EMBO Practical Course: Computational Aspects of the Protein Target Selection, Protein Production Management and Structure Analysis Pipeline

Hands-on Practical:
RCSB PDB Chemical Component Tools

September 24, 2008

The following exercises provide examples using Ligand Expo and the RCSB PDB website to find, visualize and analyze information about the small molecules in the Protein Data Bank.

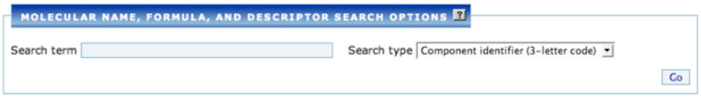
The Chemical Component Dictionary (formerly the HET Group Dictionary) is as an external reference file describing all residue and small molecule components found in PDB entries. This dictionary contains detailed chemical descriptions for standard and modified amino acids/nucleotides, small molecule ligands, and solvent molecules. Each chemical definition includes descriptions of chemical properties such as stereochemical assignments, aromatic bond assignments, idealized coordinates, chemical descriptors (SMILES & InChI), and systematic chemical names.

The chemical component dictionary is organized by the 3-character alphanumeric code that PDB assigns to each chemical component. New chemical component definitions appear in the dictionary as the entries in which they are observed are released in the PDB archive; consequently, the dictionary is updated with each weekly PDB release. Obsolete components remain in the dictionary marked with status OBS.

Users can search and browse the Chemical Component Dictionary using resources such as MSDchem and Ligand Expo.

The exercises below will explore the Ligand Expo site. The Ligand Expo *Search* page has three different query options, referred to as “first, second, third” in this document.

First Search Form:

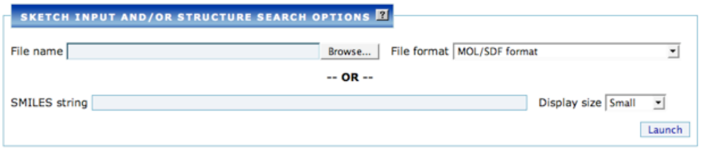


MOLECULAR NAME, FORMULA, AND DESCRIPTOR SEARCH OPTIONS

Search term

Search type Component identifier (3-letter code)

Second Search Form:



SKETCH INPUT AND/OR STRUCTURE SEARCH OPTIONS

File name

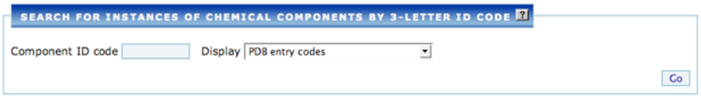
File format MOL/SDF format

-- OR --

SMILES string

Display size Small

Third Search Form:



SEARCH FOR INSTANCES OF CHEMICAL COMPONENTS BY 3-LETTER ID CODE

Component ID code

Display PDB entry codes

Exercise 1: Finding general information about a chemical component using Ligand Expo.

In this example we will explore the content of a wwPDB Chemical Component definition as presented in the Ligand Expo chemical component report. Go to the Ligand Expo search page (select *Search* from http://ligand-expo.rcsb.org/), and enter ATP (the three-letter code for adenosine 5'-triphosphate), in the first search form and select *Go*. This search will produce a report with following tabs: *Chemical Details, Geometry, Atom Nomenclature, Downloads,* and *Related Resources*. The following questions can typically be answered using the information in this report:

- **What is chemical structure and 3D structure of the molecule?**
  A simple 2D chemical sketch and a 3D Jmol depiction are presented on the top of the *Chemical Details* page. Dragging the mouse over the 3D display window while holding the left button depressed will adjust the view orientation. The right mouse button activates a variety of display features. For instance, to remove the atom labels in the 3D view, right click with the mouse and select: *Style – Labels – None*.

- **How many chiral centers are in the molecule?**
  The *Chemical Details* report contains essential chemical information about the molecule such as molecular names, synonyms, molecular formula, formula weight, and stereochemical features (including a list of chiral centers).

- **What are the SMILES and InChI strings for the molecule?**
  The *Chemical Details* report contains SMILES and InChI descriptors, which can then be Googled for more information.

