PDBe: Exploring a Protein Data Bank (PDB) entry

This course will show you how to explore the structural information in a PDB entry using the Protein Data Bank in Europe (PDBe) - EMBL-EBI’s resource for the collection, organisation and dissemination of data on biological macromolecular structures.

This course follows on from PDBe: Searching the Protein Data Bank [2].

An undergraduate level understanding of biology is required for this course. Some bioinformatics knowledge is useful but not essential.

Learning objectives:

- Navigate and extract the information from a particular PDB entry using the PDBe
- Download data in an appropriate format

What information does a PDB entry contain?

A typical PDB [3] entry consists of the 3D coordinates of proteins and/or nucleic acids along with bound molecules (e.g. sugars, lipids, inhibitors, metals etc.) and solvent (e.g. water) molecules.
The contents of each entry are separated into polymers and non-polymers:

- The proteins/DNA/RNA are categorised as polymers whereas water or any other bound small molecules [4] fall in the non-polymer category;
- Polymers or non-polymers of identical chemical composition (e.g. proteins with identical amino acid [5] sequences or sugars with identical chemical formula, bond and stereochemistry) are grouped together to form a distinct chemical entity.

In addition to the 3-dimensional (3D) atomic coordinates [6], a PDB entry can be explored for a variety of information:

- visualising interactive 3D structure;
- secondary structure [7], domains and folds present in the proteins;
- biological assembly or quaternary structures for the proteins and DNA/RNA;
- sequence information for all the proteins and nucleic acids that are present in the entry along with their mapping to UniProt [8] (protein) or GenBank (RNA);
- bound molecules or ligands and their environment;
- source and expression system of the proteins/nucleic acids;
- quality of the structure and experimental information;
- publication information.

All PDB entries are assigned a four character accession [9] code, which always starts with a number, e.g. 1xyz.

In this tutorial, we will use the PDB entries 4KGC [10], 2WS9 [11], 1G20 [12] and 2AFI [13] to explore the entry pages.

**How is the information structured?**

On PDBe webpages, information from a PDB entry is divided into different sections to make the data easier for you to browse (Figure 1). We will refer to them as PDB entry pages throughout this course.
**Figure 1** Schematic diagram of a PDB entry page divided into five major sections.

The image below shows you a snapshot of an individual PDB entry page, structured into five major sections (Figure 2).
Figure 2 A snapshot of a sample PDB entry page. The 'Details' link (highlighted in purple) in the 'Function and Biology', 'Structural analysis' and 'Experimental validation' sections allows you to explore more about specific aspects of a structure.

We will now go through each of these sections (except for 'Experiments and Validation) in more detail.

PDB entry overview

In the PDB entry overview section, you will find some basic information about that specific entry, including the source organism, primary publication and an image gallery [14].
3D visualisation of a PDB entry

On the right hand side of a PDB entry page you will see the ‘Quick links’ menu, which provides links to each of the five sections and enables you to download [15] the structure in several formats and view it in 3D (Figures 3 and 4).

Figure 3 The 3D visualisation link (highlighted in red) helps you to interactively explore the contents of the PDB entry.

Figure 4 An interactive 3D view of the PDB entry 4KGC.

Image gallery

Image gallery (I)

Every PDB [3] entry page comes with a portfolio of images. The images display the contents of the
entry from various chemical, structural and functional perspectives (Figure 5).
Figure 5 Clicking on the image highlighted in red will open up a portfolio of images.

**Image gallery (II)**

These images include:

1) Front, top and side views of the PDB entry coloured by chains - each constituent polymer [16] chain of the entry has a distinct colour (Figure 6).
Figure 6  Front view of the PDB entry 4KGC where every chain has a distinct colour.

**Image gallery (III)**

2) Front, top and side views of the PDB entry coloured by chemically distinct molecules. Protein and nucleic acid chains with identical sequences are coloured the same (Figure 7).
Figure 7 Front view of the PDB entry 4KGC coloured by chemically distinct chains.

Image gallery (IV)

3) Front, top and side views of the PDB entry where one chemically distinct molecule is highlighted at a time (Figure 8).
Figure 8 Front view of the PDB entry 4KGC where a chemically distinct DNA polymer is highlighted in blue.

