The Life and Times of the PDB Format - Looking Towards the Future with mmCIF

Gregg Crichlow[1], Catherine Lawson[2], John Westbrook[3], Stephen Burley[4], Sanja Abbott[5], Kumaran Baskaran[6], John Berrisford[7], Zukang Feng[8], Yasuyo Ikegawa[9], Masashi Yokochi[10], Jasmine Y Young[11], Jeffrey Hoch[12], Genji Kurisu[13]

[1]No affiliation given [2]Institute for Quantitative Biomedicine, Rutgers Univ, [3]No affiliation given, [4]RCSB Protein Data Bank, Rutgers University, [5]No affiliation given, [6]No affiliation given, [7]No affiliation given, [8]Rutgers University, [9]No affiliation given, [10]No affiliation given, [11]Rutgers University, [12]No affiliation given

[13]No affiliation given

gregg.crichlow@rutgers.edu

The PDB file format was created when the Protein Data Bank was established In 1971. Originally designed to use 70 columns on punched cards, the format specified fixed column positions and column widths for all data items. As the science and technology evolved, this legacy PDB format can no longer support data representation for large complex structures. This includes large entries that do not fit the PDB file format (those containing >62 chains and/or >99999 ATOM records).

Since the late 1990s, PDB files are also available in mmCIF, an extensible machine readable format that is not limited by fixed column positions, and is extensible to support new data content. The legacy PDB format has not been modified or extended to support new content since 2012. Since 2014, legacy PDB formatted files have not been produced for entries containing multi-character chain ids or atom serial numbers with greater than five digits.

Working with the community PDBx/mmCIF Working Group, wwPDB is transitioning to sunset the legacy PDB format when the PDB Chemical Component Dictionary (CCD) IDs are extended beyond three alphanumeric characters which cannot be accommodated by the PDB format. We anticipate running out of three-character CCD IDs in the next three to five years. At that time, wwPDB will start issuing four alphanumeric codes for CCD IDs in the OneDep deposition-validation-biocuration system. Entries containing these codes will not have legacy PDB format files. In 2022, wwPDB will begin the implementation of this CCD ID extension. In addition, wwPDB also plans to implement extended PDB IDs that are eight-characters with a PDB prefix, e.g. pdb_00001abc. This extended PDB ID will be included in the PDBx/mmCIF file for new entries issued with four-character PDB IDs. Once the four-character PDB IDs are all consumed, newly deposited PDB entries will only be available in PDBx/mmCIF format.

wwPDB is asking community and user software developers to review their code and remove such limitations for the future.