- **What are representative bond distances, bond angles and torsion angles for this molecule?**
  The *Geometry* report tab contains tables of covalent geometry for a representative example molecule from the PDB and for a set of ideal coordinates computed using the Corina program.

- **How does the PDB label the atoms in the molecule?**
  The *Atom Nomenclature* report tab contains a table of the atom names used by PDB. Two sets of labels are presented. The column labeled V3 contains the labels that are in current use, and these replace the V2 labels were used prior to 2007. The *Download* tab also contains 2D depictions with a variety of labeling options.

- **Where can I find more about the biological activity of the molecule?**
  The *Related Resource* tab contains a table of links to sites with additional information about the molecule. In the case of ATP, additional information on biological activity may be found in links to DrugBank, PubChem, ChEBI, KEGG, and BindingDB. For instance, follow the link for DrugBank identifier DB00171 to the *Pharmacology* and *Mechanisms of Action* sections.

- **Where can I purchase a sample of the molecule?**
  From the *Related Resources* report, select the link to search eMolecules by SMILES string. From the eMolecules report page, follow one of the links to commercial suppliers.

- **In what electronic formats is this chemical and structural information downloadable?**
  From the *Downloads* tab the example and ideal 3D coordinates may be downloaded in PDB and SDF/MOL formats. The wwPDB Chemical Component definition in mmCIF format is also provided.

Exercise 2: Using the search and browse features of Ligand Expo and RCSB PDB.

This example illustrates a variety of search features available from Ligand Expo search form, http://ligand-expo.rcsb.org/ld-search.html. A general description of the Ligand Expo search system is available in the on-line help pages at http://ligand-expo.rcsb.org/help.html.
The following list contains some example queries along with some suggestions on how to perform each search.

- **How do can I search by a common name such as adenosine triphosphate?**
  Enter *adenosine triphosphate* in the first search form, select search type *Molecule Name (similar)*, and press *GO*. The resulting report will present the list of molecules with molecular names or synonyms lexicographically similar to the target name. To view the detailed chemical component report for any of the molecules in the report, select the *Chemical Details* link in the second column of the table.

- **How many molecules in the PDB contain both platinum and chlorine, nickel and iron, or only tin?**
  Enter the target partial chemical formula (e.g. Pt Cl) in the first search form, select search type *Formula (subset),* and press *GO.* This search will match any molecular formula containing the specified constituent elements. The search can be refined using the exact subset qualifier to restrict the search to a specified stoichiometry (e.g. Pt Cl2).

- **Which PDB entry has the highest resolution and contains ibuprofen? Which PDB entry containing ibuprofen was released first?**
  Enter *ibuprofen* in the first search form, select search type *Molecular Name (exact),* and press *GO*. In the *View Options* column of the resulting report, select *Coordinate Files*. This will produce a report of the PDB structures containing the target molecule sorted by resolution.

  Using the third search form, enter the 3-letter code for ibuprofen, *IBP*, and press *GO.* The resulting report will list the PDB entries containing this molecule sorted by release date. Note that the first report lists all instances, and the second report lists the unique PDB entries.

- **Which molecules in PDB contain fused six-membered rings?**
  One way to approach this problem is to sketch the ring scaffold using the Marvin Sketch search input tool. The sketched molecule can then used as a target for a sub-graph match search. For this example, press the *Launch* button in the second search form. Wait a few seconds for the sketch tool to load into your browser. Once loaded, locate the naphthalene ring template in the lower menu bar and drag this into the sketch window. Click the *arrow icon* in the right top menu bar to close the selection. To perform the search, select search type *Relaxed (heavy atom connectivity only)* in the Substructure Search form, and press the *Search* button. The resulting report will contain the molecules containing the fused 6-membered ring scaffold.

  Other common ring systems observed in PDB molecules can be viewed from the *Browse menu* under the *Ring Systems* tab.

