PDBe-KB: An integrated knowledge base of molecular structures and their functional annotations that support large scale data mining

Mihaly Varadi, *PDBe-KB consortium and EMBL/EBI, Hinxton, UK*

Abstract

The Protein Data Bank in Europe - Knowledge Base (PDBe-KB, https://pdbe-kb.org) is a community-driven, collaborative data resource that provides manually curated and computationally derived structural and functional annotations for proteins and small molecules (1). These annotations include catalytic sites, ligand binding sites, protein flexibility propensities, post-translational modification sites, and the effects of genetic variability or mutations, to name a few. The main motivations for integrating these data into PDBe-KB are (i) to increase the visibility and reduce the fragmentation of annotations contributed by specialist data resources, (ii) to make these data more findable, accessible, interoperable and reusable (FAIR) and (iii) to place macromolecular structure data in their biological context, thus facilitating their use by the broader scientific community in fundamental and applied research.

PDBe-KB is an open consortium that currently includes 30 data resources from 11 countries. We link their annotations to PDB structure data on the chain and residue levels in a distributable graph database (https://www.ebi.ac.uk/pdbe/pdbe-kb/graph-download). This graph, implemented in Neo4j, contains over 1 billion nodes and 4 billion edges, capturing some of the complexity of the biological context of proteins and their interacting small molecule partners. We update the database weekly, in parallel with the weekly PDB releases.

Scientists can also access data programmatically through an API (https://pdbe-kb.org/graph-api), in addition to using the graph database as a stand-alone research tool (2).

Thus, researchers can access a comprehensive set of structural and functional annotations integrated with core PDB data through direct database access, API endpoints and protein-focused web pages. These data access mechanisms can support high-throughput data mining to facilitate and potentially automate research and structure-based drug discovery pipelines.

References:

1) PDBe-KB consortium, Varadi, M., Berrisford, J., Deshpande, M., Nair, S.S., Gutmanas, A., Armstrong, D., Pravda, L., Al-Lazikani, B., Anyango, S., et al. (2020) PDBe-KB: a community-driven resource for structural and functional annotations. Nucleic Acids Res., 48, D344–D353.

2) Nair,S., Váradi,M., Nadzirin,N., Pravda,L., Anyango,S., Mir,S., Berrisford,J., Armstrong,D., Gutmanas,A. and Velankar,S. (2021) PDBe aggregated API: programmatic access to an integrative knowledge graph of molecular structure data. Bioinformatics, 10.1093/bioinformatics/btab424.