JOURNAL OF APPLIED CRYSTALLOGRAPHY
ISSN 1600-5767

DCC: a Swiss army knife for structure factor analysis and validation

Huanwang Yang,[a] Ezra Peisach,[a]* John D. Westbrook,[a] Jasmine Young,[a] Helen M. Berman[a] and Stephen K. Burley[a,b,c]

[a]Research Collaboratory for Structural Bioinformatics Protein Data Bank, Department of Chemistry and Chemical Biology, Center for Integrative Proteomics Research, Rutgers, State University of New Jersey, 174 Frelinghuysen Road, Piscataway, NJ 08854, USA, [b]Institute for Quantitative BioMedicine, Rutgers, State University of New Jersey, 174 Frelinghuysen Road, Piscataway, NJ 08854, USA, and [c]San Diego Supercomputer Center and Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA 92093, USA. *Correspondence e-mail: ezra.peisach@rcsb.org

Received 25 November 2015
Accepted 15 March 2016

Edited by A. R. Pearson, Universität Hamburg, Germany

Keywords: Protein Data Bank; structure factor validation; utility programs; *DCC*.

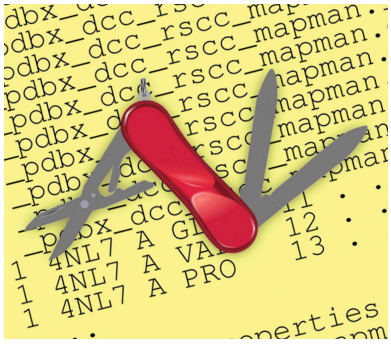
Since 2008, X-ray structure depositions to the Protein Data Bank archive (PDB) have required submission of experimental data in the form of structure factor files. RCSB PDB has developed the program *DCC* to allow worldwide PDB (wwPDB; http://wwpdb.org) biocurators, using a single command-line program, to invoke a number of third-party software packages to compare the model file with the experimental data. *DCC* functionality includes structure factor validation, electron-density map generation and slicing, local electron-density analysis, and residual *B* factor analysis. *DCC* outputs a summary containing various crystallographic statistics in PDBx/mmCIF format for use in automatic data processing and archiving pipelines.

1. Introduction

The Protein Data Bank (PDB) is the single global archive of biological structures determined by X-ray crystallography, nuclear magnetic resonance (NMR) and three-dimensional electron microscopy. The archive is managed by the Worldwide PDB collaboration (wwPDB) (Berman *et al.*, 2003). wwPDB members include the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) (Berman *et al.*, 2000), Protein Data Bank in Europe (Velankar *et al.*, 2016), Protein Data Bank Japan (Kinjo *et al.*, 2012) and the Biological Magnetic Resonance Bank (Ulrich *et al.*, 2008).

Prior to 2008, only the atomic coordinate model of the structure was required for PDB archive deposition. Subsequently, submission of experimental data (structure factors for X-ray crystallography, restraints and chemical shifts for NMR) became mandatory (http://www.wwpdb.org/news/news?year=2007#29-November-2007). At this time, numerous individual programs were available to aid in the manipulation and validation of the experimental data relative to the model, but all required expertise and familiarity with the details of each program.

DCC was created by RCSB PDB to combine and enable use of these existing programs. Some of the features include structure factor validation, electron-density map calculation, real-space *R* (RSR) calculations, detection and correction of partial *B* factors, and production of cut electron maps and scripts for display in *Jmol* (Hanson, 2010). The program name, *DCC*, comes from one of these functions and was named for electron-density correlation coefficient. These features are used daily by wwPDB biocurators.



