Protein Data Bankの新しい登録システムと構造評価ツール

New deposition system and a validation tool of Protein Data Bank

中川 敦史
Atsushi Nakagawa

大阪大学蛋白質研究所
Institute for Protein Research

https://pdbj.org/
https://wwpdb.org/
PDBx/mmCIF
What is the problem?
Problems of “PDB format”

- PDB format is more than 40 years old and does not support today's science.
- PDB Record format limitations:
  - Max. 62 chains (even with some tricks)
  - Max. 99,999 atoms
- No bond orders or chirality specified for ligands
- No real support for NMR, EM, hybrid methods, …
- Meta-data specification cumbersome and inflexible
Too big to handle…

- HIV-1 capsid (3J3Q)
  - 1,356 chains
  - 2,440,800 atoms
  - 25 PDB entries
    - 1VU5, 1VU6, ...
  - 3J3Q only as mmCIF & PDBML
More problems...

- Very complicated “REMARK” lines
- Inflexibility of data definition

For example, residue numbers (author-defined or universal?)

REMARK 2 RESOLUTION.  2.60 ANGSTROMS.

REMARK 200 EXPERIMENTAL DETAILS
REMARK 200  EXPERIMENT TYPE                : X-RAY DIFFRACTION
REMARK 200  DATE OF DATA COLLECTION        : NULL
REMARK 200  TEMPERATURE           (KELVIN) : 100
REMARK 200  PH                             : 6.5
REMARK 200  NUMBER OF CRYSTALS USED        : 1

REMARK 350 BIOMOLECULE: 1
REMARK 350  AUTHOR DETERMINED BIOLOGICAL UNIT: MONOMERIC
REMARK 350  SOFTWARE DETERMINED QUATERNARY STRUCTURE: MONOMERIC
REMARK 350  SOFTWARE USED: PISA
REMARK 350  APPLY THE FOLLOWING TO CHAINS: A
REMARK 350       BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350       BIOMT2   1  0.000000  1.000000  0.000000        0.00000
REMARK 350       BIOMT3   1  0.000000  0.000000  1.000000        0.00000

REMARK 465 MISSING RESIDUES
REMARK 465  THE FOLLOWING RESIDUES WERE NOT LOCATED IN THE
REMARK 465  EXPERIMENT.  (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN
REMARK 465  IDENTIFIER; SSSEQ=SEQUENCE NUMBER; I=INSERTION CODE.)
REMARK 465
REMARK 465       M RES C SSSEQI
REMARK 465       GLY A   123
REMARK 465       ALA A   124
REMARK 465       SER A   125
So new(ish) formats...

“The wwPDB has established PDBx/mmCIF as the new standard format for data exchange and archiving in structural biology.”
**mmCIF**

mmCIF (macromolecular Crystallographic Information Format)

- was developed under the auspices of the International Union of Crystallography (IUCr) for macromolecular structures from the Crystallographic Information Format (CIF) widely used for small molecules.

- is a STAR (Self-defining Text Archive and Retrieval) format.

- has a very strict dictionary.

mmCIF is a list of *data items*, which consist a pair of *name* and *value*.
An example of mmCIF

_cell.length_a                   98.000
_cell.length_b                   89.400
_cell.length_c                   86.700
_cell.angle_alpha                90.00
_cell.angle_beta                117.80
_cell.angle_gamma                90.00
_symmetry.space_group_name_H-M   'C 2        '

loop_
_atom_site.label_seq_id
_atom_site.group_PDB
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_comp_id
_atom_site.auth_seq_id
_atom_site.label_asym_id
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.id

1  ATOM  N  N    ALA  201  A  38.840   0.236   1.012  1.00 34.65  1
1  ATOM  C  CA   ALA  201  A  38.356  -0.999   0.357  1.00 42.26  2
1  ATOM  C  C    ALA  201  A  37.098  -1.547   1.056  1.00 41.25  3
1  ATOM  O  O     ALA  201  A  36.619  -0.946   2.028  1.00 29.44  4

data item = name + value.
Hierarchy of mmCIF

The *name*, not the data structure, of mmCIF has hierarchy.

```
_entity_poly.type       'polypeptide(L)'
_cell.length_a          87.433
```

Knowing the meaning of *category groups* would help understanding of the data in mmCIF.
Some of the category groups

- **entity** biological & chemical information about the object in the entry.
- **atom** atomic information (name, coordinate etc)
- **struct** characteristic structure (secondary structure, crosslink etc)
- **chem_comp** chemical information (ligands etc)
- **citation** citation (journals)
- **refln** reflection data (Rsym, resolution etc)
- **exptl** experimental data (crystallization condition etc)
- **refine** refinement
- **symmetry** symmetry
- **cell** unit cell
“label” & “auth”

• **label**  Mandatory term defined by PDB
  →  Strict, and easy to use in comprehensive cases

