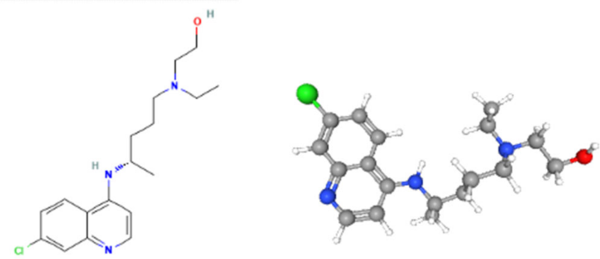
CM5239 Tutorial 2 – Chemical Databases and Pubchem Searches

(A) Access 5 popular chemical databases

1. PubChem (https://pubchem.ncbi.nlm.nih.gov)
2. ZINC (http://zinc.docking.org)
3. ChEMBL (https://www.ebi.ac.uk/chembl)
4. Chemspider (https://chemspider.com/)
5. Protein Data Bank (https://rcsb.org/)

The molecule to search through these databases (except for PDB) is hydroxylchloroquine (a possible drug for COVID-19 virus).



Answer the following questions:

(1) PubChem: Search for (S)-hydroxychloroquine. What is the canonical SMILES string?

(2) PubChem: What is the LogP value? Is this molecule hydrophilic or hydrophobic?

(3) PubChem: Find literature on coronavirus studies.

(4) ZINC: Copy the SMILES string from PubChem to perform the search under "Substance)" How many vendors?

(5) ZINC: How many clinical trials record available for this molecule?

(6) ChEMBL: How many targets are currently available in ChEMBL?

(7) ChEMBL: What are the reported drug mechanisms of hydroxychloroquine?

(8) Chemspider: Draw hydroxylchloroquine and perform search. How many similar structures with same skeleton can be found (under "Searches")?

(9) Chemspider: Find the toxicity profile and toxicity mechanism via Toxin-Target Database T3D3512 (link within Chemspider – Properties - Toxicity).

(10) PDB: Search protein 1AKE. What is the ligand molecule of the ligand-protein complex?

(11) PDB: Download the PDB textfile (i.e. 3D structure of 1AKE). How many protein atoms? How many solvent atoms? [non-hydrogen]

ChemSpider, managed by the Royal Society of Chemistry (RSC), is a free chemical structure database providing fast text and structure search access to over 100 million structures, consolidating information such as chemical structures, properties, and spectra from various internet sources.

ZINC is a free online database providing a diverse collection of commercially available chemical compounds for virtual screening in drug discovery, offering researchers information on chemical structures, suppliers, and pricing to aid in computational studies and early-stage drug design.

ChEMBL is a bioinformatics database that catalogues information on the bioactivity of small molecules, their interactions with biological targets, and therapeutic applications, serving as a valuable resource for drug discovery and chemical biology research.

(b) PubChem Similarity and Substructure Searches

This tutorial describes various searches that can be performed in PubChem. Currently PubChem has three different search interfaces:

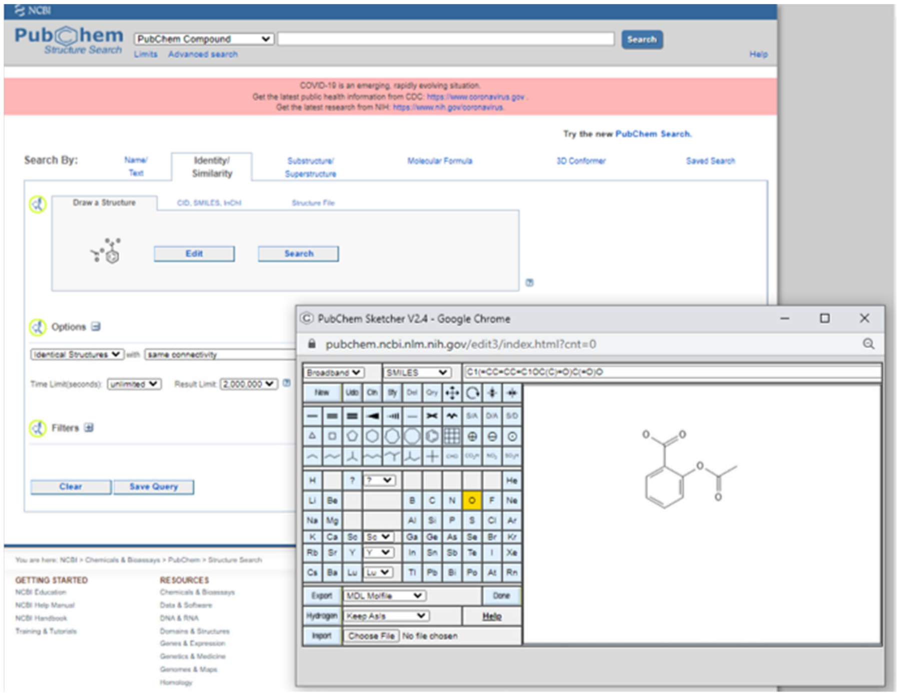
1. PubChem homepage (http://pubchem.ncbi.nlm.nih.gov)
2. PubChem Chemical Structure Search (https://pubchem.ncbi.nlm.nih.gov/search/search.cgi)
3. PubChem Search (https://pubchem.ncbi.nlm.nih.gov/search/).

