Flexible searching of small molecule and protein structural data with the CSD Python API toolkit

Juliette Pradon, Ilenia Giangreco
The Cambridge Crystallographic Data Centre (CCDC), Cambridge, CB2 1EZ, UK; Email: pradon@ccdc.cam.ac.uk

Introduction

Successful modern drug discovery research makes extensive use of structural data – from target proteins, candidate drug molecules, and complexes of the two. The value of protein-ligand structural information is well accepted. In addition, knowledge of molecular conformations and interactions derived from small molecule structures alone can have a significant impact in drug discovery. The CSD Python API scripting interface has been developed to mine the world’s two most important structural databases, the Cambridge Structural Database (CSD) [1] and the Protein Data Bank (PDB) [2], and extract data driven insight, which can inform the design, development and identification of new and better pharmaceutical products.

Overview of search capabilities in the CSD Python API

1. **Text/numeric searching** (of data associated with CSD entries only)
2. **Similarity searching** (using the CSD fingerprint of ligands)
3. **Fast (sub)structure searching** (with support for atom constraints) of ligands.
   Substructure searching also supports geometric constraints to mine 3D intermolecular interaction patterns between any of: protein binding sites, waters, metals and small molecules, with support for nucleotides.

**Note:** these three search types can be combined using ‘And’, ‘Or’ and ‘Not’, and can be constrained using search filters, e.g. filtering by maximum R-factor, organic/organometallic, disorder, etc.

4. Sequence-independent **protein cavity & sub-pocket searching** methods to identify similar protein binding sites, useful in many areas of pharmaceutical drug design (e.g. biosensoric replacement, polypharmacology, off-target prediction). Three methods, varying on speed & accuracy:
   - fast cavity graph comparison (based on Local Cliques [3] algorithm)
   - cavity graph comparison (based on CavBase [4] algorithm)
   - cavity histograms comparison (based on RAPMAD [5] algorithm)

5. **Pharmacophore searching** of CSD & PDB data simultaneously, to mine for structural motifs that bind in similar environments and generate new ideas for ligand mimicking, scaffold hopping or growing into a sub-pocket.
   Note: optional protein/small molecule components and intra/intramolecular constraints on pharmacophore features to tailor searches; and ability to filter search results based on database annotations.

Example 1 – mining for ligand chemistry that binds in a specific protein environment

How to search for ligands that bind to a Zinc ion, where it coordinates a histidine and two cysteine residues, in PDB protein-ligand binding sites?
Write a complex 3D substructure query:

![Example 1](image1)

Output the superimposed hits as mol2 (note that NAD is classified as a co-factor) and the hit ligands as 2D diagrams, highlighting the Zinc-coordinating ligand atom:

![Example 1](image2)

Use these results to generate ideas of how to modify your compound to bind in this protein environment.

Example 2 – identifying the geometric interaction preferences for a protein-ligand interaction pair

How to search for benzamidine fragment in ligand interacting with carbonyl of a residue sidechain (with no specified amino acid type) in PDB protein-ligand binding sites? Write a complex 3D substructure query:

![Example 2](image3)

Plot the centroid-centroid distance and angle of interaction between the carboxylic group of the benzamidine and the benzamidine fragment, and access the protein name (via EC_number) where interaction is seen:

![Example 2](image4)

Observe marked interaction preference for these two functional groups, with peak of distance at ~2.8 Å and angle at ~60°, in the first 500 hits.

Example 3 – identifying similar protein environments and their existing ligand binders

How to detect aromatic cages in the PDB database and identify what ligand functionality binds inside these pockets? Write a pharmacophore query from the putrescine receptor binding site, selecting three features:

Access the annotations to find out in what other protein classes do such aromatic cages occur, with hits in many unrelated proteins such as:

Can then identify existing ligand chemistry that binds in such aromatic cages by adding a small molecule heavy atom feature.

References


CSD Python API: downloads.ccdc.cam.ac.uk/documentation/API/index.html

Conclusions

The CSD Python API is essentially feature complete with regards to functionality exposed through all CCDC software components for Discovery and Materials users. This includes API access to: protein-ligand docking, virtual screening of a compound library against a pharmacophore query obtained from one or multiple overlaid ligands; generation of interaction maps around small molecules or within a protein binding site; conformer generation and molecular minimisation; calculation of probabilities for the formation of H-bonds; calculation of a crystal packing similarity metric; molecular geometry analysis... Please let us know what else you would like to see exposed in the CSD Python API!

www.ccdc.cam.ac.uk

+44 (0)1223 336408