

PDB検証レポートを利用した PDBj検索サービスの紹介

第8回 生命医学情報学連合大会 PDBjランチョンセミナー
横地政志 (Masashi Yokochi)
PDBj @ 蛋白質研究所 大阪大学

オープンデータ¹とPDBの活動²

- ・ 誰でも利用できる
- ・ 透明性の確保
- ・ 専門家と知識、経験の共有
- ・ 改善の提案、採択、実装、周知
- ・ 質と量の改善
 - ・ 包括的、正確な質の高いデータの公開
 - ・ データは平易な用語で記述
 - ・ データの質(特性)の記述
 - ・ 品質を保証する改善の継続
 - ・ 将来の世代の技術革新のためのデータ公開

- ・ FAIR原則
 -  Findable
 -  Accessible
 -  Interoperable
 -  Reusable
- ・ タスクフォースの結成
- ・ 諮問委員会(AC)、 
- ・ PDB/BMRB/EMDBコアアーカイブ
- ・ PDBx/mmCIF
- ・ 検証レポート
- ・ 検証レポートの改善 (Ligand/EM/NMR)
- ・ PDBコアアーカイブ／検証レポートのRDF化

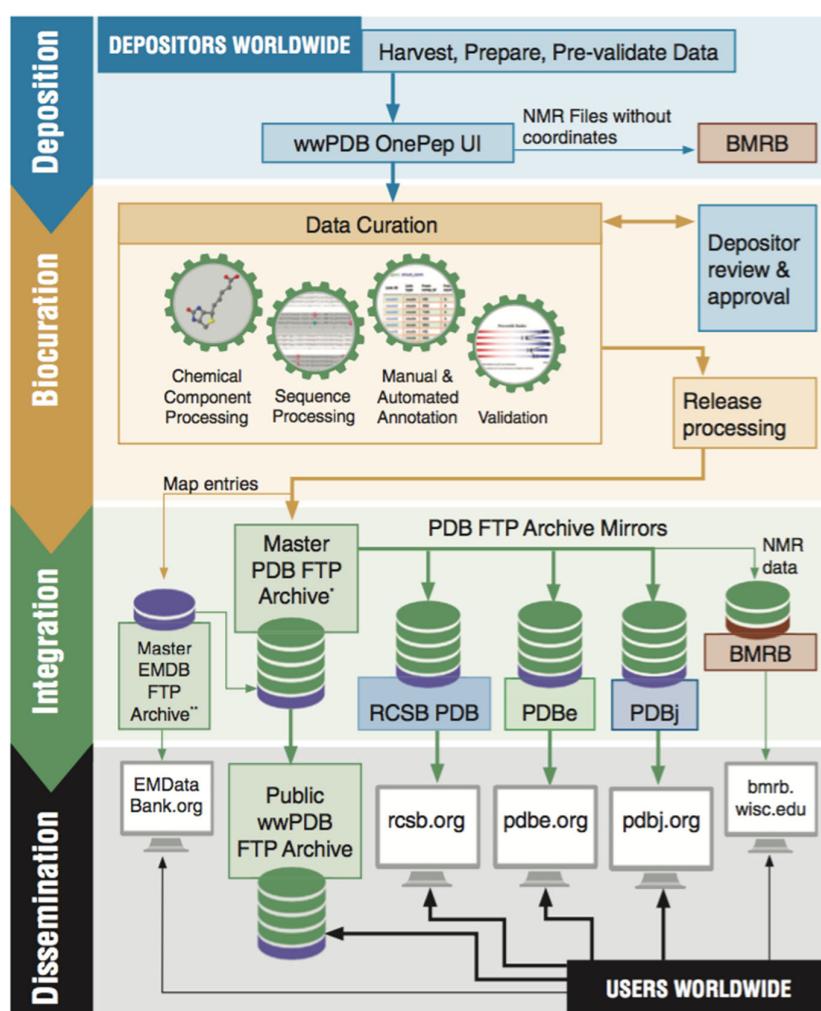
¹ オープンデータ憲章より抜粋 <http://current.ndl.go.jp/node/23799>

² <https://www.wwpdb.org>

今年3月、PDBエントリー件数は150K+



1982年 : 0.1K+
 1993年 : 1K+
 1997年 : 5K+
 2000年 : 10K+
 2008年 : 50K+
 2013年 : 100K+
 2019年 : 150K+
 2023年? : 200K+



Full wwPDB X-ray Structure Validation Report [\(1\)](#)

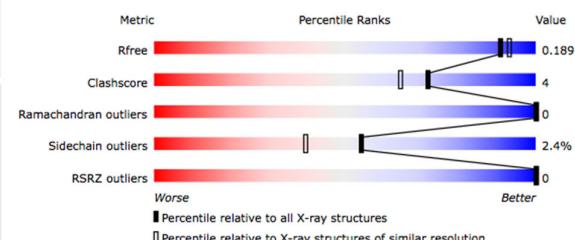
Feb 12, 2017 – 07:52 pm GMT

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)

This is a Full wwPDB X-ray Structure Validation Report for a publicly released PDB entry.

