Protein Data Bank Japan (PDBj):

Maintaining structural data archive and integration of structure data with other life sciences data resources by semantic web technologies

Haruki Nakamura, Akira R. Kinjo, Hirofumi Suzuki, Reiko Yamashita, Yasuyo Ikegawa, Takahiro Kudou, Gert-Jan Bekker, Reiko Igarashi, Yumiko Kengaku, Hasumi Cho, Junko Sato, Nahoko Haruki, Daron M. Standley, Atsushi Nakagawa

PDBj Protein Data Bank Japan
http://pdbj.org/

WORLDWIDE PDB PROTEIN DATA BANK
http://wwpdb.org/

Abstract

The Protein Data Bank Japan (PDBj, http://pdbj.org/) is a member of the worldwide Protein Data Bank (wwPDB: http://wwpdb.org/) [1, 2]. PDBj accepts and processes experimental macromolecular structure data determined by X-ray crystallography, SAXS/SANS, NMR, and Electron Microscopy as a Data-in service. While maintaining the archive in collaboration with other wwPDB partners, PDBj also provides a wide range of Data-out services and tools for analyzing structures and functions of proteins, as shown in Table 1.

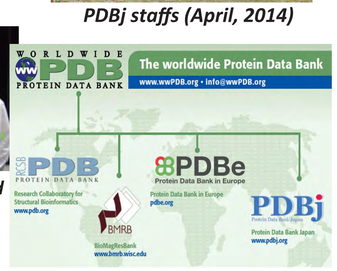
To facilitate the integration of structure data with other life sciences data resources, we have developed data in formats compatible with semantic web technologies: PDB/RDF (http://rdf.wwpdb.org/pdb/) and Chem_Comp/RDF (http://rdf.wwpdb.org/cc/), PDB and chemical compounds data in the Resource Description Framework (RDF) format, along with its ontology in Web Ontology Language (OWL) based on the PDB mmCIF Exchange Dictionary [1, 2]. NMR experimental data in BMRB database are also being described in RDF as BMRB/RDF, and those data are integrated through the linked-data technology.

Protein Data Bank Japan (*http://pdbj.org/*)

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. ***the PDB archive contains more than 100,000 entries now.***

PDBj staffs (April, 2014)

wwPDB members and their heads

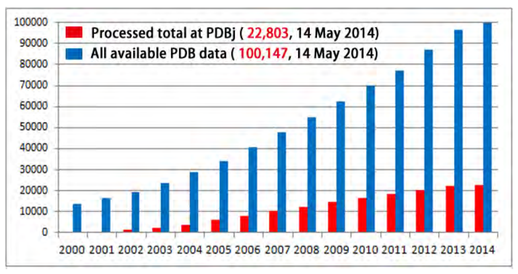


Activities/Services of each member of the wwPDB

- **"Data-in"** activitiy, 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.

Data in at PDBj and wwPDB

PDBj curates and processes about **a quarter of the deposited data**, mainly from Asian and Oceania regions.

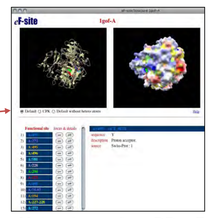


Summary page

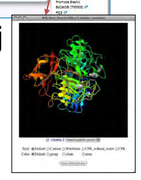
http://pdbj.org/

Data viewer at PDBj

Amino acid sequence (FASTA)



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

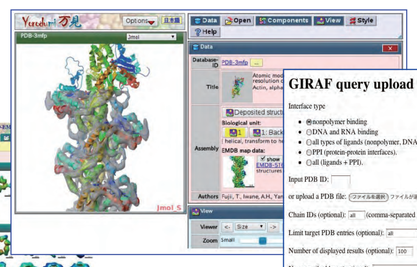


Graphic viewer: *jV*
http://pdbj.org/jV/

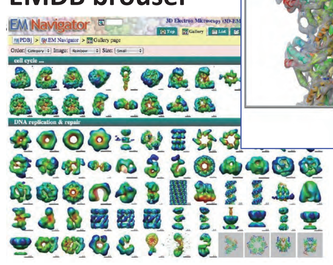
Kinjo et al. (2012) Nucl. Acids Res. 40, D453.

