# *Section III*

## **Cheminformatics - Basics:** *Searching for Chemicals on the Internet (PubChem, eMolecules)*

*Oleg Ursu, Dan C. Fara and Tudor I. Oprea*

Objective

Introduction to mining PubChem database for chemicals structures and biological activities associated with the chemical structures.

### 1. PubChem

PubChem provides information on the biological activities of small molecules. It is a component of NIH's Molecular Libraries Roadmap Initiative. PubChem includes substance information, compound structures, and bioactivity data in three primary databases, PCSubstance, PCCompound, and PCBioAssay, respectively.

- PCSubstance contains more than 38 million records. You can check the count of substance records as of today.
- PCCompound contains more than 18 million unique structures. You can check the count of compound records as of today.
- PCBioAssay contains more than 680 bioassays. Each bioassay contains a various number of data points. You can check the count of bioassay records as of today.

The Substance/Compound database, where possible, provides links to bioassay description, literature, references, and assay data points. The BioAssay database also includes links back to the Substance/Compound database. PubChem is integrated with Entrez, NCBI's primary search engine, and also provides compound neighboring, sub/superstructure, similarity structure, bioactivity data, and other searching features.

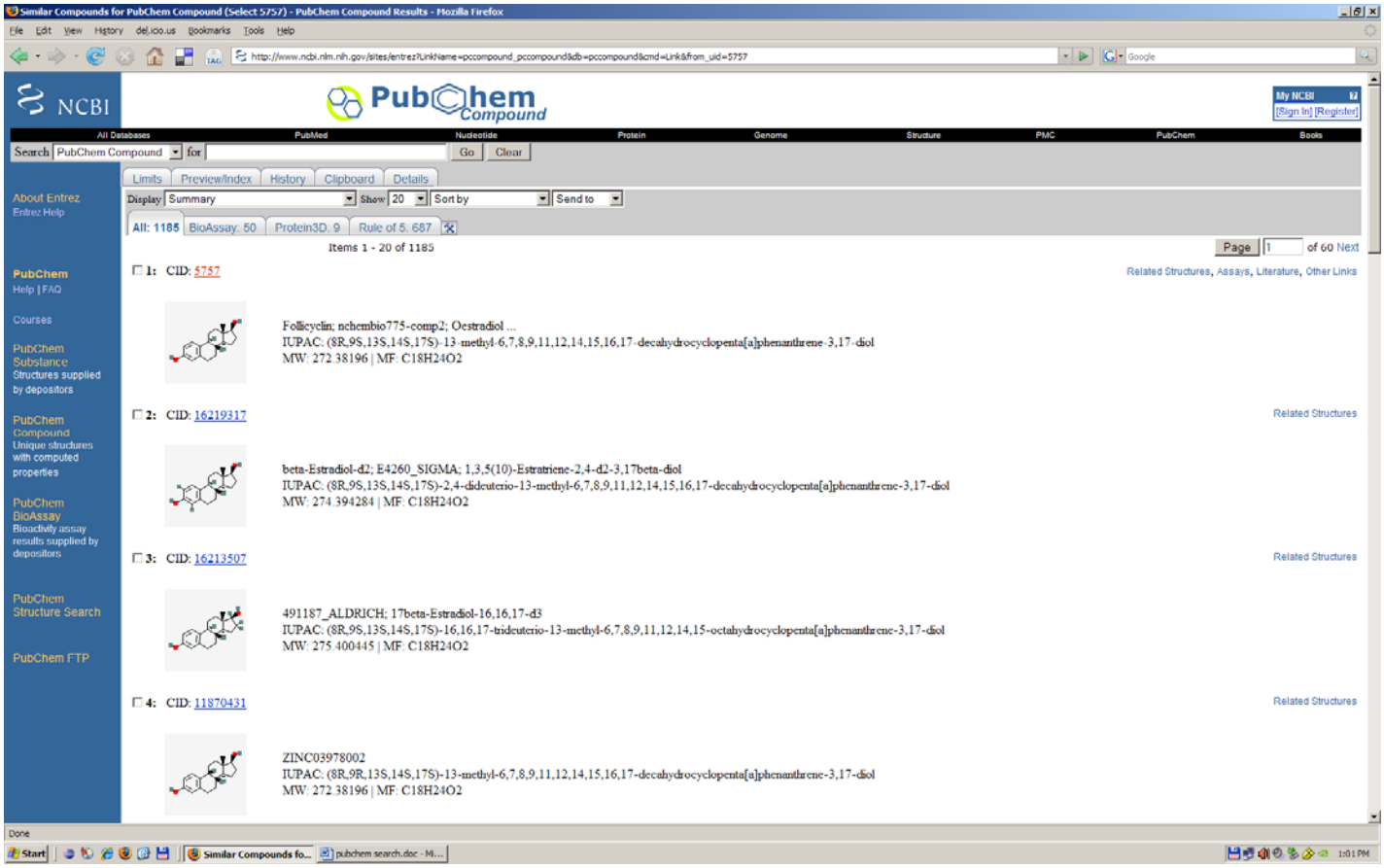
PubChem contains substance and bioassay information from a multitude of depositors. You can check the PubChem data source status as of today.

