Data Management at the RCSB PDB

John Westbrook
Rutgers, The State University of New Jersey
Data Management at the RCSB PDB

- Dictionary-driven architecture
- Pipeline of data collection spanning experimental structural biology
- Tools to automate data harvesting
- Continuous focus of data standardization and data quality
- Distribution and delivery deeply integrated with biological and medical resources
Dictionary Resources

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.

- Data files in mmCIF format can be downloaded from the RCSB PDB website or by ftp.
- Software tools are available for preparing and editing depositions.
- 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
- STAR/mmCIF: An Extensive Ontology for Macromolecular Structure and Beyond (PDF) *Bioinformatics* (2000) 16(2), 159-168.
- mmCIF Software Developers Workshop 1997
- mmCIF Dictionary Templates
- mmCIF Examples
- References

Data Dictionaries

- PDB mmCIF Exchange Dictionary | (ASCII) | (compressed) | (HTML) | XML Schema |
  Data dictionary developed as a collaboration between MSD-EBI and RCSB and used by wwPDB members for data exchange.

- mmCIF Dictionary | (ASCII) | (compressed) | (HTML)
  IUCr Standard Dictionary
  Original working group members: Paula M. Fitzgerald, Helen Berman, Phil Bourne, Brian McMahon, Keith Watenpaugh, and John Westbrook
PDBML Schemas
http://pdbml.pdb.org/

- New PDBML website
- Schemas provided for the PDB Exchange dictionary and component dictionaries
- Schemas and data dictionaries are updated synchronously

Structure Determination
Data Pipeline

TargetDB
PepcDB
BioSync

Target Selection
Sample Preparation
Data Collection
Data Processing
Structure Determination
Refinement

pdb_extract

PDB
Target Registration Database
TargetDB • http://targetdb.pdb.org/

- Targets downloadable in XML (~120K targets)
- Targets downloaded from 21 worldwide centers weekly
- Target search by:
  - Sequence (FASTA), project target ID, project site, status (selected, cloned, expressed, … in PDB), update date, protein name, source organism
- Report output in HTML, FASTA, and XML
- Integrates PDB entry sequences (~90K sequences)
- Includes PDB pre-release sequence data
- Provides links to related project sequence databases
- Summary reports of target and project progress
- 3-Tier architecture (Apache/Tomcat/MySQL)
Protein Expression Purification and Crystallization Database
PepcDB • http://pepcdb.pdb.org/

- Extends content of TargetDB
- Includes all TargetDB sequences
- All protocols for cloning, expression, purification are stored and are searchable
- Reports provide links to status history, related protocols, project, sequence and domain databases
BioSync: A Structural Biologist’s Guide to Synchrotron Facilities

Current Synchrotron Information: Updated by synchrotron personnel

Beamline Descriptions Contain: Wavelength ranges, flux and optical configurations, detectors, hardware / software

Beamline Deposition Statistics: Beamline statistics for synchrotrons around the world based on PDB release date

Coming Soon: Beamlines outside the US, more small angle x-ray scattering and x-ray spectroscopy, including cross-links to the RCSB PDB website.

Brief History: BioSync was formed in 1990 as a grassroots organization that would promote access to synchrotron radiation for scientists whose primary research is in the field of structural biology. The members were leaders of North American research groups that used synchrotron radiation for experiments in structural biology. The members of the original Steering Committee for BioSync were: Keith Watenpaugh – Chair, Hugh Huxley, Sung-Hou Kim, Keith Moffat, Janet Smith, Robert Sweet, and Edwin Westbrook.

The activities of BioSync included documenting the needs of structural biologists for synchrotron radiation, evaluating the status of existing and planned facilities, recommending funding and access policies, lobbying in Washington for improved support of synchrotron rings and beamlines, and organizing a web-based clearinghouse of beamline information.

> Read More...
Data Collection - BioSync

National Synchrotron Light Source (NSLS)
is located at Brookhaven National Laboratory in Upton, NY