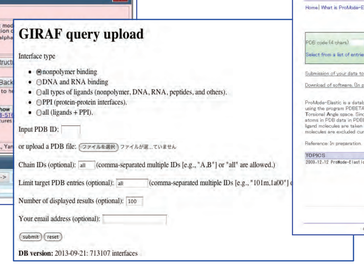
PDBj services

Yorodumi: PDB & EMDB



EM Navi: EMDB brouser

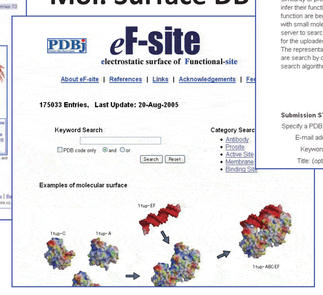




GIRAF: Similar binding sites

ProMode: Normal Modes

eF-site: Mol. Surface DB



eF-seek: Similar surfaces

PDB/RDF format for Semantic Web

Service from wwPDB by Akira R. Kinjo (PDBj) & Tom Oldfield (PDBe)

http://rdf.wwpdb.org/

Kinjo et al. (2012) Nucl. Acids Res. 40, D453-D460.

In UniProt RDF:

```
<rdf:Description rdf:about="http://rdf.wwpdb.org/pdb/1BY4">
  <rdf:type rdf:resource="http://purl.uniprot.org/core/Structure_Resource"/>
  <database rdf:resource="http://purl.uniprot.org/database/PDB"/>
  <method rdf:resource="http://purl.uniprot.org/core/X-Ray_Crystallography"/>
  <resolution rdf:datatype="http://www.w3.org/2001/XMLSchema#float">2.10</resolution>
</rdf:Description>
```

Table 1. Data-out tools and derived databases

Service	Description	URL
PDBj Mine	Simple and SQL search for PDB data	pdbj.org/mine [3]
SeSAW	Functional site prediction	pdbj.org/sesaw [4]
GIRAF	Interaction interface similarity search	pdbj.org/giraf [5-9]
EM Navigator	EM structure browser	pdbj.org/emnavi [1]
jV Molecular	graphics viewer	pdbj.org/jv [10]
Yorodumi	Integrated PDB/EMDB data browser	pdbj.org/yorodumi [1]
Spanner	Homology modeling	pdbj.org/spanner [11]
eF-site	Electrostatic surface database	pdbj.org/ef-site [12-14]
eF-seek	Search against eF-site	pdbj.org/ef-seek [13,14]
eF-surf	Building electrostatic surfaces	pdbj.org/ef-surf [13]
ProMode	**Elastic** Normal mode analysis with elastic network model	pdbj.org /promode_elastic [15]

References:
[1] A. R. Kinjo et al., Nucl. Acids Res. 40, D453 (2012) [2] H. M. Berman et al., Structure 21, 1485-1491 (2013)
[3] A. R. Kinjo et al., Database 2010, baq021 (2010) [4] D, M. Standley et al., Bioinformatics 26, 1258 (2010)
[5] A. R. Kinjo, H. Nakamura, Biophysics 3, 75 (2007) [6] A. R. Kinjo, H. Nakamura, Structure 17, 234 (2009)
[7] A. R. Kinjo, H. Nakamura, J. Mol. Biol. 399, 526 (2010) [8] A. R. Kinjo, H. Nakamura, PLoS One 7, e31437 (2012)
[9] A. R. Kinjo, H. Nakamura, Biophysics 8, 79 (2012) [10] K. Kinoshita, H. Nakamura, Bioinformatics 20, 1329 (2004)
[11] M. Lis et al., Immunome Res. 7, 1 (2011) [12] K. Kinoshita, H. Nakamura, Protein Sci. 12, 1589 (2003)
[13] K. Kinoshita et al., Nucl. Acids Res. 35, W398 (2007) [14] Y. Murakami et al., Protein Science, 22, 1379-1391 (2013)
[15] H. Wako, S. Endo, Biophys. Chem. 159, 257 (2011)

