X-tal Raw Data Archive (XRDa):
A crystallographic raw diffraction image archive in Asia

Gert-Jan Bekker, *Genji Kurisu
Institute for Protein Research, Osaka University
In 1971:

**CRYSTALLOGRAPHY**

**Protein Data Bank**

A repository system for protein crystallographic data will be operated jointly by the Crystallographic Data Centre, Cambridge, and the Brookhaven National Laboratory. The system will be responsible for storing atomic coordinates, structure factors and electron density maps and will make these data available on request. Distribution will be on magnetic tape in machine-readable form whenever possible. There will be no charge for the service other than handling costs. Files


Cytochrome c @4.5 Å in 1971 and @2.3 Å in 1975, which was the 1st Asian and 21st PDB entry in the entire archive.

*J. Biochem. (Tokyo)* **70**, 913-924 (1971)

*J. Biochem. (Tokyo)* **77**, 147-162 (1975)
Early Data-out services of PDB in Japan

The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures

The Protein Data Bank is a computer-based archival file for macromolecular structure. The Bank stores a uniform format atomic coordinates and partial bond connectivities, as derived from crystallographic studies. Text included in each data entry gives pertinent information for the structure at hand (e.g. species from which the molecule has been obtained, resolution of diffraction data, literature citations and specifications of secondary structure). In addition to atomic co-ordinates and connectivities, the Protein Data Bank stores structure factors and phase, although these latter data are not placed in any uniform format. The Bank also contains information, such as the BioMagResBank (BMRB) NMR structural database (http://www.bmrb.wisc.edu/). Table 24.1.3.1 gives a summary of the contents of the PDB archives.

Table 24.1.3.1. PDB mirror sites as of May 1999

<table>
<thead>
<tr>
<th>Country</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>University of San Luis</td>
</tr>
<tr>
<td>Australia</td>
<td>Australian National Genomic Information Service, Sydney; The Walter and Eliza Hall Institute of Medical Research, Melbourne</td>
</tr>
<tr>
<td>Brazil</td>
<td>ICB UFMG, Inst. de Ciencias Biologicas, Univ. Federal de Minas Gerais</td>
</tr>
<tr>
<td>China</td>
<td>Institute of Physical Chemistry, Peking University, Beijing</td>
</tr>
<tr>
<td>France</td>
<td>Institut de Génétique Humaine, Montpellier</td>
</tr>
<tr>
<td>Germany</td>
<td>GMD, German National Research Center for Information Technology, Saarbrücken</td>
</tr>
<tr>
<td>India</td>
<td>Bioinformatics Centre, University of Pune</td>
</tr>
<tr>
<td>Israel</td>
<td>Weizmann Institute of Science, Rehovot</td>
</tr>
<tr>
<td>Japan</td>
<td>Institute of Protein Research, Osaka University</td>
</tr>
<tr>
<td>Poland</td>
<td>ICM Interdisciplinary Centre for Modelling, Warsaw University</td>
</tr>
<tr>
<td>Taiwan</td>
<td>National Tsing Hua University, HsinChu</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>Cambridge Crystallographic Data Centre, Cambridge; EMBL Outstation, EBI, Hinxton</td>
</tr>
<tr>
<td>United States</td>
<td>Bio Molecular Engineering Research Center, Boston University, North Carolina Supercomputing Center, Research Triangle Park, University of Georgia, Athens, Georgia; PDB at Brookhaven National Laboratory</td>
</tr>
</tbody>
</table>

In the same period, the proliferation and increasing power of computers, the introduction of relatively inexpensive interactive display systems, and the advent of the World Wide Web have helped to increase the efficiency of data delivery. The PDB first appeared on floppy disks and magnetic tape. In 1990, archival copies of the data bank were made available on CD-ROM and the Internet. This page was retrieved automatically, and the molecular structures can be displayed using the public-domain molecular viewer PDB SHELL (Sussman, 1997), which is interfaced through web browsers such as Netscape Communicator and Internet Explorer. It’s a simple example of its use is illustrated in Fig. 24.1.3.1 in a journal article.

Magnetic Tape ➔ CD-ROM ➔ Internet (BNL PDB mirror site @Osaka 1998)
PDBj was founded in 2000

Data-in and Data-out services in Japan.

https://pdbj.org. in English, Japanese, Chinese and Korean
wwPDB was founded in 2003

CORRESPONDENCE

Announcing the worldwide Protein Data Bank

In recognition of the growing international and interdisciplinary nature of structural biology, three organizations have formed a collaboration to oversee the newly formed worldwide Protein Data Bank (wwPDB). The Research Collaboratory for Structural Bioinformatics (RCSB), the European Molecular Biology Laboratory (EMBL) and the Protein Data Bank Japan (PDBj) have created the wwPDB, a single archive that will provide information on proteins and nucleic acids to the global community.