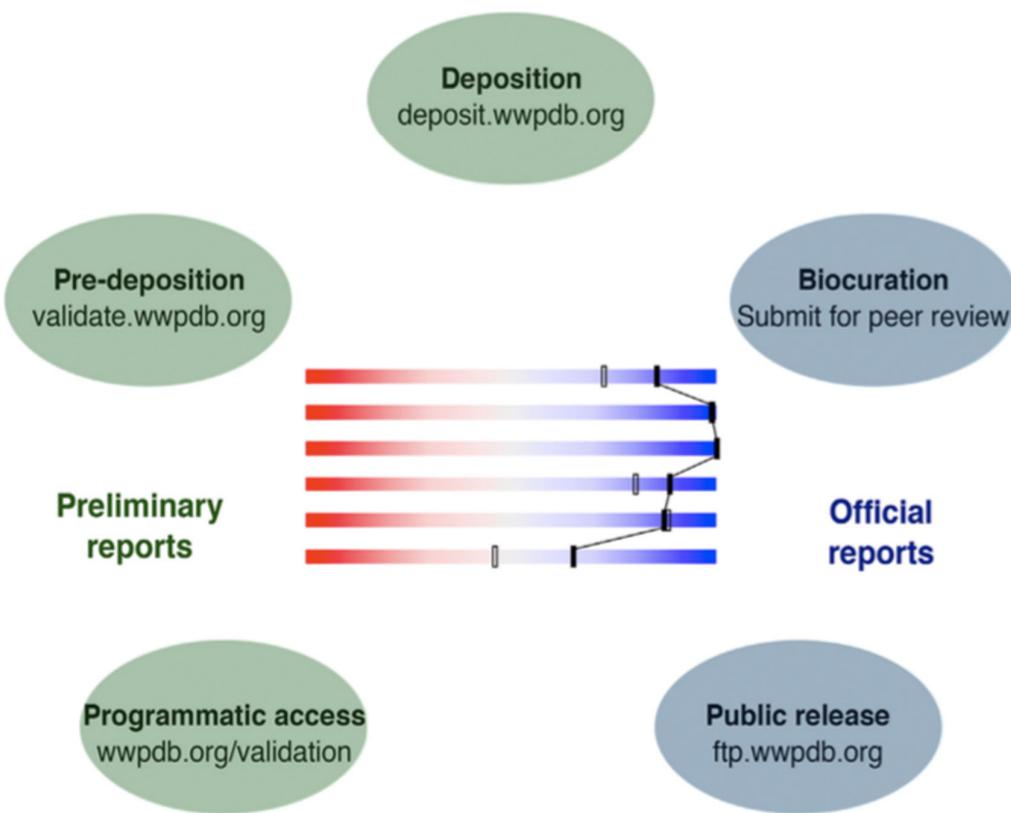
We welcome your comments at validation@mail.wwpdb.org
 A user guide is available at
<http://wwpdb.org/validation/2016/XrayValidationReportHelp>
 with specific help available everywhere you see the [\(1\)](#) symbol.

[wwPDB validation report \(PDF\)](#)



Percentile view of validation report

wwPDB validation reports



Structure 25, 1916–1927, 2017

PDB検証レポートの検証項目

X-ray/EM/NMR

- Geometric & conformational
 - bond, angle, planarity
 - protein backbone conformation
 - protein side-chain conformation
- Atomic & molecular interaction
 - all-atom contacts
 - under packing
 - hydrogen bond quality
- Non-protein
 - nucleic acids (RNA pucker, suite)
 - carbohydrates (N-glycan core)
 - ligands (CSD)
 - ions & other solvent
- Incomplete model (e.g. CA_ONLY)

X-ray

- Structure factor & electron density
 - Wilson plot outliers, tNCS
 - wrong space group
 - twinning
 - agreement (R_{free} , RSR, RSCC)

NMR

- Chemical shifts
 - completeness
 - outliers
 - estimated reference error
 - random coil index
- Structure ensembles
 - representative model (medoid)
 - domain detection