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**

PDB

```
ATOM      1  N   GLN A  39      24.690 -27.754  24.275  1.00 60.76           N
ATOM      2  CA  GLN A  39      23.581 -26.768  24.416  1.00 60.98           C
ATOM      3  C   GLN A  39      23.990 -25.379  23.905  1.00 59.98           C
ATOM      4  O   GLN A  39      25.070 -25.209  23.330  1.00 60.25           O
ATOM      5  CB  GLN A  39      23.136 -26.685  25.878  1.00 60.69           C
ATOM      6  N   VAL A  40      23.115 -24.395  24.122  1.00 59.58           N
ATOM      7  CA  VAL A  40      23.342 -23.010  23.690  1.00 57.26           C
ATOM      8  C   VAL A  40      24.000 -22.152  24.778  1.00 56.00           C
ATOM      9  O   VAL A  40      23.992 -20.920  24.692  1.00 55.53           O
ATOM     10  CB  VAL A  40      22.015 -22.337  23.275  1.00 57.32           C
```

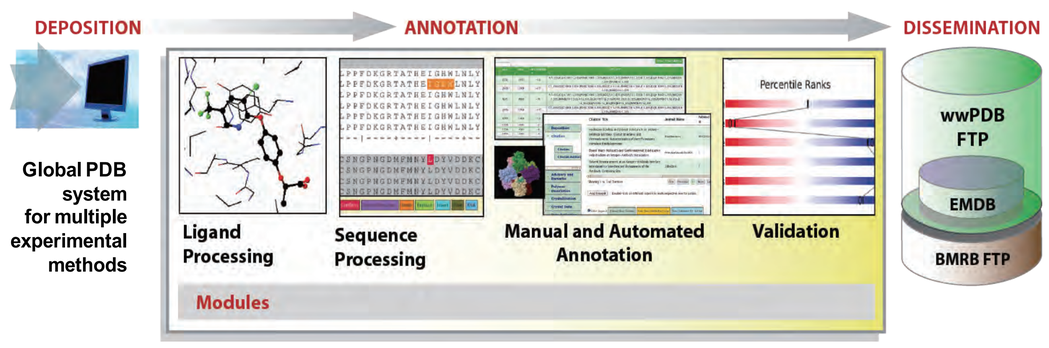
PDBx/mmCIF

```
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.auth_atom_id
_atom_site.type_symbol
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_seq_id
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.pdbx_PDB_model_num
_atom_site.occupancy
_atom_site.pdbx_auth_alt_id
_atom_site.B_iso_or_equiv
ATOM    1  N   N  GLN A 39  24.690 -27.754  24.275  1  1.000  .  60.760
ATOM    2  CA  C  GLN A 39  23.581 -26.768  24.416  1  1.000  .  60.980
ATOM    3  C   C  GLN A 39  23.990 -25.379  23.905  1  1.000  .  59.980
ATOM    4  O   O  GLN A 39  25.070 -25.209  23.330  1  1.000  .  60.250
ATOM    5  CB  C  GLN A 39  23.136 -26.685  25.878  1  1.000  .  60.690
ATOM    6  N   N  VAL A 40  23.115 -24.395  24.122  1  1.000  .  59.580
ATOM    7  CA  C  VAL A 40  23.342 -23.010  23.690  1  1.000  .  57.260
ATOM    8  C   C  VAL A 40  24.000 -22.152  24.778  1  1.000  .  56.000
ATOM    9  O   O  VAL A 40  23.992 -20.920  24.692  1  1.000  .  55.530
ATOM   10  CB  C  VAL A 40  22.015 -22.337  23.275  1  1.000  .  57.320
ATOM   11  N   N  ALA A 41  24.560 -22.804  25.797  1  1.000  .  54.570
```

PDBx/mmCIF

New Annotation System

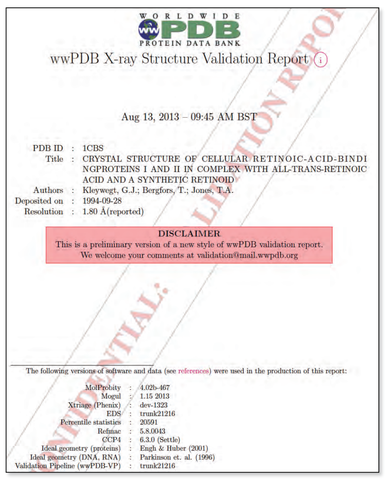
For data increase and high quality data management



-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*
  - -Fixing occasional bugs
  - -Collecting feedback to inform possible changes
  - *validation@mail.wwpdb.org*
- January 2014 -validation data for all X-ray structures has made publicly available through the wwPDB ftp sites



PDBx/mmCIF Software Support

http://mmcif.wwpdb.org/

- **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**