**NSLS Structural Biology Beamlines Overview**

<table>
<thead>
<tr>
<th>Beamline</th>
<th>Operational (general use)</th>
<th>Expts.</th>
<th>Wavelength (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X2A</td>
<td>Yes</td>
<td>MAD</td>
<td></td>
</tr>
<tr>
<td>X4A</td>
<td>Yes</td>
<td>MAD, SAD</td>
<td></td>
</tr>
<tr>
<td>X4C</td>
<td>Yes</td>
<td>Monochromatic (SAD, Native)</td>
<td></td>
</tr>
<tr>
<td>X6A</td>
<td>yes, 2003</td>
<td>Multiple-wavelength anomalous diffraction and single wavelength</td>
<td>6 to 22 keV</td>
</tr>
<tr>
<td>X6B</td>
<td>yes</td>
<td>MAD</td>
<td>0.6-6</td>
</tr>
<tr>
<td>X6A</td>
<td>yes</td>
<td>MAD</td>
<td>0.64-1.35</td>
</tr>
<tr>
<td>X6B</td>
<td>No - as of October 2006</td>
<td>X-ray absorption spectroscopy, extended fine structure X-ray absorption spectroscopy, fine structure X-ray absorption spectroscopy, near edge fine structure absorption structure</td>
<td>0.65-3.09</td>
</tr>
<tr>
<td>X12B</td>
<td>yes</td>
<td>MAD</td>
<td>0.65-1.61</td>
</tr>
<tr>
<td>X12C</td>
<td>yes</td>
<td>MAD</td>
<td>0.8-2.5</td>
</tr>
<tr>
<td>X2B</td>
<td>yes</td>
<td>MAD and Laue</td>
<td>0.4-4.0</td>
</tr>
<tr>
<td>X26C</td>
<td>yes</td>
<td>MAD/SAD and Laue</td>
<td>0.83 to 1.8</td>
</tr>
<tr>
<td>X29A</td>
<td>yes</td>
<td>MAD/SAD</td>
<td></td>
</tr>
<tr>
<td>other bio beamlines</td>
<td>Operational (general use)</td>
<td>Expts.</td>
<td>Wavelength (Å)</td>
</tr>
<tr>
<td>X1A</td>
<td>yes</td>
<td>scanning transmission X-ray microscopy and microspectroscopy (IMG)</td>
<td>40-45</td>
</tr>
<tr>
<td>X8B</td>
<td>yes</td>
<td>XAS EXAFS XANES</td>
<td>0.95-3.09</td>
</tr>
</tbody>
</table>

Links to NSLS

- General Information
- Beam time, xtal beamlines
- Operations
- Staff
- Shipping
- Training
- Travel
- Maps
- Publications
pdb_extract is a data extraction and validation tool which automates the assembling mmCIF data files from the outputs of X-ray and NMR structure determination applications.

Automated and accurate deposition of structures solved by X-ray diffraction to the Protein Data Bank. Huanwang Yang, Vladimir Guranovic, Shuchismita Dutta, Zukang Feng, Helen M. Berman and John D. Westbrook  *Acta Cryst* (2004). D60, 1833-1839
Data collection and reduction
  – HKL, SCALEPACK, d*TREK, SAINT, SCALA
Molecular replacement
  – CNS, CNX, Amore, Morep, EPMR
Heavy atom phasing
  – CNS, CNX, SOLVE, MLPHARE, SHARP/autoSHARP, SHELXD, SHELX, PHASES, SnB, BnP, Phaser
Density modification
  – CNS, CNX, DM, Solomon, RESOLVE, SHELXE
Structure refinement
  – CNS, CNX, REFMAC, RESTRAIN, SHELXL, TNT, WARP, PHENIX

NMR - CNS, CYANNA, DYANNA
**pdb_extract - Online**

The **pdb_extract** online tool extracts key information from X-ray crystallographic and NMR software applications in preparation for PDB deposition:

- Reduce the human effort required to assemble complete and validated protein structure entries ready for PDB deposition.
- Prepare an mmCIF file for deposition quickly.

**HOW TO RUN:**

- Select experimental method (X-Ray or NMR).
- Browse your workstation for the name of the coordinate file obtained from final structure refinement.
- Select the file type and program name.
- Press the RUN button to start this operation.

![Screenshot of pdb_extract interface]

**NOTE:**

- If the file size is large, it is recommended to upload gzipped (*.gz) or compressed (*.Z) file for faster loading.
- After assembling your data, you will get two mmCIF files for X-Ray or one mmCIF file for NMR. Please load them to ADIT for a complete deposition.

