

Modifications to the Protein Data Bank: A new PDB format, Data Deposition, and Validation

Genji Kurisu

Institute for Protein Research



<http://pdbj.org/>
<http://wwpdb.org/>

Protein Data Bank Japan

<http://pd bj.org/>

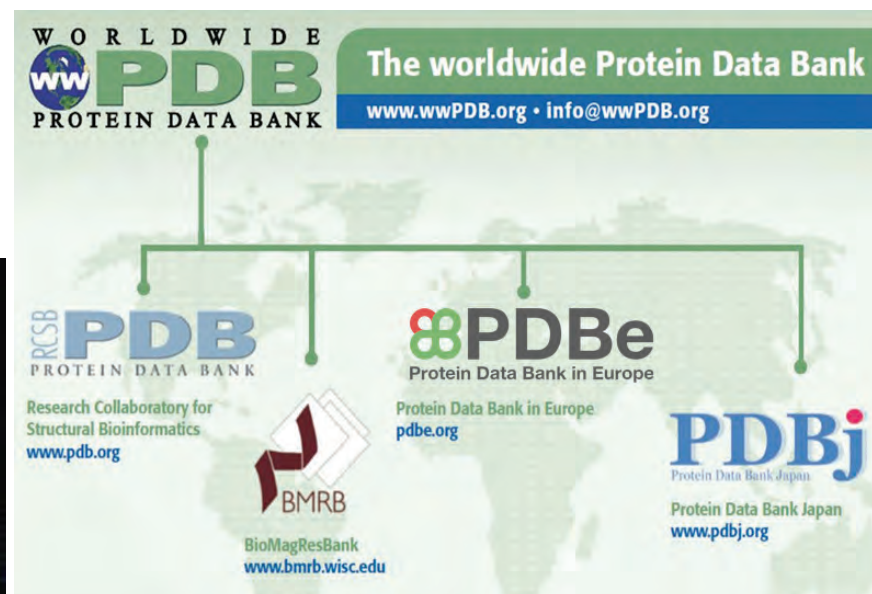
Since 2001, PDBj has been managed at **Institute for Protein Research, Osaka University** as a member of the **wwPDB**, to curate and process the deposited data for an open and single archive.



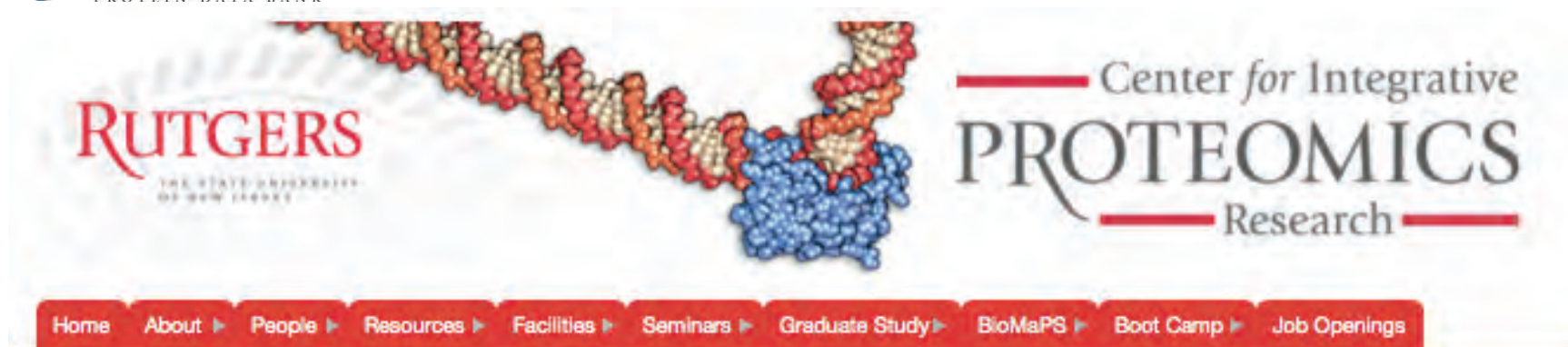
PDBj staffs (April 2014)



wwPDB.org



wwPDB members and their heads



Stephen K. Burley, M.D., D.Phil.