• **auth**  Defined by the authors
  →  problem-specific. Often those appear in the “PDB file”
<table>
<thead>
<tr>
<th>Atom</th>
<th>Type</th>
<th>Chain</th>
<th>Residue</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>Occupancy</th>
<th>Biso</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOM 187</td>
<td>N</td>
<td>N</td>
<td>SER A 27</td>
<td>21.428</td>
<td>-12.414</td>
<td>-12.809</td>
<td>1.00</td>
<td>27.27</td>
</tr>
<tr>
<td>ATOM 188</td>
<td>C</td>
<td>CA</td>
<td>SER A 27</td>
<td>20.880</td>
<td>-12.688</td>
<td>-14.140</td>
<td>1.00</td>
<td>24.00</td>
</tr>
<tr>
<td>ATOM 189</td>
<td>C</td>
<td>C</td>
<td>SER A 27</td>
<td>20.911</td>
<td>-11.408</td>
<td>-14.954</td>
<td>1.00</td>
<td>26.14</td>
</tr>
<tr>
<td>ATOM 190</td>
<td>O</td>
<td>O</td>
<td>SER A 27</td>
<td>21.257</td>
<td>-10.346</td>
<td>-14.429</td>
<td>1.00</td>
<td>28.43</td>
</tr>
<tr>
<td>ATOM 192</td>
<td>C</td>
<td>OG</td>
<td>SER A 27</td>
<td>18.601</td>
<td>-12.708</td>
<td>-12.920</td>
<td>1.00</td>
<td>31.92</td>
</tr>
<tr>
<td>ATOM 193</td>
<td>N</td>
<td>N</td>
<td>THR A 28</td>
<td>20.573</td>
<td>-11.487</td>
<td>-16.237</td>
<td>1.00</td>
<td>25.11</td>
</tr>
<tr>
<td>ATOM 194</td>
<td>C</td>
<td>CA</td>
<td>THR A 28</td>
<td>20.617</td>
<td>-10.283</td>
<td>-17.059</td>
<td>1.00</td>
<td>23.13</td>
</tr>
</tbody>
</table>
Parent-Child relationship

<table>
<thead>
<tr>
<th>Parent</th>
<th>Child</th>
</tr>
</thead>
<tbody>
<tr>
<td>loop_</td>
<td>loop_</td>
</tr>
<tr>
<td>_struct_asym.id</td>
<td>_entity_poly_seq.entity_id</td>
</tr>
<tr>
<td>_struct_asym.pdbx_blank_PDB_chainid_flag</td>
<td>_entity_poly_seq.num</td>
</tr>
<tr>
<td>_struct_asym.pdbx_modified</td>
<td>_entity_poly_seq.mon_id</td>
</tr>
<tr>
<td>_struct_asym.entity_id</td>
<td>entity_poly_seq.hetero</td>
</tr>
<tr>
<td>_struct_asym.details</td>
<td>1 1 ACE n</td>
</tr>
<tr>
<td>A N N 1?</td>
<td>1 2 SER n</td>
</tr>
<tr>
<td>B N N 2?</td>
<td>.</td>
</tr>
<tr>
<td>C N N 1?</td>
<td>2 1 VAL n</td>
</tr>
<tr>
<td>D N N 2?</td>
<td>2 2 GLU n</td>
</tr>
<tr>
<td>E N N 3?</td>
<td>.</td>
</tr>
<tr>
<td>loop_</td>
<td>loop_</td>
</tr>
<tr>
<td>_entity.id</td>
<td>_entity.id</td>
</tr>
<tr>
<td>_entity.type</td>
<td>_entity.type</td>
</tr>
<tr>
<td>_entity.src_method</td>
<td>_entity.src_method</td>
</tr>
<tr>
<td>_entity.pdbx_description</td>
<td>_entity.pdbx_description</td>
</tr>
<tr>
<td>_entity.formula_weight</td>
<td>_entity.formula_weight</td>
</tr>
<tr>
<td>entity.pdbx_number_of_molecules</td>
<td>1 polymer man 'HEMOGLOBIN (DEOXY) (ALPHA CHAIN)' 15683.402 2</td>
</tr>
<tr>
<td></td>
<td>2 polymer man 'HEMOGLOBIN (DEOXY) (BETA CHAIN)' 16153.497 2</td>
</tr>
<tr>
<td></td>
<td>3 non-polymer syn 'PROTOPOPHRIN IX CONTAINING FE' 616.498 4</td>
</tr>
<tr>
<td></td>
<td>4 water nat water 18.015 38</td>
</tr>
</tbody>
</table>
Dictionary of PDBx/mmCIF
PDBx/mmCIF Dictionary Resources

http://mmcif.wwpdb.org/
Direct comparison of PDB and mmCIF
**EXPDTA**  
X-RAY DIFFRACTION

<table>
<thead>
<tr>
<th>_exptl.method</th>
<th>'X-RAY DIFFRACTION'</th>
</tr>
</thead>
</table>

**REMARK**  
2 RESOLUTION. 1.70 ANGSTROMS.

<table>
<thead>
<tr>
<th>_refine.ls_d_res_high</th>
<th>1.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
<td>Value</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>_cell.length_a</td>
<td>62.650</td>
</tr>
<tr>
<td>_cell.length_b</td>
<td>96.300</td>
</tr>
<tr>
<td>_cell.length_c</td>
<td>62.650</td>
</tr>
<tr>
<td>_cell.angle_alpha</td>
<td>90.00</td>
</tr>
<tr>
<td>_cell.angle_beta</td>
<td>90.00</td>
</tr>
<tr>
<td>_cell.angle_gamma</td>
<td>90.00</td>
</tr>
<tr>
<td>_cell.Z_PDB</td>
<td>4</td>
</tr>
<tr>
<td>_symmetry.space_group_name_H-M</td>
<td>'P 1 21 1'</td>
</tr>
<tr>
<td>Field</td>
<td>Value</td>
</tr>
<tr>
<td>------------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>_struct_ref.id</td>
<td>1</td>
</tr>
<tr>
<td>_struct_ref.db_name</td>
<td>UNP</td>
</tr>
<tr>
<td>_struct_ref.db_code</td>
<td>GAOA_DACDE</td>
</tr>
<tr>
<td>_struct_ref.pdbx_db_accession</td>
<td>Q01745</td>
</tr>
<tr>
<td>_struct_ref_seq.seq_align_beg</td>
<td>1</td>
</tr>
<tr>
<td>_struct_ref_seq.seq_align_end</td>
<td>639</td>
</tr>
<tr>
<td>_struct_ref_seq.db_align_beg</td>
<td>42</td>
</tr>
<tr>
<td>_struct_ref_seq.db_align_end</td>
<td>680</td>
</tr>
</tbody>
</table>

