RCSB PROTEIN DATA BANK: Enhancing Data Exploration for Future Generations of PDB Users

Stephen Burley[1]

[1]RCSB Protein Data Bank, Rutgers University

Stephen.Burley@rcsb.org

The Protein Data Bank (PDB) was established in 1971 as the first open-access digital data resource in biology. Beginning with only seven protein structures, the PDB archive has ballooned to more than 190,000 experimentally-determined, three-dimensional (3D) structures of proteins, DNA, and RNA (totaling more than 1 billion atoms). Currently supported experimental methods include macromolecular crystallography (including micro-electron diffraction), nuclear magnetic resonance spectroscopy, and 3D electron microscopy. The PDB Core Archive is managed jointly by the Worldwide Protein Data Bank (wwPDB), including five Core Members [RCSB Protein Data Bank, RCSB PBD; Protein Data Bank in Europe, PDBe; Protein Data Bank Japan, PDBj; Electron Microscopy Data Bank, EMDB; and Biological Magnetic Resonance Bank, BMRB]. wwPDB will anticipates expanding over the coming years to include Protein Data Bank China (PDBc) and Protein Data Bank India (PDBi), both as Associate Members. Additional wwPDB Core Archives include two jointly-managed specialist data resources for 3D electron microscopy (EMDB) and nuclear magnetic resonance spectroscopy (BMRB).

The PDB Core Archive is universally regarded as a core data science resource of fundamental importance to the wider life-science community and long-term preservation of machine-readable biological data. PDB structures are the molecules of life. Knowledge of 3D structures (shapes) of biomolecules, how they evolve with time, and how they function in nature is essential for understanding fundamental biology, biomedicine, and energy science. PDB data impacts basic and applied research on health and disease of humans, animals, and plants; production of food and energy; and other research pertaining to global prosperity and environmental sustainability. 3D biostructure data are also important to biopharmaceutical and biotechnology companies, accelerating data-driven discovery and development of new drugs, materials, and devices. Today, powerful pulsed X-ray facilities, cryogenic electron microscopes, and new integrative/hybrid (I/H) methods for structure determination are accelerating biomedical research with functional insights into ever more complex biological systems at the atomic level. Cryo-electron tomography even allows study of macromolecular machines "caught in the act" inside frozen cells. In favorable cases, atomic-level structures can be determined using sub-volume averaging of multiple copies of the same molecular assembly imaged via tomography.

The RCSB PDB research-focused website (RCSB.org) is visited by many millions of unique users annually. To meet user needs and accommodate relentlessly growing volume and complexity of PDB data, RCSB.org website software architecture relies on modular web services and Application Programming Interfaces or APIs that enable efficient data delivery and streamlined maintenance, and facilitate enhancement of existing features and addition of new features. The new architecture also supports use of Boolean logic in searching across PDB data and metadata and additional information, which is integrated weekly from more than 50 external data resources. The RCSB PDB education/outreach-focused PDB-101 website (PDB101.RCSB.org) is visited by nearly one million unique users annually. The most popular PDB-101 feature is the Molecule of the Month article series, now exceeding 270 lavishly illustrated articles. Recently introduced RCSB.org and PDB101.RCSB.org features intended to enable the next generation of PDB data explorers will be presented.

RCSB PDB is a founding member of the Worldwide Protein Data Bank (wwpdb.org). 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 R01GM3198.