Director, Center for Integrative Proteomics Research

Associate Director, RCSB-PDB

Distinguished Professor, Department of Chemistry and Chemical Biology

Member, Cancer Institute of New Jersey



A new director of the RCSB-PDB

(Helen M. Berman is going to be a chair of the wwPDB)

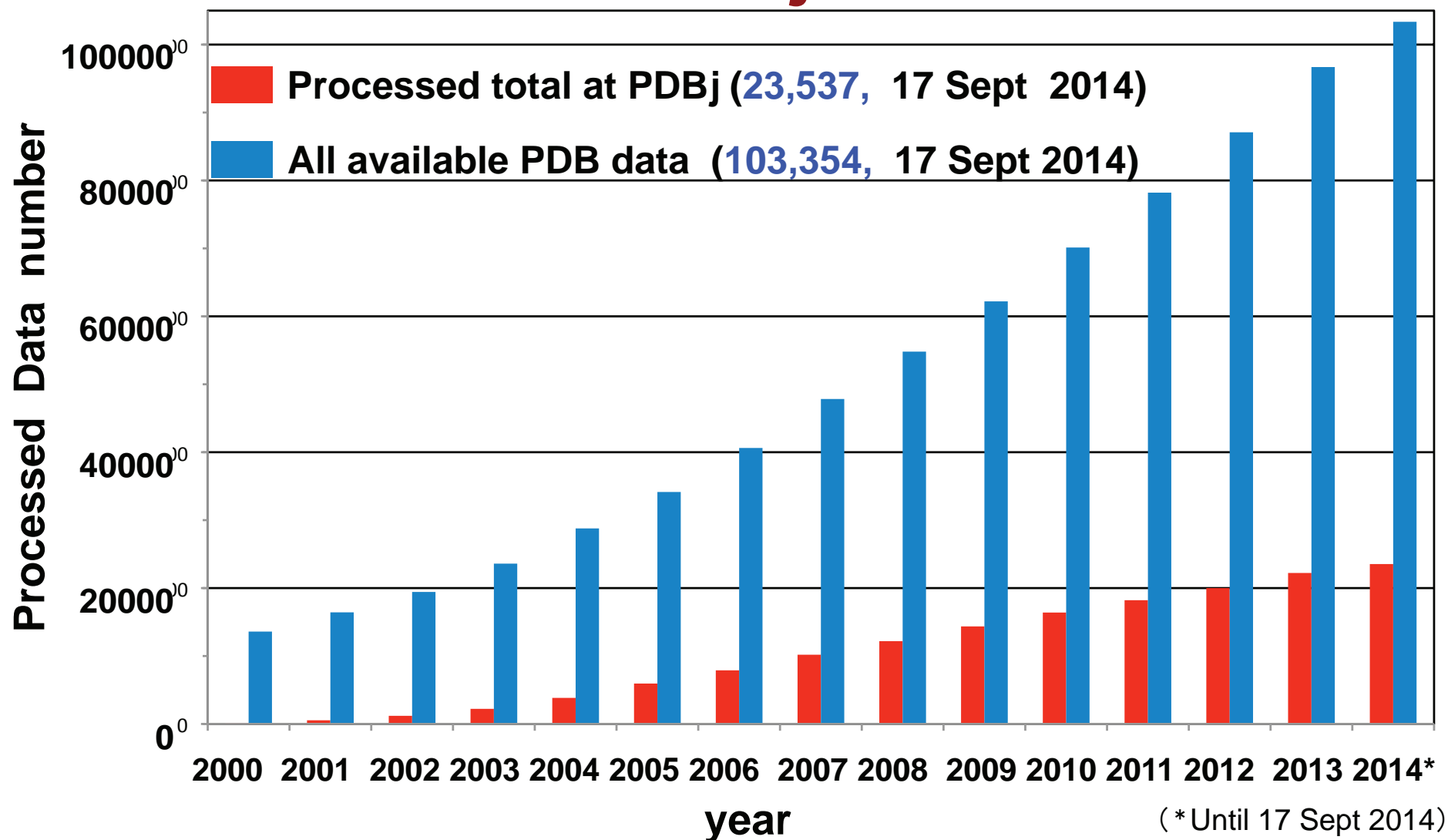
Activities/Services of each member of the wwPDB

- **“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, common archive as the ftp site and the characteristic services by each wwPDB member

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, common archive as the ftp site and the characteristic services by each wwPDB member

Data-in at PDBj and wwPDB



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

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, common archive as the ftp site and the characteristic services by each wwPDB member

