Understanding PDB Data: Looking at Structures

The PDB archive is a repository of atomic coordinates and other information describing proteins and other important biological macromolecules. Structural biologists use methods such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy to determine the location of each atom relative to each other in the molecule. They then deposit this information, which is then annotated and publicly released into the archive by the wwPDB.

The constantly-growing PDB is a reflection of the research that is happening in laboratories across the world. This can make it both exciting and challenging to use the database in research and education. Structures are available for many of the proteins and nucleic acids involved in the central processes of life, so you can go to the PDB archive to find structures for ribosomes, oncoproteins, drug targets, and even whole viruses. However, it can be a challenge to find the information that you need, since the PDB archives so many different structures. You will often find multiple structures for a given molecule, or partial structures, or structures that have been modified or inactivated from their native form.

Looking at Structures is designed to help you get started with charting a path through this material, and help you avoid a few common pitfalls. These chapters are intertwined with one another. To begin, select a topic from the right menu, or select a topic from below:

- **PDB Data**
  The primary information stored in the PDB archive consists of coordinate files for biological molecules. These files list the atoms in each protein, and their 3D location in space. These files are available in several formats (PDB, mmCIF, XML). A typical PDB formatted file includes a large "header" section of text that summarizes the protein, citation information, and the details of the structure solution, followed by the sequence and a long list of the atoms and their coordinates. The archive also contains the experimental observations that are used to determine these atomic coordinates.

- **Visualizing Structures**
  While you can view PDB files directly using a text editor, it is often most useful to use a browsing or visualization program to look at them. Online tools, such as the ones on the RCSB PDB website, allow you to search and explore the information under the PDB header, including information on experimental methods and the chemistry and biology of the protein. Once you have found the PDB entries that you are interested in, you may use visualization programs to allow you to read in the PDB file, display the protein structure on your computer, and create custom pictures of it. These programs also often include analysis tools that allow you to measure distances and bond angles, and identify interesting structural features.

- **Reading Coordinate Files**
  When you start exploring the structures in the PDB archive, you will need to know a few things about the coordinate files. In a typical entry, you will find a diverse mixture of biological molecules, small molecules, ions, and water. Often, you can use the names and chain IDs to help sort these out. In structures determined from crystallography, atoms are annotated with temperature factors that describe their...
vibration and occupancies that show if they are seen in several conformations. NMR structures often include several different models of the molecule.

- Potential Challenges

You may run into several challenges as you explore the PDB archive. For example, many structures, particular those determined by crystallography, only include information about part of the **functional biological assembly**. Fortunately the PDB can help with this. Also, many PDB entries are **missing portions of the molecule** that were not observed in the experiment. These include structures that include only alpha carbon positions, structures with missing loops, structures of individual domains, or subunits from a larger molecule. In addition, most of the crystallographic structure entries do not have information on hydrogen atoms.

Except where noted, this feature is written and illustrated by David S. Goodsell.
Looking at Structures: Methods for Determining Atomic Structures

Several methods are currently used to determine the structure of a protein, including X-ray crystallography, NMR spectroscopy, and electron microscopy. Each method has advantages and disadvantages. In each of these methods, the scientist uses many pieces of information to create the final atomic model. Primarily, the scientist has some kind of experimental data about the structure of the molecule. For X-ray crystallography, this is the X-ray diffraction pattern. For NMR spectroscopy, it is information on the local conformation and distance between atoms that are close to one another. In electron microscopy, it is an image of the overall shape of the molecule.

In most cases, this experimental information is not sufficient to build an atomic model from scratch. Additional knowledge about the molecular structure must be added. For instance, we often already know the sequence of amino acids in a protein, and we know the preferred geometry of atoms in a typical protein (for example, the bond lengths and bond angles). This information allows the scientist to build a model that is consistent with both the experimental data and the expected composition and geometry of the molecule.

When looking at PDB entries, it is always good to be a bit critical. Keep in mind that the structures in the PDB archive are determined using a balanced mixture of experimental observation and knowledge-based modeling. It often pays to take a little extra time to confirm for yourself that the experimental evidence for a particular structure supports the model as represented and the scientific conclusions based on the model.

X-ray Crystallography

Most of the structures included in the PDB archive were determined using X-ray crystallography. For this method, the protein is purified and crystallized, then subjected to an intense beam of X-rays. The proteins in the crystal diffract the X-ray beam into one or another characteristic pattern of spots, which are then analyzed (with some tricky methods to determine the phase of the X-ray wave in each spot) to determine the distribution of electrons in the protein. The resulting map of the electron density is then interpreted to determine the location of each atom. The PDB archive contains two types of data for crystal structures. The coordinate files include atomic positions for the final model of the structure, and the data files include the structure factors (the intensity and phase of the X-ray spots in the diffraction pattern) from the structure determination. You can create an image of the electron density map using tools like the Astex viewer, which is available through a link on the Structure Summary page.

X-ray crystallography can provide very detailed atomic information, showing every atom in a protein or nucleic acid along with atomic details of ligands, inhibitors, ions, and other molecules that are incorporated into the crystal. However, the process of crystallization is difficult and can impose limitations on the types of proteins that may be studied by this method. For example, X-ray crystallography is an excellent method for determining the structures of rigid proteins that form nice, ordered crystals. Flexible proteins, on the other hand, are far more difficult to study by this method because crystallography relies on having many, many molecules aligned in exactly the same orientation, like a repeated pattern in wallpaper. Flexible portions of protein will often be invisible in crystallographic electron density maps, since their electron density will
be smeared over a large space. This is described in more detail on the page about missing coordinates.