The wwPDB represents a milestone in the evolution of the Protein Data Bank (PDB), which was established in 1971 at Brookhaven National Laboratory as the sole international repository for three-dimensional structural data of biological macromolecules. Since July 1, 1999, the PDB has been managed by three member institutions of the RCSB: Rutgers, The State University of New Jersey, the San Diego Supercomputer Center at the University of California, San Diego, and the Center for Advanced Research in Biotechnology at the National Institute of Standards and Technology.

The wwPDB recognizes the importance of providing equal access to the database—whether on desktops, web-based services and via diverse platforms from different regions of the world. Therefore, the wwPDB members will ensure that datasets are deposited from diverse locations and will make sure that these datasets are fully translatable. The diverse regions will include the sequence, annotation, and structure data. The diversity will be kept public so that the distribution into databases will ensure the PDB archive is available to all researchers around the world.

The wwPDB is a repository that stores and distributes data on PDB resources. Its role will be to provide a single archive that will provide information on proteins and nucleic acids to the global community.

In addition, the legacy PDB format will be kept available to users. The distribution of the PDB archive will be maintained at regular intervals and will be accessible to users around the world. The legacy PDB format will be supported for at least 10 years, and it will be the standard format for the wwPDB.

The wwPDB is the sole provider of the legacy PDB format and is responsible for ensuring the integrity and accessibility of the data. The wwPDB will also provide a mechanism to ensure that data is available to users around the world.

Recently, the wwPDB has announced a new service called the wwPDB API, which allows users to access the wwPDB data through an application programming interface (API). The API is available to researchers, developers, and other organizations who want to integrate wwPDB data into their own applications.

The wwPDB is committed to ensuring that the legacy PDB format remains accessible to all users. The wwPDB API is another step in this effort and will enable developers to integrate wwPDB data into their applications.

Helen Berman, Kim Henrick
Haruki Nakamura

Nature Structural Biology 10, 980 (2003) doi: 10.1038/NSB1203-980
PDBj is in charge of processing and annotating the depositions from Asia

The history of PDBj from 2000

- **PDBj founded**
- **Launch of Promode service at PDBj**
- **Launch of eF-site service at PDBj**
- **SF become mandatory**
- **CS become mandatory**
- **PDBx/mmCIF become mandatory**
- **20,000 depositions at PDBj**
- **50,000 depositions at PDBj**
- **20th Anniversary of PDBj**

**Key Events**

- **PDBj founded**
- **2000**: PDBj founded
- **2001**: Launch of Promode service at PDBj
- **2002**: Launch of eF-site service at PDBj
- **2003**: SF become mandatory
- **2005**: CS become mandatory
- **2010**: PDBx/mmCIF become mandatory
- **2010**: 20,000 depositions at PDBj
- **2020**: 50,000 depositions at PDBj
- **2023**: 20th Anniversary of PDBj

**Important Structures**

- **The 1st MX structure processed by PDBj was 1GC0 (L-Met γ-lyase), deposited from Japan**
- **The 1st NMR structure processed by PDBj was 1EJQ (Syndecan-4), deposited from Korea**
Structural Biology Data Archiving in Asia

1. NMR Spectroscopy

Spectral Data
- Chemical shifts
- Coupling constants
- Time-domain data
- Spectral peak lists

Relaxation Data
- R1/T1, R2/T2, R1rho/T1rho, NOEs

Coordinates (XYZ)
- PDB Core Archive with the metadata including the links to other databases & source of proteins
Structural Biology Data Archiving in Asia

2. Cryo Electron Microscopy

2D Images → 3D Maps → Coordinates (XYZ)

EMPIAR

Raw 2D images of Cryo-EM including motion correction movies (~5 TB/entry)

EMDB

EMDB

PDBj

PDBj

Core Archives with the metadata including the links to other databases, experimental condition & source of proteins
Structural Biology Data Archiving in Asia

3. X-ray Crystallography

2D raw images $\rightarrow$ Structure Factors $\rightarrow$ Coordinates (XYZ)

PDB Core Archive with the metadata including the links to other databases, experimental condition & source of proteins

- ProteinDiffraction.org - https://proteindiffraction.org
- SBGrid - https://sbgrid.org
- CXIDB - http://www.cxidb.org

No public database for X-ray images available in Asia
Archive for X-ray Diffraction Images

https://xrda.pdbj.org
Now 81 entries are released
Archive for X-ray Diffraction Images
https://xrda.pdbj.org