Data-out from PDBj

<http://pd bj.org/>

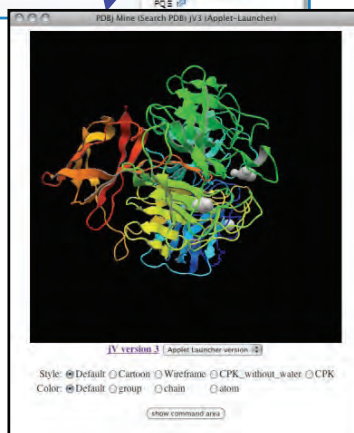
Amino acid sequence (FASTA)

```
>1GOF: GALACTOSE OXIDASE
ASAPIGSAISRNNAVTCDSAGSNECKAIDGNKDTFHTFYGANGDKFPHTYIDMK
TTQNVNGLSMLPRQDQNGWIGRBEVYLSSDGTWNGSPVAGSGWADSTTKYSNFPETP
ARIVRLVAITEARGQPWTSIAEINVQASSYTAPQFGLGRWGPTIDLPVFAAAALEPTS
GRVLNHSYVNDAPGGSGGIIITSSWSPGTQIVSDRTVTYKMDPCGISHDGOQIV
VTGONDAKTSLYSSSDSWIPGDMQVARGYQSSATHSDGRTVTIGGSGGVEKNGE
VYSFSKRTSLPNAKVNPHLTADKQGLYRSDNHAWLPGWKGSGVQAGSPSTAHNNYTS
GSGDVSAGKRSNPGVAPDAMCGNAVYDAVKGKILTFGGSPDYQSDATNAHILITLG
EPGTSPTNVFASNGLYFARTHTSVLPDGGSTPTGGQRRGIFEDSTPTVTPEIIVPEO
DTPYKQNPRIYVYHISILLPGSRVPMGGGLGCGCTNMFDAQIFPPVYLTNNGHL
ATRFKITRTSTQSEVVGGRITISTDSSISKALRYGTATHTVNTDQRRPILZLTNNGH
SYSEQVPSDGVALPGYVHMLFVHNSAGVPSVASTIRVTO
```

Functional site	focus & details	ACT SITE
1) A:495	<input type="checkbox"/> on <input type="checkbox"/> off	sequence Y
2) A:272	<input type="checkbox"/> on <input type="checkbox"/> off	description Proton acceptor
3) A:495	<input type="checkbox"/> on <input type="checkbox"/> off	source Swiss-Prot: 1
4) A:496	<input type="checkbox"/> on <input type="checkbox"/> off	
5) A:581	<input type="checkbox"/> on <input type="checkbox"/> off	
6) A:228	<input type="checkbox"/> on <input type="checkbox"/> off	
7) A:590	<input type="checkbox"/> on <input type="checkbox"/> off	
8) A:272	<input type="checkbox"/> on <input type="checkbox"/> off	
9) A:495	<input type="checkbox"/> on <input type="checkbox"/> off	
10) A:75-87	<input type="checkbox"/> on <input type="checkbox"/> off	
11) A:194	<input type="checkbox"/> on <input type="checkbox"/> off	
12) A:227-228	<input type="checkbox"/> on <input type="checkbox"/> off	
13) A:272	<input type="checkbox"/> on <input type="checkbox"/> off	

Data viewer at PDBj

Graphic viewer: jv
<http://pd bj.org/jv/>



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

Kinjo et al. NAR 40, D453 (2012)

New standard PDB format: PDBx/mmCIF

- Current 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**
- **PDBML (XML) and RDF format files are available.**
- **Start in 2014 and the current PDB format will be phased out in 2016.**