Biological molecule crystals are finicky: some form perfect, well-ordered crystals and others form only poor crystals. The accuracy of the atomic structure that is determined depends on the quality of these crystals. In perfect crystals, we have far more confidence that the atomic structure correctly reflects the structure of the protein. Two important measures of the accuracy of a crystallographic structure are its resolution, which measures the amount of detail that may be seen in the experimental data, and the R-value, which measures how well the atomic model is supported by the experimental data found in the structure factor file.

For a lively tutorial on how to look critically at atomic structures, see Gerard Kleywegt's practical on Model Validation.

**NMR Spectroscopy**

NMR spectroscopy may be used to determine the structure of proteins. The protein is purified, placed in a strong magnetic field, and then probed with radio waves. A distinctive set of observed resonances may be analyzed to give a list of atomic nuclei that are close to one another, and to characterize the local conformation of atoms that are bonded together. This list of restraints is then used to build a model of the protein that shows the location of each atom. The technique is currently limited to small or medium proteins, since large proteins present
Some of the restraints used to solve the structure of a small monomeric hemoglobin are shown here, using software from the BioMagResBank\(^1\). The protein (1vre and 1vrf) is shown in green, and restraints are shown in yellow.

A major advantage of NMR spectroscopy is that it provides information on proteins in solution, as opposed to those locked in a crystal or bound to a microscope grid, and thus, NMR spectroscopy is the premier method for studying the atomic structures of flexible proteins. A typical NMR structure will include an ensemble of protein structures, all of which are consistent with the observed list of experimental restraints. The structures in this ensemble will be very similar to each other in regions with strong restraints, and very different in less constrained portions of the chain. Presumably, these areas with fewer restraints are the flexible parts of the molecule, and thus do not give a strong signal in the experiment.

In the PDB archive, you will typically find two types of coordinate entries for NMR structures. The first includes the full ensemble from the structural determination, with each structure designated as a separate model. The second type of entry is a minimized average structure. These files attempt to capture the average properties of the molecule based on the different observations in the ensemble. You can also find a list of restraints that were determined by the NMR experiment. These include things like hydrogen bonds and disulfide linkages, distances between hydrogen atoms that are close to one another, and restraints on the local conformation and stereochemistry of the chain.
The tail of the T4 bacteriophage has been examined by combining electron microscopy and atomic structures. The image shows a surface rendering of the EM data (emd-1048) with atomic coordinates from PDB entries 1pdf, 1pdi, 1pdl, 1pdm, 1pdp, and 2f18.

Electron Microscopy

Electron microscopy is also used to determine structures of large macromolecular complexes. A beam of electrons is used to image the molecule directly. Several tricks are used to obtain 3D images. If the proteins can be coaxed into forming small crystals or if they pack symmetrically in a membrane, electron diffraction can be used to generate a 3D density map, using methods similar to X-ray diffraction. If the molecule is very symmetrical, such as in virus capsids, many separate images may be taken, providing a number of different views. These views are then aligned and averaged to extract 3D information. Electron tomography, on the other hand, obtains many views by rotating a single specimen and taking several electron micrographs. These views are then processed to give the 3D information.

For a few particularly well-behaved systems, electron diffraction produces atomic-level data, but typically, electron micrographic experiments do not allow the researcher to see each atom. Electron micrographic studies often combine information from X-ray crystallography or NMR spectroscopy to sort out the atomic details. Atomic structures are docked into the electron density map to yield a model of the complex. This has proven very useful for multimolecular structures such as complexes of ribosomes, tRNA and protein factors, and muscle actomyosin structures.
Looking at Structures: Resolution

Resolution is a measure of the quality of the data that has been collected on the crystal containing the protein or nucleic acid. If all of the proteins in the crystal are aligned in an identical way, forming a very perfect crystal, then all of the proteins will scatter X-rays the same way, and the diffraction pattern will show the fine details of crystal. On the other hand, if the proteins in the crystal are all slightly different, due to local flexibility or motion, the diffraction pattern will not contain as much fine information. So resolution is a measure of the level of detail present in the diffraction pattern and the level of detail that will be seen when the electron density map is calculated. High-resolution structures, with resolution values of 1 Å or so, are highly ordered and it is easy to see every atom in the electron density map. Lower resolution structures, with resolution of 3 Å or higher, show only the basic contours of the protein chain, and the atomic structure must be inferred. Most crystallographic-defined structures of proteins fall in between these two extremes. As a general rule of thumb, we have more confidence in the location of atoms in structures with resolution values that are small, called "high-resolution structures".

Electron density maps for structures with a range of resolutions are shown. The first three show tyrosine 103 from myoglobin, from entries 1a6m (1.0 Å resolution), 106m (2.0 Å resolution), and 108m (2.7 Å resolution). The final example shows tyrosine 120 from hemoglobin (chain b), from entry 160h (3.0 Å resolution). In the pictures, the blue and yellow contours surround regions of high electron density, and the atomic model is shown with sticks. The electron density was imaged using the Astex viewer that is available on the Structure Summary page for each entry (just click the "IDS" link in the "Experimental Method" field).