Citation information

Citation page

The citation page, accessed by clicking on the citation title, shows publications which are related to the entry.

A particular PDB [3] entry can be related to different types of citation (Figure 9).
Figure 9 Different types of citation information can be related to a particular PDB entry.

Figures are shown for open access publications

Both figures and figure legends are displayed for papers published in open access [17] publications.
Figure 10 PDB Entry 4KGC with its primary open access publication - the red box highlights figures from the same publication.

Clicking on the highlighted red box in Figure 10 will provide an image gallery display of the figures and legends from the corresponding publication (Figure 11).

Figure 11 Snapshot of a figure legend from an open access publication.

Function and Biology

What is the function of the protein?

The 'Function and Biology' section of the entry highlights the:

- biochemical function;
- biological process;
- cellular component;
- sequence;
- domain and structural domain annotations of the proteins.

If a protein functions as an enzyme then the Enzyme Class information ([EC number](#)) and the reaction that it catalyses are also shown (Figure 12).
Figure 12 The 'Function and Biology' section highlights key information about proteins from the corresponding PDB entry.

Clicking on the highlighted 'Details' link (Figure 12) will give you an in depth analysis of the functional and biological processes associated with the proteins present in that entry (Figure 12). It will also help you answer the questions that appear in the next section of this course.

Further exploring the function and biology of the protein

Is the protein an enzyme?

If yes, then what type of reaction does it catalyse? Information about the enzyme-catalysed reactions are obtained from the ExPASy [19] bioinformatics resource portal and are shown at the top of the page (Figure 13).
**EC 1.18.6.1: Nitrogenase**

Reaction catalysed:

\[ 8 \text{ reduced ferredoxin} + 8 \text{ H}(+) + N(2) + 16 \text{ ATP} + 16 \text{ H}(2)O = 8 \text{ oxidized ferredoxin} + H(2) + 2 \text{ NH}(3) + 16 \text{ ADP} + 16 \text{ phosphate}. \]

Comments:

- Composed of two proteins that can be separated but are both required for nitrogenase activity.
- Dinitrogen reductase is a [4Fe-4S] protein, which, with two molecules of ATP and ferredoxin, generates an electron.
- The electron is transferred to the other protein, dinitrogenase (molybdoferredoxin).
- Dinitrogenase is a molybdenum-iron protein that reduces dinitrogen in three successive two-electron reductions from nitrogen to two molecules of ammonia; the molybdenum may be replaced by vanadium or iron.
- The reduction is initiated by formation of hydrogen in stoichiometric amounts.
- Acetylene is reduced to ethylene (but only very slowly to ethane), azide to nitrogen and ammonia, and cyanide to methane and in the absence of a suitable substrate, hydrogen is slowly formed.
- Ferredoxin may be replaced by flavodoxin (see EC 1.19.6.1).
- Formerly EC 1.18.2.1.

Systematic name:

Reduced ferredoxin: dinitrogen oxidoreductase (ATP-hydrolyzing)

---

Figure 13 Information about the reaction mechanisms of enzymes present in a PDB entry.

**What specific biological processes are associated with the proteins?**

The Gene Ontology [20] (GO) annotations associated with the proteins help us answer the following questions (Figure 14):

1. What are the biochemical functions associated with the protein?
2. Which biological processes are carried out by these proteins?
3. Which cellular component does the protein belong to?
Figure 14 Snapshot of the functional annotation (GO terms) of the proteins present in the PDB entry. Clicking on the links for each function will return a list of PDB entries containing proteins with the same function.

**Which sequence family does the protein belong to?**

The sequence family information for each protein present in the entry is derived from the Pfam [21] and the InterPro [22] databases (Figure 15). The Pfam database provides information about domain [23] architectures present in various protein families.
Figure 15 Pfam domains. PF00142 is highlighted in light green for Nitrogenase iron protein 1 and PF00148 is highlighted in blue and olive green for Nitrogenase molybdenum-iron protein alpha and beta chain respectively.

