Ligand Validation for the Protein Data Bank

Stephen K. Burley, M.D., D.Phil.  
Director, RCSB Protein Data Bank  
University Professor and Henry Rutgers Chair  
Founding Director, Institute for Quantitative Biomedicine  
Member, Rutgers Cancer Institute of New Jersey  
Rutgers, The State University of New Jersey

AsCA, Auckland, New Zealand December 4th 2018

rcsb.org
Protein Data Bank History (1971-2018)

- PDB 1st Open Access digital data resource in all of biology
- Single global archive for protein, DNA, and RNA experimental structures
- Today, Open Access to >146,000 structures
- wwPDB collaboration US (RCSB PDB), EU (PDBe), Japan (PDBJ), and BMRB
Outline

- wwPDB Chemical Component Dictionary
- Validating and Biocurating PDB Ligands
- Detecting PDB Ligand Outliers
- Correcting PDB Structures
- Impact of PDB on Drug Approvals
wwPDB Chemical Component Dictionary
wwPDB Chemical Component Dictionary

- wwPDB maintained library of all chemical components present in PDB archive
  - >26,400 chemical component definitions
  - 400 additional definitions of amino acid protonation variants
- ~2300 new components added in 2017

ftp.wwpdb.org/pub/pdb/data/monomers/components.cif
Validating and Biocurating PDB Ligands
wwPDB Deposition/Validation/Biocuration

- 13,049 new structures deposited in 2017
  - RCSB PDB processed 6,208 structures (~85% from US/Canada)
- Workload balanced among wwPDB Partners
  - RCSB PDB: Americas/Oceania
  - PDBe: Europe/Africa
  - PDBj: Asia/Middle East

Young et al. (2017) *Structure* 25, 536-545.
OneDep Ligand Validation and Biocuration

a). Automatic comparison of deposited component and the Chemical Component Dictionary (CCD)
- number of heavy atoms
- number of chiral centers
- chirality
- aromatic flags
- bond order

<table>
<thead>
<tr>
<th>INSTANCE</th>
<th>TOP HIT</th>
<th>MATCH STATUS</th>
<th>SELECTION</th>
<th>ALL</th>
<th>COMPOSITE SCORE</th>
<th>ASSIGNMENT STATUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1_C_MAN.1078_</td>
<td>BMA</td>
<td>close match</td>
<td></td>
<td></td>
<td>100 / 100 / 100 / n/a / 100</td>
<td>Not Assigned</td>
</tr>
<tr>
<td>1_C_NAG.1074_</td>
<td>NAG</td>
<td>close match</td>
<td></td>
<td></td>
<td>100 / 100 / 100 / n/a / 100</td>
<td>Not Assigned</td>
</tr>
<tr>
<td>1_C_NAG.1076_</td>
<td>NAG</td>
<td>close match</td>
<td></td>
<td></td>
<td>100 / 100 / 100 / n/a / 100</td>
<td>Not Assigned</td>
</tr>
<tr>
<td>1_E_MES.1083_</td>
<td>MES</td>
<td>passed</td>
<td></td>
<td></td>
<td>100 / n/a / n/a / n/a / 100</td>
<td>(MES)</td>
</tr>
<tr>
<td>1_E_MES.1084_</td>
<td>MES</td>
<td>passed</td>
<td></td>
<td></td>
<td>100 / n/a / n/a / n/a / 100</td>
<td>(MES)</td>
</tr>
</tbody>
</table>

SUPPLEMENTAL INFO: Chiral center C3 has sp2 hybridization instead of sp3

b). 2D and 3D ligand comparison panel

Deposited Ligand

Closest CCD Match

REA in entry 1CBS with LLDF=1.31 (RSR=0.10, CC=0.95)

TMP in entry 3HW4 with LLDF=6.77 (RSR=0.41, CC=0.70)

Table of local electron density map for non-polymer chemical components

Table of local electron density map for non-polymer chemical components

Young et al. (2017) Structure 25, 536-545.
Principal Ligand Validation Metrics

- **Agreement with Known Chemical Geometry**
  - Bond Lengths: RMSZ, # |Z|>2 (ref. CCDC Mogul)
  - Bond Angles: RMSZ, # |Z|>2 (ref. CCDC Mogul)
  - Plus analyses of Chirality, Torsions, and Rings
    (N.B.: Depends critically on choice of restraints.)

