

Intro to PDB Module

Purpose: To introduce students interested in the life sciences to bioinformatic and computational tools which are frequently used throughout many fields of Biology

Goal: To develop a strong understanding of what PDB is capable of and when it may be used

With the increasing dependence on computational power in the world of Biology, it is more important now than ever to understand the vast number of **bioinformatic** tools available. From annotating a **genome** to developing entire simulations of protein binding, the pool of opportunities for biological discoveries has grown infinitely large with the continued development of computational methods. Over the course of this module, you will explore one of the most common bioinformatic tools which is readily available to any scientist.

What is bioinformatics?

What is computational biology and how is it different from bioinformatics?

What is a genome?

Protein Data Bank (PDB)

The PDB is an open access resource full of 3D structure data for over 180,000 different macromolecules.

1. Navigate to the following webpage: <https://www.rcsb.org/>
2. Hover over the dropdown menu at the top of the page labelled "About"
3. Click on "About RCSB PDB"

Why is it important for a database of macromolecule structural data to exist?

Briefly describe the vision statement of the PDB.

About how much data is stored in the PDB?

- At the top of the page, type “1GFL” into the search bar and click the magnifying glass button to the right of the search bar

You should now see the following page:

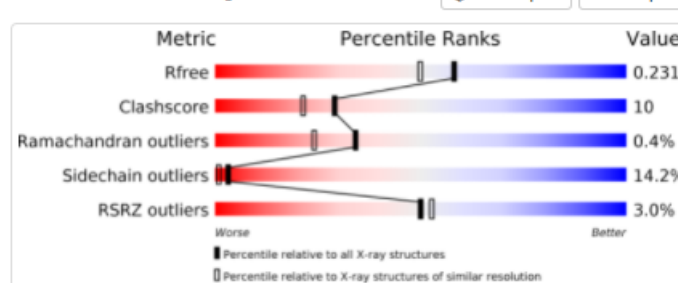
The screenshot shows the PDB 1GFL page. At the top, there are logos for PDB-101, PDB, EMDatabank, and others. Below the logos is a navigation bar with tabs: Structure Summary, 3D View, Annotations, Experiment, Sequence, Genome, and Versions. The main content area is divided into two columns. The left column shows a 3D ribbon diagram of the protein structure, labeled 'Biological Assembly 1'. The right column contains the following information:
1GFL
STRUCTURE OF GREEN FLUORESCENT PROTEIN
DOI: 10.2210/pdb1GFL/pdb
Classification: FLUORESCENT PROTEIN
Organism(s): Aequorea victoria
Expression System: Escherichia coli
Mutation(s): No
Deposited: 1996-08-23 Released: 1997-01-11
Deposition Author(s): Yang, F., Moss, L.G., Phillips Jr., G.N.
Experimental Data Snapshot
Method: X-RAY DIFFRACTION
Resolution: 1.90 Å
R-Value Free: 0.262
R-Value Work: 0.214
R-Value Observed: 0.214
wwPDB Validation
3D Report Full Report
Metric Percentile Ranks Value
Rfree 0.231
Clashscore 10
Ramachandran outliers 0.4%
Sidechain outliers 14.2%
RSRZ outliers 3.0%

From this initial “Structural Summary” page, we can retrieve some very important information about this protein. First, we see that this protein is a green fluorescent protein (often denoted GFP) and that its overall classification is fluorescent protein. In addition, we see that this specific protein was extracted from the organism *Aequorea victoria*. On the left we can see a 3D representation of this protein. Let us begin by focusing on the “Experimental Data Snapshot” portion of this summary:

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 1.90 Å
R-Value Free: 0.262
R-Value Work: 0.214
R-Value Observed: 0.214

