Activities of PDBj and wwPDB:
A new PDB format, Data Deposition, Validation, and Data Integration

Haruki Nakamura

Institute for Protein Research
http://pdbj.org/
http://wwpdb.org/
wwPDB
wwPDB
Advisory Committee
on Sept 27, 2013
Rutgers Univ.
Since 2001, PDBj has been managed at Institute for Protein Research, Osaka University as a member of the wwPDB, to curate, edit and process the deposited data for an open, public, and single archive of the wwPDB.
Activities/Services of each member of the wwPDB

• “Data-in” activity, common in all the wwPDB members with high quality control. For that purpose, new format, data deposition, and validation system are developed

• “Data-out” services, characteristic at each wwPDB member site except common ftp-site data.
PDBj curates and processes about a Quarter of the deposited data, mainly from Asian and Oceania regions.
New Annotation System
For data increase and high quality data management

- Enables workload balancing and increased productivity
- Better quality assurance of ligand chemistry and polymer sequences
- PDBx/mmCIF is the master file format
- Validation suites based on recommendations from expert task forces; X-ray validation pipeline is available as a stand-alone server
- System will support all accepted experimental methods
Validation Report

- **Version 1.0** in production use since August 2013
  - [http://www.wwpdb.org/validation.html](http://www.wwpdb.org/validation.html)
  - Fixing occasional bugs
  - Collecting feedback to inform possible changes
    validation@mail.wwpdb.org

- January 2014 – validation data for all X-ray structures will be made publicly available through the wwPDB ftp sites

---

**Disclaimer**

This is a preliminary version of a new style of wwPDB validation report. We welcome your comments at validation@mail.wwpdb.org

The following versions of software and data (see references) were used in the production of this report:

- MolProbity: 4.0.2b-467
- Mogo: 1.16-2013
- Xtriage (Phenix): dev-1323
- EDS: trunk21216
- Percentile statistics: 20501
- Refmac: 5.8.0043
- CCP4: 6.3.0 (Settle)
- Ideal geometry (proteins): Engh & Huber (2001)
- Ideal geometry (DNA, RNA): Parkinson et. al. (1996)
- Validation Pipeline (wwPDB-VP): trunk21216
Validation Report

• Summary
  – Quality vs. all PDB X-ray
  – Quality vs. entries at similar resolution
  – Overview of residue-based quality for every polymer
  – Table of ligands that may need attention

The reported resolution of this entry is 1.80 Å.
Percentile scores (ranging between 0-100) for global validation metrics of the entry are shown in the following graphic. The table shows the number of entries on which the scores are based.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Whole archive (#Entries)</th>
<th>Similar resolution (#Entries, resolution range(Å))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rfree</td>
<td>65580</td>
<td>5922 (1.84-1.76)</td>
</tr>
<tr>
<td>Clashcore</td>
<td>76988</td>
<td>5640 (1.82-1.78)</td>
</tr>
<tr>
<td>Ramachandran outliers</td>
<td>75395</td>
<td>6528 (1.84-1.76)</td>
</tr>
<tr>
<td>Sidechain outliers</td>
<td>75377</td>
<td>6529 (1.84-1.76)</td>
</tr>
<tr>
<td>RSR2 outliers</td>
<td>65576</td>
<td>5522 (1.84-1.76)</td>
</tr>
</tbody>
</table>

The table below summarizes the geometric issues observed across the polymeric chains and their fit to the electron density. The red, orange, yellow and green segments on the lower bar indicate the fraction of residues that contain outliers for >=3, 2, 1 and 0 types of geometric quality criteria. The upper red bar (where present) indicates the fraction of residues that have poor fit to the electron density.

<table>
<thead>
<tr>
<th>Mol</th>
<th>Chain</th>
<th>Length</th>
<th>Quality of chain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>137</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mol</th>
<th>Chain</th>
<th>Length</th>
<th>Quality of chain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>371</td>
<td></td>
</tr>
</tbody>
</table>

The following table lists non-polymeric compounds that contain outliers for geometric or electron-density-fit criteria:

<table>
<thead>
<tr>
<th>Mol</th>
<th>Type</th>
<th>Chain</th>
<th>Res</th>
<th>Geometry</th>
<th>Electron density</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>NAG</td>
<td>A</td>
<td>401</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td>NAG</td>
<td>C</td>
<td>401</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
Validation Report

• Residue quality
  Red dots: poor density ($\text{RSR-Z} > 2$, as in EDS)