The InterPro annotation of a protein family

The InterPro [22] database provides information about characteristic signatures present in the proteins which are derived from the protein families, domains and important sites (Figure 16).
What are the structural folds/domains present in the proteins from this entry?

The structural domain information for the protein is derived from the CATH [24] and SCOP [25] databases. In the following example, a CATH alpha beta domain is detected in the Nitrogenase molybdenum-iron protein (Figure 17). There are a further two CATH domains present in this entry (not shown in the figure). The CATH domains are highlighted in blue, green and yellow.

Figure 16 InterPro annotation (IPR005977) showing protein domains present in nitrogenase iron protein 1.
Figure 17 CATH annotations (3.40.50.1980) for signature domains indicate presence of 'Rossmann fold' in Nitrogenase molybdenum-iron protein.

The SCOP classification of protein folds

Similarly, the SCOP [25] annotation [26] for the same protein indicates that it belongs to the SCOP defined class of alpha and beta proteins. The SCOP domains are highlighted in green within the protein (Figure 18).
Structure analysis

The 'Structure analysis' section of the entry overview page provides a quick summary about the macromolecules present in the entry (Figure 19). This section includes:

- type and number of chemically distinct polymers ('Entry contents'). In this example, the entry '4KGC' contains four unique proteins and two unique DNA molecules. The protein histone H4 is one of the four unique proteins present in the entry. There are two copies of this molecule, chains B and F;
- type of assemblies that can be formed by the macromolecules;
- length of each chemically distinct protein/nucleic acid;
- source of each of the chemically distinct macromolecules;
- relevant UniProt [8] accession [9] number for proteins (where sequence information is available in the UniProt);
- sequence domain [23] information as annotated by Pfam [21].
Figure 19 Structure analysis section of PDB entry 4KGC. The 'Entry contents' section reveals the chemically distinct macromolecules present in the entry. You can explore each macromolecule further by clicking on 'Molecule details' link.

Detailed structure analysis

Clicking on the 'Details' link at the top of Figure 19 [27] takes you to the 'Structure analysis' page for the entry. This page provides you with:

- detailed analysis of the macromolecules that are present in the entry along with the assemblies they form;
- further options to explore the detailed 3D structures of all the macromolecules present in the entry interactively.

In the structure analysis pages, all the images are displayed in a context dependent manner. For example, under the 'Macromolecules' section, only images highlighting the relevant macromolecule are displayed (Figure 20).
Clicking on the 'Molecule details' link on the 'Structure analysis' summary/page (Figure 21) loads the 1D, 2D and 3D visualisation option for the relevant macromolecules.

Interactive exploring: What’s new?

Interactive visualisation enables you to view each protein chain in the following layout (Figure 22):

1. One dimensional (1D) layout of the sample sequence (dark green line), the built model (grey line), model build quality (mixture of green and yellow line), secondary structure [7] (purple line), Pfam [21] (red line). Where available CATH [24] annotation [26] and SCOP [25] domain [23] annotation will also be shown;
2. Two dimensional (2D) topology diagram of the protein chain;
3. Three dimensional (3D) visualisation of the same protein using the Jsmol viewer;
4. Clicking on a particular region of the protein chain on any of the 1D/2D/3D diagrams will highlight the same region in all the three layouts.

Figure 22 Interactive exploring of proteins using the 1D, 2D and 3D layout of the protein.

Try it for yourself....

Have a go at interactively exploring the histone H4 protein [28] present in PDB entry 4KGC.

Which amino acids from the protein chain participate in binding interactions with ligands?

Figure 23 below shows you how to explore a ligand binding site in the protein.

- The green dots on the 1D view highlight the position where a ligand interacts with the protein chain.
- Clicking on the protein chain at that position highlights the amino acid (red square box) present in all the three (1D/2D/3D) layouts.
- The zoomed 3D view clearly indicates the amino acid Arginine is in close proximity to the ligand.
Figure 23 Interactive exploring of a ligand binding site present in the protein.

What is the build quality of the protein chain?

Figure 24 provides you with a visual representation of the build quality of the protein chain.