Validation software used in wwPDB validation report

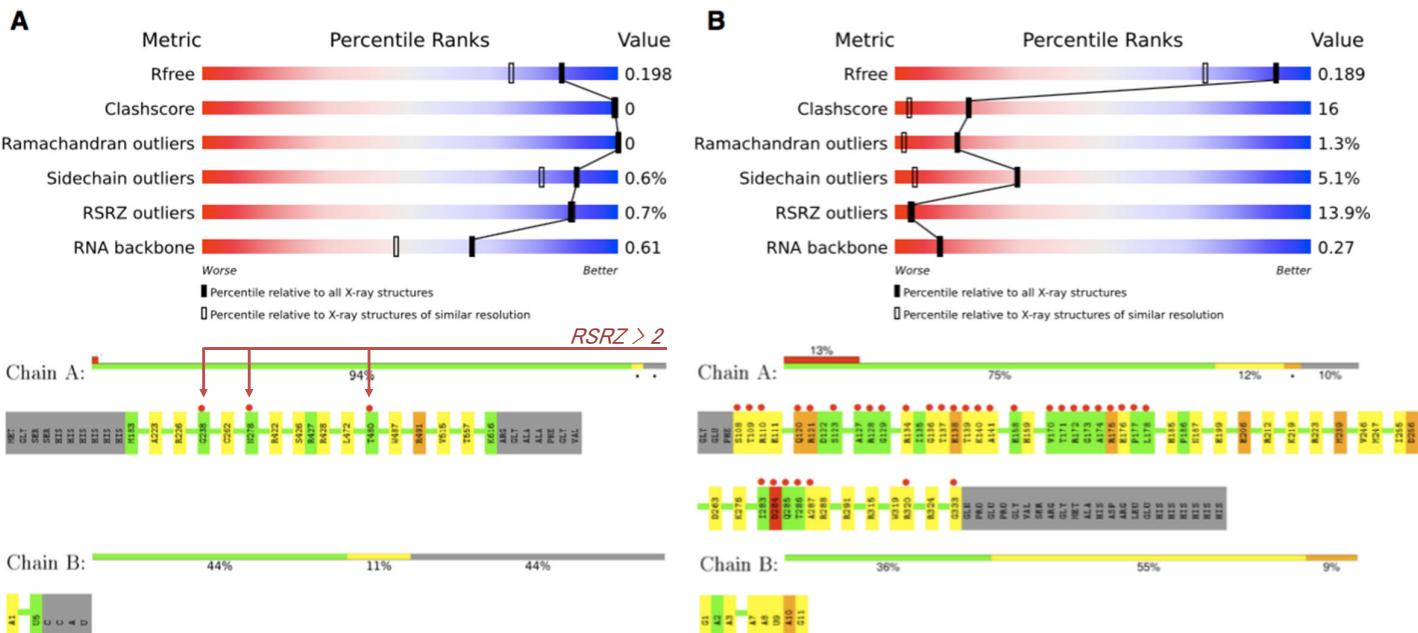
Table 3. Component Software Packages Included in the 2017 Version of the Validation Pipeline

Software Package	Which Section and Metric of the Report the Package Is Used for	Reference
MolProbity	model geometry: bond lengths and bond angles of standard protein residues and nucleotides, too-close contacts, Ramachandran outliers, rotamer outliers, RNA suiteness	Chen et al., 2010
MAXIT	model geometry: symmetry-related too-close contacts, stereochemistry issues, identification of <i>cis</i> -peptides	Maxit (Z.F., https://sw-tools.rcsb.org/apps/MAXIT/index.html)
Mogul	model geometry: bond-length and bond-angle outliers in small molecules	Bruno et al., 2004
Xtriage (Phenix)	crystallographic data and refinement statistics: signal-to-noise, twinning	Adams et al., 2010
DCC	crystallographic data and refinement statistics: R , R_{free} fit to crystallographic data: R_{free}	Yang et al., 2016
EDS	fit to crystallographic data: real-space R outliers	Kleywegt et al., 2004
Cyrange	NMR ensemble composition: identification of well-defined protein cores	Kirchner and Güntert, 2011
RCI	NMR chemical shifts: prediction of protein backbone order parameter from chemical shifts	Berjanskii and Wishart, 2005
PANAV	NMR chemical shifts: suggested referencing corrections in chemical shift assignments	Wang et al., 2010

Structure 25, 1916–1927, 2017

PDB検証レポート (PDF版要約)

Summary quality metrics in wwPDB validation reports



PDB検証レポート (PDF版テーブル)

• Standard geometry

Mol	Chain	Bond lengths RMSZ # Z >5	Bond angles RMSZ # Z >5
1	A	0.47 0/1107	0.71 0/1491

There are no bond length outliers.

There are no bond angle outliers.

There are no chirality outliers.

There are no planarity outliers.

