An introduction to the Protein Data Bank

Matthew Conroy
Protein Data Bank in Europe
PDBe.org; conroy@ebi.ac.uk

This talk will give an overview of the Protein Data Bank, what data are available, and how to assess its quality. In addition, how the wider community use PDB data will be reviewed

History

The Protein Data Bank (PDB) is the first open access digital archive in biology having been formed in 1971 (Protein Data Bank *Nature New Biol.* **233**, 223 (1971)). It is the single worldwide repository for experimentally determined structures of biological macromolecules and their complexes. Beginning with fewer than ten structures, and predating the internet, data was distributed by mail on magnetic tape. Forty five years later, the PDB contains over 140000 structures of more than 42000 unique biological macromolecules (Fig. 1).

The structures contained in the PDB are not limited to those of proteins. Structures of proteins, nucleic acids and complexes of these are present. Around 75% of structures contain at least one small molecule ligand.

The vast majority of structures are determined by X-ray crystallography, but around 10% of the data are derived from nuclear magnetic resonance experiments. In 1991, the PDB saw its first deposition of a structure solved by electron microscopy (EM). While the overall fraction of such structures in the archive remains small (less than 1% of the archive), recent advances in the EM field mean that this proportion is rising. 3% of structures released in the first half of 2016 were solved by EM.

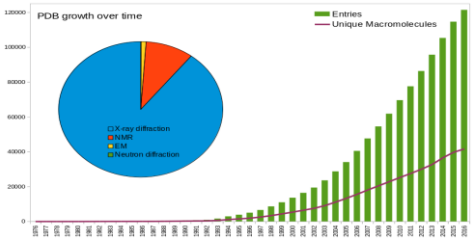


Figure 1. The growth of the PDB archive over its history and the content by experimental method.

In addition to atomic coordinates, experimental data must also be deposited to the PDB to support the model and enable its validation. For diffraction studies, these data are the structure factors and for NMR, chemical shifts and restraints. It is now mandatory for the map to be deposited in EMDB (Tagari *et al* 2002, Lawson *et al* 2016) for structures determined by electron microscopy.

Originally a collaboration between Brookhaven National Laboratory and CCDC in Cambridge, today, the PDB archive is managed by the Worldwide Protein Data Bank organization (wwPDB; http://wwpdb.org) (Berman *et al* 2003), which currently includes three founding regional data centers, located in the US (RCSB Protein Data Bank or RCSB PDB; http://rcsb.org), Japan (Protein Data Bank Japan or PDBj; http://pdbj.org), and Europe (Protein Data Bank in Europe or PDBe; http://pdbe.org), plus a global NMR specialist data repository BioMagResBank, composed of deposition sites in the US (BMRB; http://www.bmrb.wisc.edu) and Japan (PDBj-BMRB; http://bmrbdep.pdbj.org). Together,

these wwPDB partners collect, annotate, validate, and disseminate standardized PDB data to the public without any limitations on its use.

Deposition of atomic models and experimental data is via a web-based interface with depositions organised on a geographic basis (Fig. 2)

Protein Data Bank in Europe (pdbe.org):	Europe & Africa
Protein Data Bank Japan (pdbj.org):	Asia
RCSB PDB (rcsb.org):	The Americas & Oceania

All sites release data into the central FTP repository ("The PDB") which users can access from each of these partner sites. The sites compete in the way data is presented and services offered. Since its inception, data in the PDB archive have been freely accessible to the user community.

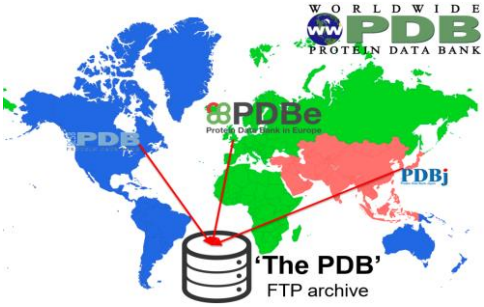


Figure 2. Data is deposited to the PDB via three partner sites based on geographical location

PDBe has a simple, easy to use search interface which enables searching via a variety of fields and subsequent faceting of the results. Results can be viewed in an 'entry-specific' manner or organised by other criteria such as unique macromolecules. Results can be sorted by quality, so that the best representative structure in a search result can be identified. (Velankar *et al* 2016). Given the obviously visual nature of macromolecular data, PDBe uses many images and online viewers to present the structures to users. Most recently, we have implemented an online 3D viewer incorporating electron density (Fig. 3)

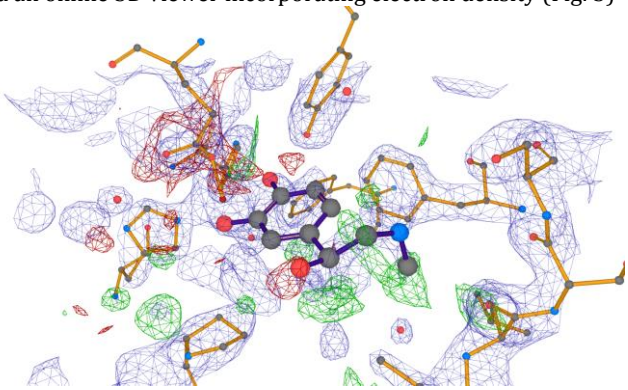


Figure3. Screen shot from PDB entry 3pah at PDBe (pdbe.org/3pah, Erlandsen et al 1998) showing electron density (2Fo-Fc and Fo-Fc) for the bound ligand.