**COMPND** 4 EC: 1.1.3.9;

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>_entity.id</td>
<td></td>
</tr>
<tr>
<td>_entity.type</td>
<td></td>
</tr>
<tr>
<td>_entity.src_method</td>
<td></td>
</tr>
<tr>
<td>_entity.pdbx_description</td>
<td></td>
</tr>
<tr>
<td>_entity.formula_weight</td>
<td></td>
</tr>
<tr>
<td>_entity.pdbx_ec</td>
<td></td>
</tr>
<tr>
<td>1 polymer</td>
<td>man 'GALACTOSE OXIDASE' 68579.250 1.1.3.9</td>
</tr>
</tbody>
</table>
HETNAM     CU COPPER (II) ION
HETNAM     NA SODIUM ION
HETNAM     ACY ACETIC ACID
FORMUL    2   CU    CU  2+
FORMUL    3   NA    NA  1+
FORMUL    4   ACY   2(C2 H4 O2)
FORMUL    6   HOH   *316(H2 O)

loop_
_entity.id
_entity.type
_entity.src_method
_entity.pdbx_description
_entity.formula_weight
_entity.pdbx_number_of_molecules
_entity.details
_entity.pdbx_mutation
_entity.pdbx_fragment
_entity.pdbx_ec
1 polymer   man 'GALACTOSE OXIDASE' 68579.250 1  ? ? ? 1.1.3.9
2 non-polymer syn 'COPPER (II) ION'  63.546   1  ? ? ? ?
3 non-polymer syn 'SODIUM ION'       22.990   1  ? ? ? ?
4 non-polymer syn 'ACETIC ACID'       60.052   2  ? ? ? ?
5 water      nat water              18.015  316 ? ? ? ?
HELIX    1    1    TYR    A    136    TYR    A    140    5

loop_
_struct_conf.conf_type_id
_struct_conf.id
_struct_conf.pdbx_PDB_helix_id
_struct_conf.beg_label_comp_id
_struct_conf.beg_label_asym_id
_struct_conf.beg_label_seq_id
_struct_conf.pdbx_beg_PDB_ins_code
_struct_conf.end_label_comp_id
_struct_conf.end_label_asym_id
_struct_conf.end_label_seq_id
_struct_conf.pdbx_end_PDB_ins_code
_struct_conf.beg_auth_comp_id
_struct_conf.beg_auth_asym_id
_struct_conf.beg_auth_seq_id
_struct_conf.end_auth_comp_id
_struct_conf.end_auth_asym_id
_struct_conf.end_auth_seq_id
_struct_conf.pdbx_PDB_helix_class
_struct_conf.details
_struct_conf.pdbx_PDB_helix_length

HELX_P    HELX_P1    1    TYR    A    14    ?    TYR    A    18    ?    TYR    A    136    TYR    A    140    5
### PDBj

<table>
<thead>
<tr>
<th>CISPEP</th>
<th>PRO A 51</th>
<th>PRO A 52</th>
<th>0</th>
<th>2.53</th>
</tr>
</thead>
<tbody>
<tr>
<td>CISPEP</td>
<td>GLY A 162</td>
<td>PRO A 163</td>
<td>0</td>
<td>-1.03</td>
</tr>
</tbody>
</table>

```json
loop_
_struct_mon_prot_cis.pdbx_id
_struct_mon_prot_cis.label_comp_id
_struct_mon_prot_cis.label_seq_id
_struct_mon_prot_cis.label_asym_id
_struct_mon_prot_cis.label_alt_id
_struct_mon_prot_cis.pdbx_PDB_ins_code
_struct_mon_prot_cis.auth_comp_id
_struct_mon_prot_cis.auth_seq_id
_struct_mon_prot_cis.auth_asym_id
_struct_mon_prot_cis.pdbx_label_comp_id_2
_struct_mon_prot_cis.pdbx_label_seq_id_2
_struct_mon_prot_cis.pdbx_label_asym_id_2
_struct_mon_prot_cis.pdbx_PDB_ins_code_2
_struct_mon_prot_cis.pdbx_auth_comp_id_2
_struct_mon_prot_cis.pdbx_auth_seq_id_2
_struct_mon_prot_cis.pdbx_auth_asym_id_2
_struct_mon_prot_cis.pdbx_PDB_model_num
_struct_mon_prot_cis.pdbx_omega_angle
1 PRO 51 A . ? PRO 52 A 1 2.53
2 GLY 162 A . ? GLY 162 A PRO 163 A 1 -1.03
```
To Provide Format Compatibility

- Adopt a *PDB friendly* PDBx/mmCIF style -
  - All records on a single text line
  - Columns presented in standard column order
  - Tabular presentation with leading record names
    (e.g. ATOM, CELL, REFINE)
  - Method-independent features in left-most columns
    (e.g. identifiers & coordinates)
  - Method-specific features in the right-most columns
    (e.g. ADPs, NMR order/disorder parameters)
  - Continue to support PDB nomenclature semantics
    (e.g. PDB style chains, residue numbering, and insertion codes)

