Establishing the Next Generation of the Protein Data Bank


HELEN M. BERMAN 1,2

1. Board of Governors Professor of Chemistry and Chemical Biology and Director of the RCSB Protein Data Bank (RCSB PDB).
2. Department of Chemistry and Chemical Biology and the Center for Integrative Proteomics Research Rutgers, The State University of New Jersey

The Protein Data Bank (PDB) archive was officially launched at Brookhaven National Laboratory (BNL) in 1971 (Protein Data Bank 1971), thirteen years after the first crystal structure of the globular protein myoglobin was published (Kendrew et al. 1958). The PDB was initially established as a repository for the 3D structures of biological macromolecules that were beginning to be determined using X-ray crystallography. In those early days, first under the leadership of Walter Hamilton, and then Tom Koetzle, the PDB served as an archive for the structural biologists who would submit and/or request access to data. The available protein structure coordinates were announced in a newsletter, and scientists who wanted access to these data would send blank magnetic tapes to BNL. Data would be loaded onto these tapes and mailed back to the users; in the 1990’s, CD-ROM distribution was introduced. In 1988, when there were almost 300 structures in the PDB, a total of 259 tapes were distributed. A similar number of CD-ROMs were distributed in 1992. As usage of the World Wide Web became more common, a PDB website simplifying access to the data was created under the leadership of Joel Sussman (Peitsch et al. 1995; Stampf, Felder, and Sussman 1995). The PDB user community began to grow and diversify.

In 1998, the National Science Foundation issued a Request For Proposals for management of the PDB archive, and a new organization responded with an ambitious proposal. The Research Collaboratory for Structural Bioinformatics (RCSB) group included experts in creating databases for structural biology from Rutgers, The State University of New Jersey, University of California San Diego, and the National Institute of Standards and Technology. Following a competitive review, the RCSB received the award to manage the PDB and the new RCSB PDB was launched at rcsb.org in 1999. This management change was not universally applauded, and there were many misconceptions about what changes may take place. The article published in 2000 in Nucleic Acids Research (Berman et al. 2000) aimed to announce the new systems developed by the RCSB PDB for data deposition, validation, and access, and to clarify any misconceptions.

One development described in the article was the use of the Macromolecular Crystallographic Information File dictionary (Fitzgerald et al. 2005) to standardize and process data. PDBx/mmCIF contains specific definitions for every data item and does not have any restrictions with respect to the size and complexity of a structure. Some users were concerned that the PDB file format, created in the
1970s (Bernstein et al. 1977) and used by so many software applications, would no longer be available. However, data now processed and annotated data using the PDBx/mmCIF dictionary could be easily translated into the PDB file format. In fact, data represented in the PDB file format became far more standardized and reliable due to the checking enabled by the use of the formalized data dictionary. Today, the limitations of the original PDB file format are obvious, given its inability to support the increasingly large and complex structures that are being deposited to the archive. As a result, there is community acceptance for PDBx/mmCIF as the official format for PDB data.

The Nucleic Acids Research article also described specific validation procedures that were incorporated to assess the quality of the data deposited to the archive. In 1999, some PDB users felt that the archive should only contain data as directly submitted by the authors, since validation procedures could lead to a misrepresentation of author intent. While structural biologists may have expert understanding of the data represented in a PDB entry, this point of view did not take into account the growing number of non-structural biologist users and students to whom data quality is very important. Indeed in the 2000’s, there was a huge push for more powerful and standardized validation methods in response to concerns about misconduct and misuse of the data. Now the need for uniform and reliable data across the archive is well recognized.

The 2000 article also described partnerships established with the European Bioinformatics Institute and Osaka University for data processing and exchange so as to ensure a single archive. At the time, some believed that multiple data processing centers would only lead to a disorganized data archive. However, the collaboration between data deposition and processing centers only continued to strengthen over the years. These partnerships were formalized in 2003 with the creation of the Worldwide Protein Data Bank (wwPDB; wwwpdb.org), whose new mission was to ensure standard representation and processing standards in the PDB archive (Berman, Henrick, and Nakamura 2003). The wwPDB collaboration has worked to standardize data across the archive through targeted “remediation” efforts (Lawson et al. 2008; Henrick et al. 2008) and developed a new system for data deposition and annotation that will be used by all wwPDB data centers (Quesada et al. 2011; Gore, Velankar, and Kleywegt 2012; Young et al. 2013). In addition, the existence of several data centers makes it possible to keep up with an increasing data load.

The RCSB PDB was also responsible for the development and support of resources for distribution of the data archive. As noted in the article, the PDB archive contained 10,714 released entries of proteins and nucleic acids as studied by X-ray crystallography, nuclear magnetic resonance (NMR), and other methods in September 1999. This included the distribution of the archive of “flat” data files via FTP, which was still copied onto CD-ROMs and sent via postal mail on a quarterly basis. In addition, the RCSB PDB established a website offering many functionalities. At its core was an integrated system of databases, including a relational database, an object oriented database, a text search engine, and a database of crystallization information. The RCSB PDB website supported a variety of searches for structures across the archive, the production of a large variety of reports, and the ability to visualize and analyze individual structures. Links to related resources were provided. Online mirrors of this website were also maintained among the RCSB PDB partner sites and with international collaborators.

The 2000 Nucleic Acids Research article described a Protein Data Bank that was poised to evolve with new science, utilize new technologies to serve the data, and work with its user community to provide the resources required to enable scientific research. The citation is used to reference access through rcsb.org to the ~100,000 entries now available in the PDB archive, as well as to the specialized resources that the RCSB PDB provides for analysis and visualization. A review of citations referencing the 2000 article shows that the vast majority of citations are from users who do not themselves deposit data. These articles are from the wide spectrum of the biological community including biochemists and cell, molecular, and evolutionary biologists. Mathematicians, statisticians and physicists also cite use of the RCSB PDB. Website usage of the RCSB PDB site continues to increase. On average, RCSB.org was accessed by about 319,000 unique visitors from about 190 countries, transferring 1924 GB of data from the website each month in 2013. The initial services provided by the RCSB PDB website continue
to be developed and improved in order to support the expanding community of users (Rose et al. 2013; Rose et al. 2011). Increasingly, students and teachers use the RCSB PDB for the educational resources provided, such as a regular *Molecule of the Month* column. A goal of the RCSB PDB is to create a structural view of biology. That goal is within reach.

**ACKNOWLEDGEMENTS**

The RCSB PDB is funded by the NSF (DBI 0829586), NIH, and DOE. RCSB PDB is a member of the Worldwide Protein Data Bank.

**REFERENCES**


Quesada, Martha, John Westbrook, Tom Oldfield, Jasmine Young, Jawahar Swaminathan, Zukang Feng, Sameer Velankar, Takanori Matsuura, Eldon Ulrich, Steve Madding, Gerard J. Kleywegt, John L. Markley, Haruki Nakamura, and Helen M. Berman. 2011. "The wwPDB common tool for deposition and


Stampf, DR, CE Felder, and JL Sussman. 1995. “PDBBrowse - a graphics interface to the Brookhaven Protein Data Bank.” Nature no. 374:572-574. doi: 10.1038/374572a0

Young, Jasmine Y., Zukang Feng, Dimitris Dimitropoulos, Raul Sala, John Westbrook, Marina Zhuravleva, Chenghua Shao, Martha Quesada, Ezra Peisach, and Helen M. Berman. 2013. “Chemical annotation of small and peptide-like molecules at the Protein Data Bank.” Database no. 2013:bat079.