Improving the quality of 3D structure data in the Protein Data Bank with coordinate versioning supported by OneDep

J. Y. Young

Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

jasmine.young@rcsb.org

The Protein Data Bank (PDB) [1] was established as the first open-access digital data resource in biology in 1971 with just seven X-ray crystallographic structures of proteins. Today, the single global archive houses more than 177,000 experimentally determined 3D structures of biological macromolecules that are made freely available to millions of users worldwide with no limitations on usage. This information facilitates basic and applied research and education across the sciences, impacting fundamental biology, biomedicine, biotechnology, bioengineering, and energy sciences. The PDB archive is managed jointly by the Worldwide Protein Data Bank [2-3] (wwPDB, wwpdb.org) which is committed to making PDB data Findable-Accessible-Interoperable–Reusable (*FAIR*) [4].

To ensure the highest quality structure data, the wwPDB OneDep system for structure deposition [5], validation [6], and biocuration [7] (deposit.wwpdb.org) provides enhanced validation reports. During 2019-2020, wwPDB implemented depositor-initiated coordinate versioning that enables the Depositor of Record (or Principal Investigator) to replace previously released x,y,z atomic coordinates without obsoleting the original PDB entry or changing the PDB ID. This feature was developed in response to feedback from PDB depositors, who were reluctant to update their structures because the newly issued PDB ID would differ from that reported in original structure publication. We thank all of many PDB depositors, who have proactively corrected their structures using the new versioning feature within OneDep. To date, more than 200 PDB structures have been updated with newly versioned atomic coordinates.

Since the early days of the COVID-19 pandemic, PDB data have informed our understanding of SARS-CoV-2 protein structure, function, and evolution, and facilitated structure-guided discovery and development of anti-coronaviral drugs, vaccines, and neutralizing monoclonal antibodies. More than 1,000 SARS-CoV-2 related protein structures are now freely available from the PDB, reflecting enormous efforts made by the structural biology community in fighting the pandemic. Occasionally, rapid PDB data deposition and publication of coronavirus structural studies driven by an understandable sense of urgency has resulted in public release of PDB structures containing minor errors. The wwPDB coordinate versioning feature described above has enabled rapid correction of SARS-CoV-2 related structures archived in the PDB.

The wwPDB strongly encourages PDB depositors to update structures as needed using OneDep. Doing so will improve the quality of data stored in the archive, while preserving original PDB IDs and maintaining connections to the scientific literature.

[1] Protein Data Bank. (1971). Crystallography: Protein Data Bank. Nature (London), New Biol. 233:223-223.

[2] Berman, H., Henrick, K., Nakamura, H. (2003). Announcing the worldwide protein data bank. Nat Struct Biol. 10:980.

[3] wwPDB consortium. (2019). Protein Data Bank: The single global archive for 3d macromolecular structure data. Nucleic Acids Res. 47:D520-D528.

[4] Wilkinson, MD, Dumontier, M, Aalbersberg, IJ, Appleton, G, Axton, M, Baak, A, Blomberg, N, Boiten, JW, da Silva Santos, LB, Bourne, PE, et al. (2016). The FAIR guiding principles for scientific data management and stewardship. Sci Data. 3:1-9

[5] Young, J. Y., Westbrook, J. D., Feng, Z., Sala, R., Peisach, E., Oldfield, T. J., Sen, S., Gutmanas, A., Armstrong, D. R., Berrisford, J. M., et al. (2017). OneDep: Unified wwpdb system for deposition, biocuration, and validation of macromolecular structures in the pdb archive. Structure. 25:536-545.

[6] Gore, S., Sanz Garcia, E., Hendrickx, P. M. S., Gutmanas, A., Westbrook, J. D., Yang, H., Feng, Z., Baskaran, K., Berrisford, J. M., Hudson, B. P., et al. (2017). Validation of structures in the protein data bank. Structure. 25:1916-1927.

[7] Young, J. Y., Westbrook, J. D., Feng, Z., Peisach, E., Persikova, I., Sala, R., Sen, S., Berrisford, J. M., Swaminathan, G. J., Oldfield, T. J., et al. (2018). Worldwide protein data bank biocuration supporting open access to high-quality 3d structural biology data. Database. 2018:bay002.

Keywords: Protein Data Bank; PDB; worldwide Protein Data Bank; structure data quality; biocuration; validation; OneDep; structure biology; macromolecular structure

RCSB PDB is funded by the National Science Foundation (DBI-1832184), the US Department of Energy (DE-SC0019749), and the National Cancer Institute, National Institute of Allergy and Infectious Diseases, and National Institute of General Medical Sciences of the National Institutes of Health under grant R01GM133198.