Questions, comments, and suggestions? **Contact Us.**
Deposition Services

RCSB PDB Data Validation and Deposition Services

NEW Beta-ADIT at RCSB | ADIT at RCSB or PDBj | pdb_extract
NEW Beta-ADIT NMR | Validation Server at RCSB or PDBj | Ligand Depot

The following data deposition tools and instructions can make your structure deposition quick, easy, complete and accurate:

- **Prepare** your structural data for deposition using pdb_extract and/or the desktop version of ADIT
- **Validate** your structure using the Validation Server at RCSB PDB or the Validation Server at PDBj
- **Deposit** your structure using the structure deposition tool beta-ADIT at RCSB PDB or ADIT at RCSB PDB or ADIT at PDBj, or using **AutoDep** at MSD-EBI.
  - The RCSB PDB, PDBj, and MSD-EBI are members of the wwPDB.
    - Instructions for X-ray crystallography structure depositions.
    - Instructions for NMR structure depositions.
    - Instructions for EM structure depositions.
    - More information on deposition of structures determined by other methods (including Electron diffraction, Fiber diffraction, Theoretical modeling).

- **Search** for
  - Your ligand using **Ligand Depot**
  - Appropriate sequence database references for proteins or nucleic acids in your structure (e.g. using **BLAST**)
Data Query Support Strategy

Uniform Data
Extended Annotation

Improved Query Functionality
wwPDB Remediation Project

The wwPDB has collaborated to remediate the PDB archive and create a more consistent set of files.

E-MSD is supported by grants from the Wellcome Trust, the EU (TEMBLOR, NMRQUAL and IIMS), CCP4, the BBSRC, the MRC and EMBL.

PDBj is supported by grant-in-aid from the Institute for Bioinformatics Research and Development, Japan Science and Technology Agency (BIRD-JST), and the Ministry of Education, Culture, Sports, Science and Technology (MEXT).

The BMRB is supported by NIH grant LM05799 from the National Library of Medicine.

The RCSB PDB is supported by grants from the National Science Foundation, National Institute of General Medical Sciences, the Office of Science-Department of Energy, the National Library of Medicine, the National Cancer Institute, the National Center for Research Resources, the National Institute of Biomedical Imaging and Bioengineering, the National Institute of Neurological Disorders and Stroke, and the National Institute of Diabetes & Digestive & Kidney Diseases.
• Sequence and taxonomy
  – Resolved anomalies relative to UniProt (61K sequences)
  – Resolved anomalies between chemical and macromolecular sequence
• Atom nomenclature
  – Conforms to IUPAC for standard amino acids and nucleotides
• Ligands and monomers
  – New dictionary with full chemical description (7800+ definitions)
  – All files annotated relative to this dictionary (163K non-polymers + 101K polymer residues)
• Biological assembly defined
  – Viruses now uniformly annotated (250 entries)
Nomenclature Standardization

- IUPAC-compliant H-atom names for standard amino acids and nucleotides (exceptions: OXT and HXT)
- DNA and RNA now differentiated (Adenine is DA for DNA; A for RNA)
- Modified nucleotides expressed as 3-letter codes
  - (replacing +T, +C … etc.)
- PDB asterisks replaced by single quotes in atom labels
  - (O2* becomes O2’)
- Unusual names made more conventional
  - (AC8 becomes C8A)
- White-space characters removed
  - (N 2 becomes N2)
Chemical Dictionary

- Stereochemical assignments
- Aromatic bond assignments
- Nomenclature
  - All atom labels begin with element type symbol
  - Retention of all prior names as an alternate identifier
- Model and idealized coordinates
- Chemical descriptors (SMILES & InChI)
- Systematic chemical names
- Redundant chemical components obsoleted
- Additional definitions for protonated forms

Worldwide Protein Data Bank
www.wwpdb.org
Protonation Variants

• Companion dictionary of 220 definitions
• Additional definitions provide nomenclature for protonation states of:
  – N, C -terminal and free amino f
  – common side chain variants
• IUPAC compliant names
• Covers BMRB & CCPN variants
Virus Structures
wwPDB Remediation Test Site


The Chemical Component Dictionary and smaller test sets of data files are also available for testing.

PDB users are encouraged to test the remediated data files between April and July 2007. As of July, this site containing remediated data files. The final transition date will be announced on this website.

Comments about the files should be sent to info@wwpdb.org. Major announcements will be made at the websites.

About the wwPDB Remediation Project

The evolution of experimental methods, functional knowledge of proteins, and methods used to process archive. The wwPDB has remediated these data to create a more uniform archive.