Further development
1) Semi-automatic data upload
2) Automatic data quality check
Please deposit your experimental data to PDBj

https://xrda.pdbj.org

https://deposit-bmrnb.pdbj.org

https://empiar.pdbj.org

All three experimental data types are now archived in Asia!
**wwPDB validation report**

- **Model Quality**
  - Bond lengths and angles (outlier info, RMS-Z)
  - Chirality, planarity
  - Close contacts (including worst clashes, MolProbity clash score)
  - Torsion angles (Ramachandran statistics, protein rotamers)
  - Ligand geometry (Mogul analysis)

- **Residue Plots**
  - Residues with model-quality outliers (0, 1, 2, >2)
  - Residues with RSR-Z > 5 are highlighted
  - Residues not observed

*Overall Quality Summary*

<table>
<thead>
<tr>
<th>Metric</th>
<th>Percentile Ranks</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rfree</td>
<td></td>
<td>0.185</td>
</tr>
<tr>
<td>Clashscore</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Ramachandran outliers</td>
<td></td>
<td>0.001</td>
</tr>
<tr>
<td>Sidechain outliers</td>
<td></td>
<td>1.54</td>
</tr>
<tr>
<td>RSR,Z outliers</td>
<td></td>
<td>1.26</td>
</tr>
</tbody>
</table>

*Residue Plots*

*Chain A:*

- [Diagram of residue plots]
wwPDB/RDF-validation graph

PDBj-BMRB Data Server:
common open representations of BMRB NMR-STAR data in XML, RDF and JSON formats

Query examples

Category holders
1. Select all category holders of datablock class of BMRB entry 15400. Show
2. Select all category holders of datablock class of Metabolomics entry 16000400. Show

Entry statistics
3. Count entries per submission year and experimental method (subtypes). Show

Assembly descriptions
4. Select all assembly names, asyn IDs, entity IDs, polymer types, formula weights and functions in a assembly. Show

Entity descriptions
5. Select all entity names and sequences of polymer entities expressed using one-letter code. Show
6. Select all original source information of molecular entities and external links to NCBI Taxonomy. Show
7. Select all biological systems to produce molecular entities and external links to NCBI Taxonomy. Show

Citation information
8. Select citation information of all entries together with external links to PubMed and DOL, if available. Show
Example: Entry search based on wwPDB/RDF-validation

"4CK1", "INTEGRASE", "(4-CARBOXY-1,3-BENZODIOXOL-5-YL)METHYL-[(2-[(4-METHOXYPHENYL)METHYLCARBAMOYL]PHENYL)METHYL]AZANIUM", "OM1", "0.081"
Depositor-Initiated Atomic Coordinate Versioning

- Improve the quality of PDB structures
- Preserve original PDB IDs and maintain connections to the scientific literature
- Reasoning captured
- The latest versioned file can be accessed at FTP
- The latest minor version of each major version is available at versioned FTP

ftp-versioned.wwpdb.org
Depositor-Initiated Atomic Coordinate Versioning

- Improve the quality of PDB structures
- Preserve original PDB IDs and maintain connections to the scientific literature
- Reasoning captured
- The latest versioned file can be accessed at FTP
- The latest minor version of each major version is available at versioned FTP

ftp-versioned.wwpdb.org
Deliverables
Reproducibility of data sets is paramount.

The team aim for a single point of contact for definitive molecular models, namely at the PDBj and its XRDa, the X-ray Diffraction Data Archive based at the Institute for Protein Research in Japan.

We aim to avoid dispersed multiple versions of a protein model derived from a single raw diffraction data set. Controlled versioning procedure of PDB entries should be tightly linked.

A critical deliverable is to realise metrics of ‘definitive reusability’ which would then be applicable to the individual diffraction data sets held in the XRDa. These metrics and the definitive diffraction data files are a bedrock of interoperability.
A glimpse of the variation of X-ray diffraction resolution limit choice involving the commonly used metrics in macromolecular crystallography:-

Resolution cut off estimates:-

resolution of all data : 1.913
based on CC(1/2) >= 0.33 : 1.946
based on mean(I/sigma) >=2.0 : 3.037
based on R-merge < 0.5 : 2.411
based on R-meas < 0.5 : 2.497
based on completeness >=90% : 2.335
based on completeness >=50% : 2.155

Via the Diederichs and Karplus method, using the XRDa entry the resolution limit should be 2.4 Å.
The depositor, Sato et al (Biochem. J. 478, 1023–1042) used 2.40 Å.
Acknowledgements

Diffraction Data Case Study for GOSC
John R Helliwell (UK)
Loes Kroon-Batenburg (The Netherlands)
CCD ID becomes 5-letters

- 3-letter ID codes of the Chemical Component Dictionary (CCD) will run out in 2022.
- Only the PDBx/mmCIF format will be provided for CCD IDs with 5-letters.