PDB-CAT: Classification and Analysis Tool for PDBx/mmCIF

UNIVERSITAT ROVIRA i VIRGILI
CHEMINFORMATICS & NUTRITION

Ariadna Llop-Peiró, Gerard Pujadas, Santiago Garcia-Vallvé
ariadna.llop@urv.cat, gerard.pujadas@gmail.com, santi.garcia-vallve@urv.cat

PDB - CAT
PDB CLASSIFICATION AND ANALYSIS TOOL

Introduction

The increment of structure data presents challenges for certain applications, particularly in the context of virtual screening setup, where the classification based on interactions within ligand-protein complexes is required. However, sorting through large numbers of PDB files to find the desired structures can be a time-consuming task. **PDB-CAT** is a program that categorizes a group of protein structures into three types: ligand-free, covalently bonded, and non-covalently bonded. The input comprises a dataset of PDB files in PDBx/mmCIF format (which will be mandatory for PDB files in the short term). The output consists of a CSV file with structural information and a set of folders where the input PDB files are classified. **PDB-CAT** is easy to use, offering two different running options: (1) classification and (2) classification with mutation detection.

Methods

This program requires the following packages: *biopython*, *pdbecif*, *pandas*, *re*, *os,* and *shutil,* and uses Python 3.10.9.

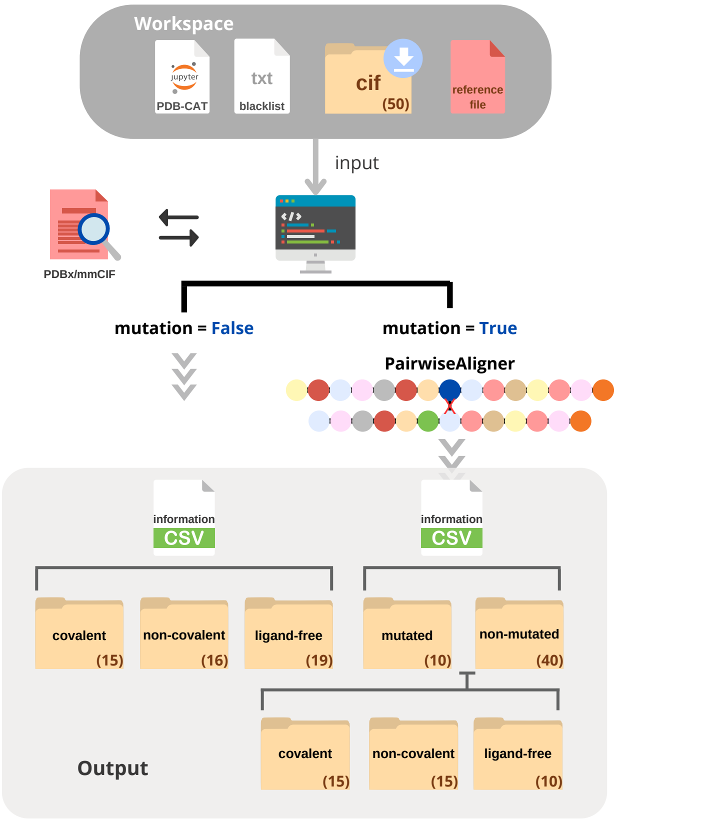
Results

The PDBx/mmCIF files underwent thorough analysis and mutation categorization. Out of the 50 M-pro structures downloaded from the PDB, **40** were identified as **non-mutated**. These were further classified into **10 ligand-free** structures, **12 covalent** complexes, and **18 non-covalent** complexes. Therefore, **10 mutations** were analyzed, extracting information about the exact mutated residue, the identity percentage, and the gaps in the sequence compared to the reference sequence.

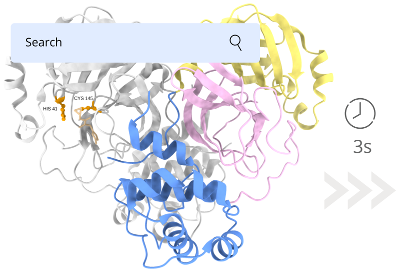
Also, a CSV was created to collect all crucial information.

Aim

PDB-CAT is a Jupyter Notebook that aims to simplify this process by automatically categorizing the structures based on the type of interaction between atoms in the protein and the ligand, and additionally checking for any mutations in the sequence.



PDB-CAT Flowchart with M-Pro Example



SARS-CoV-2 Main Protease

M-pro example CSV Output Overview

PDB_ID	Chain_ID	Num_Res	Complex	Ligand	Peptide_Like	Covalent_Bond	Mutation	Location	Identity	Gaps
5R7Y	A	306	Yes	JFM	No	No	0		100.00	0
6LU7	A	306	Yes	PRD_002214	Yes	Yes	0		100.00	0
7TQ5	A,B	309	Yes	IRW ITX	No	Yes	0		98.39	5
8DRR	A,B,C	306	No		No	No	3	C145A, G302A, F305L	98.37	2

PDB-CAT presents a useful tool:

1. To classify PDBx/mmCIF structures into ligand-free structures, covalent, and non-covalent complexes
2. To detect mutations, and gaps between different structures of the same protein
3. To facilitate the virtual screening setup (e.g. choosing the best target structure to conduct a protein-ligand docking)
4. To contribute to the format transition from PDB to PDBx/mmCIF

Limitations

PDB-CAT's limitation lies in its dependence on the author's description in the PDBx/mmCIF files. The accuracy and effectiveness of the program are contingent on the information provided by the file's author.

Check our GitHub

This program is available at @URV-cheminformatics

Check our webpage

This work was supported by the project PID2022-138327OB-I00 financed by MCIN/AEI/10.13039/501100011033/FEDER, UE

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
AGENCIA ESTATAL DE INVESTIGACIÓN
FEDER
Fondo Europeo de Desarrollo Regional
UNIÓN EUROPEA
"Una manera de hacer Europa"