• Too close contacts

Mol	Chain	Non-H	H(model)	H(added)	Clashes	Symm-Clashes
1	A	1091	0	1106	7	0
2	A	22	0	27	2	0
3	A	100	0	0	2	0
All	All	1213	0	1133	9	0

• Protein backbones

Mol	Chain	Analysed	Favoured	Allowed	Outliers	Percentiles
1	A	135/137 (98%)	132 (98%)	3 (2%)	0	100 100

• Protein sidechains

Mol	Chain	Analysed	Rotameric	Outliers	Percentiles
1	A	123/123 (100%)	120 (98%)	3 (2%)	52 38

• Ligand geometry

Mol	Type	Chain	Res	Link	Bond lengths			Bond angles		
					Counts	RMSZ	# Z > 2	Counts	RMSZ	# Z > 2
2	REA	A	200	-	19,22,22	1.05	1 (5%)	26,30,30	1.02	2 (7%)

All (1) bond length outliers are listed below:

Mol	Chain	Res	Type	Atoms	Z	Observed(Å)	Ideal(Å)
2	A	200	REA	C1-C6	2.25	1.56	1.53
2	A	200	REA	C18-C5-C6	2.08	126.83	127.31

All (2) bond angle outliers are listed below:

Mol	Chain	Res	Type	Atoms	Z	Observed(°)	Ideal(°)
2	A	200	REA	C11-C10-C9	-2.40	123.89	127.31
2	A	200	REA	C18-C5-C6	2.08	126.83	124.51

There are no chirality outliers.

There are no torsion outliers.

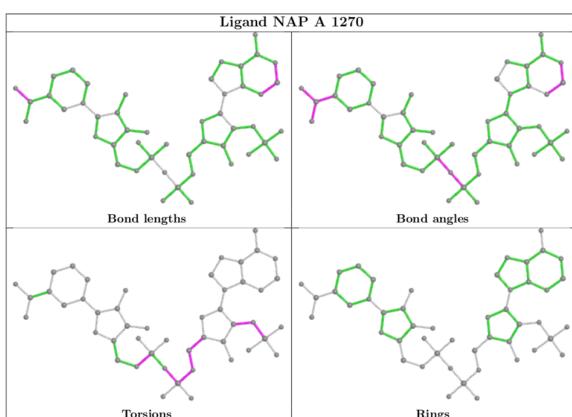
There are no ring outliers.

1 monomer is involved in 2 short contacts:

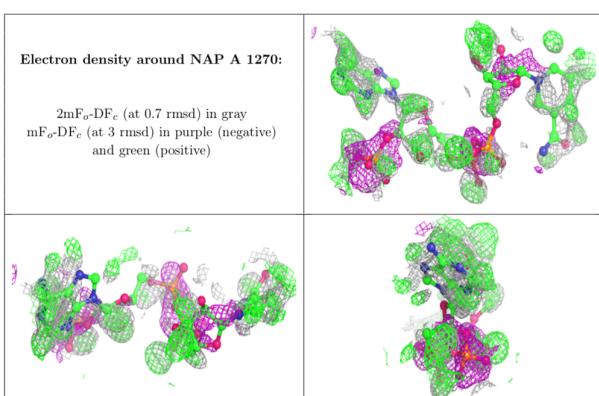
Mol	Chain	Res	Type	Clashes	Symm-Clashes
2	A	200	REA	2	0

PDB検証レポートの改善 — リガンドの検証項目の可視化と電子密度マップの公開

Geometric analysis provided by CCDC Mogul will be highlighted on a 2D diagram of the ligand, as shown below.



In addition to geometric validation for ligands, for X-ray diffraction PDB entries the wwPDB validation report also presents images displaying the ligand and the surrounding electron density map.



今年6月より、リガンド可視化情報をwwPDBの検証レポートに組み込みました。このリガンド可視化情報は、登録者が「Ligand of Interest」として選択したリガンドか、分子量が250より大きく、かつ構造化学的に標準値から外れた値を持つリガンドに対して表示されます。

全ての妥当性検証項目を含む検証レポート(XML版)の入手法

The screenshot shows the validation report page for entry 1cbs. It lists several files for download:

- PDBx/mmCIF: 1cbs.cif.gz (36.71 KB)
- PDBML: 1cbs.xml.gz (49.11 KB), 1cbs-noatom.xml.gz (11.63 KB), 1cbs-extatom.xml.gz (27.12 KB)
- PDBMLplus: 1cbs-plus.xml.gz (51.85 KB), 1cbs-plus-noatom.xml.gz (14.36 KB), 1cbs-add.xml.gz (2.73 KB)
- RDF: 1cbs.rdf.gz (23.62 KB)
- 構造因子: 1cbssf.ent.gz (149.64 KB)
- 生物学的単位 (PDB形式): 1cbs.pdb1.gz (25.94 KB) (A) *author defined assembly, 1 molecule(s) (monomeric)
- PDF: 1cbs_validation.pdf.gz (411.76 KB)
- PDF-full: 1cbs_full_validation.pdf.gz (411.94 KB)
- 検証レポート (XML): 1cbs_validation.xml.gz (8.17 KB) (highlighted with a red box)
- PNG: 1cbs_multipercentile_validation.png.gz (140.86 KB)
- SVG: 1cbs_multipercentile_validation.svg.gz (904 B)

On the right side, there is a sidebar with links to other databases and services like FSSP, SCOP, VAST, PISA, UniProt, PFam, e-site, and Promote Elastic.

