Growth Of the PDB Archive Requires Transition to Pdbx/Mmcif Format Files

Sutapa Ghosh[1], Zukang Feng[2], Yuhe Liang[3], Ezra Peisach[4], Irina Persikova[5], Chenghua Shao[6], Jasmine Y Young[7], wwPDB Team[8], Stephen K Burley[9]

[1]RCSB Protein Data Bank, Rutgers, The State University of New Jersey, [2]RCSB Protein Data Bank, Rutgers, The State University of New Jersey, [3]RCSB Protein Data Bank, Rutgers, The State University of New Jersey, [4]RCSB Protein Data Bank, Rutgers, The State University of New Jersey, [5]RCSB Protein Data Bank, Rutgers, The State University of New Jersey , [6]RCSB Protein Data Bank, Rutgers, The State University of New Jersey, [7]RCSB Protein Data Bank, Rutgers, The State University of New Jersey, [8]RCSB Protein Data Bank, Rutgers, The State University of New Jersey, PDBe, EMBL-European Bioinformatics Institute, Hinxton, PDBj, Institute for Protein Research, Osaka University, Osaka, Osaka, Japan. EMDB, EMBL- European Bioinformatics Institute, BMRB, UConn Health, [9]RCSB Protein Data Bank, Rutgers, The State University of New Jersey, RCSB Protein Data Bank, San Diego Supercomputer Center, University of California

sutapa@rcsb.rutgers.edu

The Protein Data Bank (PDB) was established in 1971 as the first open-access digital data resource in biology with just seven X-ray crystallographic structures of proteins. Today, the single global PDB archive houses more than 200,000 experimentally-determined three- dimensional (3D) structures of biological macromolecules that are made freely available to many millions of users worldwide with no limitations on usage. 3D biostructure information facilitates basic and applied research and education across the sciences, impacting fundamental biology, biomedicine, biotechnology, and energy sciences. The Worldwide Protein Data Bank partnership (wwPDB, wwpdb.org) currently includes five Full Members (RCSB PDB, PDBe, PDBj, BMRB, and EMDB) and one Associate Member (PDBc), which together manage the PDB, EMDB, and BMRB Core Archives. wwPDB Members are committed to ensuring that structural biology data are Findable, Accessible, Interoperable, and Reusable (FAIR).

Growth of the PDB and development of new structure determination methods are changing how 3D biostructures are represented. For many years, fixed-format PDB files, based on the 80-column Hollerith punch card, were fit for purpose. In the modern era, many larger, more complex 3D biostructures that cannot be represented using the legacy format are being contributed to the PDB. To overcome this challenge, PDBx/mmCIF (Protein Data Bank exchange/macromolecular CIF) data standard/dictionary was adopted in 2014 as the official master format for the PDB Core Archive. It is flexible, fully-extensible, both human- and machine-readable, and can accommodate 3D biostructures of any size and composition. As its name implies, PDBx/mmCIF is based on CIF (Crystallographic Information File or Crystallographic Information Framework) developed for chemical crystallography by the International Union of Crystallography in 1990. Importantly, PDBx/mmCIF format files may include additional metadata that cannot be stored in legacy PDB format files (e.g., XFEL experimental details, ligands of interest, detailed revision history).

In the very near future (Figure 1), the wwPDB will exhaust available three-character Chemical Component Dictionary (CCD) IDs (e.g, 4W8, the CCD ID for nirmatrelvir, the active ingredient of Pfizer's SARS-CoV-2 Main Protease inhibitor). Although not as pressing, the wwPDB will eventually exhaust available four-character PDB identifiers (IDs; e.g., 7RFR, co- crystal structure of the SARS-CoV-2 Main Protease bound to nirmatrelvir). To address these challenges, newly defined CCD compounds will be assigned 5-character IDs, once 3- character IDs run out. Existing 3-character CCD IDs will remain as before (e.g., ALA remains ALA).

PDB IDs will be extended to eight characters (e.g., 7RFR will become PDB_00007RFR). While best efforts are currently made by wwPDB partners to provide legacy format files for most 3D biostructures stored in the PDB Core Archive, those containing small-molecule ligands represented using extended CCD IDs will only be made available in PDBx/mmCIF format.

Herein, we explain the limits of the fixed-format legacy PDB file and showcase the power of the fully-extensible PDBx/mmCIF data standard/dictionary, with emphasis on (a) understanding the basics of PDBx/mmCIF data dictionary and file format, (b) software tools for generating and working with PDBx/mmCIF files, and (c) accessing 3D biostructure data in PDBx/mmCIF format.

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.

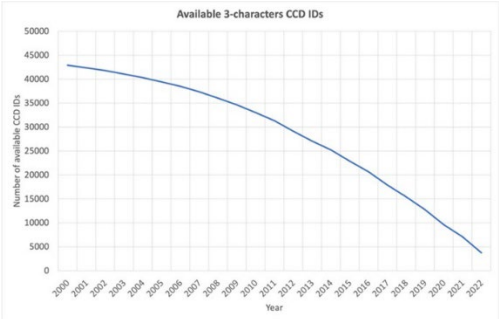

Figure-1: The number of available 3-character CCD IDs annually

Figure 1