exp. Data:		
BMRB ↗		
Structure Factors ↗		
EDS ↗		
Resolution[Å]: 	1.80	
R-Value:	0.176 (obs.)	
R-Free:	n/a	
Space Group:	P 21 21 21 ↗	
Unit Cell:		
<u>Length [Å]</u>	<u>Angles [°]</u>	
a = 50.84	α = 90.00	
b = 42.77	β = 90.00	
c = 28.95	γ = 90.00	

Figure 3: Experimental Details Section of Entry 1UBQ (Ubiquitin)

PDB ID: 1UBQ. Vijay-Kumar, S., Bugg, C.E., Cook, W.J. *J. Mol. Biol.* 194 (1987): 531-544. DOI: 10.2210/pdb1ubq/pdb.

PyMOL

Website: <http://www.pymol.org>

MIT licensed full version: <https://ist.mit.edu/pymol/all>

Registration form for free student license: <http://pymol.org/edu/>

Pymol Wiki: http://pymolwiki.org/index.php/Main_Page (contains extensive documentation)

User Guide: http://www.pymolwiki.org/index.php/Practical_Pymol_for_Beginners

There are many additional User Guides online (see end of document for links).

PyMOL is a Python-based program to visualize protein structures and render figures. The full version of PyMOL is licensed to MIT for educational and academic research use. Students can also download and install a free educational version of PyMOL after registration at the link above. Alternatives to PyMOL are Coot (great for model building and analyzing structures, but not for making figures) and Chimera (open source, powerful visualizer).

Starting PyMOL opens two windows: an external GUI with command line and an internal GUI with a viewer (Figure 4). The viewer allows for interactive manipulation of the structure using the mouse. The external GUI and command line allow for powerful manipulations using PyMOL code.

You can load a structure either from your hard drive, via the GUI (File>open) or the command line (load path/filename.pdb), or by directly retrieving it from the PDB via the command line (fetch XXXX , where XXXX is the PDB ID).