• Model/data fit ligands etc.
  – “LLDF” – Local Ligand Density Fit
    = Z-score of ligand RSR relative to nearby polymeric residues
Data-out from PDBj
http://pdbj.org/ and http://legacy.pdbj.org/

PDBj-BMRB

Duplicated portal
Summary page

http://pdbj.org/

Data viewer at PDBj

Amino acid sequence (FASTA)

Graphic viewer: jV
http://pdbj.org/jV/

Molecular surface DB: eF-site
http://ef-site.hgc.jp/eF-site/

Kinjo et al. NAR 40, D453 (2012)
PDBj services

**EM Navi**: EMDB browser

**Yorodumi**: PDB & EMDB

**GIRAF**: Similar binding site

**ProMode**: Normal Mode

**eF-site**: Mol. SurfaceDB

**eF-seek**: Similar surface
EM Navigatorとは？

- 生体分子や生体組織の3次元電子顕微鏡データを、気軽にわかりやすく眺めるためのウェブサイトです。
- EMDBとPDBのデータを利用しています（統計情報）
- 分子構造生物学の専門家にも、初心者や専門外のうちにも利用いただけるサイトを目指しています。
- PDBが運営しています。

お知らせ

- 2013-09-11: 公開データ
- その他の最新データ: EMD更新, PDB更新

詳細はこちら
New format PDB: PDBx/mmCIF

• PDB format is almost 40 years old and does not support today’s science.

• PDB Record format limitations
  – Max. 62 chains
  – Max. 99,999 atoms
  – No bond orders or chirality specified for ligands
  – No support for NMR, EM, hybrid methods, …
  – Meta-data specification cumbersome and inflexible

• Preserve backward compatibility where possible
• web service to create PDB format data files
• Start in early 2014
Current Supported Archival Formats

protein structure format universe

PDB (ca. 1974)
PDBx/mmCIF (ca. 1997)
PDBML (ca. 2005)
RDF (ca. 2011)

In managing the formats, PDBx is the master format.
### PDB Format Example

**REMARK 3** DATA USED IN REFINEMENT.

- **RESOLUTION RANGE HIGH (ANGSTROMS)**: 1.57
- **RESOLUTION RANGE LOW (ANGSTROMS)**: 23.00
- **DATA CUTOFF (SIGMA(F))**: 0.000
- **COMPLETENESS FOR RANGE (%)**: NULL
- **NUMBER OF REFLECTIONS**: 43316

**REMARK 3** FIT TO DATA USED IN REFINEMENT.

- **CROSS-VALIDATION METHOD**: NULL
- **FREE R VALUE TEST SET SELECTION**: NULL
- **R VALUE (WORKING + TEST SET)**: NULL
- **R VALUE (WORKING SET)**: 0.191
- **FREE R VALUE**: 0.221
- **FREE R VALUE TEST SET SIZE (%)**: NULL
- **FREE R VALUE TEST SET COUNT**: 2189

#### ATOM Format

<table>
<thead>
<tr>
<th>Number</th>
<th>Type</th>
<th>Residue</th>
<th>Chain</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>Temperature</th>
<th>Occupancy</th>
<th>Element</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N</td>
<td>VAL</td>
<td>A</td>
<td>363</td>
<td>22.741</td>
<td>-1.397</td>
<td>11.729</td>
<td>1.00</td>
<td>33.32</td>
</tr>
<tr>
<td>2</td>
<td>CA</td>
<td>VAL</td>
<td>A</td>
<td>363</td>
<td>21.557</td>
<td>-0.831</td>
<td>11.024</td>
<td>1.00</td>
<td>32.13</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>VAL</td>
<td>A</td>
<td>363</td>
<td>20.954</td>
<td>-1.757</td>
<td>9.943</td>
<td>1.00</td>
<td>31.73</td>
</tr>
<tr>
<td>4</td>
<td>O</td>
<td>VAL</td>
<td>A</td>
<td>363</td>
<td>19.737</td>
<td>-1.906</td>
<td>9.845</td>
<td>1.00</td>
<td>30.94</td>
</tr>
<tr>
<td>5</td>
<td>CB</td>
<td>VAL</td>
<td>A</td>
<td>363</td>
<td>21.883</td>
<td>0.552</td>
<td>10.391</td>
<td>1.00</td>
<td>33.45</td>
</tr>
</tbody>
</table>

- Record-oriented with fixed column format
- Metadata in semi-structured remarks
- Documentation by example
- Most widely used and supported archival format
PDBx/mmCIF Format Example