The results of these efforts can be found in the remediated coordinate files. Highlights include:

| Sequence | Updated references to databases and taxonomies
Resolved differences between chemical and macromolecular sequences |
| Citation | Verified and updated primary citation assignments |
| Assembly and virus information | Improved representation of deposited and experimental coordinate frame |
| Nucleic acid labelling | Deoxy and ribose nucleotides assigned separate chemical definitions. The RNA forms remain labeled as A, G, C, U, I. |
| Beamline data | Beamline and synchrotron facility names have been made consistent with the Journal of Structural Biology |
| Atom nomenclature | Standardized to reflect changes in Chemical Component Dictionary (see below) |

Properties

Name: ADENOSINE-5'-TRIPHOSPHATE
Formula: C10 H16 N5 O13 P3
Formal charge: 0
Molecular weight: 567.181 g/mol
Component type: NON-POLYMER

Chemical features

Atom count: 47
Chiral atom count: 6
Chiral atoms: PB PA C4' C3' C2' C1'
Bond count: 49
Aromatic bond count: 10
Testing and Roll-out

- Remediated data files and chemical dictionary previewed by 60 software developers and database maintainers (November 2006)
- Example files and chemical components dictionary released for public review (January 2007)
- Full remediated archive released for public review in April 2007
- Planned review period continues through July 2007

Images of remediated PDB entries 1tk0, 2bbv & 407d obtained using Jmol, Chimera & OpenRasmol
Welcome to the RCSB PDB

The RCSB PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

The RCSB is a member of the wwPDB whose mission is to ensure that the PDB archive remains an international resource with uniform data.

This site offers tools for browsing, searching, and reporting that utilize the data resulting from ongoing efforts to create a more consistent and comprehensive archive.

Information about compatible browsers can be found here.

A narrated tutorial illustrates how to search, navigate, browse, generate reports and visualize structures using this new site. [This requires the Macromedia Flash player download.]

Comments? info@rcsb.org

Molecule of the Month: Exosomes

Our genetic information is stored safely inside the nucleus of each cell. However, most of the action in a typical cell occurs outside the nucleus. Proteins are built in the cytoplasm, energy is produced in the mitochondria, and interactions with the environment occur at the cell surface. So, the nucleus needs a way to communicate with the rest of the cell. RNA molecules perform this job. They are the messengers that deliver genetic information from the nucleus to places where it is needed for synthesis and control.

More ...

Previous Features
Structure Summary Page

**Title**: RIBONUCLEASE A-URIDINE VANADATE COMPLEX: HIGH RESOLUTION RESOLUTION X-RAY STRUCTURE (1.3 Å)

**Authors**: Ladner, J.E., Wladkowsk, B., Svensson, L.A., Sjolin, L., Gilliland, G.L.


**History**: Deposition 1995-07-27 Release 1997-04-01

**Experimental Method**: Type: X-RAY DIFFRACTION  Data: N/A

**Parameters**
- Resolution: 1.25 Å
- R-Value: 0.197 (work)
- R-Free: n/a
- Space Group: P 2₁ (P 1 2₁ 1)

**Unit Cell**
- Length [Å]:
  - a: 29.80
  - b: 90.00
  - c: 30.20
  - α: 90.00
  - β: 106.10
  - γ: 90.00

**Molecular Description Asymmetric Unit**
- Polymer: 1
- Molecule: RIBONUCLEASE A
- Chain: _
- EC no.: 3.1.27.5

**Classification**
- **Source**: Polymer: 1
- **Scientific Name**: Bos taurus
- **Common Name**: Bovine
- **Type**: Bovine

**Chemical Component**
- **Identifier**: UVC
- **Name**: URIDINE-2',3'-VANADATE
- **Formula**: C₉H₁₂N₂O₅Y
- **Drug Similarity**: [View]

**SCOP Classification**
- **Class**: Alpha and beta proteins (a+b)
- **Family**: Ribonuclease A-like
- **Domain**: Ribonuclease A (also ribonuclease B, S)
- **Species**: Cow (Bos taurus)

**CATH Classification**
- **Domain**: 1ruv09
- **Class**: Alpha beta
- **Architecture**: Roll
- **Topology**: P-30
- **Homology**: P-30

**PFAM Classification**
- **Chain**: N
- **PFAM Accession**: PF00074
- **PFAM ID**: RNaseA
- **Description**: Pancreatic ribonuclease
- **Type**: Domain
- **Clan ID**: n/a
Search Options

Searching the RCSB PDB

- **Advanced Search**: Allows searches of all types - database fields, browsable ontologies, and text searches
- **Latest Release**: Shows the structures loaded this week
- **Sequence**: To search using a sequence, or by similarity to the sequence of a given PDB structure
- **Ligands**: To search based on ligand or ligand substructure
- **Models**: To search for modeled structures (those structures which are NOT experimentally determined)
- **Unreleased Entries**: Searches based on structures that are not yet released (minimal data available: ID, title, authors, and possibly sequence)
- **Structural Genomics Targets**: Links to TargetDB and a listing of Structural Genomics Target Info

