RCSB PDB Next-generation Data Delivery and Search Services

C Zardecki[1,3], J Duarte[2], C Bi[2], C Bhikadiya[1, 3], S Bittrich[2], L Chen[1], D Guzenko[2], R Lowe[3], J Segura[2], Y Valasatava[2], J Westbrook[3], S Burley[1,2,3]

[1]Rutgers Proteomics, Piscataway, NJ, [2]Protein Data Bank, CA, [3]RCSB Protein Data Bank, NJ

zardecki@gmail.com

RCSB Protein Data Bank (PDB) provides tools for analysis and visualization of 3D structures of biological macromolecules stored in the PDB archive. Recently-introduced Search and Data Delivery APIs offer comprehensive functionality and high performance at RCSB.org. The new services represent a complete overhaul of the software/data management architecture, transforming a monolithic application into a micro-service-oriented and cloud-ready resource. The data model is based on the PDBx/mmCIF dictionary (http://mmcif.wwpdb.org/) with extensions that facilitate usage and delivery for the RCSB PDB website and web services. For Data delivery (https://data.rcsb.org), a GraphQL interface allows arbitrary retrieval of data across the entire data model. To the best of our knowledge, this represents a first in Structural Bioinformatics. Search services (https://search.rcsb.org) are supported by a powerful Search API with a JSON-based Domain Specific Language (DSL). Arbitrary boolean logic search is now possible across all fields available in our data model. Importantly, a search aggregator layer seamlessly combines text searches from the Elasticsearch engine with specialized bioinformatics algorithms that perform searches against macromolecular sequence and/or atomic coordinate data. Examples of the searches integrated by the aggregator are mmseqs2 sequence search (1), BioZernike structure shape search (2), and sequence motif search. Users of existing services are strongly encouraged to migrate to the new APIs before November 2020, when legacy RCSB PDB APIs (REST search and fetch) will be discontinued. 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. 1) Mirdita M, Steinegger M and Soeding J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. Bioinformatics, doi: 10.1093/bioinformatics/bty1057 (2019). Dmytro Guzenko, Stephen K. Burley, Jose M. Duarte. Real time structural search of the Protein Data Bank (2020) bioRxiv doi: https://doi.org/10.1101/845123 2) Dmytro Guzenko, Stephen K. Burley, Jose M. Duarte. Real time structural search of the Protein Data Bank (2020) bioRxiv doi: https://doi.org/10.1101/845123

Figure 1