OPEN ACCESS

2. Methods

2.1. Program function

DCC is a Python wrapper for a number of third-party software programs, including *SFCHECK* (Vaguine *et al.*, 1999), *PHENIX* (Adams *et al.*, 2002), *REFMAC* (Murshudov *et al.*, 1996), *MAPMAN* (Kleywegt *et al.*, 2004) and *CNS* (Brünger *et al.*, 1998). Through a command-line interface, *DCC* converts structure factor files from any recognized format, creates the specific input files required for each of these programs and then runs the required programs (Table 1). *DCC* will also utilize whatever metadata are present in the atomic coordinate model file, including ***TLS*** records and wavelength and twinning information, to produce suitable input data for third-party packages. For instance, a virus structure in which strict non-crystallographic symmetry (NCS) refinement has been used may not include atomic coordinates for the entire asymmetric unit in the model file. In this case, *DCC* will expand the coordinates using the NCS operators for use with third-party programs.

One challenge in producing files for input to refinement programs is how best to represent ligands. Refinement programs require a full definition of all chemical components present in the system, including bond order and connectivity. However, for unreleased components, and prior to processing, such information is not available. Therefore, *DCC* treats all ligands as individual atoms for presentation to the refinement programs.

For structure factor validation, the user may specify which refinement package to use, or an automatic mode may be invoked that will use the package specified in the model file (Table 1). A zero-cycle (static) refinement is used and the resulting calculated $R_{\rm work}$ and $R_{\rm free}$ and other statistics are captured. Based on the statistical analysis of the calculated data items, errors and warnings will be included in the output file. Sample output is depicted in Fig. 1.

When ***TLS*** restraints are used in refining a structural model with *REFMAC*, authors occasionally deposit structures containing only partial *B* factors without including the isotropic ***TLS*** contribution (Touw & Vriend, 2014). *DCC* detects these partial *B* factors and then uses *TLSANL* (Howlin *et al.*, 1993) to produce full *B* factors before performing validation.

DCC uses *REFMAC* (Murshudov *et al.*, 1996) to produce electron-density maps. For local density analysis of both polymer and non-polymer residues, both *EDSTAT* (Tickle, 2012) and *MAPMAN* (Kleywegt *et al.*, 2004) are used to calculate RSR factors, density correlations and the real-space difference density *Z* score. *MAPMASK* (Winn *et al.*, 2011) is used to produce sliced maps for use with *Jmol* visualization.

The results of any analysis, and any additional calculations performed by *DCC*, are captured and stored in a PDBx/mmCIF formatted file. This feature allows *DCC* to be utilized as a component by other programs for further analysis. This capability also allows for the generation of tabular reports for review during PDB archive biocuration and facile loading to relational databases.

Table 1
The list of command-line options available in *DCC*.

The basic command '`dcc -pdb xyzfile -sf sffile`' performs the default functionalities described in the text using *REFMAC*. Any metadata in the model file are utilized in the calculations. If such information is not found in the file, parameters are optimized so that the calculated statistics best match those reported.