wwPDB Validation



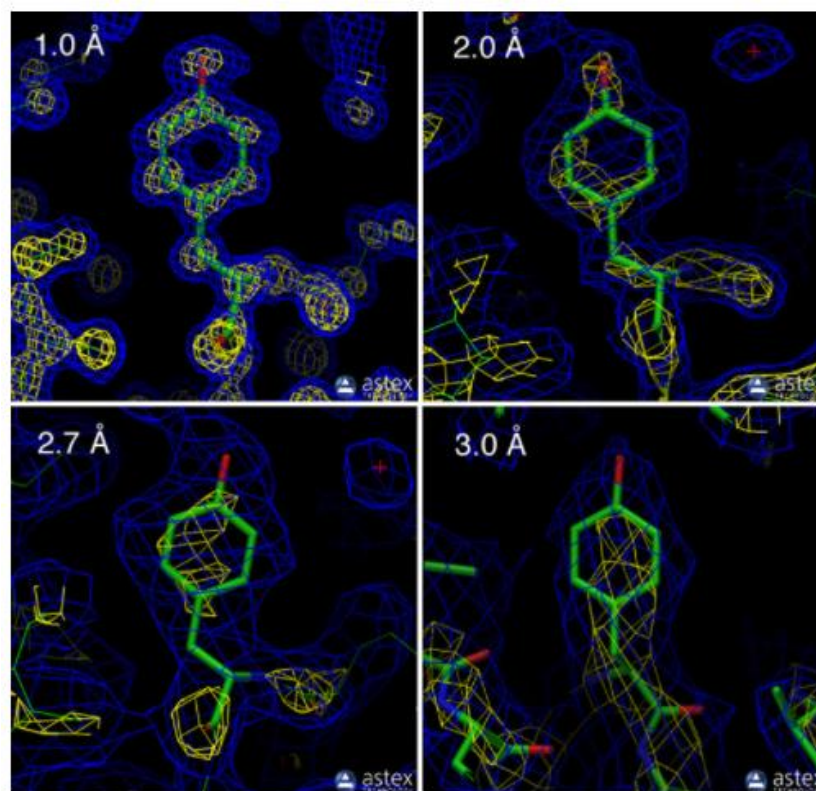
This section allows us to assess the quality of the structure we are viewing. We will be focusing on the data provided on the left of the above image.

Method

Provides the specific method that was used to capture the protein crystal structure. In this case, the crystal structure of this GFP was found using X-Ray Diffraction where X-Rays are fired at a crystallized protein and the diffraction of these waves as they bounce off the crystal is captured to produce a structural model of the protein.

Resolution

A measure of the quality of data obtained from the crystallized structure. A high resolution (low numerical value) is commonly associated with a nearly perfect crystal free of excess motion whereas a low resolution (high numerical value) represents a more disordered crystal. As can be seen in the below image, a resolution of 1.0 Å allows us to draw conclusions about a structure with less guesswork than is needed with a resolution of 3.0 Å (the symbol Å stands for angstrom which is a unit of length measuring one-ten billionth of a meter!). The resolution of our GFP is 1.90 Å.



R-Value Free, R-Value Work, R-Value Observed

When a scientist creates a model for the structure of a protein, this model is tested against a simulation to determine how close the predicted and simulated models are to each other. The R-Value is a measure of how similar the two are. An R-Value of 0.0 would mean that no differences were observed between the model and the simulation and that the model was perfect. An R-Value of around 0.7 would indicate to us that the model and the simulation did not agree well at all and that the quality of that specific model is low. An R-Value of around 0.2 is accepted to be typical. So now, what do we mean by “Free” and why is this value different? Well, since one goal of structure prediction and model construction is to have the predicted

model and computational simulation agree with each other it is very easy for scientists to be biased (whether accidentally or otherwise). The “Free” R-Value helps to remove some of this bias by decreasing the amount of experimental data used when simulating. This way, there is less of a chance of a model and simulation agreeing by chance.

Is this a quality 3D structure of this GFP given the data in the “Experimental Data Snapshot”? Why or why not?

Why do you think the R-Value Free is higher than the Observed or Work values?

We now know one way to assess whether we are looking at a quality structure or not. Now, we are going to look at some other useful information provided to us by this entry.

5. Scroll down on the “Structural Summary” page until you see the following

Macromolecule Content

- Total Structure Weight: 53.78 kDa ⓘ
- Atom Count: 3950 ⓘ
- Modelled Residue Count: 460 ⓘ
- Deposited Residue Count: 476 ⓘ
- Unique protein chains: 1

This box provides us with some basic structural information about the structural model for this specific protein. We are first provided with the total weight of the structure (measured in kiloDaltons which is roughly 1.66×10^{-27} kilograms!). Next, we are given the total number of atoms in the published model. You will notice this number is quite a bit higher than the residue counts provided below it. This is because a residue is a term used for an entire amino acid for example. Therefore, an alanine would count as one residue, but we know that an individual alanine is composed of more than one atom (alanine has 13!). The final piece of information we get from this box is that this model is composed of only one unique polymer. What this tells us is that when we see the two structures in the image on the “Structural Summary” page, we are seeing two of the exact same structure.