Format

For many years, data was distributed in PDB format. This ASCII text file was both human and computer readable, but, with origins in the punched card data storage of the 1970s it

struggles to cope with the size and complexity of macromolecular structure data generated today. For example, the maximum number of atoms which can be stored in a PDB file is 99999. This number was improbably high to crystallographers of the era, but is routinely exceeded today! Now, PDBx format is the master format for the archive, a format extended from mmCIF (Westbrook *et al* 2005) This format has none of the size limitations inherent in PDB format and is widely extensible.

Atomic coordinate data is arranged in a hierarchy which is implicit by the polymeric nature of the molecules which the PDB archives: atoms form residues which are linked to form chains. Identical chains are then classified together as an entity. There are currently around 7000 data items which can be used to capture the experimental metadata, for example, crystallisation conditions, field strength of NMR magnet used, species from which the protein originates or quaternary structure of the complex.

Small molecules in the PDB

Small molecules in an entry are queried against the PDB chemical components dictionary and matched such that all instances of a given molecule are represented in the same way. For small molecules new to the PDB, a dictionary is created including machine-readable chemical descriptors, such as SMILES strings, InChi and InChi keys (Sen *et al* 2014). Small molecules are validated using Mogul from CCDC.

These metadata enable small molecules to be linked to other chemical resources such as ChEMBL (ebi.ac.uk/chembl/), DrugBank (drugbank.ca) or ChEBI (ebi.ac.uk/chebi/). For small molecules which are present in both the PDB and Cambridge Structural Database (CSD) from CCDC (Groom *et al* 2016), the corresponding CSD data are available via the respective FTP areas of wwPDB member organisations.

Other PDB metadata

Much of the metadata for an entry is harvested during data deposition or added during the curation process. Cross links are added to other data resources such as PubMed, UniProtKB, the NCBI taxonomy or EMDB, to capture the biological, chemical and structural context of the entry. For structures determined by X-ray crystallography, convention dictates that only the asymmetric unit is deposited in the PDB, however this may not represent the true quaternary structure of the complex. For instance PDB entry 2dgc (pdbe.org/2dgc; Keller *et al* 1995) contains one protein and one DNA chain in the ASU, but by applying symettry one can see that the quaternary structure is twice this. (see Fig. 4). Assembly information is therefore added to the entry at the time of annotation. External data resources sometimes change their accessions so the SIFTS project (Velankar *et al* 2013) exists to maintain up-to-date cross references between PDB and UniProt data resources.

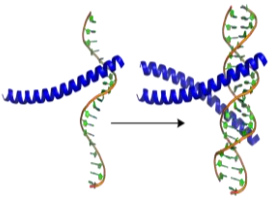


Figure 4. The asymmetric unit of PDB entry 2dgc and the quaternary structure

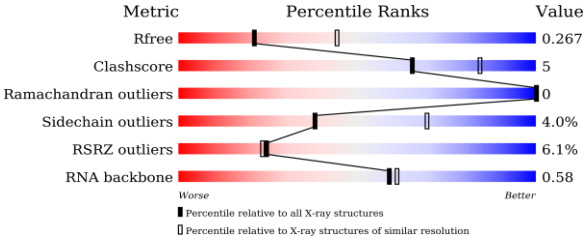
Validation

The PDB as an archive does not reject a structure based on its quality. However, following the recommendations of community 'task forces' with different specialisms (Read *et al* 2011, Henderson *et al* 2012, Montelione *et al* 2013, Adams *et al* 2016) validation reports are produced for all entries. The validation for X-ray entries is the most established (Gore *et al* 2012) but this validation, and that for other methodologies, continues to evolve.

In addition to a summary graphic indicating the quality of a structure relative to others in the archive (Fig. 5), a detailed report listing individual outliers is available in both PDF and

xml format. The latter is used to display outliers in the structure in viewers such as Coot and PyMOL (via PDBe's plugin) and on PDBe's website in the LiteMol viewer (Fig 5).

These reports are available for publicly released data but are also supplied at the time of data deposition and the wwPDB recommends that they are made available to referees. A stand-alone validation server is available so that the quality of a structure can be assessed during the refinement process.