- Large entries will be internally converted to PDB format files (but you will meet some problem because of the limitation of the format...
<table>
<thead>
<tr>
<th>Atom</th>
<th>Type</th>
<th>Charge</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>occupancy</th>
<th>B_iso_or_equiv</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 N</td>
<td>N</td>
<td>1</td>
<td>24.690</td>
<td>-27.754</td>
<td>24.275</td>
<td>1.00</td>
<td>60.76</td>
</tr>
<tr>
<td>2 CA</td>
<td>C</td>
<td>1</td>
<td>23.581</td>
<td>-26.768</td>
<td>24.416</td>
<td>1.00</td>
<td>60.98</td>
</tr>
<tr>
<td>3 C</td>
<td>C</td>
<td>1</td>
<td>23.990</td>
<td>-25.379</td>
<td>23.905</td>
<td>1.00</td>
<td>59.98</td>
</tr>
<tr>
<td>4 O</td>
<td>O</td>
<td>1</td>
<td>25.070</td>
<td>-25.209</td>
<td>23.330</td>
<td>1.00</td>
<td>60.25</td>
</tr>
<tr>
<td>5 CB</td>
<td>C</td>
<td>1</td>
<td>23.136</td>
<td>-26.685</td>
<td>25.878</td>
<td>1.00</td>
<td>60.69</td>
</tr>
<tr>
<td>6 N</td>
<td>N</td>
<td>1</td>
<td>23.115</td>
<td>-24.395</td>
<td>24.122</td>
<td>1.00</td>
<td>59.58</td>
</tr>
<tr>
<td>7 CA</td>
<td>C</td>
<td>1</td>
<td>23.342</td>
<td>-23.010</td>
<td>23.690</td>
<td>1.00</td>
<td>57.26</td>
</tr>
<tr>
<td>8 C</td>
<td>C</td>
<td>1</td>
<td>24.000</td>
<td>-22.152</td>
<td>24.778</td>
<td>1.00</td>
<td>56.00</td>
</tr>
<tr>
<td>9 O</td>
<td>O</td>
<td>1</td>
<td>23.992</td>
<td>-20.920</td>
<td>24.692</td>
<td>1.00</td>
<td>55.53</td>
</tr>
<tr>
<td>10 CB</td>
<td>C</td>
<td>1</td>
<td>22.015</td>
<td>-22.337</td>
<td>23.275</td>
<td>1.00</td>
<td>57.32</td>
</tr>
<tr>
<td>11 N</td>
<td>N</td>
<td>1</td>
<td>24.560</td>
<td>-22.804</td>
<td>25.797</td>
<td>1.00</td>
<td>54.57</td>
</tr>
</tbody>
</table>
wwPDB Service site for a new format

http://mmcif.pdbj.org/
Many programs now accept PDBx/mmCIF

- Refinement
  - Phenix
  - refmac5
  - Buster

- Graphics / Model Building
  - Coot
  - PyMOL
  - UCSF Chimera
  - CCP4MG
  - Rasmol
When you have any difficulties/troubles, please do not hesitate to ask us!!
新しい登録システム

OneDep System

wwpdb.org
Since mid-2016 Depositors have been directed to the appropriate Regional Data Center

Processing by Data Center

- RCSB PDB *: 45%
- PDBj: 19%
- PDBe: 36%

* including Group depositions at RCSB PDB

Depositions by Geography

- Americas: 40%
- Oceania: 2%
- Europe: 19%
- Africa: 36%
- Asia: 0%
- Unknown**: 3%

** Commercial depositions at legacy system
OneDep Processing Times

(a) ~1hr: Simple structures without issues

(b) ~4 hrs: More complex structures without issues

(c) ~15 hrs: Structures with issues, including Depositor response time

If the replacement will be taken place, it will take the (almost) same time.
Recalculation of Validation Reports

- Archive Snapshot taken Dec 31, 2015
- Statistics recalculated

- March 2016: X-ray Validation Reports updated

- May 2016: NMR and EM Validation Reports released

Statistics from Dec 31, 2013

<table>
<thead>
<tr>
<th>Metric</th>
<th>Percentile Ranks</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rfree</td>
<td>worse</td>
<td>0.213</td>
</tr>
<tr>
<td>Clashscore</td>
<td>worse</td>
<td>3</td>
</tr>
<tr>
<td>Ramachandran outliers</td>
<td>worst</td>
<td>0.1%</td>
</tr>
<tr>
<td>Sidechain outliers</td>
<td>worst</td>
<td>1.9%</td>
</tr>
<tr>
<td>RSRZ outliers</td>
<td>worse</td>
<td>0.4%</td>
</tr>
</tbody>
</table>

Statistics from Dec 31, 2015

<table>
<thead>
<tr>
<th>Metric</th>
<th>Percentile Ranks</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rfree</td>
<td>worse</td>
<td>0.216</td>
</tr>
<tr>
<td>Clashscore</td>
<td>worse</td>
<td>3</td>
</tr>
<tr>
<td>Ramachandran outliers</td>
<td>worst</td>
<td>0.1%</td>
</tr>
<tr>
<td>Sidechain outliers</td>
<td>worst</td>
<td>1.9%</td>
</tr>
<tr>
<td>RSRZ outliers</td>
<td>worse</td>
<td>0.4%</td>
</tr>
</tbody>
</table>

(~20% increase in Archive)
Key Features of NMR Reports

**Chemical Shifts:**
- Referencing
- Assignment Completeness
- Statistical Outliers
- Random Coil Index

**Ensemble Analyses:**
- Well-Defined vs. Ill-Defined
- Polymer Segments

### Chemical Shifts

<table>
<thead>
<tr>
<th>Nucleus</th>
<th># values</th>
<th>Correction ± precision, ppm</th>
<th>Suggested action</th>
</tr>
</thead>
<tbody>
<tr>
<td>$^1$H</td>
<td>88</td>
<td>$0.48 \pm 0.16$</td>
<td>None needed (&lt; 0.5 ppm)</td>
</tr>
<tr>
<td>$^1$C$_\alpha$</td>
<td>86</td>
<td>$0.07 \pm 0.16$</td>
<td>None needed (&lt; 0.5 ppm)</td>
</tr>
<tr>
<td>$^1$C$_\eta$</td>
<td>80</td>
<td>$2.86 \pm 0.11$</td>
<td>Should be applied</td>
</tr>
<tr>
<td>$^{15}$N</td>
<td>80</td>
<td>$-0.01 \pm 0.34$</td>
<td>None needed (&lt; 0.5 ppm)</td>
</tr>
</tbody>
</table>