• Name – value pairs

```plaintext
_exptl.entry_id          1XBB
_exptl.method           'X-RAY DIFFRACTION'
_exptl.crystals_number  1
```

• Tables or loop’s

```plaintext
loop_
_database_PDB_rev.num   _database_PDB_rev.date     _database_PDB_rev.date_original _database_PDB_rev.mod_type   _database_PDB_rev.replaces _database_PDB_rev.status
1  2004-11-02  2004-08-30  0  1XBB  ?
2  2005-03-22  ?          ?  1  1XBB  ?
3  2009-02-24  ?          ?  1  1XBB  ?
```

• Simple syntax
• Named data items
• Data semantics defined in the PDBx data dictionary
• Software support in most popular languages
Three flavors of XML files:
- fully marked-up files
- files without atom records
- files with a more space efficient encoding of atom records

Follows naming and semantics of the PDBx data dictionary
• Entry point for semantic web and reasoning systems
• Translates data items in PDBx/mmCIF schema into triples with URL identifiers
• Follows naming and semantics of the PDBx data dictionary
• http://pdbj.org/rdf/<pdbID>/<categoryName>/<pkey1>,...
• For example, http://pdbj.org/rdf/1GOF/entity/1

<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:PDBo="http://pdbj.org/schema/pdbx-v40.owl#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <PDBo:PDBID rdf:about="http://pdbj.org/pdb/1GOF">
    <rdfs:label>1GOF</rdfs:label>
  </PDBo:PDBID>
  <rdf:Description rdf:about="http://pdbj.org/rdf/1GOF/entity/1">
    <PDBo:entity.formula_weight>68579.250</PDBo:entity.formula_weight>
    <PDBo:entity.id>1</PDBo:entity.id>
    <PDBo:entity.pdbx_description>GALACTOSE OXIDASE</PDBo:entity.pdbx_description>
    <PDBo:entity.pdbx_ec>1.1.3.9</PDBo:entity.pdbx_ec>
    <PDBo:entity.pdbx_number_of_molecules>1</PDBo:entity.pdbx_number_of_molecules>
    <PDBo:entity.src_method>man</PDBo:entity.src_method>
    <PDBo:entity.type>polymer</PDBo:entity.type>
    <PDBo:link_to_enzyme rdf:resource="http://purl.uniprot.org/enzyme/1.1.3.9"/>
    <PDBo:of_datablock rdf:resource="http://pdbj.org/rdf/1GOF"/>
    <PDBo:referenced_by_entity_keywords rdf:resource="http://pdbj.org/rdf/1GOF/entity_keywords/1"/>
    <PDBo:referenced_by_entity_poly rdf:resource="http://pdbj.org/rdf/1GOF/entity_poly/1"/>
    <PDBo:referenced_by_entity_src_gen rdf:resource="http://pdbj.org/rdf/1GOF/entity_src_gen/1"/>
    <PDBo:referenced_by_struct_asym rdf:resource="http://pdbj.org/rdf/1GOF/struct_asym/A"/>
    <PDBo:referenced_by_struct_ref rdf:resource="http://pdbj.org/rdf/1GOF/struct_ref/1"/>
    <rdf:type rdf:resource="http://pdbj.org/schema/pdbx-v40.owl#entity"/>
  </rdf:Description>
</rdf:RDF>
Transitional Home for Large Structures

Large single entries are now stored separately on the wwPDB ftp site, and PDB internally produces divided/split PDB format files.

ftp://ftp.wwpdb.org/pub/pdb/data/large_structures/mmCIF/

HIV-1 Capsid 3J3Q –
• 1356 chains
• >2M atoms
• 25 – PDB format entries

3J3Y

3J3Q
Providing Format Compatibility

• Adopt a *PDB friendly* mmCIF/PDBx style -
  – All records on a single text line
  – Columns presented in standard column order.
  – Tabular presentation with leading record names (e.g. ATOM, CELL, REFINE)
  – Method independent features in left-most columns (e.g. identifiers & coordinates)
  – Method specific features in the right-most columns (e.g. ADPs, NMR order/disorder parameters)
  – Continue to support PDB nomenclature semantics (e.g. PDB style chains, residue numbering, and insertion codes)