- The build quality of each protein chain is derived from the wwPDB validation report and is displayed in the 1D layout as ‘Quality’.
- The green, yellow, orange and red segments indicate residues with varied number of outliers in the geometry of the molecule. For example, the sections coloured in green have no geometric outliers while yellow, orange and red segments indicate a progressive increase in the number of such outliers.
- The additional red segment above the ‘Quality’ line (if present) indicates residues that have a poor fit to the experimental data.
- The 2D topology diagram for the chain is coloured based on the build quality of the model chain.
Figure 24 The 2D layout is coloured according to the model quality. The model quality is also displayed in the 1D layout - issues with the model quality in the region are highlighted in red.

What are the signature domains/folds present in the protein chain?

Pfam domain

- The Pfam [21] annotation [26] of a particular protein is mapped against the primary sequence of the protein present in the entry.
- The 2D topology diagram provides you with an option to highlight the Pfam domain [23](s) present in a protein chain by selecting ‘Pfam’ from the pulldown menu (Figure 25).
The Pfam domain present in chain C of PDB entry 4KGC is highlighted in its 1D, 2D and 3D layout.

**CATH domain**

- The presence of any [CATH](24) [domain](23) in the protein [secondary structure](7) is indicated in the 1D layout of the protein chain.
- The 2D topology diagram can be coloured by CATH domain by selecting the pulldown option at the bottom of the topology diagram.
- Clicking on the CATH domain information in the 1D layout (circled in red in Figure 26) highlights the presence of the same domain in the 2D and 3D layout.
The CATH domain present in chain H of PDB entry 1G20 is highlighted in its 1D, 2D and 3D layout.

### SCOP domain

- The SCOP [25] classification of a structural fold is also displayed in the 1D/2D/3D layout of the protein chain.
- The 2D topology diagram can also be coloured by SCOP domain [23] by selecting the pull-down option at the bottom of the topology diagram.
- Likewise, clicking on the SCOP domain information in the 1D layout (circled in red in Figure 27) highlights the presence of the same domain in the 2D and 3D layout.
What are the probable quarternary structures and biological assemblies?

The assembly pages provide detailed information about probable quaternary structures and biological assemblies that can be generated using crystallographic symmetry information (Figure 28).

Clicking on the images (highlighted in red) will open up an image gallery where only the relevant assembly images are displayed as shown in Figure 29.
**Figure 28** Generation of virus capsid (highlighted in red) for PDB entry 2WS9.

**Figure 29** Image gallery displaying biological assemblies and probable quaternary structures that can be generated based on crystallographic symmetry information. The virus assembly image here is coloured by chemically distinct molecules.

**Things to consider:**

- The contents of a PDB entry do not always represent the biological assembly.
- For X-ray entries, quaternary structures can be generated by expanding the crystal
symmetry.
- For structures that are not determined by X-ray methods, contents of the PDB entry are likely to represent the biological assembly of the macromolecule.
- It is possible to generate more than one quaternary structure by expanding crystal contacts for a particular PDB entry.
- Some of these quaternary structures can be stable in the solution and can represent the biological assembly.

Unravelling the symmetry of a virus capsid using assembly information

Viruses are special in the PDB [3] because not all the chains that are present in the so called crystallographic asymmetric unit are deposited. Instead, only the minimum numbers of chains that can uniquely describe an icosahedral repeat [30] unit are deposited in the PDB.

- The elements of icosahedral symmetry involve 6 five-fold, 10 three-fold, and 15 two-fold rotation axes.
- Symmetry operations are then used to generate the full viral capsid.
- PDB entry 2WS9 has four chains in the file. But the actual assembly (viral capsid) is a 240-mer.
- The hetero 240-mer assembly in the image displays the full virus capsid for the entry 2WS9.
- The assembly images of the hetero 20-mer and hetero 24-mer reveal the five-fold and the (2 x 3) fold rotation axes of the icosahedral repeat.

An example of unravelling the symmetry of a viral capsid using assembly information is shown in Figure 30.
Figure 30 The entire virus capsid is made up of 240 chains. The PDB entry 2WS9 contains only four chains, but these are sufficient to describe whole virus capsid via the 5-fold (hetero 20-mer) and (2 x 3) fold (hetero 24-mer) rotations of the icosahedral repeat.
Ligands and Environments

The Ligands and Environments section

The 'Ligands and Environments' section lists all the chemically distinct ligands and their binding sites that are present in a particular PDB [3] entry (Figure 31).