The basic command:	
`dcc -pdb xyzfile -sf sffile`	Where `xyzfile` is the coordinate file in either PDB or PDBx(mmCIF) format, and `sffile` is the structure factor file, which can be in any of the following formats [mtz, mmCIF, CIF (for small molecules; IUCr), *CNS/Xplor*, *HKL2000*/*SCALEPACK*, *Dtrek*, *SHELX*, *SAINT*, *EPMR*, *XSCALE*, *XPREP*, *TNT*, *XTALVIEW*, *X-GEN*, *XENGEN*, *MULTAN* and *MAIN*].
The options below can be added to the above command to perform additional tasks:	
`-o`	Followed by an output file name to hold the calculated statistics. If not given, the default name (`pdbfile + _rcc_sum.cif`) will be used.
`-diags`	Followed by a log file name to hold error/warning messages.
`-verb`	Add to keep the intermediate files during computations.
`-rsr_all`	Add to calculate electron-density statistics (RSR, RSRZ, RSCC) by groups [residual, main chain, side chain, phosphate (if RNA/DNA)].
`-edstat`	Add to use the *EDSTAT* program to calculate electron-density statistics (RSR, RSRZ, RSCC, RSZD, RSDO) by groups [residual, main chain, side chain, phosphate (if RNA/DNA)].
`-sfcheck`	Add to validate X-ray data by *SFCHECK*.
`-refmac`	Add to validate X-ray data by *REFMAC* (default).
`-phenix_x`	Add to validate X-ray data by *PHENIX* (`model_vs_data`).
`-phenix_n`	Add to validate neutron data by *PHENIX* (`model_vs_data`).
`-phenix_xn`	Add to validate neutron and X-ray hybrid data by *PHENIX* (`model_vs_data`). The structure factor file (`sffile`) must be in mmCIF format. The first data block must be the X-ray data and the second data block must be the neutron data.
`-cns`	Add to validate X-ray data by *CNS*/*Xplor*.
`-all`	Add to validate X-ray data by all the programs (*SFCHECK*, *REFMAC*, *Phenix*). The calculated statistics such as $R/R_{\rm free}$ will be listed by the programs.
`-auto`	Add to validate X-ray data by the program used for refinement in the coordinate file (`xyzfile`). If the program fails then other programs will be used.
`-map`	Add to calculate maps ($mF_{\rm o} - DF_{\rm c}$, $2mF_{\rm o} - DF_{\rm c}$) in *CCP4* format.
`-ligmap`	Add to produce all the files (ligand density maps, tables and html files) and *Jmol* scripts for displaying the ligand density in a browser.
`-omitmap`	Add to calculate residual electron-density statistics (RSR, RSRZ, RSCC) after omitting all the ligands.
`-omit`	Followed by an identifier to calculate the omit map. For example, the command `dcc -pdb xyzfile -sf sffile -omit A_3:5` calculates a map omitting residue numbers from 3 to 5 of chain *A*.
`-fem`	Add to calculate density statistics and the map using the feature-enhanced map in *PHENIX*.
`-bfull`	Convert residual to full *B* factors using the command `dcc -bfull xyzfile`.

```
data_4NL7
#
_pdbx_dcc_mapman.pdbid  4NL7
_pdbx_dcc_mapman.details
;Items below are the local density correlation using mapman and refmac(Dcc).
correlation: Dcc=(<xy>-<x><y>)/[sqrt(<x**2>-<x>**2)*sqrt(<y**2>-<y>**2)]

Real spaceR: RSR = sum(|x-y|/x+y|) sum over all grid around residue
             x=Do (observed density 2mFo-dFc); y=Dc (caculated denisty Fc)

real_space_Zscore:  (RSR-<RSR>)/sigma
Biso_mean:  occupancy-weighted average B = (SUM B*Q)/(SUM Q)
occupancy_mean:  the average occupancy of each residue = S_occ / Nuniq
;

#
loop_
 _pdbx_dcc_rscc_mapman.model_id
 _pdbx_dcc_rscc_mapman.pdb_id
 _pdbx_dcc_rscc_mapman.auth_asym_id
 _pdbx_dcc_rscc_mapman.auth_comp_id