• Large entries will be internally converted to divided/split PDB format files.
<table>
<thead>
<tr>
<th>Atom</th>
<th>Type</th>
<th>Residue</th>
<th>Chain</th>
<th>Position</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>Occupancy</th>
<th>B-factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N</td>
<td>GLN</td>
<td>A</td>
<td>39</td>
<td>24.690</td>
<td>-27.754</td>
<td>24.275</td>
<td>1.00</td>
<td>60.76</td>
</tr>
<tr>
<td>2</td>
<td>CA</td>
<td>GLN</td>
<td>A</td>
<td>39</td>
<td>23.581</td>
<td>-26.768</td>
<td>24.416</td>
<td>1.00</td>
<td>60.98</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>GLN</td>
<td>A</td>
<td>39</td>
<td>23.990</td>
<td>-25.379</td>
<td>23.905</td>
<td>1.00</td>
<td>59.98</td>
</tr>
<tr>
<td>4</td>
<td>O</td>
<td>GLN</td>
<td>A</td>
<td>39</td>
<td>25.070</td>
<td>-25.209</td>
<td>23.330</td>
<td>1.00</td>
<td>60.25</td>
</tr>
<tr>
<td>5</td>
<td>CB</td>
<td>GLN</td>
<td>A</td>
<td>39</td>
<td>23.136</td>
<td>-26.685</td>
<td>25.878</td>
<td>1.00</td>
<td>60.69</td>
</tr>
<tr>
<td>6</td>
<td>N</td>
<td>VAL</td>
<td>A</td>
<td>40</td>
<td>23.115</td>
<td>-24.395</td>
<td>24.122</td>
<td>1.00</td>
<td>59.58</td>
</tr>
<tr>
<td>7</td>
<td>CA</td>
<td>VAL</td>
<td>A</td>
<td>40</td>
<td>23.342</td>
<td>-23.010</td>
<td>23.690</td>
<td>1.00</td>
<td>57.26</td>
</tr>
<tr>
<td>8</td>
<td>C</td>
<td>VAL</td>
<td>A</td>
<td>40</td>
<td>24.000</td>
<td>-22.152</td>
<td>24.778</td>
<td>1.00</td>
<td>56.00</td>
</tr>
<tr>
<td>9</td>
<td>O</td>
<td>VAL</td>
<td>A</td>
<td>40</td>
<td>23.992</td>
<td>-20.920</td>
<td>24.692</td>
<td>1.00</td>
<td>55.53</td>
</tr>
<tr>
<td>10</td>
<td>CB</td>
<td>VAL</td>
<td>A</td>
<td>40</td>
<td>22.015</td>
<td>-22.337</td>
<td>23.275</td>
<td>1.00</td>
<td>57.32</td>
</tr>
</tbody>
</table>

PDBx/mmCIF
PDBx/mmCIF Software Support

- Phenix and Refmac – produce native PDBx files for deposition
- MMDB - macromolecular object library in CCP4
- iotbx.cif/ucif - CCTBx C++/Python IO library with dictionary validation
- CCIF – CCP4 C++ library with FORTRAN support and dictionary validation
- CBFLib - ANSI-C library for CIF & imgCIF files
- mmLIB - Python toolkit supporting CIF & mmCIF
- BioPython - Python toolkit for computational biology
- PyCifRW - Python CIF/mmCIF parsing tools
- BioJava - Java mmCIF IO package
- STAR::Parser – Perl mmCIF parser and molecular object library
- RCSBTools - C++/Python parsing and dictionary validation tools plus many other supporting format conversion and data management applications
- Visualization - Chimera, Jmol, OpenRasMol
  PDB actively working with community developers to help fill in missing functionalities. Two workshops scheduled in Fall 2013 …
Converter Program: CIFTr

http://mmcif.pdb.org

Dictionary Resources

The Protein Data Bank (PDB) uses macromolecular Crystallographic Information File (mmCIF) data dictionaries to describe the information content of PDB entries. The PDB Exchange data dictionary consolidates content from a variety of crystallographic dictionaries including: the IUCr Core, mmCIF, Image and symmetry dictionaries. The PDB Exchange Dictionary also includes extensions describing NMR, Cryo-EM, and protein production data. PDB data processing, data exchange, annotation, and database management operations all make heavy use of the data format and the content of the PDB Exchange Dictionary. Software tools are used to convert mmCIF data files to the older PDB format and to PDBML/XML.

- mmCIF/PDBx example files for PDB entries containing large structures
- PDBx/mmCIF tutorial for the protein structure modules of BioJava
- Data files in mmCIF format can be downloaded from the RCSB PDB website or by ftp.
- Software tools are available for preparing and editing deposits.
- Software tools are available for converting mmCIF data files to PDB and PDBML formats
- A complete list of PDB software tools for managing PDB data in mmCIF format can be found here.