7 bound ligands:

- Ca$^{+2}$
- Mg$^{+2}$
- HCA
- CFN
- CLF
- SF4
- ADP

No modified residues

Figure 31 The 'Ligands and Environments' section.

Further information about individual ligands from a PDB entry

Clicking on the individual type of ligands provides further information about the ligand [31] including formula, molecular weight, SMILES [32] string, and also displays images highlighting the particular ligand and its immediate binding environment (Figure 32).
Figure 32 Chemical information about the ligand CLF which is an 8Fe-7S metal cluster. The image highlights the position of the ligand (coloured in blue) in the protein.

The image gallery

The image gallery for the ligand CLF from the entry 2AFI (Figure 33) highlights:

- the position of the ligand (coloured in blue) in the context of the rest of the entry (coloured in grey);
- the binding environment for the ligand and provides you with the 2D image of the ligand.
Figure 33 The image gallery for the ligand CLF from the PDB entry 2AFI.

**Understanding the interactions between the ligand and its environment**

In addition to the 3D views, 2D plots of individual ligand environments are also available (Figure 34). These will help you to distinguish between the covalent and non-covalent interactions.
CLF 5498 bound to chain J

Environment details

Figure 34 2D plots and 3D views of the ligand CLF (8Fe-7S metal cluster).

**Downloading data from a PDB entry page**

**What is available to download?**

There are several different download options and file formats available from a PDB [3] entry page (Figure 35).
**Figure 35** The red box highlights all the data download options available from the front page of a PDB entry.

**Files from the PDB archive:**

- PDB format file;
- mmCIF format file;
- experimental data file;
- validation reports.

**Other files distributed by PDBe [33]**

- updated mmCIF format file where the PDB archive mmCIF format file has undergone some standardisation of vocabularies and connectivity information for every chemical compound present in the PDB entry has been added;
- individual assemblies generated by the PDB entry and distributed in mmCIF format;
- sequence for the proteins/DNA/RNA in FASTA [37] format.

**Additional download options**

Context-dependent download options are available to you when exploring a particular section of the PDB [3] entry page (e.g. assembly information, chain information etc).

**Assembly information**

When exploring the assembly [38] information of a particular entry, you can click on the 'Downloads' link on the assembly page to see all the available download options (Figure 36).
Figure 36 The assembly information page for PDB entry 2WS9 along with the assembly files that are available for download from this page.

Molecule details

When exploring the 'Molecule details' [27] page for a particular polymer [16] entity, the 'Downloads' option allows you to download the sequence of that particular chain along in FASTA [37] format, along with other related information (Figure 37).

Figure 37 The 'Molecule details' page for chain 1 of PDB entry 2WS9.

Summary

- A typical PDB [3] entry contains 3D coordinates of the macromolecular structures that are experimentally derived using X-ray/NMR/electron microscopy [39].
- It also contains information about secondary structure [7], biological assemblies, sequence
mapping information (UniProt [8] for protein and GenBank for RNA) for the proteins and nucleic acids that are present in the entry.

- The PDB entry can be further explored for binding environment of ligands and other small molecules [4].
- The information from a PDB entry is structured into five major sections on the PDBe [33] website to make the data easier to browse.
- Every PDB entry page comes with a portfolio of images [40] - displaying the contents of the entry from various structural, functional and chemical perspectives.
- The citation page [41] not only provides information about primary and associated publications related to the entry, it also displays figures and figure legends for open access [17] publications.
- The 'Structure analysis [27]' section enables you to interactively explore proteins in 1D/2D/3D layout.
- The 'Ligands and Environments [42]' section lists all the chemically distinct ligands that are present in a particular PDB entry and shows their binding sites.
- There are several different download options [15] and file formats present from a PDB entry page. In addition to the PDB archive files, every entry page provides download options for biological assembly files (which can be instantly viewed using any visualisation tool like Chimera, pymol [43] etc.), sequence files and SIFTS [34] information.