- **Which PDB molecules match my SMILES string? Which PDB entries contain this molecule?**
  Enter a SMILES string (e.g. CCC(CC)O[C@@H]1C=C(C[C@@H]([C@H]1NC(=O)C)N)C(=O)O) in first search form, select search type *SMILES* or *Chemically similar to SMILES,* and press *GO*. In the *View Option* column of the resulting report, select *Coordinate Files*. This will produce a report containing each instance of the molecule in PDB. Notice that the molecule may appear multiple times within a single entry. The rightmost column of this report presents the options for downloading coordinates for each instance in PDB, mmCIF and SDF/MOL formats.

- **What modified amino acids and nucleotides are found in the PDB?**
  Ligand Expo provides a browser for modified amino acids and nucleotides. From the Ligand Expo home page select the *Browse* menu option. Tabs are provided for amino acids, nucleotides, popular pharmaceuticals and common ring systems. Select the *Amino Acid* tab and use the pull down menu to select the amino acid of interest. Press the *Browse* button to display a report of modified amino acids. The report will include protonation variants with extended identifier codes. For instance, the report for tryptophan will include TRP_LFZW-- the zwitterionic form of this amino acid.

- **What is the environment around the terminal carbonyl group of atorvastatin in PDB entry 1HWK?**
  From the Ligand Expo search page enter the 3-letter code for atorvastatin (*117*), in the third search form, select display *PDB entry codes + coordinates files*, and press *GO*. In the instance summary report, select one of the links to *Launch Viewer*. After a few moments, the AstexViewer applet window will display the 3D view of entry 1HWK and a window showing a 2D chemical depiction of neighboring residues, ligands and solvent. Click on the carbonyl portion of atorvastatin in the chemical diagram. This will re-center and zoom the 3D molecular diagram. Clicking on atoms in the 3D view will display atom and residue information. AstexViewer has a wide range of display and analysis options that are described at http://ligand-expo.rcsb.org/applets/astex/ViewerDocumentation.html.

Select the PDB ID *1HWK* either from the first column of the report menu, or enter it at http://www.pdb.org. Scroll down the resulting Structure Explorer main page to the Ligand Chemical Components section. Select the rightmost *View* option for atorvastatin, code *117,* to launch the Ligand Explorer application. After a view moments this application will be downloaded to your local system using Java Web Start. The Ligand Explorer window will display a cartoon representation of the protein molecule and a ball-and-stick representation of the ligand. To study the environment around one of the atorvastatin molecules select *Hydrophilic protein-ligand interactions* in the left menu and press the *Apply* button. This will highlight and label a variety of specific contacts.

Exercise 3: Dictionary and data file downloads from Ligand Expo.

Ligand Expo provides a variety of download options for chemical information about molecules in the PDB and associated coordinate data. This information is updated weekly after each PDB release.

- **The wwPDB Chemical Components dictionary is available in which formats?**
  The Download page provides links the Chemical Components dictionary in mmCIF and PDBML formats. The ideal coordinates in the chemical definitions may also be downloaded in a single SDF/MOL format file.

- **Why are there so many different SMILES strings?**
  SMILES are calculated using the OpenEye OECHEM library. The SMILES strings implement a canonical ordering algorithm and support stereochemical assignments. SMILES are also calculated using the CACTVS software system. Both software systems are widely used and are not algorithmically equivalent, particularly with respect to atom ordering. Tab delimited files are provided for SMILES and InChI descriptors for each PDB chemical component.

- **I want to compare the structures of all of the instances of coumarin in PDB. Is there an easy way to get these data files?**
  All ligands and non-standard amino acids and nucleotides instances are available for download in tar file bundles and in single data files. Data downloads are available in PDB, mmCIF, SDF/MOL and PDBML formats. More details on the organization of the file bundles are provided on the *Download* page.

- **What do the cryptic names for the coordinate files mean?**
  Each chemical component data files in Ligand Expo is assigned a file name that uniquely identifies the particular file within the PDB archive. The includes the PDB ID, model number, chain id, residue number, mmCIF asym_id, mmCIF sequence number, and a flag to indicate disorder. More details on the file name conventions are provided on the *Download* page.