Dictionary Content and Representation

- Background and Introduction about mmCIF
Converter Program: CIFTr

http://mmcif.pdb.org

CIFTr is an application program for translating files in mmCIF format into files in PDB format. CIFTr is currently available in source form and binary form for several UNIX platforms.

CIFTr can be used to translate the files released that are in in mmCIF format at ftp://ftp.rcsb.org/pub/pdb/data/structures/.

Installation and usage notes for the source distribution of CIFTr

Click here to download source distribution of CIFTr.

Installation and usage notes for the binary distribution of CIFTr

Click here to download binary (Linux, SGI, OSF, MACOSX, or SUN) distribution of CIFTr.
PDB/RDF format for Semantic Web
Service from wwPDB by Akira R. Kinjo (PDBj) & Tom Oldfield (PDBe)

http://rdf.wwpdb.org/


In UniProt RDF:

```xml
<rdf:Description rdf:about="http://rdf.wwpdb.org/pdb/1BY4">
  <database rdf:resource="http://purl.uniprot.org/database/PDB"/>
  <resolution rdf:datatype="http://www.w3.org/2001/XMLSchema#float">2.10</resolution>
</rdf:Description>
```
BMRB/XML and /RDF for Semantic Web
(developed by PDBj and BMRB)

NMR-STAR v3

BMRB/XML

BMRB/ RDF

Format | NMR-STAR v3 | BMRB/XML | BMRB/RDF
--- | --- | --- | ---
Ontology | NMR-STAR v3 dictionary | BMRB/XML schema | BMRB/OWL
Validation | ADIT-NMR & manual validation | Schema validation | RDF validator
Data type | Text file (.str) | XML file (.xml) | RDF file (.rdf)

BMRB/XML and BMRB/RDF for Semantic Web (developed by PDBj and BMRB)

Yokochi, M. et al., in preparation
PDBj: Protein Data Bank Japan (2013)

- **Head**
  - Nakamura, Haruki, Ph. D. (Prof., IPR, Osaka Univ.)

- **Group for PDB Database Curation**
  - Nakagawa, Atsushi, Ph. D. (Group Leader, Prof., IPR, Osaka Univ.)
  - Matsuda, Makoto, Ph. D. (IPR, Osaka Univ.)
  - Igarashi, Reiko (IPR, Osaka Univ.)
  - Kengaku, Yumiko (IPR, Osaka Univ.)
  - Cho, Hasumi, Ph. D. (IPR, Osaka Univ.)
  - Ikegawa, Yasuyo (IPR, Osaka Univ.)
  - Sato, Junko (IPR, Osaka Univ.)

- **Group for Development of new tools and services**
  - Kinjo, Akira R., Ph. D. (IPR, Osaka Univ.)
  - Iwasaki, Kenji, Ph. D. (IPR, Osaka Univ.)
  - Suzuki, Hirofumi, Ph. D. (IPR, Osaka Univ.)
  - Yamashita, Reiko (IPR, Osaka Univ.)
  - Kudou, Takahiro (IPR, Osaka Univ.)
  - Nishikawa, Ken, Ph. D. (Guest Prof., IPR, Osaka Univ.)
  - Bekker, Gert-Jan (IPR, Osaka Univ.)

- **Group for BMRB**
  - Fujiwara, Toshimichi, Ph. D. (Group Leader, Prof. Osaka Univ.)
  - Akutsu, Hideo, Ph. D. (Prof., IPR, Osaka Univ.)
  - Kojima, Chojiro, Ph. D. (IPR, Osaka Univ.)
  - Kobayashi, Naohiro, Ph. D. (IPR, Osaka Univ.)
  - Iwata, Takeshi (IPR, Osaka Univ.)
  - Takahashi, Ami (IPR, Osaka Univ.)
  - Yokochi, Masashi (IPR, Osaka Univ.)

- **Collaboratory Researchers**
  - Wako, Hiroshi, Ph. D. (Prof., Waseda Univ.) (for Pro Mode)
  - Ito, Nobutoshi, Ph. D. (Prof., Tokyo Medical and Dental Univ.)
  - Kinoshita, Kengo, Ph.D. (Prof., Tohoku Univ.) (for e F-site)
  - Standley, Daron, Ph. D. (IFReC, Osaka Univ.) (for SeqNavi, StructNavi, SeSAW, and ASH)
  - Katoh, Kazutaka, Ph. D. (IFReC, Osaka Univ.) (for MAFFTash)

- **Secretary**
  - Haruki, Nahoko (IPR, Osaka Univ.)