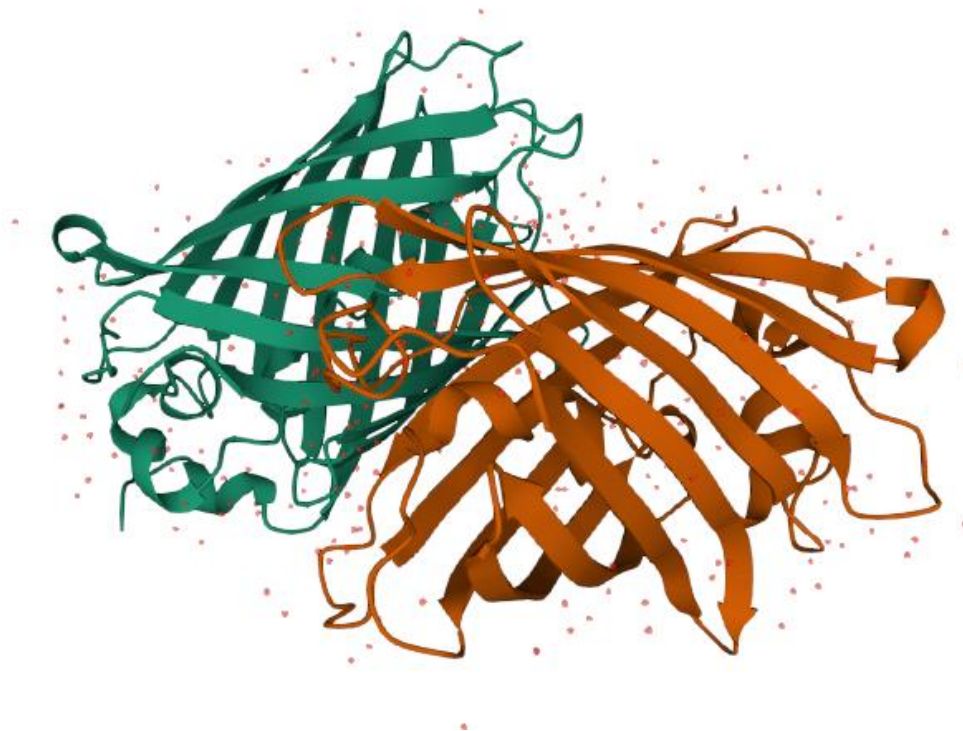
Now, let us look at how we can view and interact with this molecule!

6. Under the image of the protein, next to the bolded “Explore in 3D” text, click on structure

One of the first valuable pieces of information on this page is the protein sequence. This can be found near the top of the page when selecting the “3D View” tab and the sequence of this GFP can be seen below:

Sequence of	1GFL STRUCT...	Chain	1: GREEN FLU...	A	
1	21	41	61	81	101
ASKGEELFTGVVPI	LVELDGVN	GKFSVS	GEGEDATY	GKLT	FKICTTGKLP
VFVWPTLV	TF	SYGVQCF	SRYPDH	MKR	HDFK
SAMPE	STVQERT	IFFKDD	GNVKT	RAEV	
121	141	161	181	201	221
KFEGDTLVNRIEL	KGIDFK	EDGNIL	GKLEYN	YN	SHNVYIM
ADKQ	KNGIKV	NFKIR	HNIED	GSVQ	LADHYQ
QNTPI	GDG	FVLL	PDNHY	L	STQSALS
KDPNE	KRDH	MVLL	EFV		
TAAGITHG	MDELYK				

Right below this sequence panel, we are able to see an interactive version of the structural model for this protein.



First, let us talk about some general conventions you will see used in most structural models of proteins. Each individual protein chain will be colored a different color. So in this case we see two protein chains (remember from the previous section that while there are two presented, they are identical), one in green and one in orange. In addition, you may notice arrows throughout the model. These arrows represent the direction of the protein chain from **N-Terminus** to **C-Terminus**. Next, the dots which you may notice scattered around the model are representative of the **solvent** in which the protein exists. Cells are filled with a variety of solvents (very often water) and thus proteins exist naturally within these solutions. The last thing you may notice is that some regions of the protein chain appear to be organized in a specific pattern. These special regions indicate specific secondary structures, such as the **alpha helix** and **beta sheet**, as well as help to visualize certain tertiary structures, such as a **beta barrel**.