### Ensemble Analyses

#### Well-Defined vs. Ill-Defined

#### Polymer Segments
# OneDep Publication Plan

<table>
<thead>
<tr>
<th>Component</th>
<th>Focus Topics</th>
<th>Submission Timeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full System</strong></td>
<td><em>High-level over of Deposition, Biocuration and Validation – Primary Reference</em></td>
<td>Sep 2016</td>
</tr>
<tr>
<td>Validation</td>
<td>VTFs, supported methods, content, benefits to the depositors, annotators, and users, what have been implemented, limitations, and future improvements</td>
<td>Nov-Dec 2016</td>
</tr>
<tr>
<td>Full Biocuration Pipeline</td>
<td>Work Flow Manager, Ligand and Sequence annotation, checks, and communication</td>
<td>Jan-Feb 2017</td>
</tr>
<tr>
<td>Full Deposition Pipeline</td>
<td>More complete data, more checks, allow multiple file replacement, more efficient and better data quality</td>
<td>Mar-Apr 2017</td>
</tr>
<tr>
<td>3DEM</td>
<td>Changes/enhancements made, dictionary, extended data items, richer content for EMDB and PDB</td>
<td>Mar-Apr 2017</td>
</tr>
<tr>
<td>NMR Validation</td>
<td>VTF, CS, restraints validation, and NEF</td>
<td>TBD (dependency: VTF and NEF WG)</td>
</tr>
<tr>
<td>Enhanced Ligand Validation</td>
<td>Following implementation of Ligand Validation Workshop recommendations</td>
<td>TBD</td>
</tr>
</tbody>
</table>
# Planned 2016/2017 Deliverables

<table>
<thead>
<tr>
<th><strong>Core Infrastructure Support:</strong></th>
<th>Upgrade Security; Enable Use of External Computing Resources; Encrypt Traffic; Implement Depositor of Record Versioning; Management of User Credentials</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>New Content:</strong></td>
<td>Migration of Legacy Entries → OneDep system; Begin Carbohydrate Remediation; Inclusion of NMR-SAXS Hybrid Method; Capture Experimental Assembly Data</td>
</tr>
<tr>
<td><strong>Enhance Depositor Experience:</strong></td>
<td>Data File Re-upload at Deposition; Conditional Controlled Vocabulary; Support Ligand Validation; Support NMR Exchange Format Files; Support Depositor Assembly with Experimental Evidence; Implement EM MAP Validation</td>
</tr>
<tr>
<td><strong>Enhance Validation:</strong></td>
<td>Implement Ligand Validation Workshop Recommendations; Support Validation vs. NMR Restraints and CS in NMR Exchange Format</td>
</tr>
<tr>
<td><strong>Enhance Bioculator Experience:</strong></td>
<td>Improve WorkFlow for Large Structures; Increase Reprocessing Automation; Improve CIF Editor Usability</td>
</tr>
</tbody>
</table>
Starting deposition
Deposition process

wwPDB Deposition: D_1300002147 -- Requested ID: PDB

Navigation
- Instructions
- Communication
- File upload

Log out

Deposition instructions

We recommend that you take a moment to review the wwPDB policies and procedures before you start a deposition for the first time.

Navigation menus (left panel)
- This panel allows you to move to different pages.
- Clicking on a folder will open and close the folder.
- Green tick icons on the navigation panel indicate completed pages.
- Red exclamation icons on the navigation panel indicate pages with errors or uncompleted.

Hints on using the deposition tool
- It is recommended that you complete as many fields as possible.
- A field with a red box and warning text must be corrected before you can submit.
- A field with a yellow box and warning text is an unusual value and should be checked.
- The current deposition system does not support multiple simultaneous depositions. Please do one at a time.

Communication with annotators
- To send a message to the curators click on 'Communication' in the navigation menu.
- All messages should be in connection to this entry.
- All communications will be archived permanently.

Feedback welcome
- If you need help with deposition issues, please contact us via the communication interface within the deposition system.

Continue to file upload

Version: V2.7.1
First week depositions

70%以上が1日で登録

87%以上が1週間以内に登録
95%以上が30日以内に登録
Deposition process

 wwPDB Deposition: D_1300002147 -- Requested

We recommend that you take a moment to review the wwPDB policies and procedures before you start a deposition for the first time.

Navigation menus (left panel)

- This panel allows you to move to different pages.
- Clicking on a folder will open and close the folder.
- Green checkmarks on the navigation panel indicate completed pages.
- Red exclamation icons on the navigation panel indicate pages with errors or uncompleted.

Hints on using the deposition tool

- It is recommended that you:
  - A field is required.
  - A field is optional.
  - The current field is:

Communications

- To send a message:
  - All messages sent:
  - All comments received:

Feedback questions

- If you have a question:

Continue to file upload

Log out

3ヶ月間何もしないとセッションは消えます

セッションにログインすればOK
構造の間違いをチェック

Validation
これまでの経緯

- 1970年代〜1980年代 - 「構造情報の蓄積のみ」「構造は絶対的に正しい」
これまでの経緯

- 1970年代〜1980年代 – 「構造情報の蓄積のみ」「構造は絶対的に正しい」
- 1980年代後半 – 間違った構造の登録が見つかってきた
- 1990年代初頭 – 構造の誤りのチェック（validation）が始まる
- 1990年代後半 – 構造のPDBへの登録が必須に
- 2000年代半ば – 実験データ（回折データ）の登録にはまだ抵抗があった
- … データのねつ造が見つかった! X線解析の信頼性の失墜!
- 2008年2月 – 実験データの登録が必須に（化学シフトは2010年12月）
- 2008年4月 – wwPDB X-ray Validation Task Force (VTF)が組織される
A New Generation of Crystallographic Validation Tools for the Protein Data Bank

