Synergies between the Protein Data Bank and the community

The Protein Data Bank (PDB) is a community resource. But how do we define community, and how has it changed over the last 50 years since the PDB was founded? How did the community influence the evolution of the PDB, and how did the PDB influence both the science and the behavior of the community?

Helen M. Berman

The PDB community is large and heterogeneous. It consists of structural biologists who deposit their data, scientists, educators and students who use the data, and journals that publish articles about the deposited structures and the corresponding analyses. The community also includes professional societies whose members deposit and use the PDB data and funding agencies that have ensured that the PDB has the resources to operate. Each sector has its own needs and requirements that must be considered for the PDB to be a useful and effective resource.

From the beginning, the PDB archive was a global operation, and since 2003, it has been managed by an international group called the Worldwide Protein Data Bank (wwPDB) that consists of data centers in the US, Europe and Asia[1]. Since its inception, the wwPDB has tried to ensure that the various community interests are reflected in its data management policies.

Community efforts that began in the 1960s came together in 1971 when Walter Hamilton at Brookhaven National Laboratory (BNL) in New York and Olga Kennard at the Crystallographic Data Centre (Cambridge, UK) agreed to set up the PDB as an international resource for structures obtained using X-ray crystallography[2]. The archive was launched with just seven structures. At first, protein crystallographers needed to be convinced to add their data to this newly formed resource. Tom Koetzle, who became the head of the PDB after Hamilton's untimely death in 1973, searched the literature for newly published structures and wrote personal letters to protein crystallographers requesting that they deposit their data. Several researchers were notable champions of the effort. Michael Rossmann put his considerable energy into lobbying his colleagues to deposit data. At the time, sharing crystallographic data was a challenging endeavor, requiring shipment of computer punch cards or magnetic tape reel to the PDB. But at a minimum, deposition ensured that the data would not be lost. Although structure validation tools were minimal as compared to now, the depositor could be assured that errors were caught by the PDB staff, especially Frances Bernstein at BNL, who worked with the PDB for almost twenty-five years. Before long, structures determined using nuclear magnetic resonance (NMR) spectroscopy[3] and then three-dimensional electron microscopy (3DEM) were also deposited[4]. As the PDB grew in size and in scope, it became clear to some members of the structural biology community—most notably Dick Dickerson and Fred Richards—that the deposition of data should be a prerequisite for the publication of any structure in a scientific journal. The International Union of Crystallography (IUCr) formed a committee to recommend best practices and, in 1989, guidelines were published that articulated the timing for depositions of coordinates and structure factors and that placed "holds" on data release until publication[5]. Not surprisingly, the first journal to require deposition was *Acta Crystallographica D*. Now virtually all journals require the deposition of structural data into the PDB.

PDB holdings have steadily increased over the years. In 1976, the number of structures available was 13; in 2021, it is more than 175,000. As the PDB grew, the user community expanded along with it. Computational biologists began to use PDB data for protein classification[6,7] and protein structure prediction[8]. New specialty data resources were created, with deep curation for particular groups of structures; there are now more than 200 related resources that are regularly updated with PDB data. A whole new field called structural bioinformatics was born[9]. The PDB also became an essential resource for drug development efforts, as exemplified during the HIV–AIDS epidemic of the 1980s and the current COVID-19 pandemic. As modern web and visualization tools were brought online, PDB data were used more and more for educational purposes[10], and the PDB began to create specific resources for education[11]. Right now, more than 2.5 million coordinate sets are downloaded every day by a very diverse user community.

