Harvesting Data Collection Information for RCSB PDB Depositions

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

http://mmcif.pdb.org/

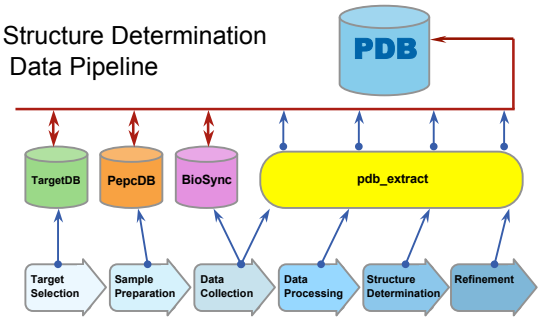
PDBML Schemas

http://pdbml.pdb.org/

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

PDBML: the representation of archival macromolecular structure data in XML. John Wesbrook, Nobutoshi Ito, Haruki Nakamura, Kim Henrick and Helen M. Berman, *Bioinformatics*, 21(7), 988-992, 2005.

Structure Determination Data Pipeline



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)

RCSB PDB PROTEIN DATA BANK www.pdb.org • info@rcsb.org

Protein Expression Purification and Crystallization Database

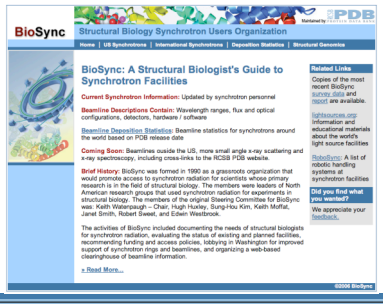
PepcDB • http://pepcdb.pdb.org/

- 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

RCSB PDB PROTEIN DATA BANK www.pdb.org • info@rcsb.org

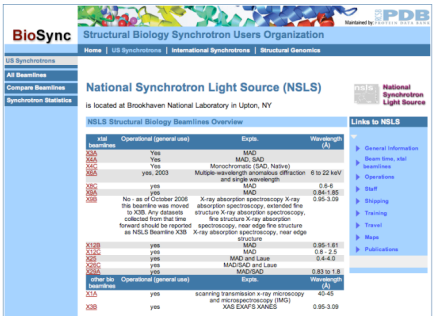
Data Collection - BioSync

http://biosync.rcsb.org/



RCSB PDB PROTEIN DATA BANK www.pdb.org • info@rcsb.org

Data Collection - BioSync



RCSB PDB PROTEIN DATA BANK www.pdb.org • info@rcsb.org

Data Collection - BioSync

Table 6: Structural Genomics structures in the PDB for which data were collected at NSLS

Synchrotron Beamline	Total Synchrotron Structures	Total SG Structures	PSI Structures	Non-PSI Structures in North America	SGC Structures	NPPSFA Structures	European SG Structures
X12B	185	7	3	3	1	0	0
X12C	461	67	63	4	0	0	0
X25	530	36	30	3	3	0	0
X29C	213	6	5	1	0	0	0
X29A	180	85	76	8	1	0	0
X3A	5	4	4	0	0	0	0
X4A	675	214	214	0	0	0	0
X4C	9	3	3	0	0	0	0
X8A	31	5	3	2	0	0	0
X8C	336	40	10	28	1	0	1
X9A	132	61	60	1	0	0	0
X9B	275	12	11	1	0	0	0
UNKNOWN	5	0	0	0	0	0	0
Total	2961	521	465	49	6	0	1