Your feedback

Please tell us what you thought about this course. Your feedback is invaluable and helps us to improve our courses and thus enhance your learning experience.

Learn more

Find out more

- PDBe Quips [44] (Quite Interesting PDB Structures) are short interactive articles that explore 'quite interesting' structures from the PDB archive.

Recommended online courses

- PDBe: Searching the Protein Data Bank [2] - this course will show you how to search and navigate the PDBe resource;
- PDBe webinar [45] - an introduction to search and entry pages;
- PDBeFold [46]: Searching for structural homologues of a protein;
- PDBePISA [47]: Identifying and interpreting the likely biological assemblies of a protein structure;
- PDBeChem [48]: Searching for small molecules and small molecule fragments.

Recommended reading

Get help and support on PDBe

EMBL-EBI's PDBe Team [50] develops and maintains the EBI's Protein Data Bank in Europe [51].

Support

- Try the PDBe tutorials [52] to find out more about macromolecular structures and how to navigate the PDBe website.
- For general enquiries about the PDB [3], email the PDB help desk [53].
- For deposition enquiries, email the PDB deposition [54] or Electron Microscopy Data Bank [55] deposition addresses.
- For general enquiries about the Electron Microscopy Data Bank, email the EMDB help desk [56].

Collaborators

PDBe collaborates with the X-ray crystallography [57], Nuclear Magnetic Resonance [58] (NMR) spectroscopy and cryo-Electron Microscopy (EM) communities. To keep abreast of new developments in the NMR community, PDBe has participated in EU projects and continues to contribute to the Collaborative Computational Project for the NMR community [59] (CCPN). PDBe also operates EMDB [60], the international repository for density maps, which are created using high-resolution biological transmission electron microscopy in collaboration with RCSB [61] and Baylor College of Medicine [62]. EMDB contains both macromolecular images and structures reconstructed using the single-particle method and images of sub-cellular regions from electron tomography [63].

Funding

Core support for PDBe is provided by the European Molecular Biology Laboratory [64] and The Wellcome Trust [65]. Specific projects are funded by the European Union [66], the UK Biotechnology and Biosciences Research Council [67](BBSRC) and the National Institutes of Health [68] (NIH).

Contributors

Sanchayita Sen [1]
EMBL-EBI
Project Leader PDB/EMDB curation - Kleywegt team: Protein Data Bank in Europe
Abhik Mukhopadhyay [69]

EMBL-EBI
Scientific Database Curator - Kleywegt team: Protein Data Bank in Europe

Source URL: http://www.ebi.ac.uk/training/online/course/pdbe-exploring-protein-data-bank-pdb-entry

Links
[1] http://www.ebi.ac.uk/training/online/trainers/ssen
[4] http://www.ebi.ac.uk/training/online/glossary/small-molecules
[8] http://www.ebi.ac.uk/training/online/glossary/uniprot
[9] http://www.ebi.ac.uk/training/online/glossary/accession
[12] http://www.pdbe.org/1g20
[16] http://www.ebi.ac.uk/training/online/glossary/polymer
[17] http://www.ebi.ac.uk/training/online/glossary/open-access
[18] http://www.ebi.ac.uk/training/online/glossary/ec-number
[19] http://www.ebi.ac.uk/training/online/glossary/expasy
[21] http://www.ebi.ac.uk/training/online/glossary/pfam
[22] http://www.ebi.ac.uk/training/online/glossary/interpro
[23] http://www.ebi.ac.uk/training/online/glossary/domain
[26] http://www.ebi.ac.uk/training/online/glossary/annotation
[28] http://www.ebi.ac.uk/pdbe/entry/pdb/4kgc/protein/2
[29] http://www.ebi.ac.uk/training/online/glossary/crystallographic-symmetry
[31] http://www.ebi.ac.uk/training/online/glossary/ligand
[32] http://www.ebi.ac.uk/training/online/glossary/smiles
[33] http://www.ebi.ac.uk/training/online/glossary/pdbe
[34] http://www.ebi.ac.uk/training/online/glossary/sifts