As the PDB began to be widely used, it became more and more critical that the data were reliable. Coordinate data needed not only to yield good geometry but also had to fit the experimental data. Both the PDB[12] and its users[13] urged that structure factors be deposited. In 2007, erroneous structures deposited in the PDB were detected by independent analyses of the structures and structure factors. Structural biologists were dismayed and deeply concerned by these events[14]. Until 2008, the PDB only mandated the deposition of coordinates; deposition of the underlying experimental data that would allow checking of the structure against the data was optional. In 2008, structure factor deposition became mandatory, opening the door to the establishment of a rigorous validation pipeline. An X-ray Validation Task Force[15] was commissioned by the wwPDB to identify the criteria to be used. Task Forces were also created for NMR[16], 3DEM[17], small angle scattering[18,19] and integrative modeling methods[20,21]. Through these Task Forces, researchers from the respective communities provide recommendations regarding data standards as well as best practices for structure curation and validation. The wwPDB implements recommendations of the Task Forces in the form of validation reports that are produced by the wwPDB data management system, called OneDep[22]. The reports are now an integral part of the data deposition process and are required by many journals as part of the review process[23]. wwPDB validation reports are made publicly available at the time of data release.

Each step in the evolution of the PDB involved collaborations and conversations among community members, with the goal of building a consensus as to how exactly the archive should be operated. The passionate involvement of the various

sectors is one reason why the PDB has endured for 50 years and why it will continue to be a key resource for many years to come. ❐

Helen M. Berman [1,2] ✉

[1]*Department of Chemistry and Chemical Biology, Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), Rutgers, the State University of New Jersey, Piscataway, NJ, USA.*
[2]*The Bridge Institute, Michelson Center for Convergent Bioscience, University of Southern California, Los Angeles, CA, USA.*
✉*e-mail: berman@rcsb.rutgers.edu*

Published online: 7 May 2021
https://doi.org/10.1038/s41594-021-00586-6

References

1. Berman, H., Henrick, K. & Nakamura, H. *Nat. Struct. Biol.* **10**, 980 (2003).
2. *Nature New Biol.* **233**, 223 (1971).
3. Williamson, M. P., Havel, T. F. & Wüthrich, K. *J. Mol. Biol.* **182**, 295–315 (1985).
4. Henderson, R. et al. *J. Mol. Biol.* **213**, 899–929 (1990).
5. International Union of Crystallography. *Acta Crystallogr. A* **45**, 658 (1989).
6. Andreeva, A., Kulesha, E., Gough, J. & Murzin, A. G. *Nucleic Acids Res.* **48**, D376–D382 (2020).
7. Sillitoe, I., Dawson, N., Thornton, J. & Orengo, C. *Biochimie* **119**, 209–217 (2015).
8. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. *Proteins* **87**, 1011–1020 (2019).
9. Gu, J. & Bourne, P. E. (eds) *Structural Bioinformatics* 2nd edn. (Wiley-Blackwell, 2009).
10. Martz, E. *Trends Biochem. Sci.* **27**, 107–109 (2002).
11. Goodsell, D. S., Zardecki, C., Berman, H. M. & Burley, S. K. *Biochem. Mol. Biol. Educ.* **48**, 350–355 (2020).
12. Jiang, J., Abola, E. & Sussman, J. L. *Acta Crystallogr. D Biol. Crystallogr.* **55**, 4 (1999).
13. Wlodawer, A. *Acta Crystallogr. D Biol. Crystallogr.* **63**, 421–423 (2007).
14. Borrell, B. *Nature* **462**, 970 (2009).
15. Read, R. J. et al. *Structure* **19**, 1395–1412 (2011).
16. Montelione, G. T. et al. *Structure* **21**, 1563–1570 (2013).
17. Henderson, R. et al. *Structure* **20**, 205–214 (2012).
18. Trewhella, J. et al. *Structure* **21**, 875–881 (2013).
19. Trewhella, J. et al. *Acta Crystallogr. D Struct. Biol.* **73**, 710–728 (2017).
20. Sali, A. et al. *Structure* **23**, 1156–1167 (2015).
21. Berman, H. M. et al. *Structure* **27**, 1745–1759 (2019).
22. Young, J. Y. et al. *Structure* **25**, 536–545 (2017).
23. *Nat. Struct. Mol. Biol.* **23**, 871 (2016).

Competing interests

The author declares no competing interests.