The PubChem homepage provides a search interface for all three primary databases (Substance, Compound, and BioAssay). However, the search box on the PubChem homepage can accepts textual keywords only, and it is difficult to input non-textual queries (such as chemical structures). The PubChem Chemical Structure Search allows users to perform various searches using both textual and non-textual queries. This search interface is integrated with PubChem Sketcher, which enables users to provide the 2-D structure of a molecule as a query for chemical structure search. While the PubChem Chemical Structure Search is limited to search for chemical structures, the PubChem Search allows users to search for bioassays, bioactivities, patents, and targets as well as chemical structures. In this tutorial, you will use the Chemical Structure Search for chemical structure search.

(i) Identity and Similarity Search

Identity search is to locate a particular chemical structure that is "identical" to the query chemical structure. Although identity search seems conceptually straightforward, one should keep in mind that the word "identical" can have different notions, e.g. (i) multiple tautomeric forms in equilibrium, (ii) molecule has a chiral stereo center (*R*- and *S*-forms), and (iii) isotopically substituted species. Depending on how to deal with these nuances of chemical structures, identical search will return different results. The identity search in the PubChem Chemical Structure Search allows users to choose a desired degree of "sameness" from several predefined options. To see these options, one need to expand the options section by clicking the "plus" button next to the "option" section heading.

To perform an identity search for aspirin, go to Chemical Structure Search Page (https://pubchem.ncbi.nlm.nih.gov/search/search.cgi) and select the "Identity/Similarity" tab and expand the Options sections by clicking the "plus" button next to the "Options". Select the "Identical Structures" with "same connectivity. Provide the 2-D structure of aspirin using the PubChem Sketcher (i.e., choose "Draw a Structure") as a query for chemical structure search. How many records are returned? Why there are different "identical" structures?



COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.

Try the new PubChem Search.

Identical Structures with same connectivity

Time Limit(seconds): unlimited Result Limit: 2,000,000

pubchem.ncbi.nlm.nih.gov/edit3/index.html?cnt=0

Broadband SMILES C1(=CC=CC=C1OC(C)=O)C(=O)O

Export MDL Molfile

Hydrogen Keep As Is

Import Choose File No file chosen

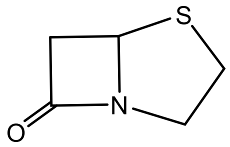
You are here: NCBI > Chemicals & Bioassays > PubChem > Structure Search

To perform a similarity search for aspirin, select the "Similar Structures" and "95%" from the top-down menu. In this search, use SMILES search (CC(=O)OC1=CC=CC=C1C(=O)O). Repeat the search with the following similarity search threshold: 90%, 85% and 80%. How many records are returned for each search? What is the method used by PubChem to compare 2-D similarity? How is similarity score quantified in PubChem?

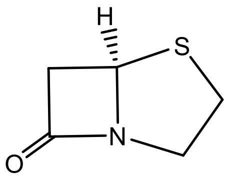
(ii) Substructure Search

To perform a substructure search for the core structure of penicillin (i.e. β-lactam ring + thiazolidine ring, see diagram below), go to Chemical Structure Search Page and select the "Substructure/Superstructure" tab and expand the Options sections by clicking the "plus" button next to the "Options". Select the Substructure" with match stereochemistry "Ignore". Provide the 2-D structure of aspirin's core structure (see structure (a) below) using the PubChem Sketcher (i.e., choose "Draw a Structure") as a query for chemical structure search. How many records are returned?

Perform a similar substructure search including the chiral carbon centre using the SMILES search (C1S[C@@H]2CC(=O)N2C1, see structure (b) below). Expand the Options sections by clicking the "plus" button next to the "Options". Select the "Substructure" with match stereochemistry "Exact". How many records are returned?



(a)



(b)