RCSB PDB PROTEIN DATA BANK www.pdb.org • info@rcsb.org

Data Collection - BioSync

The Advanced Photon Source (APS)

is located at Argonne National Laboratory in Argonne, IL

APS Structural Biology Beamlines Overview

xtal beamlines	Collaborative Access Team (CAT)	Operational (general use)	Expts.	Wavelength (Å)
19-ID	Argonne National Laboratory Structural Biology Center	yes	MAD, SAD	
22-BM	SER-CAT	yes	Macromolecular, MAD	0.886 - 1.908 Å
22-ID	SER-CAT	yes	Monochromatic	0.79-1.9
23-BM	GM/CA-CAT Sponsored by NCIARF & NCI	Available early 2007	MAD	0.7-1.9
23-ID-B	GM/CA-CAT Sponsored by NCIARF & NCI	yes	MAD/SAD	
23-ID-D	GM/CA-CAT Sponsored by NCIARF & NCI	yes	MAD/SAD	
24-ID-C	Northwestern Collaborative Access Team	yes	Macromolecular Crystallography, MAD/SAD	Designed for 5-23keV. Being comissioned for lower and higher energies...
31-ID	SGX-CAT	Yes	MAD / SAD Macromolecular crystallography	
18-ID	BioCAT	Available to general users	SAXS/WAXS, XAS, Fiber Diffraction	

National Synchrotron Light Source (NSLS)

is located at Brookhaven National Laboratory in Upton, NY

NSLS Structural Biology Beamlines Overview

xtal beamlines	Operational (general use)	Expts.	Wavelength (Å)
X4A	Yes	MAD, SAD	
X4C	Yes	Monochromatic (SAD, Native)	
X8C	yes, 2003	Multiple-wavelength anomalous diffraction and single-wavelength	6 to 22 keV
X25	yes	MAD	0.6-6
X9B	No - as of October 2006 this beamline was moved to X3B. Any datasets collected from that time forward should be reported as NSLS Beamline X3B	X-ray absorption spectroscopy, X-ray absorption spectroscopy, extended fine structure X-ray absorption spectroscopy, fine structure X-ray absorption spectroscopy, near edge fine structure X-ray absorption spectroscopy, near edge fine structure	0.95-3.09
X12B	yes	MAD	0.95-1.61
X12C	yes	MAD	0.8 - 2.5
X29	yes	MAD and Laue	0.4-4.0
X29C	yes	MAD/SAD and Laue	
X29A	yes	MAD/SAD	0.83 to 1.8

other bio beamlines	Operational (general use)	Expts.	Wavelength (Å)
X1A	yes	scanning transmission x-ray microscopy and microspectroscopy (IMG)	40-45
X3B	yes	XAS EXAFS XANES	0.95-3.09

RCSB PDB PROTEIN DATA BANK www.pdb.org • info@rcsb.org

pdb_extract

http://pdb-extract.pdb.org/
http://sw-tools.pdb.org/

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

www.pdb.org • info@rcsb.org

pdb_extract
http://pdb-extract.pdb.org/
http://sw-tools.pdb.org/

- 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

www.pdb.org • info@rcsb.org

Current Data Item Coverage

Data Item	PDB_EXTRACT # pop	PDB_EXTRACT # total	PDB_EXTRACT %	ALL # pop	ALL # total	ALL %
_refins.number_all	1049	2239	46.8	16540	37906	43.6
_refins.number_obs	2229	2239	99.5	34154	37906	90.1
_refins.pdbx_Rsym_value	693	2239	30.9	11225	37906	29.6
_refins.percent_possible_obs	2166	2239	96.7	32683	37906	86.2
_refins.pdbx_redundancy	1694	2239	75.6	26547	37906	70.0
_refins.pdbx_Rmerge_I_obs	1848	2239	82.5	23562	37906	62.1
_refins.pdbx_Rsym_value	693	2239	30.9	11225	37906	29.6
_refins.pdbx_netI_over_av_sigmaI	1782	2239	79.5	24282	37906	64.0