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

PDB

```

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

wwPDB Service site for a new format

<http://mmcif.wwpdb.org/>

PDBx/mmCIF Home Dictionaries Documentation Downloads Contact Us

PDBx/mmCIF Dictionary Resources

This site provides information about the format, dictionaries and related software tools used by the Worldwide Protein Data Bank ([wwPDB](http://wwpdb.org/)) to define data content for deposition, annotation and archiving of PDB entries.

[Browse the current dictionary »](#)

Dictionaries

- [Browse the current dictionary»](#)
- [Download/view all dictionaries »](#)
- [Search dictionaries»](#)

Documentation

- [PDB -> PDBx/mmCIF correspondences »](#)
- [PDBx/mmCIF for large structures »](#)
- [Software resources »](#)
- [C++ » and Python » programming examples](#)
- [File syntax » and dictionary organization »](#)
- [Atomic » and molecular » descriptions](#)
- [References »](#)
- [Glossary »](#)

FAQs

Questions about PDBx/mmCIF format, and data content, or software tools? Check out the [FAQ»](#)

wwPDB Service site for a new format

http://mmcif.pdbj.org/

PDBx/mmCIF
[トップ](#)
[辞書](#)
[文書](#)
[ダウンロード・サービス](#)
[お問い合わせ](#)


[検索](#)

PDBx/mmCIF 辞書関連情報

このサイトでは、PDBエントリーの登録、アノテーション、データ保管を行う際に 国際蛋白質構造データバンク ([wwPDB](#)) で利用されているファイルの書式、辞書、ソフトウェア・ツールに関する情報を提供しています。

[現行の辞書を見る](#)

辞書

- [現行の辞書を見る](#)
- [全ての辞書をダウンロードする](#)
- [辞書を検索する](#)

文書

- [PDB→PDBx/mmCIFの対応](#)
- [巨大構造のためのPDBx/mmCIF](#)
- [ソフトウェア](#)
- [C++とPythonのプログラム例](#)
- [ファイルの書式と辞書の構成](#)
- [原子と分子の記述方法](#)
- [参考文献](#)
- [用語集](#)

よくある質問

PDBx/mmCIFフォーマット、データ内容、ソフトウェアツールについての質問は、[よくある質問](#)を参照下さい。

このページの内容は wwPDB から提供されている原文を PDBj で翻訳し、適宜加筆修正したものです。

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** - **UCSF Chimera, Jmol, OpenRasMol, jV, molmil**

PDB actively working with community developers to help fill in missing functionalities.

Transitional Home for Large Structures



Large structure entries

Other lists

Entry ID	Structure Name	Authors	Deposit date	Release date	Modification date	Cite
3J3Q	ATOMIC-LEVEL STRUCTURE OF THE ENTIRE HIV-1 CAPSID	Zhao, G., Fenila, J.R., Yufenyay, E.L., Meng, X., Chen, B., Ning, J., Ann, J., Gronenborn, A.M., Schulten, K., Aiken, C., Zhang, P.	2013-04-12	2013-05-29	2013-06-12	Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. Nature, 497, 2013
3J3Y	ATOMIC-LEVEL STRUCTURE OF THE ENTIRE HIV-1 CAPSID (186 HEXAMERS + 12 PENTAMERS)	Zhao, G., Fenila, J.R., Yufenyay, E.L., Meng, X., Chen, B., Ning, J., Ann, J., Gronenborn, A.M., Schulten, K., Aiken, C., Zhang, P.	2013-05-06	2013-05-29	2013-06-12	Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. Nature, 497, 2013