At the bottom right, the URL <https://pdbj.org> is shown.

mmCIF化された次世代のPDB検証レポート

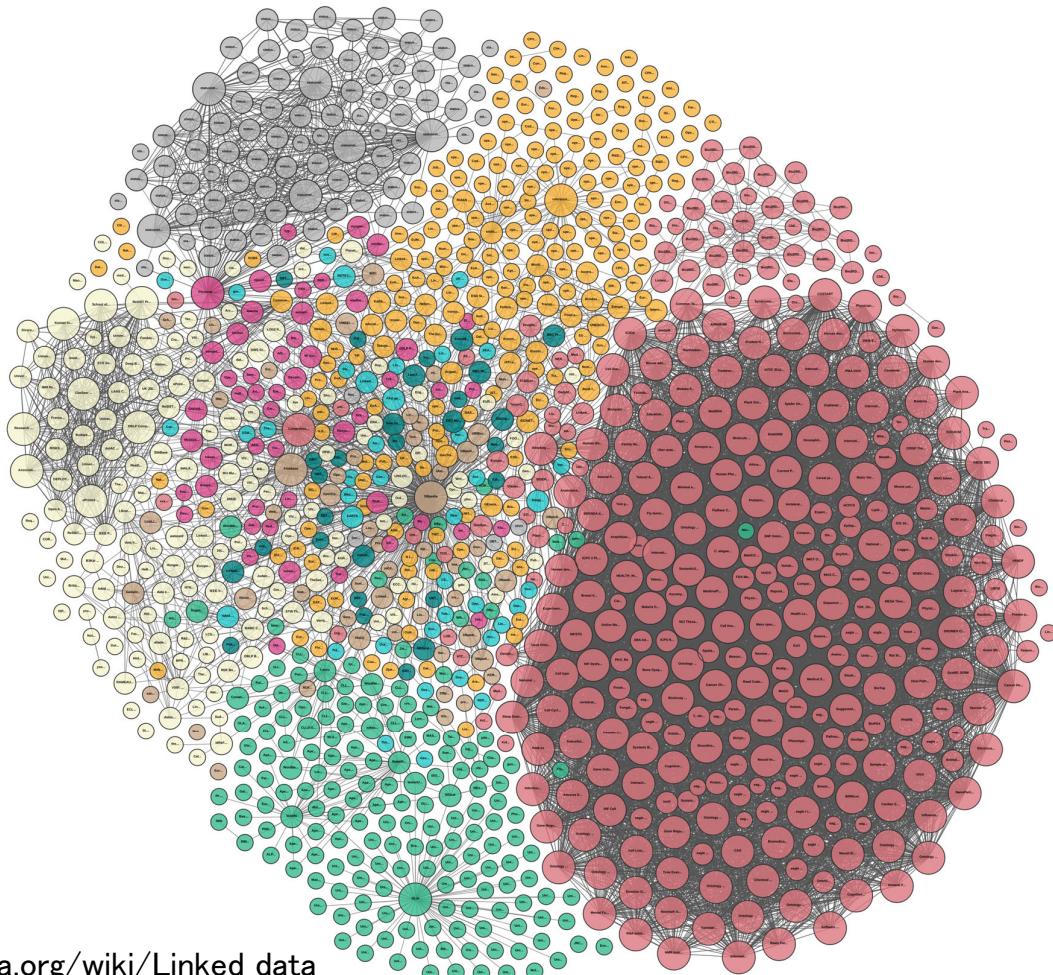
```
2505 lines (2504 sloc) | 163 KB
1 data_5B1L-validation-alt
2 #
3 _entry.id 5B1L
4 #
5 _pdbx_percentile_list.entry_id 5B1L
6 _pdbx_percentile_list.name all,2.35,xray
7 #
8 loop_
9 _pdbx_percentile_view.entry_id
10 _pdbx_percentile_view.conditions_id
11 _pdbx_percentile_view.type
12 _pdbx_percentile_view.rank
13 5B1L 1 all_atom_clashscore 51.6
14 5B1L 2 all_atom_clashscore 55.1
15 5B1L 3 Ramachandran_outlier_percent 100.0
16 5B1L 4 Ramachandran_outlier_percent 100.0
17 5B1L 5 rotamer_outliers_percent 94.3
18 5B1L 6 rotamer_outliers_percent 97.9
19 5B1L 7 R_value_R_free 48.5
20 5B1L 8 R_value_R_free 68.3
21 5B1L 9 RSRZ_outliers_percent 72.0
22 5B1L 10 RSRZ_outliers_percent 81.2
23 #
24 loop_
25 _pdbx_percentile_conditions.id
26 _pdbx_percentile_conditions.number_entries_total
27 _pdbx_percentile_conditions.ls_d_res_high
28 _pdbx_percentile_conditions.ls_d_res_low
29 1 112137 ? ?
30 2 1626 2.34 2.38
31 3 110173 ? ?
32 4 1605 2.34 2.38
33 5 110143 ? ?
34 6 1606 2.34 2.38
35 7 100710 2.34 2.38
```

