Multivariate Analyses of Quality Metrics for Crystal Structures in the PDB Archive

The Protein Data Bank archive (PDB) is the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies. PDB was established in 1971 to store the results of crystallographic studies, and expanded to support NMR and electron microscopy. The Worldwide PDB (wwPDB) organization manages the PDB archive and ensures that PDB data are freely and publicly available to the global community.[1]

While most entries are described by scientific publications, corresponding citations are not required. PDB entries are considered publications themselves and are assigned DOIs.

wwPDB produces Validation Reports that provide an assessment of structure quality using widely accepted standards and criteria recommended by method-specific Validation Task Forces.[2-4] These Reports are important tools for communicating crystallographic results across the pipeline. For Depositors, the validation reports provide an opportunity to ensure the data submitted best represent the experimental data. For Users, Reports communicate data quality, even to Users who are not experts in crystallography.

A recent study[5] by the RCSB Protein Data Bank[6] compared data quality in depositions before and after the introduction of the wwPDB Validation Reports. Quality improvements were found relating to pairwise atom-atom clashes, sidechain torsion angle rotamers, and local agreement between the atomic coordinate structure model and experimental electron density data for proteins. These improvements are largely independent of resolution limit, sample molecular weight, and other confounding factors, thus manifest the pronounced impact on data quality of the introduction of community-accepted validation tools and improving visibility of the potential data issues in the biocuration process. On the other hand, no significant improvement in the quality of associated ligands was observed from the study. Efforts are currently underway to improve presentation of ligand quality metrics.[7]

The relation among different quality measures was also studied by correlation and Principal Component Analysis (PCA). Our results revealed that three primary quality measures (Clashscore, percent of Ramachandran Outliers, and percent of Sidechain Rotamer Outliers) are strongly correlated with one another, leading to the construction of a combined molecular geometry quality measure. The combined geometry measure, together with Rfree and Real Space R-factor Z-score, can be used to summarize the overall structure quality in a easy-to-display 3-dimensional quality metrics space, which can in turn be reduced to a 1-dimensional overall quality metric readily interpretable by all PDB archive users, and thus an important tool for communicating crystallographic results.

RCSB PDB is funded by a grant (DBI-1338415) from the National Science Foundation, the National Institutes of Health, and the US Department of Energy.

Chenghua Shao RCSB PDB; Rutgers, The State University of New Jersey
Huanwang Yang RCSB PDB; Rutgers, The State University of New Jersey
John D Westbrook RCSB PDB; Rutgers, The State University of New Jersey
Jasmine Young RCSB PDB; Rutgers, The State University of New Jersey
Christine Zardecki RCSB PDB; Rutgers, The State University of New Jersey
Stephen K. Burley RCSB PDB; Rutgers, The State University of New Jersey

References

1. H. M. Berman, K. Henrick, H. Nakamura. (2003) Announcing the worldwide Protein Data Bank. *Nat Struct Biol* 10: 980.
2. R. Henderson, A. Sali, M. L. Baker, B. Carragher, B. Devkota, K. H. Downing, E. H. Egelman, Z. Feng, J. Frank, N. Grigorieff, W. Jiang, S. J. Ludtke, O. Medalia, P. A. Penczek, P. B. Rosenthal, M. G. Rossmann, M. F. Schmid, G. F. Schroder, A. C. Steven, D. L. Stokes, J. D. Westbrook, W. Wriggers, H. Yang, J. Young, H. M. Berman, W. Chiu, G. J. Kleywegt, C. L. Lawson. (2012) Outcome of the first electron microscopy validation task force meeting. *Structure* 20: 205-214.
3. G. T. Montelione, M. Nilges, A. Bax, P. Guntert, T. Herrmann, J. S. Richardson, C. D. Schwieters, W. F. Vranken, G. W. Vuister, D. S. Wishart, H. M. Berman, G. J. Kleywegt, J. L. Markley. (2013) Recommendations of the wwPDB NMR Validation Task Force. *Structure* 21: 1563-1570.
4. R. J. Read, P. D. Adams, W. B. Arendall, 3rd, A. T. Brunger, P. Emsley, R. P. Joosten, G. J. Kleywegt, E. B. Krissinel, T. Lutteke, Z. Otwinowski, A. Perrakis, J. S. Richardson, W. H. Sheffler, J. L. Smith, I. J. Tickle, G. Vriend, P. H. Zwart. (2011) A new generation of crystallographic validation tools for the protein data bank. *Structure* 19: 1395-1412.
5. C. Shao, H. Yang, J. D. Westbrook, Y. Y. Young, C. Zardecki, S. K. Burley. (2017) Multivariate analyses of quality metrics for crystal structures in the PDB archive. *Structure* http://dx.doi.org/10.1016/j.str.2017.01.013.
6. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne. (2000) The Protein Data Bank. *Nucleic Acids Res* 28: 235-242.
7. P. D. Adams, K. Aertgeerts, C. Bauer, J. A. Bell, H. M. Berman, T. N. Bhat, J. M. Blaney, E. Bolton, G. Bricogne, D. Brown, S. K. Burley, D. A. Case, K. L. Clark, T. Darden, P. Emsley, V. A. Feher, Z. Feng, C. R. Groom, S. F. Harris, J. Hendle, T. Holder, A. Joachimiak, G. J. Kleywegt, T. Krojer, J. Marcotrigiano, A. E. Mark, J. L. Markley, M. Miller, W. Minor, G. T. Montelione, G. Murshudov, A. Nakagawa, H. Nakamura, A. Nicholls, M. Nicklaus, R. T. Nolte, A. K. Padyana, C. E. Peishoff, S. Pieniazek, R. J. Read, C. Shao, S. Sheriff, O. Smart, S. Soisson, J. Spurlino, T. Stouch, R. Svobodova, W. Tempel, T. C. Terwilliger, D. Tronrud, S. Velankar, S. C. Ward, G. L. Warren, J. D. Westbrook, P. Williams, H. Yang, J. Young. (2016) Outcome of the First wwPDB/CCDC/D3R Ligand Validation Workshop. *Structure* 24: 502-508.