3J6B	STRUCTURE OF THE YEAST MITOCHONDRIAL LARGE RIBOSOMAL SUBUNIT	Amunts, A., Brown, A., Bai, X.C., Llacer, J.L., Hussain, T., Emsley, P., Long, F., Murshudov, G., Scheres, S.H., Ramakrishnan, V.	2014-01-22	2014-04-09		Structure of the yeast mitochondrial large ribosomal subunit. Science, 343, 2014
4BP7	ASYMMETRIC STRUCTURE OF A VIRUS-RECEPTOR COMPLEX	Derk, K.C., Thompson, R., Baker, A.M., Hiscor, J.A., Barr, J.N., Stockley, P.G., Ranson, N.A.	2013-07-17	2013-09-11		The Asymmetric Structure of an Icosahedral Virus Bound Its Receptor Suggests a Mechanism for Genome Release. Structure, 21, 2013
4BTS	THE CRYSTAL STRUCTURE OF THE EUKARYOTIC 40S RIBOSOMAL SUBUNIT IN COMPLEX WITH EIF1 AND EIF1A	Weisser, M., Voigts-Hofmann, F., Rabi, J., Leibundgut, M., Ban, N.	2013-05-24	2013-07-17		The Crystal Structure of the Eukaryotic 40S Ribosomal Subunit in Complex with EIF1 and EIF1A. Nat. Struct. Mol. Biol., 2013
4CTF	THE LIMITS OF STRUCTURAL PLASTICITY IN A PICORNAVIRUS CAPSID REVEALED BY A MASSIVELY EXPANDED EQUINE RHINITIS A VIRUS PARTICLE	Baker, S.E., Groppelli, E., Pearson, A.R., Stockley, P.G., Rowlands, D.J., Ranson, N.A.	2014-04-02	2014-05-21	2014-05-28	Limits of Structural Plasticity in a Picornavirus Capsid Revealed by a Massively Expanded Equine Rhinitis A Virus Particle. J. Virol., 88, 2014
4CTG	THE LIMITS OF STRUCTURAL PLASTICITY IN A PICORNAVIRUS CAPSID REVEALED BY A MASSIVELY EXPANDED EQUINE RHINITIS A VIRUS PARTICLE	Baker, S.E., Groppelli, E., Pearson, A.R., Stockley, P.G., Rowlands, D.J., Ranson, N.A.	2014-05-01	2014-05-21	2014-05-28	Limits of Structural Plasticity in a Picornavirus Capsid Revealed by a Massively Expanded Equine Rhinitis A Virus Particle. J. Virol., 88, 2014
4NWR	COMPUTATIONALLY DESIGNED TWO-COMPONENT SELF-ASSEMBLING TETRAHEDRAL CAGE T33-28	King, N.P., Sale, J.B., Shetter, W., McLamane, D.E., Gonen, S., Gonen, T., Yeates, T.O., Baker, D.	2013-12-06	2014-05-28		Accurate design of co-assembling multi-component protein nanomaterials. Nature, 2014
4O9Y	CRYSTAL STRUCTURE OF TCDA1	Meusch, D., Getzoglanni, C., Bhemov, R.G., Lang, A.E., Hofhage, O., Vetter, I.R., Aldones, K., Rausner, S.	2014-01-03	2014-02-26	2014-03-05	Mechanism of Tc toxin action revealed in molecular detail. Nature, 2014

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

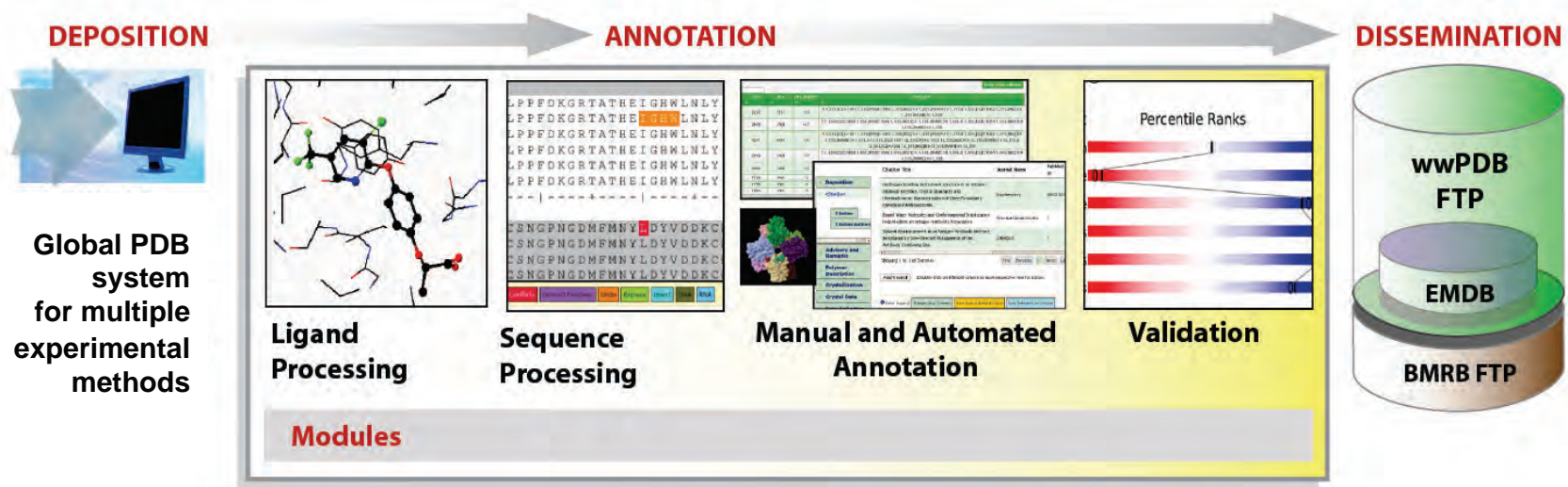


HIV-1 Capsid 3J3Q

- 1356 chains
- >2M atoms
- 25 – PDB format entries

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 will be made publicly available through the wwPDB ftp sites



wwPDB X-ray Structure Validation Report i

Aug 13, 2013 – 09:45 AM BST

PDB ID : 1CBS