 _pdbx_dcc_rscc_mapman.auth_seq_id
 _pdbx_dcc_rscc_mapman.label_alt_id
 _pdbx_dcc_rscc_mapman.label_ins_code
 _pdbx_dcc_rscc_mapman.correlation
 _pdbx_dcc_rscc_mapman.real_space_R
 _pdbx_dcc_rscc_mapman.real_space_Zscore
 _pdbx_dcc_rscc_mapman.Biso_mean
 _pdbx_dcc_rscc_mapman.occupancy_mean
 _pdbx_dcc_rscc_mapman.id
1 4NL7 A GLY   11 . .   0.702  0.264   -0.06    39.62  1.000   1
1 4NL7 A VAL   12 . .   0.201  0.467    2.37    54.77  1.000   2
1 4NL7 A PRO   13 . .   0.673  0.303    0.41    45.20  1.000   3
;...
#
#Overall properties from mapman:
_pdbx_dcc_rscc_mapman_overall.pdbid  4NL7
_pdbx_dcc_rscc_mapman_overall.correlation  0.7910
_pdbx_dcc_rscc_mapman_overall.correlation_sigma  0.0972
_pdbx_dcc_rscc_mapman_overall.real_space_R  0.2688
_pdbx_dcc_rscc_mapman_overall.real_space_R_sigma  0.0837


####################Overall values#########################
#
_pdbx_dcc_density.DCC_version    '2.14  (2015-09-10)'
_pdbx_dcc_density.pdbid      4NL7
_pdbx_dcc_density.pdbtype    ''
_pdbx_dcc_density.unit_cell  '107.101   62.292   57.069  90.00  95.07  90.00'
_pdbx_dcc_density.space_group_name_H-M                  'C 1 2 1'
_pdbx_dcc_density.space_group_pointless                 'C 1 2 1'
_pdbx_dcc_density.ls_d_res_high                         3.000
_pdbx_dcc_density.ls_d_res_high_sf                      3.00
_pdbx_dcc_density.ls_d_res_low_sf                       56.85
_pdbx_dcc_density.R_value_R_work                        0.2955
_pdbx_dcc_density.R_value_R_free                        0.2955
_pdbx_dcc_density.working_set_count                     6702
_pdbx_dcc_density.free_set_count                        6702
_pdbx_dcc_density.occupancy_min                         1.0
_pdbx_dcc_density.occupancy_max                         1.0
_pdbx_dcc_density.occupancy_mean                        1.0
_pdbx_dcc_density.Biso_min                              0.0
_pdbx_dcc_density.Biso_max                              86.25
_pdbx_dcc_density.Biso_mean                             26.878
_pdbx_dcc_density.B_wilson                              24.02
_pdbx_dcc_density.B_wilson_scale                        0.13
_pdbx_dcc_density.mean_I2_over_mean_I_square            2.528
_pdbx_dcc_density.mean_F_square_over_mean_F2            0.764
_pdbx_dcc_density.mean_EZ_I_abs                         0.809
_pdbx_dcc_density.Padilla-Yeates_L_mean                 0.421
_pdbx_dcc_density.Padilla-Yeates_L2_mean                0.248
_pdbx_dcc_density.Padilla-Yeates_L2_mean_pointless      0.328
_pdbx_dcc_density.Z_score_L_test                        6.312
_pdbx_dcc_density.twin_type                             ?
_pdbx_dcc_density.twin_operator_xtriage                 ?
_pdbx_dcc_density.twin_fraction_xtriage                 ?
_pdbx_dcc_density.twin_Rfactor                          ?
_pdbx_dcc_density.I_over_sigI_resh                      1.883706
_pdbx_dcc_density.I_over_sigI_diff                      -0.2
_pdbx_dcc_density.I_over_sigI_mean                      ?
_pdbx_dcc_density.ice_ring                              ?
_pdbx_dcc_density.anisotropy                            6.608e-01
_pdbx_dcc_density.Z-score                               8.64
_pdbx_dcc_density.prob_peak_value                       7.717e-03
_pdbx_dcc_density.translational_pseudo_symmetry         Y
_pdbx_dcc_density.wavelength                            1.0
_pdbx_dcc_density.B_solvent                             74.833