Browsing the RCSB PDB

**Gene Ontology**: The GO Annotation project has mapped PDB IDs and corresponding chain IDs to the GO terms below.

- **Biological Process**
- **Cellular Component**
- **Molecular Function**

**Enzyme Classification**: Browse based on Swiss-Prot/GenBank/KEGG/author specified mapping of the enzyme to EC number

**Medical Subject Headings**: Browse based on MeSH terms, the National Library of Medicine’s controlled vocabulary

**Source Organism**: Browse based on NCBI Taxonomy

**Genome Location**: Browser based on Swiss-Prot/GenBank accession numbers associated with the structures and the loci

**SCOP Classification**: Browse based on Structural Classification of Proteins data

**CATH Classification**: Browse based on clustering at four major levels, Class(C), Architecture(A), Topology(T) and Homologous superfamily (H)
CATH is a novel hierarchical classification of protein domain structures, which clusters proteins at four major levels, Class(C), Architecture(A), Topology(T) and Homologous superfamily (H). Class, derived from secondary structure content, is assigned for more than 90% of protein structures automatically. Architecture, which describes the gross orientation of secondary structures, independent of connectivities, is currently assigned manually. The topology level clusters structures according to their topological connections and numbers of secondary structures. The homologous superfamilies cluster proteins with highly similar structures and functions. The assignments of structures to toplogy families and homologous superfamilies are made by sequence and structure comparisons.

Here you can browse by CATH structural classifications.
Advanced Search
Report Options
### Crystallization Description Report

<table>
<thead>
<tr>
<th>Structure ID</th>
<th>Crystallization Method</th>
<th>Temp °C</th>
<th>pH Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1H2A</td>
<td>VAPOR DIFFUSION, SITTING DROP</td>
<td>298.00</td>
<td>4.70</td>
</tr>
<tr>
<td>1DBF</td>
<td>VAPOR DIFFUSION, HANGING DROP</td>
<td>298.00</td>
<td>3.00</td>
</tr>
<tr>
<td>1JOG</td>
<td>VAPOR DIFFUSION, HANGING DROP</td>
<td>293.00</td>
<td>7.30</td>
</tr>
<tr>
<td>1JOV</td>
<td>VAPOR DIFFUSION, HANGING DROP</td>
<td>n/c</td>
<td>8.50</td>
</tr>
<tr>
<td>1LQA</td>
<td>VAPOR DIFFUSION, HANGING DROP</td>
<td>295.00</td>
<td>8.50</td>
</tr>
<tr>
<td>1M15</td>
<td>VAPOR DIFFUSION, HANGING DROP</td>
<td>293.00</td>
<td>8.50</td>
</tr>
<tr>
<td>1M2G</td>
<td>VAPOR DIFFUSION, HANGING DROP</td>
<td>293.00</td>
<td>8.50</td>
</tr>
<tr>
<td>1M13</td>
<td>VAPOR DIFFUSION, HANGING DROP</td>
<td>295.00</td>
<td>7.50</td>
</tr>
<tr>
<td>1N0D</td>
<td>VAPOR DIFFUSION, HANGING DROP</td>
<td>298.00</td>
<td>8.00</td>
</tr>
<tr>
<td>1NMP</td>
<td>VAPOR DIFFUSION, HANGING DROP</td>
<td>298.00</td>
<td>7.50</td>
</tr>
<tr>
<td>1PDD</td>
<td>VAPOR DIFFUSION, HANGING DROP</td>
<td>295.00</td>
<td>8.50</td>
</tr>
<tr>
<td>1R4G</td>
<td>VAPOR DIFFUSION, HANGING DROP</td>
<td>295.00</td>
<td>7.50</td>
</tr>
<tr>
<td>1R4B</td>
<td>VAPOR DIFFUSION, HANGING DROP</td>
<td>295.00</td>
<td>7.50</td>
</tr>
</tbody>
</table>
Access

RCSB PDB
  • http://www.pdb.org/

wwPDB
  • http://www.wwpdb.org/

wwPDB Remediation
  • http://remediation.wwpdb.org/

Dictionary & Schema Resources

TargetDB & PepcDB

RCSB Software Download Site
  • http://sw-tools.pdb.org
  • CVS server rcsb-cvs.rcsb.org - anonymous access
Acknowledgements

Operated by two members of the RCSB:

The RCSB PDB is a member of the

Supported by:

www.pdb.org • info@rcsb.org