セマンティック拡張されたPDB検証レポート

	PDF	XML	mmCIF	PDBML	PDB/RDF
可読性	✓	-	-	-	-
検証情報	要約、可視化	全項目	完全 (エンティティを含む)	完全 (エンティティを含む)	全項目
検索可能性	-	✓ (XQuery)	✓ (SQL)	✓ (SQL, XQuery)	✓ (SPARQL)
PDBx/mmCIF と互換性	-	-	~90%	~90%	~90%
オープンデータ 対応	-	-	-	-	✓
目的	論文査読、 個別研究	データ交換	データ交換、ア ーカイブ	データ交換、 高速検索	集合知、 機械学習

↔ 生命科学のオープンデータ

URIsで結ばれる生命科学系オープンデータ (2017年時点)



SPARQL endpoint contains wwPDB/RDF-validation graph

<https://bmrbpub.pdbj.org>

PDBj-BMRB Data Server:
common open representations of BMRB NMR-STAR data in XML, RDF and JSON formats

Home Search Examples Download Resources NEWS

Virtuoso SPARQL Query Editor

About | Namespace Prefixes | Inference rules

Default Data Set Name (Graph IRI)
<https://rdf.wwpdb.org/pdb-validation>

Query Text

```
select distinct ?Concept where {} a ?Concept} LIMIT 100
```

(Security restrictions of this server do not allow you to retrieve remote RDF data, see [details](#).)

Results Format:

Execution timeout: milliseconds (values less than 1000 are ignored)

Options: Strict checking of void variables

(The result can only be sent back to browser, not saved on the server, see [details](#))

Query examples

Category holders

1. Select all category holders of datablock class of BMRB entry 15400: [Show](#)
2. Select all category holders of datablock class of Metabolomics entry bmse000400: [Show](#)

Entry statistics

3. Count entries per submission year and experimental method (subtype): [Show](#)

Assembly descriptions

4. Select all assembly names, asym IDs, entity IDs, polymer types, formula weights and functions in a assembly: [Show](#)

Entity descriptions

5. Select all entity names and sequences of polymer entities expressed using one-letter code: [Show](#)
6. Select all original source information of molecular entities and external links to NCBI Taxonomy: [Show](#)
7. Select all biological systems to produce molecular entities and external links to NCBI Taxonomy: [Show](#)

Citation information

8. Select citation information of all entries together with

Linked Open Data (LOD) の検索入門

検索: リガンドのRSR (実験的に求めた電子密度と構造モデルの電子密度の残差) が
10%より小さい全ての酵素-リガンド複合体を選択

```
SELECT ?PDB_ID ?enzyme ?ligand ?comp_id MIN(?RSR AS ?minRSR)
FROM <https://rdf.wwpdb.org/pdb-validation>
WHERE {

    ?entity PDBov:link_to_enzyme ?link_to_enzyme ;
        PDBov:entity:pdbx_description ?enzyme ;
        PDBov:of_datablock ?datablock .

    BIND (SUBSTR(STR(?datablock),38,4) AS ?PDB_ID)
    BIND (IRI(CONCAT(?datablock, "/pdbx_entity_nonpolyCategory")) AS ?entity_nonpoly_category)

    ?entity_nonpoly_category PDBov:has_pdbx_entity_nonpoly ?entity_nonpoly .

    ?entity_nonpoly PDBov:pdbx_entity_nonpoly.name ?ligand ;
        PDBov:pdbx_entity_nonpoly.entity_id ?entity_id ;
        PDBov:pdbx_entity_nonpoly.comp_id ?comp_id .

    FILTER (?ligand!="water" && !STRENDS(?ligand, "ION"))
    BIND (IRI(CONCAT(?datablock, "/pdbx_nonpoly_schemeCategory")) AS ?nonpoly_scheme_category)

    ?nonpoly_scheme_category PDBov:has_pdbx_nonpoly_scheme ?nonpoly_scheme .

    ?nonpoly_scheme PDBov:pdbx_nonpoly_scheme.pdb_strand_id ?asym_id ;
        PDBov:pdbx_nonpoly_scheme.pdb_seq_num ?seq_id ;
        PDBov:pdbx_nonpoly_scheme.entity_id ?entity_id ;
        PDBov:pdbx_nonpoly_scheme.mon_id ?comp_id .

    BIND (IRI(CONCAT(?datablock, "/pdbx_dcc_mapCategory")) AS ?dcc_map_category)

    ?dcc_map_category PDBov:has_pdbx_dcc_map ?dcc_map .

    ?dcc_map PDBov:pdbx_dcc_map.auth_asym_id ?asym_id ;
        PDBov:pdbx_dcc_map.auth_comp_id ?comp_id ;
        PDBov:pdbx_dcc_map.RSR ?RSR .

    FILTER (xsd:float(?RSR) < 0.1)
```