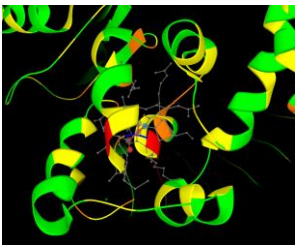


Figure 5. Left: Example of the validation summary produced by wwPDB. The quality of a structure relative to all structures in the PDB solved by X-ray diffraction, and relative to those at a similar resolution is indicated. Both geometric and fit-to-density metrics are displayed.
Right: Detail of a ribbon diagram of a protein coloured according to geometric outliers, from green (no outliers), through to red (three or more outliers for a given residue).

PDB as a resource

It was apparent relatively early in the field of structural biology (eg Rao & Rossmann 1973) that the 3 dimensional folds of proteins are conserved. Several databases such as CATH (Sillitoe *et al* 2015) and SCOP (Andreeva *et al* 2004) classify proteins based on their three dimensional fold. Several servers such as PDBeFold (pdbe.org/fold; Krissinel & Henrick 2004) exist which can search the PDB archive based purely on the 3D structure of a protein. This often reveals surprising similarities and relationships between proteins of disparate sequences (Fig. 6) and can shed light on protein function.

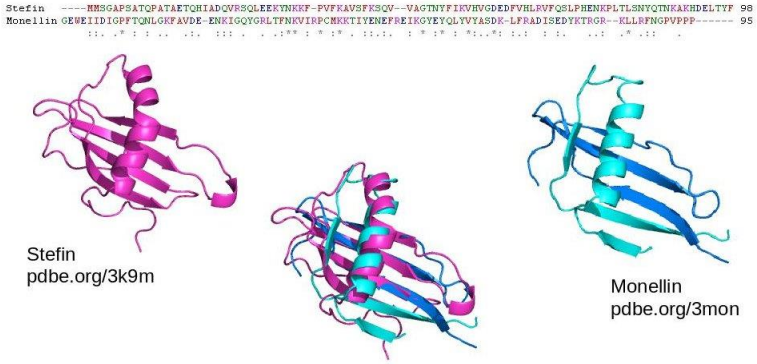


Figure 6. The sequences for two proteins, Stefin, a protease inhibitor, and Monellin, which is intensely sweet, show little similarity. However their structures (Stefin in pink and monellin in blue) overlay extremely well.

Mining the PDB enables structure prediction servers such as PHYRE2 (Kelley *et al* 2015) which uses homology detection methods to build 3D models, or resources such as Genome3D (Lewis *et al* 2015) to enable structural annotations to protein sequences. Working in the opposite direction, ProFunc (Laskowski *et al* 2005) helps identify the likely biochemical function of a protein from its three-dimensional structure. We will look at such resources in the lecture.

Challenges for the future

Structural biology is a rapidly evolving field which presents challenges for data archiving. Ever larger quantities of data can be collected and processed in shorter spaces of time in an automated manner (eg Schiebel_*et al* 2016). This presents challenges for data deposition and processing.

More and more data in structural biology is being collected by so called 'hybrid methods' where multiple techniques are utilised to obtain the structure of a macromolecule. This presents a series of challenges for archiving and validating such structures which the wwPDB are actively addressing (Sali *et al* 2014)

References

Adams *et al* (2016) *Structure* **24** 502
Andreeva *et al* (2004) *Nucleic Acids Res.* **32** D226
Berman *et al* (2003) *Nat Struct Biol.***10** 980
Erlandsen *et al (1998) Biochemistry* **37** 15638
Gore *et al* (2012) *Acta Cryst. D* 68 **478** 83
Groom *et al* (2016) Acta Cryst B **72** 171
Henderson *et al* (2012) Structure **20** 205
Keller *et al* (1995) *J. Mol. Biol.* **254** 657
Kelley *et al* (2015) *Nat Protoc.* **10** 845
Krissinel & Henrick (2004) *Acta Cryst.* D **60**, 2256
Laskowski *et al* (2005). *Nucleic Acids Res.* **33** W89
Lawson *et al* (2016) *Nucleic Acids Res.* **44** D396
Lewis *et al* (2015) *Nucleic Acids Res.* **43** D382
Montelione *et al* (2013) *Structure* **21** 1563
Rao & Rossmann (1973) *J. Mol. Biol.* **76** 241
Read *et al* (2011) *Structure* **19** 1395
Sali *et al* (2014) *Structure* **23** 1156
Schiebel_*et al* (2016) Structure **24** 1398
Sen *et al* (2014) *Database* bau116
Sillitoe *et al* (2015) *Nucleic Acids Res.* **43** D376
Tagari *et al* (2002) *Trends Biochem. Sci.* **27** 589
Velankar *et al* (2013) *Nucleic Acids Res.* **41** D483
Velankar *et al* (2016) *Nucleic Acids Res.* **44** D385
Westbrook *et al* (2005) *International Tables for Crystallography p 195.*