PDB-CAT: Classification and Analysis Tool for PDB files

Ariadna Llop-Peiró[1], Gerard Pujadas[1], Santiago Garcia-Vallvé[1]

[1]Departament de Bioquímica i Biotecnologia, Universitat Rovira i Virgili, Research group in Cheminformatics & Nutrition, 43007 Tarragona, Catalonia, Spain.,

Protein-ligand docking play a crucial role in identifying potential antiviral candidates. The initial step in protein-ligand is to search, if any, for crystallized structures of the therapeutic target. The Protein Data Bank is the public database for experimentally determined protein structures and its data continues to increase daily. To streamline the classification of PDB files, we introduce PDB-CAT, a Jupyter Notebook that automates the categorization of PDBx/mmCIF structure files based on protein-ligand bond type.

PDB-CAT efficiently navigates the extensive Protein Data Bank, through rapid classification of PDB files into different folders classifying them into ligand-free, covalent complexes, and non-covalent complexes. It operates transparently, providing access to all information related with each structure and ligand in a comprehensive CSV output. Furthermore, mutation analysis can be performed prior to classification.

PDB-CAT has been validated using the PDBBind v.2020 dataset [1], successfully classifying all complex structures to identify their ligands and gathering information about the protein, and its ligands, as well as any non-standard residues bonded to it. The validation was conducted on over 19,000 complexes, executed in just 20 minutes. PDB-CAT was also tested with SARS-CoV-2 main protease, 5R7Y protein structure was select as reference sequence to perform the mutation analysis. Among the 1,335 PDBx/mmCIF files found in Protein Data Bank, 1,128 were non-mutated, categorized into ligand-free (100), covalent complexes (399), and non-covalent complexes (629). The output CSV file also compiled essential data of the analysis of 207 mutations, including mutated residue details, identity percentage, and sequence gaps compared to the reference sequence.

This work was supported by the project PID2022-138327OB-I00 financed by MCIN/AEI/10.13039/501100011033/FEDER, UE and Martí Franquès/INVESTIGO grant 2022PMF-INV-14.

Bibliography :

[1] Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., Nie, W., Liu, Y., & Wang, R. (2015). PDB-wide collection of binding data: current status of the PDBbind database. Bioinformatics (Oxford, England), 31(3), 405–412. https://doi.org/10.1093/bioinformatics/btu626