RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education

Christine Zardecki1,2, S.K. Burley1,2,3,4,5

1RCSB Protein Data Bank, 2Institute for Quantitative Biomedicine, and 3Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
4Cancer Institute of New Jersey, Rutgers, The State University of New Jersey, New Brunswick, NJ 08903, USA
5RCSB Protein Data Bank, San Diego Supercomputer Center, University of California San Diego, La Jolla, CA 92093, USA

Christine.zardecki@rcsb.org

Keywords: Protein Data Bank, biomolecular structure, database design,

The Protein Data Bank (PDB) was established as the 1st open-access digital data resource in biology and medicine in 1971 with seven X-ray crystal structures of proteins. Today, the PDB houses ~200,000 experimentally-determined, atomic-level, three-dimensional (3D) structures of proteins and nucleic acids and their complexes with one another and small molecules (e.g., approved drugs, enzyme cofactors). The US-funded Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) and other members of the Worldwide Protein Data Bank (wwPDB) partnership jointly manage the PDB archive and support >55,000 data depositors worldwide, ensuring quality for the ever-growing body of experimentally-determined 3D biostructure information. Within the wwPDB, data processing responsibilities are distributed geographically. As the US wwPDB Data Center, RCSB PDB rigorously-validates and expertly biocurates new structures coming into the archive from the Americas and Oceania. RCSB PDB also serves as the wwPDB-designated PDB Archive Keeper, responsible for safeguarding ~1 Tb of data (amassed during five decades of continuous operations) and updating the PDB on a weekly basis.

The RCSB PDB research-focused web portal (RCSB.org) supports many millions of users worldwide, representing a broad range of expertise and interests. In addition to retrieving 3D structure data at no charge and with no usage limitations, PDB data consumers access comparative data, and external annotations, such as information about point mutations and genetic variations. Alongside ~200,000 PDB structures, RCSB.org now provides access to >1,000,000 Computed Structure Models (CSMs) generated using artificial intelligence/machine learning methods. For the avoidance of doubt, experimentally-determined PDB structures and CSMs are clearly identified as to provenance and reliability.

Specialized RCSB.org data delivery tools include:

- Structure Motif Search: Finds structures with a small number of specific amino acids in specific 3D configurations.
- Pairwise Structure Alignment calculates alignments using different trusted methods and displays sequence alignments and superposed 3D visualization. Comparisons can be made for any protein in the PDB archive and/or uploaded data files including computed structure models.
- For result sets containing multiple structures representing highly similar proteins, a new Grouping Option generates a non-redundant search result set based on sequence identity, UniProt ID, or Group depositions.
- Ligand structure quality assessment metrics are available for PDB structures determined by X-ray crystallography. Quality indicators (e.g., RSR, RSCC) have been aggregated into a ranking score that can be used for comparison across the archive and allow any user to quickly review ligand structure quality.

Additionally, Training and outreach materials are hosted at PDB101.rcsb.org to help users learn how to utilize PDB data and tell structural biology stories. Guide to Understanding PDB Data covers topics such as PDBx/mmCIF format; exploring carbohydrates in the PDB; and Computed Structure Models. The Molecule of the Month series introduces the structure and function of interesting and notable PDB structures, such as Respiratory Supercomplexes and SARS-CoV-2 structures.

RCSB PDB Core Operations are funded by National Science Foundation (DBI-1832184), US Department of Energy (DE-SC0019749), and 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.