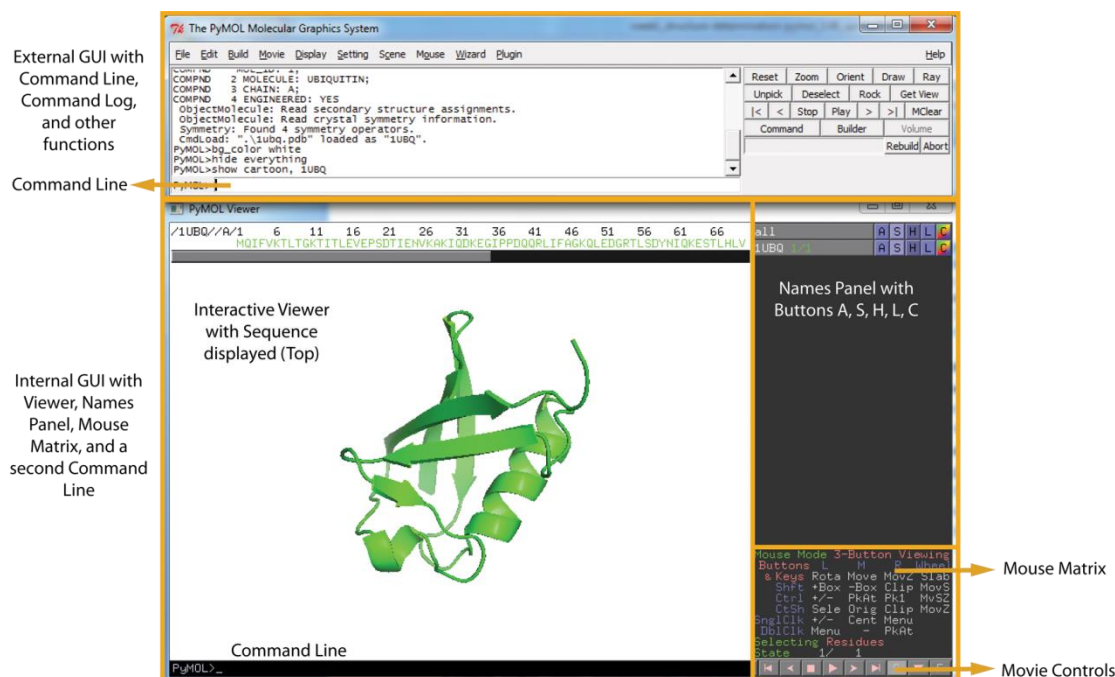


Figure 4: PyMOL Windows

© Schrodinger, LLC. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use>.

PyMOL Controls. PyMOL can be controlled both using the mouse and using the command line. A combination of both is most powerful. A basic 3-button mouse is ideal for PyMOL. **Table 2** summarizes combinations of mouse commands in the basic mode. The mode can be changed in the mouse matrix (lower right corner). In general, you want to be careful with your clicks.

Table 2: PyMOL Mouse Commands.

Keyboard Modifier	Left Button	Right Button	Middle Button
none	rotate camera, select atoms on click	move camera in Z (zoom in/out)	move camera in XY (plane of screen)
shift		move clipping planes	
shift+ctrl			set origin of rotation

There are three main ways to select residues.

- 1) **Left click** on an atom in a residue. You can change the selection mode (atom, residue, chain, etc) in the mouse matrix (lower right corner).
- 2) **Shift+Left click** on atoms to add them to the current selection.
- 3) **Command line:** typing **select selection_name, (resid 1-10)** will select all residues with numbers 1 through 10. Adding more modifiers can make the selection more or less restrictive. Modifiers include atom name, secondary structure, molecule, and atom type. For example, typing **select lys, 1UBQ and (resid 10-40) and (resn lys) and (name ca)** will select the C α atoms of all lysine residues between numbers 10 and 40 in the molecule 1UBQ and store them in a selection called “lys”.

Manipulating molecules and selections. Molecules and selections can be manipulated in the GUI via the names panel or via the command line. For the names panel, the five buttons (A, S, H, L, C) allow for the following manipulations:

- 1) **A – Action:** rename, delete, change, center, and others
- 2) **S – Show:** show a given representation (sticks, spheres, cartoon, ribbon, etc)
- 3) **H – Hide:** hide a given representation (sticks, spheres, cartoon, ribbon, etc)
- 4) **L – Label:** Put various labels on the selection.
- 5) **C – Color:** change color schemes

In this recitation, we will analyze the structure of ubiquitin. Ubiquitin is a small (8.5 kDa, 76 amino acids), essential, eukaryotic protein that plays a critical role in protein homeostasis and degradation. Ubiquitin can be attached to substrate proteins in a process called ubiquitination. Attachment of a chain of multiple ubiquitin molecules to a certain protein signals for degradation of the protein by the proteasome, as will be discussed later in this class. Ubiquitination can also be involved in regulation of membrane trafficking, translation, inflammation, and other processes.

We will load the structure of ubiquitin, change the representation, analyze the secondary structure, identify the lysine residues involved in ubiquitination, and generate a high-resolution figure.

Links to additional PyMOL tutorials:

http://bioquest.org/nimbios2010/wp-content/blogs.dir/files/2010/07/pymol_tutorial3.pdf

Step-by-step tutorial.

http://www.doe-mpi.ucla.edu/CHEM125/pymol_tutorial_060418.pdf

A little more advanced but quite comprehensive. This PDF has been relocated, but you can access it by using

<http://archive.org/web/>.

http://www.mrc-lmb.cam.ac.uk/rlw/text/MacPyMOL_tutorial.html

Intro to PyMOL for Mac.

MIT OpenCourseWare
<https://ocw.mit.edu>

5.08J Biological Chemistry II
Spring 2016

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>