www.pdb.org • info@rcsb.org

Current Data Item Coverage

Data Item	PDB_EXTRACT # pop	PDB_EXTRACT # total	PDB_EXTRACT %	ALL # pop	ALL # total	ALL %
_refins_shell.meanI_over_sigI_obs	1348	2239	60.2	19285	37906	50.8
_refins_shell.number_measured_all	265	2239	11.8	283	37906	0.7
_refins_shell.number_measured_obs	450	2239	20.1	490	37906	1.2
_refins_shell.number_unique_all	1136	2239	50.7	9646	37906	25.4
_refins_shell.number_unique_obs	286	2239	12.7	296	37906	0.7
_refins_shell.percent_possible_all	1893	2239	84.5	28230	37906	74.4
_refins_shell.percent_possible_obs	553	2239	24.7	700	37906	1.8
_refins_shell.Rmerge_I_obs	1812	2239	80.9	19318	37906	50.9
_refins_shell.pdbx_redundancy	1506	2239	67.2	17988	37906	47.4
_refins_shell.pdbx_Rsym_value	654	2239	29.2	9664	37906	25.4
_exptl_crystal.density_Matthews	2201	2239	98.3	19680	37906	51.9
_exptl_crystal.density_percent_sol	2200	2239	98.2	36750	37906	96.9

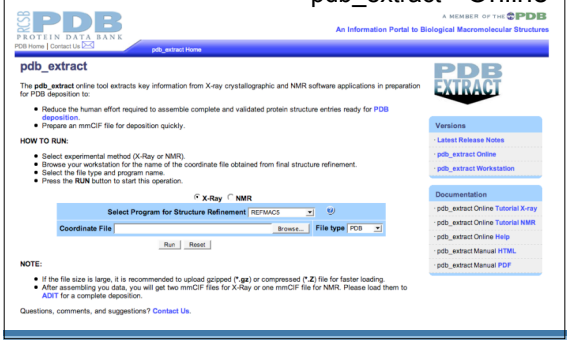
www.pdb.org • info@rcsb.org

Current Data Item Coverage

Data Item	PDB_EXTRACT # pop	PDB_EXTRACT # total	PDB_EXTRACT %	ALL # pop	ALL # total	ALL %
_diffrn.ambient_temp	2139	2239	95.5	32538	37906	85.8
_diffrn_detector.details	680	2239	30.3	10345	37906	27.2
_diffrn_detector.detector	2142	2239	95.6	33018	37906	87.1
_diffrn_detector.type	2112	2239	94.3	32289	37906	85.1
_diffrn_detector.pdbx_collection_date	2187	2239	97.6	34362	37906	90.6
_diffrn_radiation.monochromator	1206	2239	53.8	16397	37906	43.2
_diffrn_radiation.rcsb_diffrn_protocol	2239	2239	100.0	33669	37906	88.8
_diffrn_radiation_wavelength.wavelength	2164	2239	96.6	33233	37906	87.6
_diffrn_refins.number	88	2239	3.9	103	37906	0.2
_diffrn_refins.av_sigmaI_over_netI	11	2239	0.4	14	37906	0.0
_diffrn_source.type	2218	2239	99.0	32563	37906	85.9
_diffrn_source.pdbx_synchrotron_beamline	1835	2239	81.9	22849	37906	60.2
_diffrn_source.pdbx_synchrotron_site	1835	2239	81.9	23005	37906	60.6
_diffrn_source.source	2228	2239	99.5	32333	37906	85.3

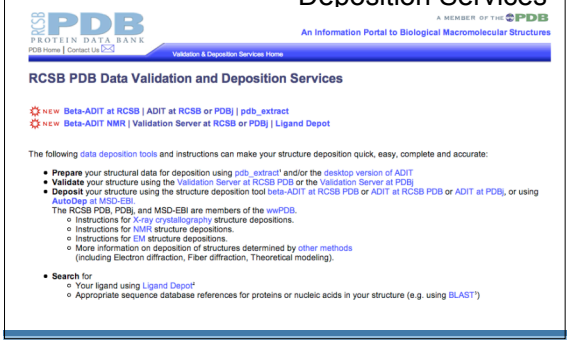
www.pdb.org • info@rcsb.org

pdb_extract - Online



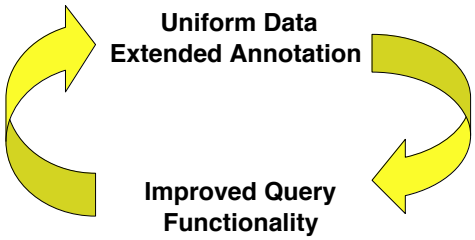
www.pdb.org • info@rcsb.org

Deposition Services



RCSB PDB PROTEIN DATA BANK — www.pdb.org • info@rcsb.org

Data Query Support Strategy



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.