Randy J. Read,1,* Paul D. Adams,2 W. Bryan Arendall, III,3 Axel T. Brunger,4 Paul Emsley,5 Robbie P. Joosten,6,7 Gerard J. Kleywegt,8,9 Eugene B. Krissinel,9,10 Thomas Lütteke,6,11 Zbyszek Otwinowski,12 Anastassis Perrakis,7 Jane S. Richardson,3 William H. Sheffler,13 Janet L. Smith,14 Ian J. Tickle,15 Gert Vriend,6 and Peter H. Zwart2

1CIMR, University of Cambridge, Cambridge CB2 0XY, UK
2Lawrence Berkeley Laboratory, Berkeley, CA 94720-8235, USA
3Department of Biochemistry, Duke University, Durham, NC 27710, USA
4Howard Hughes Medical Institute and Departments of Molecular and Cellular Physiology, Neurology and Neurological Sciences, Structural Biology, and Photon Science, Stanford University, James H. Clark Center, Stanford, CA 94305-5432, USA
5Department of Biochemistry, University of Oxford, Oxford OX1 3QU, UK
6CMBI, NCLMS, Radboud University Nijmegen Medical Centre, 6525 GA Nijmegen, The Netherlands
7Department of Biochemistry, NKI, 1066 CX Amsterdam, The Netherlands
8Department of Cell and Molecular Biology, Uppsala University, Biomedical Centre, SE-751 24 Uppsala, Sweden
9European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, UK
10STFC Rutherford Appleton Laboratory, Chilton, Didcot OX11 0QX, UK
11Justus-Liebig University Gießen, Institute of Veterinary Physiology and Biochemistry, 35392 Gießen, Germany
12UT Southwestern Medical Center, Dallas, TX 75390-8816, USA
13Department of Biochemistry, University of Washington, Seattle, WA 98195, USA
14Life Sciences Institute, Department of Biological Chemistry, University of Michigan, Ann Arbor, MI 48109, USA
15Astex Therapeutics, Cambridge CB4 0QA, UK

*Correspondence: rj27@cam.ac.uk
DOI 10.1016/j.str.2011.08.006

Read et al., Structure, 19, 1395-1412 (2011)
wwPDB X-ray Validation Pipeline

Validation pipeline 1.0

MolProbity  Xtriage  EDS  Mogul  Percentiles  PDF maker

Validation XML file

External reference files (e.g., Engh & Huber)

Deposited data (coordinates & reflections)

Distributions

PDF report for depositor & referees - Statistics and plots for the entry, per chain, per residue, and list of unusual features

Validation Reports

- Front cover
- Deposition info
- Software info
Validation Reports

- **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**

The following experimental techniques were used to determine the structure:

**X-RAY DIFFRACTION**

The reported resolution of this entry is 2.00 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.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Percentile Ranks</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rfree</td>
<td>1</td>
<td>0.229</td>
</tr>
<tr>
<td>Clashscore</td>
<td>2</td>
<td>0.6%</td>
</tr>
<tr>
<td>Ramachandran outliers</td>
<td>2</td>
<td>0.6%</td>
</tr>
<tr>
<td>Sidechain outliers</td>
<td>1</td>
<td>1.5%</td>
</tr>
<tr>
<td>RSRZ outliers</td>
<td>1</td>
<td>1.5%</td>
</tr>
</tbody>
</table>

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. A grey segment represents the fraction of residues that are not modelled. The numeric value for each fraction is indicated below the corresponding segment, with a dot representing fractions <=5%.

The upper red bar (where present) indicates the fraction of residues that have poor fit to the electron density. The numeric value is given above the bar.

<table>
<thead>
<tr>
<th>Mol</th>
<th>Chain</th>
<th>Length</th>
<th>Quality of chain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>202</td>
<td>91%</td>
</tr>
<tr>
<td>1</td>
<td>B</td>
<td>202</td>
<td>98%</td>
</tr>
</tbody>
</table>

The following table lists non-polymeric compounds, carbohydrate monomers and non-standard residues in protein, DNA, RNA chains that are outliers for geometric or electron-density-fit criteria:
2 Entry composition

There are 5 unique types of molecules in this entry. The entry contains 3605 atoms, of which 0 are hydrogens and 0 are deuteriums.

In the tables below, the ZeroOcc column contains the number of atoms modelled with zero occupancy, the AltConf column contains the number of residues with at least one atom in alternate conformation and the Trace column contains the number of residues modelled with at most 2 atoms.

- Molecule 1 is a protein called Glutathione s-transferase S2.

<table>
<thead>
<tr>
<th>Mol</th>
<th>Chain</th>
<th>Residues</th>
<th>Atoms</th>
<th>ZeroOcc</th>
<th>AltConf</th>
<th>Trace</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>196</td>
<td>Total</td>
<td>1619</td>
<td>1053</td>
<td>262</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>298</td>
<td></td>
<td>6</td>
</tr>
</tbody>
</table>

- Molecule 2 is GLUTATHIONE (three-letter code: GSH) (formula: C_{10}H_{17}N_{3}O_{6}S).