_pdbx_dcc_density.K_solvent                             0.246
_pdbx_dcc_density.TLS_refinement_reported               N
_pdbx_dcc_density.partial_B_value_correction_attempted  N
_pdbx_dcc_density.partial_B_value_correction_success    N
_pdbx_dcc_density.reflection_status_archived            Y
_pdbx_dcc_density.reflection_status_used                Y
_pdbx_dcc_density.iso_B_value_type                      FULL
_pdbx_dcc_density.reflns_twin                           ?
_pdbx_dcc_density.twin_by_xtriage                       N
_pdbx_dcc_density.twin_operator     '1: H,K,L 2: -H,-K,L '
_pdbx_dcc_density.twin_fraction     '1: 0.6782 2: 0.3218 '
_pdbx_dcc_density.tls_group_number                      0
_pdbx_dcc_density.ncs_group_number                      0
_pdbx_dcc_density.mtrix_number                          0
_pdbx_dcc_density.Matthew_coeff                         3.22
_pdbx_dcc_density.solvent_content                       61.51
_pdbx_dcc_density.Cruickshank_dpi_xyz                   ?
_pdbx_dcc_density.dpi_free_R                            0.1277
_pdbx_dcc_density.fom                                   0.6936
_pdbx_dcc_density.correlation_overall                   0.7910
_pdbx_dcc_density.real_space_R_overall                  0.2688
_pdbx_dcc_density.mFo-DFc-3sigma_positive               ?
_pdbx_dcc_density.mFo-DFc-6sigma_positive               ?
_pdbx_dcc_density.mFo-DFc-3sigma_negative               ?
_pdbx_dcc_density.mFo-DFc-6sigma_negative               ?
_pdbx_dcc_density.Bmean-Bwilson                         2.858
_pdbx_dcc_density.Rfree-Rwork                           0.0000
_pdbx_dcc_density.error
;
Warning: B factor problems (B=  0.00 :ATOM   1961  OH  TYR A 272)
Warning: Number of Bfactor problems =1
Warning: (4nl7) Too few reflections for the free set(nfree=312, 4.66%).
Warning: Large difference of R_work: reported (0.2955), calculated (0.3493).
Note: It is suggested to do validation by phenix for this entry.
Warning: Large difference of R_free: reported (0.2955) calculated (0.3343).
Warning: Too small difference between the reported R_free (0.2955) and R_work(0.2955).
Warning: R_free(0.3343) is smaller than R_work(0.3493).
Warning: There may be problem with free set. Please try free set 1 and test again.
Warning: Large difference of completeness reported(88.97), calculated(68.7705).
Note: translational pseudo symmetry is detected by xtriage.

;
#
# the final items
#
loop_
_pdbx_dcc_density_corr.ordinal
_pdbx_dcc_density_corr.program
_pdbx_dcc_density_corr.ls_d_res_high
_pdbx_dcc_density_corr.ls_d_res_low
_pdbx_dcc_density_corr.ls_R_factor_R_all
_pdbx_dcc_density_corr.ls_R_factor_R_work
_pdbx_dcc_density_corr.ls_R_factor_R_free
_pdbx_dcc_density_corr.ls_number_reflns_obs
_pdbx_dcc_density_corr.ls_percent_reflns_obs
_pdbx_dcc_density_corr.ls_number_reflns_R_free
_pdbx_dcc_density_corr.correlation_coeff_Fo_to_Fc
_pdbx_dcc_density_corr.real_space_R
_pdbx_dcc_density_corr.correlation
_pdbx_dcc_density_corr.details
1   PHENIX  3.000 13.996      ? 0.2955 0.2955   6702 88.97     6702      ? 0.2688 0.7910 'PDB reported'
2   REFMAC  3.000 13.996      ? 0.3736 0.3877   6390 87.9758    312 0.5908    ?      ? 'without TLS correction'
3   REFMAC  2.755 14.430      ? 0.3493 0.3343   6394 68.7705    312 0.6101 0.2688 0.7910 'Use EXPE+TWIN '
4   REFMAC  2.755 14.430      ? 0.3493 0.3343   6394 68.7705    312 0.6101 0.2688 0.7910 'Best solution'
```