↳ selection of enzyme

↳ ligand selection

↳ RSR < 0.1

Linked Open Data (LOD) の検索入門

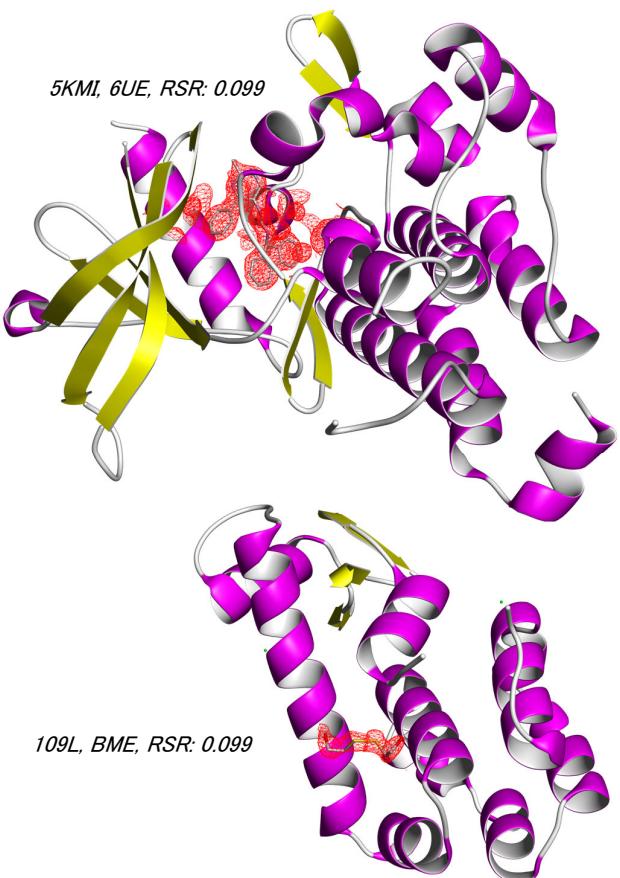
検索: リガンドのRSR (実験的に求めた電子密度と構造モデルの電子密度の残差) が 10%より小さい全ての酵素-リガンド複合体を選択

15000件余の酵素-リガンドの組み合わせがヒット

PDB_ID	enzyme	comp_id	RSR
5KMI	High affinity nerve growth factor receptor	6UE	0.099
109L	T4 LYSOZYME	BME	0.099
5CST	Ribosome inactivating protein	GOL	0.099
4MSN	cAMP and cAMP-inhibited cGMP 3',5'-cyclic phosphodiesterase 10A	2ZQ	0.099
5OX3	Glycogen phosphorylase, muscle form	PLP	0.099

⋮

4FON	Proteinase K	TE	0.016
2OLC	Methylthioribose kinase	HO	0.014
4NSG	Lysozyme C	QPT	0.012
1IXB	SUPEROXIDE DISMUTASE	MH2	0.008



Linked Open Data (LOD) の検索入門

検索: 蛋白質配列中のRSRの異常値 (RSRZ>2、残基毎のRSRのZスコアが 2σ より大きい) の割合が1%より小さい全ての酵素-リガンド複合体を選択

PREFIX PDBov: <<https://rdf.wwpdb.org/schema/pdbx-validation-v1.owl#>>

```
SELECT ?PDB_ID ?enzyme (GROUP_CONCAT(?ligand; SEPARATOR=",") AS ?ligands) ?RSRZ_outliers_percent
FROM <https://rdf.wwpdb.org/pdb-validation>
WHERE {
```

```
?map_overall PDBov:pdbx_dcc_map_overall.entry_id ?PDB_ID ;
PDBov:pdbx_dcc_map_overall.RSRZ_outliers_percent ?RSRZ_outliers_percent .
```