 Title : CRYSTAL STRUCTURE OF CELLULAR RETINOIC-ACID-BINDING PROTEINS I AND II IN COMPLEX WITH ALL-TRANS-RETINOIC ACID AND A SYNTHETIC RETINOID
 Authors : Kleywegt, G.J.; Bergfors, T.; Jones, T.A.
 Deposited on : 1994-09-28
 Resolution : 1.80 Å (reported)

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	: 4.02b-467
Mogul	: 1.15 2013
Xtriage (Phenix)	: dev-1323
EDS	: trunk21216
Percentile statistics	: 20591
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

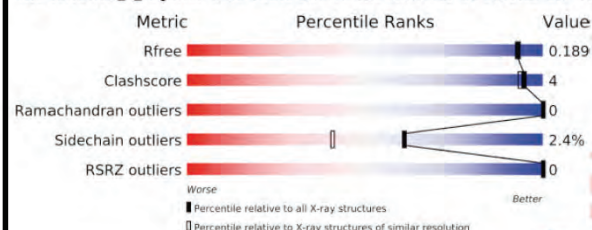
Validation Report

- **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

1 Overall quality at a glance ⁽ⁱ⁾

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.



Metric	Whole archive (#Entries)	Similar resolution (#Entries, resolution range(Å))
R_{free}	65580	5522 (1.84-1.76)
Clashscore	76988	5040 (1.82-1.78)
Ramachandran outliers	75395	6528 (1.84-1.76)
Sidechain outliers	75377	6529 (1.84-1.76)
RSRZ outliers	65576	5522 (1.84-1.76)

The table below summarises 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.

Mol	Chain	Length	Quality of chain
1	A	137	<div><div></div></div>

Mol	Chain	Length	Quality of chain
1	A	371	<div><div></div></div>
1	C	371	<div><div></div></div>

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

Mol	Type	Chain	Res	Geometry	Electron density
2	NAG	A	401	-	X
2	NAG	C	401	-	X

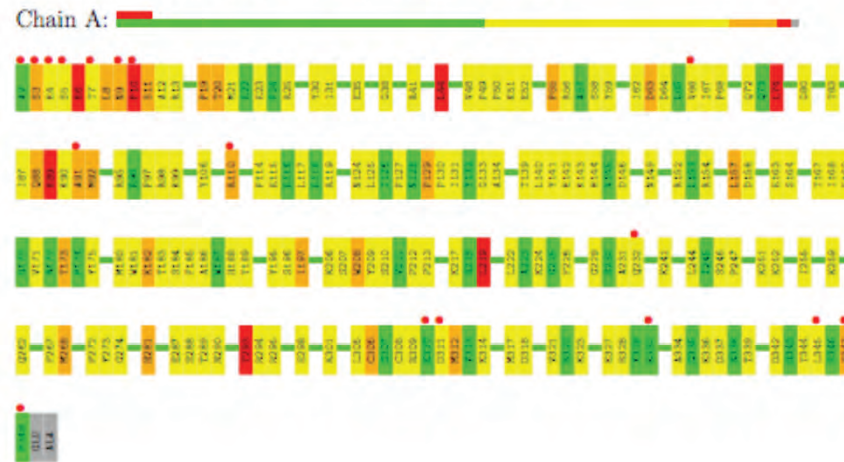
Validation Report

- Residue quality
 - Red dots: poor density ($RSR-Z > 2$, as in EDS)
- Model/data fit ligands etc.
 - “*LLDF*” – *Local Ligand Density Fit* = Z-score of ligand RSR relative to nearby polymeric residues

3 Residue-property plots

The first graphic for a chain summarises the proportions of errors displayed in the second graphic. The second graphic shows the sequence annotated by issues in geometry and electron density. Residues are color-coded according to the number of different types of geometric errors found. Green signifies no errors, yellow, orange and red 1, 2, and 3 or more errors respectively. A red dot above a residue indicates a problem with electron density. Regions of sequence for which no errors are detected are indicated by a green connector.

- Molecule 1: Jumonji domain-containing protein 2A



PDBj members in 2014

- **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.)
 - Bekker, Gert-Jan (IPR, Osaka Univ.)
- **Group for BMRB**
 - Fujiwara, Toshimichi, Ph. D. (Group Leader, Prof. Osaka Univ.)
 - Akutsu, Hideo, Ph. D. (Guest Prof., IPR, Osaka Univ.)
 - Kojima, Chojiro, Ph. D. (IPR, Osaka Univ.)
 - Kobayashi, Naohiro, Ph. D. (IPR, Osaka Univ.)
 - Iwata, Takeshi (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.)