Worldwide Protein Data Bank
www.wwpdb.org

Data Remediation

http://remediation.wwpdb.org

- 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)

Worldwide Protein Data Bank
www.wwpdb.org

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)

Worldwide Protein Data Bank
www.wwpdb.org

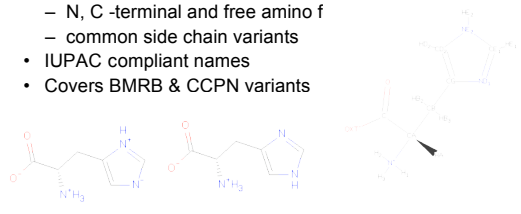
Chemical Dictionary

- Stereochemical assignments
- Aromatic bond assignments
- Nomenclature
  - IUPAC nomenclature for standard amino acids and nucleotides (Pure & Applied Chem., 70, 117-142, 1998)
  - 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

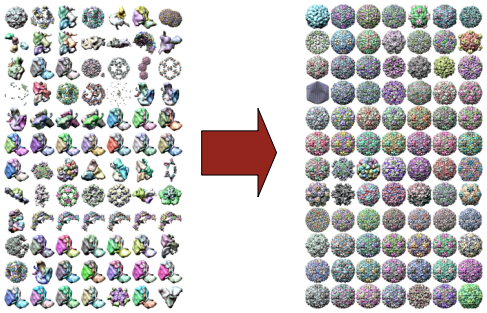
Protonation Variants

- 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



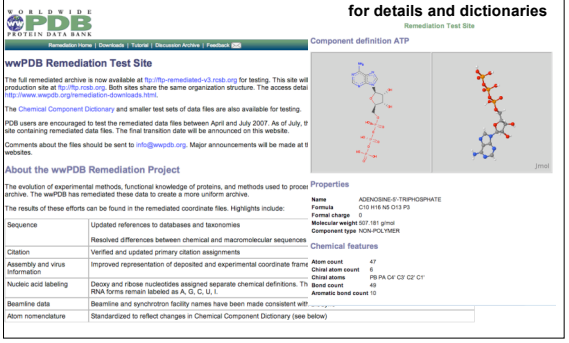
Worldwide Protein Data Bank
www.wwpdb.org

Virus Structures



Worldwide Protein Data Bank
www.wwpdb.org

http://remediation.wwpdb.org
for details and dictionaries

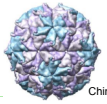


Worldwide Protein Data Bank
www.wwpdb.org

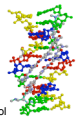
Testing and Roll-out

- Remediated data files and chemical dictionary previewed by 60 software developers and database maintainers (November 2006)
- Example files 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
- Further details of the review updated at http://www.pdb.org, http://remediation.wwpdb.org and wwPDB sites.


Jmol


Chimera

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


Rasmol

www.pdb.org • info@rcsb.org

Access

RCSB PDB
- http://www.pdb.org/

wwPDB
- http://www.wwpdb.org/

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

Dictionary & Schema Resources
- http://mmcif.pdb.org/ & http://pdbml.pdb.org

TargetDB & PepcDB
- http://targetdb.pdb.org & http://pepcdb.pdb.org

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

www.pdb.org • info@rcsb.org

The RCSB PDB Team

www.pdb.org • info@rcsb.org

Acknowledgements

Operated by two members of the RCSB: Rutgers, UCSD SDSC, Skaggs School of Pharmacy and Pharmaceutical Sciences

The RCSB PDB is a member of the Worldwide PDB

Supported by: NSF, NIGMS, Office of Science, National Library of Medicine, National Cancer Institute, National Center for Research Resources, NIBIB, National Institute of Neurological Disorders and Stroke, NIDDK