↗ % of outliers in RSR < 1%

```
FILTER (xsd:float(?RSRZ_outliers_percent) < 0.1)
```

```
BIND (IRI(CONCAT("https://rdf.wwpdb.org/pdb-validation/", ?PDB_ID, "/entityCategory")) AS ?entity_category)
```

```
?entity_category PDBov:has_entity ?entity .
```

↗ selection of enzyme

```
?entity PDBov:link_to_enzyme ?link_to_enzyme ;
PDBov:entity.pdbx_description ?enzyme .
```

```
BIND (IRI(CONCAT("https://rdf.wwpdb.org/pdb-validation/", ?PDB_ID, "/pdbx_entity_nonpolyCategory")) AS ?entity_nonpoly_category)
```

```
?entity_nonpoly_category PDBov:has_pdbx_entity_nonpoly ?entity_nonpoly .
```

↗ ligand selection

```
?entity_nonpoly PDBov:pdbx_entity_nonpoly.name ?ligand .
```

```
FILTER (?ligand!="water" && !STRENDS(?ligand, " ION"))
```

```
}
```

Linked Open Data (LOD) の検索入門

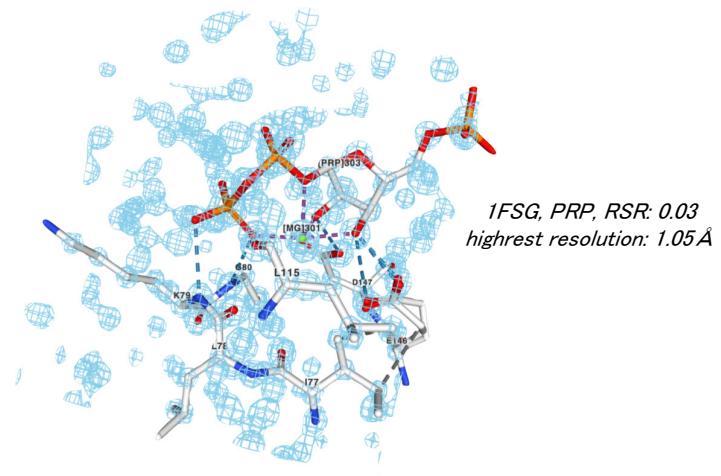
検索: 蛋白質配列中のRSRの異常値 (RSRZ>2、残基毎のRSRのZスコアが 2σ より大きい) の割合が1%より小さい全ての酵素-リガンド複合体を選択

5000件余の酵素-リガンドの組み合わせ、さらに0%に制限した場合は、1000件余

PDB_ID	Enzyme	RSRZ_outliers_percent
1BUL	NMC-A BETA-LACTAMASE	0
5A1G	S-ADENOSYLMETHIONINE SYNTHASE ISOFORM TYPE-2	0
2DRS	Xylanase Y	0
2AS1	Cytochrome c peroxidase, mitochondrial	0
1H4W	TRYPSIN IVA	0
142L	T4 LYSOZYME	0
4CIK	PLASMINOGEN	0
4L4O	Endo-1,4-beta-xylanase	0
4G5P	Epidermal growth factor receptor	0
3GA6	Exodeoxyribonuclease	0
2HB3	Protease	0
5IHG	Lysozyme C	0
1LNF	THERMOLYSIN	0
3IUY	Probable ATP-dependent RNA helicase DDX53	0
163L	T4 LYSOZYME	0
5BMP	Phosphoglucomutase	0
1LB3	T4 LYSOZYME	0
1GWO	PEROXIDASE C1A	0
5PA5	Catechol O-methyltransferase	0

酵素側のRSRの異常の割合が0%であり、リガンドのRSRが10%以下に制限した場合、670件

PDB_ID	enzyme	comp_id	RSR
1C1M	PROTEIN (PORCINE ELASTASE)	XE	0.025
5KPU	Beta-lactamase TEM	XE	0.026
1BT3	PROTEIN (CATECHOL OXIDASE)	C2O	0.027
1FSG	HYPOXANTHINE-GUANINE PHOSPHORIBOSYLTRANSFERASE	PRP	0.03
181L	T4 LYSOZYME	BNZ	0.032



PDB検証レポートの改善の継続

- 2019年10月、PDBj Mine検索サービス上でPDB検証レポートの内容が検索可能
- 2019年末までに、EMマップ、NMR制限情報に関する検証項目の追加
- 2020~2021年、PDB検証レポートはmmCIFをマスター/フォーマットに変更