<table>
<thead>
<tr>
<th>Mol</th>
<th>Chain</th>
<th>Residues</th>
<th>Atoms</th>
<th>ZeroOcc</th>
<th>AltConf</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>A</td>
<td>17</td>
<td>Total</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11</td>
<td></td>
</tr>
</tbody>
</table>

- Molecule 3 is DI(HYDROXYETHYL)ETHER (three-letter code: PEG) (formula: C_{4}H_{10}O_{3}).
Validation Reports

- Residue quality
  - One plot per polymer
  - Coloured by number of *types* of geometric outliers
  - Grey if not modelled
  - Red dots: poor density (RSR-Z > 2, as in EDS)
Validation Reports

4 Data and refinement statistics

- “Table 1”
- Xtriage

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space group</td>
<td>P 2 1 21 21</td>
<td>Depositor</td>
</tr>
<tr>
<td>Cell constants</td>
<td>45.65Å 47.56Å 77.61Å</td>
<td>Depositor</td>
</tr>
<tr>
<td>a, b, c, α, β, γ</td>
<td>90.00° 90.00° 90.00°</td>
<td>Depositor</td>
</tr>
<tr>
<td>Data completeness (%)</td>
<td>90.3 90.5</td>
<td>EDS</td>
</tr>
<tr>
<td>$R_{merge}$</td>
<td>(Not available)</td>
<td>Depositor</td>
</tr>
<tr>
<td>$R_{sym}$</td>
<td>(Not available)</td>
<td>Depositor</td>
</tr>
<tr>
<td>$&lt; I/\sigma(I) &gt;$</td>
<td>3.77 (at 1.79Å)</td>
<td>Xtriage</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>8.00° – 1.80°</td>
<td>EDS</td>
</tr>
<tr>
<td>Refinement program</td>
<td>X-PLOR</td>
<td>Depositor</td>
</tr>
<tr>
<td>$R$, $R_{free}$</td>
<td>0.200 0.237</td>
<td>DCC</td>
</tr>
<tr>
<td>Wilson B-factor (Å²)</td>
<td>14.8</td>
<td>Xtriage</td>
</tr>
<tr>
<td>Anisotropy</td>
<td>0.434</td>
<td>Xtriage</td>
</tr>
<tr>
<td>Bulk solvent $k_{sol}$ (Å³), $B_{sol}$ (Å²)</td>
<td>0.41, 58.87</td>
<td>EDS</td>
</tr>
<tr>
<td>Estimated twinning fraction</td>
<td>0.027 for k,h,l</td>
<td>Xtriage</td>
</tr>
<tr>
<td>L-test for twinning</td>
<td>$&lt;</td>
<td>L</td>
</tr>
<tr>
<td>Outliers</td>
<td>0 of 14678 reflections</td>
<td>Xtriage</td>
</tr>
<tr>
<td>$F_o$,$F_c$ correlation</td>
<td>0.95</td>
<td>EDS</td>
</tr>
<tr>
<td>Total number of atoms</td>
<td>1213</td>
<td>wwPDB-VP</td>
</tr>
<tr>
<td>Average B, all atoms (Å²)</td>
<td>16.0</td>
<td>wwPDB-VP</td>
</tr>
</tbody>
</table>

Xtriage’s analysis on translational NCS is as follows: The largest off-origin peak in the Patterson function is 9.26% of the height of the origin peak. No significant pseudo-translation is detected.
Validation Reports

- Model quality
  - Bond lengths and angles
  - Torsion angles (Ramachandran, rotamers)
  - Clashes
  - Separately for standard residues, non-standard residues, ligands, carbohydrates

- Generally: information about distribution, outlier stats, percentile scores, list of up to 5 (worst) outliers
Validation Reports

- Geometry validation of ligands and non-standard entities
  - Mogul (CCDC)

- wwPDB will get CSD coordinates for new and existing compounds
Validation Reports

- Model/data fit proteins, DNA, RNA
- RSR and RSR-Z (EDS)

6.1 Protein, DNA and RNA chains

In the following table, the column labelled ‘#RSRZ > 2’ contains the number (and percentage) of RSRZ outliers, followed by percent RSRZ outliers for the chain as percentile scores relative to all X-ray entries and entries of similar resolution. The OWAB column contains the minimum, median, 95th percentile and maximum values of the occupancy-weighted average B-factor per residue. The column labelled ‘Q< 0.9’ lists the number of (and percentage) of residues with an average occupancy less than 0.9.

<table>
<thead>
<tr>
<th>Mol</th>
<th>Chain</th>
<th>Analysed</th>
<th>&lt;RSRZ&gt;</th>
<th>#RSRZ&gt;2</th>
<th>OWAB(Å²)</th>
<th>Q&lt;0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>371/371 (100%)</td>
<td>-0.00</td>
<td>0/100</td>
<td>2, 37, 96, 164</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>C</td>
<td>371/371 (100%)</td>
<td>0.12</td>
<td>4/1%</td>
<td>81/65</td>
<td>2, 37, 96, 164</td>
</tr>
<tr>
<td>All</td>
<td>All</td>
<td>742/742 (100%)</td>
<td>0.06</td>
<td>4/0%</td>
<td>88/79</td>
<td>2, 37, 96, 164</td>
</tr>
</tbody>
</table>

All RSRZ outliers are listed below:

<table>
<thead>
<tr>
<th>Mol</th>
<th>Chain</th>
<th>Res</th>
<th>Type</th>
<th>RSRZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C</td>
<td>255</td>
<td>PHE</td>
<td>2.8</td>
</tr>
<tr>
<td>1</td>
<td>C</td>
<td>269</td>
<td>ILE</td>
<td>2.6</td>
</tr>
<tr>
<td>1</td>
<td>C</td>
<td>302</td>
<td>LEU</td>
<td>2.3</td>
</tr>
<tr>
<td>1</td>
<td>C</td>
<td>16</td>
<td>THR</td>
<td>2.2</td>
</tr>
</tbody>
</table>
RSR-Z (in EDS)

- RSR dependent on residue type and resolution
- Define RSR-Z (RSR, aa, d) = (RSR - <RSR(aa,d)>) / σ(RSR(aa,d))
  - aa = residue type
  - d = resolution (in bins of 0.2Å)
- Calculated using 10,000s of EDS entries
- Example: Trp between 2.4 and 2.6Å:
  - 2012: 58321 observations, <> = 0.1419, σ = 0.0537
  - 2008: 26794 observations, <> = 0.1602, σ = 0.0660
  - RSR=0.25 → RSR-Z=2.0 (2008: 1.4)
Validation Reports

- Model/data fit ligands etc.
- RSR as usual
- Can’t usually compute RSR-Z due to few/no occurrences in PDB
Full wwPDB X-ray Structure Validation Report

Nov 9, 2016 - 11:44 AM JST

PDB ID : IHEL
Title : Structure of protophycocyanin from Synechococcus sp. strain PCC 7002
Deposited on : 2016-11-07
Resolution : 1.40 (A) (reported)

This is a Full wwPDB X-ray Structure Validation Report.