- **Agreement with Experimental Data (|Fobs|)**
  - RSR-Real Space R-factor
  - RSCC-Real Space Correlation Coefficient
  - Plus analyses of B-factors and Occupancy<0.9

Young *et al.* (2017) *Structure* 25, 536-545.
Detailed validation report based on community norms and input

Overall quality sliders for 5 key metrics

“Table 1” summary

Tabulations of geometrical and experimental issues:
  • Macromolecules
  • Ligands

Now required by some journals for manuscript review

Overall Quality Summary

Residue Plots

Version 1 Helped Improve Structure Quality

- Clashscores, Rotamers, Ramachandran violations, RSR Z-scores all improved!
- Median Rfree rose slightly

Version 1.0 No Impact on Ligand Quality

Overall Ligand Quality: *Legacy* 2012-2013 (yellow) vs. *New* 2014-2015 (green)

NADP in PDB 1ZK4: 2|Fo|-|Fc| map at 1σ
Schlieben et al., 2005

NADP in PDB 2FZD: 2|Fo|-|Fc| map at 1σ
Steuber et al., 2006

Meeting Objectives: To bring together co-crystal structure determination experts from Academe and Industry with Crystallography and Computational Chemistry Software Developers to discuss, develop, and recommend:

- Best practices PDB deposition/validation of co-crystal structures
- Editorial/Refereeing/Publication standards for co-crystal structures
- Improvements in ligand representation across the PDB Archive
- Version 2 coming in 2019
Detecting PDB Ligand Outliers
Identifying Outliers for Chemical Geometry

- For most PDB ligand structures, agreement with known chemical geometry depends almost entirely on refinement restraints and weighting schemes.

- wwPDB Validation Report uses CCDC Mogul to identify outliers.

- Available options:
  - Exact matches with the Cambridge Structural Database (~10% of the CCD)
  - Phenix AM1
  - CCP4 Acedrg
  - Higher level semi-empirical QM calculations
Identifying Outliers for RSR and RSCC

RSR and RSCC are Resolution Dependent!
109,368 Organic Ligands; MW=240-1000/Occup.>0.9

Analyzing RSR for Outliers (1.9-2.0Å)

- Distribution of RSR values is not normal (skewed and bounded)
- Standard analyses using % cut-off, IQR, or Sigma cut-off entirely inappropriate
- Most-Probable-Range (95% blue) more appropriate

95% MPR(RSR)=0.040-0.295 (>10,000 ligands)

Analyzing RSR for Outliers (1.0-3.5Å)

- 95% MPR(RSR) ranges vary with resolution
- This approach would have detected some notorious cases*
  1. 4lro(SPD): RSR=0.34
  2. 4zz6(CTP): RSR=0.48
  3. 3iaz(AIN): RSR=0.68
  4. 4lwx(MU2): RSR=0.41

Correcting PDB Structures
Coordinate Replacement by Depositor

- PDB archive is now fully versioned
- Starting in 2019, the Depositor of Record will be able to replace atomic coordinates for co-crystal structures using OneDep
- Your original PDB ID will be preserved!
- Numbers will be limited initially to ensure that RCSB PDB, PDBe, and PDBj do not get overwhelmed with corrected structures
Impact of PDB on Drug Approvals
Impact of PDB on Drug Approvals

210 NEW DRUGS
approved
2010-2016

>$100 BILLION
of NIH funding
contributed to these approvals
(>95% on targets)

5,914 PDB Structures
contributed to 184 of these drug approvals

B-Raf Kinase complex with Vemurafenib
PDB ID 3og7

Acknowledgements

Funding
RCSB PDB is funded by a grant (DBI-1338415) from the National Science Foundation, the National Cancer Institute, the National Institute of General Medical Sciences and the US Department of Energy.

Management
The RCSB PDB is managed by:

The RCSB PDB is a member of the Worldwide Protein Data Bank partnership (wwPDB; wwpdb.org).
TIRED OF THE RAIN?
Join RCSB PDB at UCSD!

Postdoctoral Fellows

The Challenge:
Develop innovative 3D Visualization and Analysis and Bioassembly/Machine Learning tools to help accelerate research and training in biology, medicine, and related disciplines.

Further inquiries:
rcsb.org (More > Careers)
info@rcsb.org