Figure 1
Example output from running *DCC* on the PDB entry 4n7l (Saer *et al.*, 2014) invoked with the command `dcc -pdb 4n7l.cif -sf r4n7lsf.ent` using publicly available data from the PDB archive. Ellipses represent sections of the file that have been removed for brevity. The first section is the local real-space density statistics determined by the programs *REFMAC* and *MAPMAN*. The second section is a combination of the model file and the results of `phenix.xtriage`, and the third section is the result of structure factor validation in *REFMAC*. Various error/warning messages are presented by the PDBx/mmCIF data item (`pdbx_density.error`). There are more than 200 possible error/warning messages in the *DCC* program. Different structures will export different messages.

3. Results and discussion

The wrapper program *DCC* was developed as a command-line tool that can perform a variety of tasks to aid in the validation of structure factors and atomic coordinate models and the biocuration of PDB depositions. It supports format conversion and generates appropriate input files for a number of third-party programs. By the creation of a simple-to-use front end, biocurators and users are provided access to a variety of software packages without having to know the intricacies of each.

The versatility of a tool such as *DCC* is shown by its use in wwPDB validation reports. In 2008, the wwPDB formed an X-ray Validation Task Force (Read *et al.*, 2011). To develop validation reports based on their recommendations, the wwPDB created a validation suite for X-ray structures (Gore *et al.*, 2012) that uses *DCC* to validate deposited structure factors.

Another use case arose during the 2011 wwPDB remediation effort to identify X-ray structures in which partial *B* factors were present in the atomic coordinate model file. Based on the output of *DCC*, annotators corrected ***TLS*** information in the entries and furnished an indicator that only partial *B* factors were present.

4. Conclusions

The program *DCC* is a versatile tool that is used daily by wwPDB biocurators. The usage of PDBx/mmCIF allows *DCC* to be employed in automatic pipelines. It is available for download from http://sw-tools.rcsb.org.

Acknowledgements

The authors thank Chenghua Shao for extensive testing and feedback during software development. This work was funded by a grant (No. DBI-1338415) from the US National Science Foundation. The RCSB PDB is managed by two members of the RCSB, Rutgers and UCSD, and is a member of the wwPDB organization.

References

Adams, P. D., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. & Terwilliger, T. C. (2002). *Acta Cryst.* D**58**, 1948–1954.
Berman, H. M., Henrick, K. & Nakamura, H. (2003). *Nat. Struct. Biol.* **10**, 980.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Gore, S., Velankar, S. & Kleywegt, G. J. (2012). *Acta Cryst.* D**68**, 478–483.
Hanson, R. M. (2010). *J. Appl. Cryst.* **43**, 1250–1260.
Howlin, B., Butler, S. A., Moss, D. S., Harris, G. W. & Driessen, H. P. C. (1993). *J. Appl. Cryst.* **26**, 622–624.
Kinjo, A. R., Suzuki, H., Yamashita, R., Ikegawa, Y., Kudou, T., Igarashi, R., Kengaku, Y., Cho, H., Standley, D. M., Nakagawa, A. & Nakamura, H. (2012). *Nucleic Acids Res.* **40**, D453–D460.
Kleywegt, G. J., Harris, M. R., Zou, J., Taylor, T. C., Wählby, A. & Jones, T. A. (2004). *Acta Cryst.* D**60**, 2240–2249.
Murshudov, G., Dodson, E. & Vagin, A. (1996). *Proceedings of the CCP4 Study Weekend: Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 93–104. Warrington: CCLRC Daresbury Laboratory.
Read, R. J. *et al.* (2011). *Structure*, **19**, 1395–1412.
Saer, R. G., Pan, J., Hardjasa, A., Lin, S., Rosell, F., Mauk, A. G., Woodbury, N. W., Murphy, M. E. & Beatty, J. T. (2014). *Biochim. Biophys. Acta*, **1837**, 366–374.
Tickle, I. J. (2012). *Acta Cryst.* D**68**, 454–467.
Touw, W. G. & Vriend, G. (2014). *Protein Eng. Des. Sel.* **27**, 457–462.
Ulrich, E. L. *et al.* (2008). *Nucleic Acids Res.* **36**, D402–D408.
Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *Acta Cryst.* D**55**, 191–205.
Velankar, S. *et al.* (2016). *Nucleic Acids Res.* **44**, D385–D395.
Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.