What is the N-Terminus of a protein chain?

What is the C-Terminus of a protein chain?

What is solvent? Provide one specific example of a solvent that may be found within a cell.

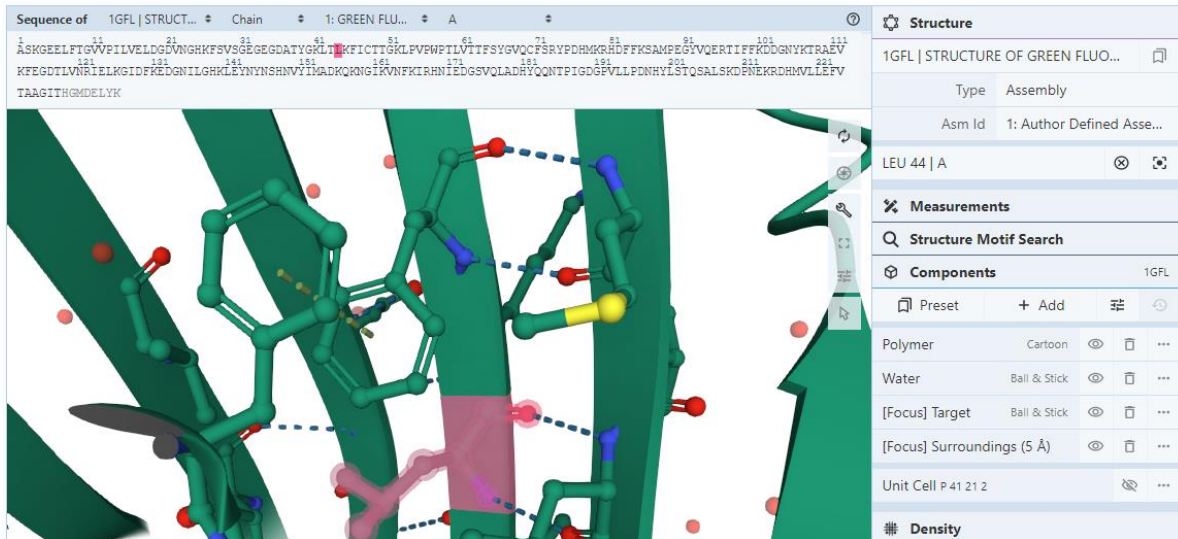
Describe the general structure of an alpha helix.

Describe the general structure of a beta sheet.

Describe the general structure of a beta barrel.

To interact with this model you may click and drag to rotate and scroll in and out to zoom in and out respectively. One final very interesting property of the models on PDB is that you can mouse over an amino acid represented in the sequence and it will highlight its location in the model. This can prove to be an invaluable tool when searching for a specific location in a protein given sequence information.

Let us say however that you want a closer look at where a specific amino acid is located and what its exact position is within the protein. Simply click on the amino acid you are interested and the window will zoom in for you and display the following information to the right:



We can clearly see that the pink highlighted region of the picture screen is the amino acid Leucine and that it is the 44th amino acid from the N-Terminus of this protein!

We have now taken a brief look at some key features of a molecule entry on PDB. While there are other very useful tools provided on this website, we will stop here to recap what we have learned. Let us put everything together and see if we can learn about another protein!

7. Locate the PDB search bar, type in “1TUB”, and click the magnifying glass search button

What protein is depicted in this structural model?

What is the resolution of this structure? Would we consider this to be a high or low resolution structural model?

What is the total weight of this structure?

How many unique protein chains can be found in this structure? What is the name for a protein with this number of unique protein chains?

What method was used to develop the crystalline structure of this protein? Briefly describe how this method works.

What secondary structures are present in this model?

What secondary structure is the 6th amino acid of the alpha subunit (Serine, S) located within?

Compare this protein structure to what you see when you search “3J2U”. Both share a protein, but one has a much lower resolution. Why do you think one of the structures has a lower resolution?
