Title: RCSB Protein Data Bank: Connecting genes to structures to ecosystems

Authors: John D. Westbrook[1], Jasmine Young,[1] Robert Lowe,[1] Christine Zardecki,[1] Jose Duarte,[2] and **Stephen K. Burley***[1,2] (Stephen.Burley@rcsb.org)

Institutions: [1]Rutgers, The State University of New Jersey, Piscataway, NJ; [2]University of California San Diego, La Jolla, CA

Website URL: http://www.rcsb.org

Project Goals: The Vision of the RCSB PDB is to enable open access to the accumulating knowledge of 3D structure, function, and evolution of biological macromolecules, expanding the frontiers of fundamental biology, biomedicine, and biotechnology.

Abstract

Protein Data Bank (PDB) was established as the 1st open access digital data resource in biology and medicine in 1971. Today, the PDB is one of two global resources for experimental data central to biological science as a public good (the other key Primary Data Archive being the International Nucleotide Sequence Database Collaboration). PDB currently houses >185,000 atomic level biomolecular structures determined by macromolecule crystallography, NMR spectroscopy, and 3D electron microscopy. It is managed by the Worldwide Protein Data Bank partnership (wwPDB; wwpdb.org) according to the FAIR principles (*i.e.,* Findability, Accessibility, Interoperability, and Reusability).

Through an internet information portal and downloadable data archive, researchers and educators can access 3D structure data for large biological molecules, such as proteins, DNA, and RNA. These are the molecules of life, found in all organisms on the planet. Knowing the 3D structure or shape of a biological macromolecule is essential for understanding the role the molecule plays in health and disease of humans, animals, and plants, food and energy production, and other topics of concern to global prosperity and sustainability.

The RCSB PDB (rcsb.org) operates the US data center for the wwPDB, serves as Archive Keeper for the global PDB Core Archive, and makes PDB data available at no charge to all data consumers worldwide with no limitations on usage. Studies of website usage, bibliometrics, and economics demonstrate the powerful impact of the PDB data on basic and applied research, clinical medicine, education, and the US economy.

During calendar 2021, >1,320 million structure data files were downloaded from the RCSB PDB by Data Consumers working worldwide. During this same period, the RCSB PDB processed >5,680 new atomic level biomolecular structures plus experimental data and metadata coming into the archive from Data Depositors working in the Americas and Oceania. In addition, RCSB PDB served many millions RCSB.org users worldwide with PDB data integrated with >50 external data

resources providing rich structural views of fundamental biology, biomedicine, and energy sciences, and supported >690,000 PDB101.rcsb.org educational website users around the globe.

RCSB PDB provides a rich collection application programmable interfaces (APIs) that enable searching of and access to PDB archival data content. These APIs enable search by key citation, biological and structural features, and retrieval of individual PDB structure entries, reports, and chemical and molecular reference data. Building on this functionality in 2022, RCSB PDB will deploy new API services that enable parallel delivery of both experimental structures and computed structure models (coming from AlphaFold2, RoseTTAFold, ModelArchive) fully integrated with sequence information and functional annotations spanning all features of the RCSB.org research focused web portal. Considerable emphasis will be placed on delivering computed structure models across entire proteomes of interest to the US Department of Energy. 3D structure data (both experimental from the PDB and computed structure models from AlphaFold2, RoseTTAFold, ModelArchive) will also be made available for delivery through the KBase portal (KBase.us).

Access to PDB data and services contribute to patent applications, drug discovery and development, publication of scientific studies, innovations that can lead to new product development and company formation, and STEM education.

References/Publications

1. Burley et al. (2021) RCSB Protein Data Bank: Powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering, and energy sciences. *Nucleic Acids Research 49*, D437-D451. DOI: 10.1093/nar/gkaa1038
2. Rose et al. (2021) RCSB Protein Data Bank: Architectural advances towards integrated searching and efficient access to macromolecular structure data from the PDB archive. *Journal of Molecular Biology 433*, 166704. DOI: 10.1016/j.jmb.2020.11.003
3. Burley et al. (2021) Proteome-scale Protein Structure Prediction with Artificial Intelligence. *New England Journal of Medicine 385*, 2191-2194. DOI: 10.1056/NEJMcibr2113027
4. Goodsell and Burley (2021) RCSB Protein Data Bank Resources for Structure-facilitated Design of mRNA Vaccines for Existing and Emerging Viral Pathogens. Structure. DOI: 10.1016/j.str.2021.10.008.

Funding Statement: *RCSB PDB is funded grants to SKB from 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. Additional funding has been provided by a KBase subcontract from Argonne National Laboratory (P.I.: Chris Henry).*