This report is produced by the wwPDB biocuration pipeline after annotation of the structure.

We welcome your comments at validation@mail.wwpdb.org
A user guide is available at
with specific help available everywhere you see the symbol.

The following versions of software and data (see references ①) were used in the production of this report:

MolProbity : 4.02b-467
Molsoft : 1.7.1 (RC1), CSD as537bc (2016)
XrayLog (Phenix) : 1.9-1692
EDS : rb-2002520
Percentile statistics : 2015/1230, v0.1 (using entries in the PDB archive December 30th 2015)
Refmac : 5.8.0.135
CCP4 : 6.5.0
Ideal geometry (proteins) : Engel & Huber (2001)
Ideal geometry (DNA, RNA) : Parkinson et al. (1996)
Validation Pipeline (wwPDB-VP) : rb-2002520
Validation Task Force

Method-specific Validation Task Forces have been convened to collect recommendations and develop consensus on additional validation that should be performed, and to identify software applications to perform validation tasks.

- X-ray Validation Task Force
- NMR Validation Task Force
- EM Validation Task Force
Validation Task Force
http://wwpdb.org/

**X-ray**

**A New Generation of Crystallographic Validation Tools for the Protein Data Bank**

*Structure 19: 1395-1412 (2011).*

**NMR**

**Recommendations of the wwPDB NMR Validation Task Force**

*Structure 21: 1563-1570 (2013).*

**EM**

**Outcome of the First Electron Microscopy Validation Task Force Meeting**

*Structure 20: 205-214 (2012).*
Validation Server

- **Stand-alone** wwPDB X-ray validation server
  
  http://wwpdb.org/
## wwPDB Validation Service

### Existing validation

- **Validation ID**
- **Password**

**Log in**

**Forgot Password**

### Deposition server

Deposit your data to PDB, BMRB and EMDG at [http://deposit.wwpdb.org](http://deposit.wwpdb.org)

### wwPDB news and announcements

**Why can’t I login to an existing session?**

On September 12, 2010 we disabled an old validation server. If your session ID looks like D_9XXX, it is no longer possible to access it. Please start a new session on this page.

### Start a new validation

Welcome to the wwPDB validation system.

This server runs the same validation as you would observe during the deposition process. This service is designed to help you check your model and experimental files prior to start of deposition.

To continue with an existing validation, please login on the left.

To start a new validation, please complete the form below. Upon completion, you will be emailed login information specific to your new validation.

- **Your e-mail address**
- **Password (optional, or we will provide one)**
  - This is a shared “group password” (6 to 10 alphanumeric characters)
- **Country**
  - Select...
- **Experimental method**
  - X-Ray Diffraction
  - Electron Microscopy
  - Solution NMR
  - Neutron Diffraction
  - Electron Crystallography
  - Solid-state NMR
  - Fiber Diffraction

Please copy this code: 84847

### Start validation session
If you will have any questions, please ask us any time.
PDBj スタッフ

- 統括責任者
  - 栗栖 源嗣 (大阪大学蛋白質研究所・教授)

- PDBjデータベース構築グループ
  - 中川 敦史 (大阪大学蛋白質研究所・教授)
  - 見学 有美子 (大阪大学蛋白質研究所・特任研究員)
  - 張 羽澄 (大阪大学蛋白質研究所・特任研究員)
  - 池川 恭代 (大阪大学蛋白質研究所・特任研究員)
  - 佐藤 純子 (大阪大学蛋白質研究所・特任研究員)
  - 西山 夏代子 (大阪大学蛋白質研究所・特任研究員)

- PDBj国際的な運営高度化グループ
  - 中村 春木 (大阪大学蛋白質研究所・教授)
  - 金城 玲 (大阪大学蛋白質研究所・准教授)
  - 岩崎 憲治 (大阪大学蛋白質研究所・准教授)
  - 鈴木 博文 (大阪大学蛋白質研究所・特任助教)
  - Bekker, Gert-Jan (大阪大学蛋白質研究所・特任助教)
  - 藤 博幸 (関西学院大学理工学部・教授)
  - 山下 鈴子 (大阪大学蛋白質研究所・特任技術専門職員)
  - 工藤 高裕 (大阪大学蛋白質研究所・特任研究員)

- BMRBデータベース管理運営グループ
  - 藤原 敏道 (大阪大学蛋白質研究所・教授)
  - 児嶋 長次郎 (横浜国立大学工学部・教授)
  - 小林 直宏 (大阪大学蛋白質研究所・特任准教授)
  - 岩田 武史 (大阪大学蛋白質研究所・特任研究員)
  - 横地 政志 (大阪大学蛋白質研究所・特任研究員)

- 研究協力者
  - 輪湖 博 (早稲田大学社会科学総合学術院・教授) for ProMode
  - 猿渡 茂 (北里大学理学部・准教授) for ProMode
  - 伊藤 暢聡 (東京医科歯科大学医学部・教授)
  - 木下 賢吾 (東北大学大学院情報科学研究科・教授) for eF-site
  - Standley, M.Daron (大阪大学微生物病研究所・教授) for SeqNavi, StructNavi, SeSAW, ASH, MAFFTash, Spanner and SFAS
  - 加藤 和貴 (大阪大学免疫学フロンティア研究センター・特任准教授) for MAFFTash

- 事務職員
  - 清水 朋子 (大阪大学蛋白質研究所・特任事務職員)
Thank you!