How to

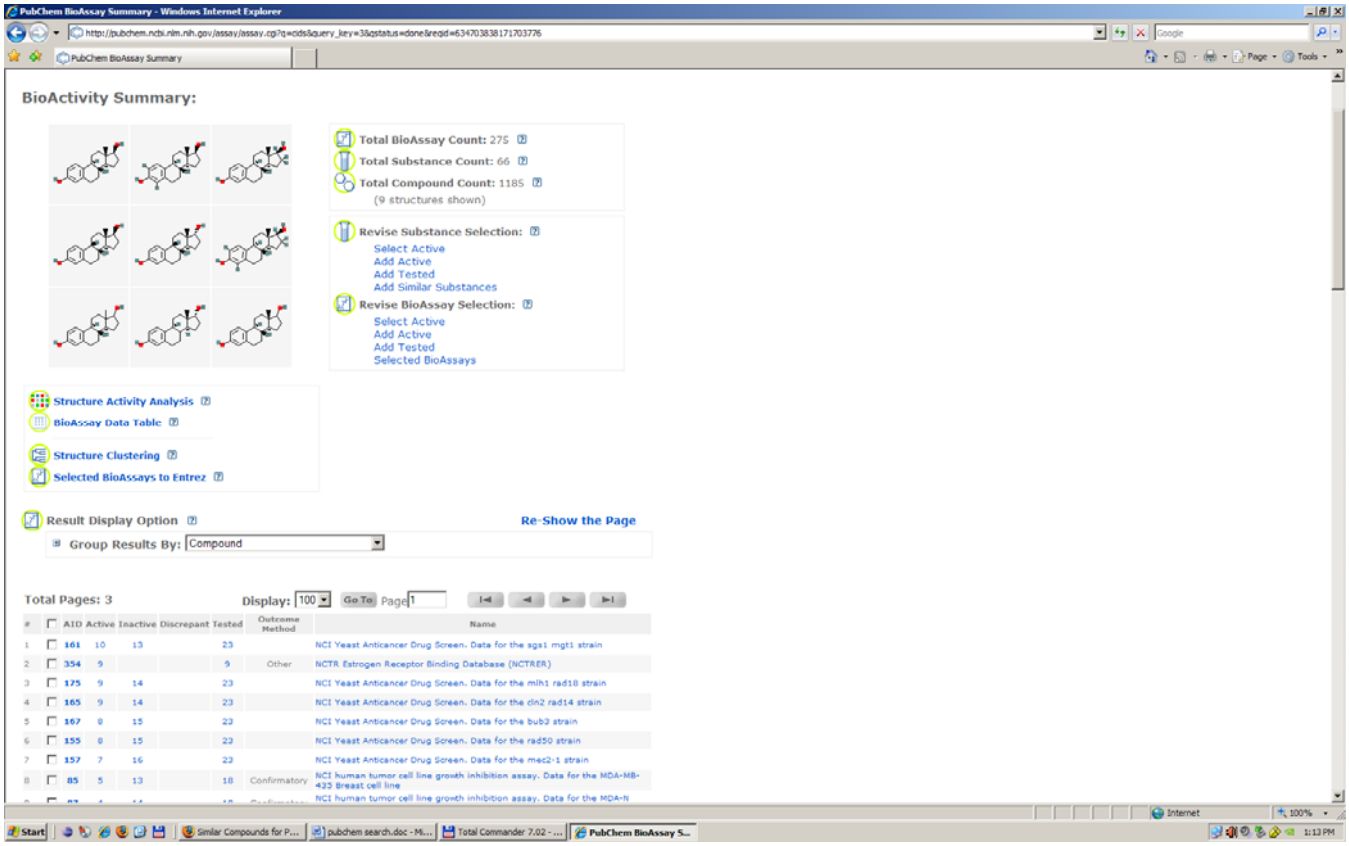
- Go to http://pubchem.ncbi.nlm.nih.gov/
- Under PubChem Text Search section select PubChem Compound
- Type beta-estradiol and click Go


- Click on the first CID (compound ID link 5757)

- The top of page opened contains the chemical structure of the Estradiol, Bioactivity Summary (assays in which the Estradiol was tested, found active, inactive and inconclusive), Protein Structures complexes with Estradiol, Substances associated with the chemical structure of Estradiol, Related Compounds and Similar compounds links to all tautomers, isomers, and similar chemical structures deposited in PubChem. For example, clicking on Similar Compounds will return a list of steroid like structures deposited in PubChem.

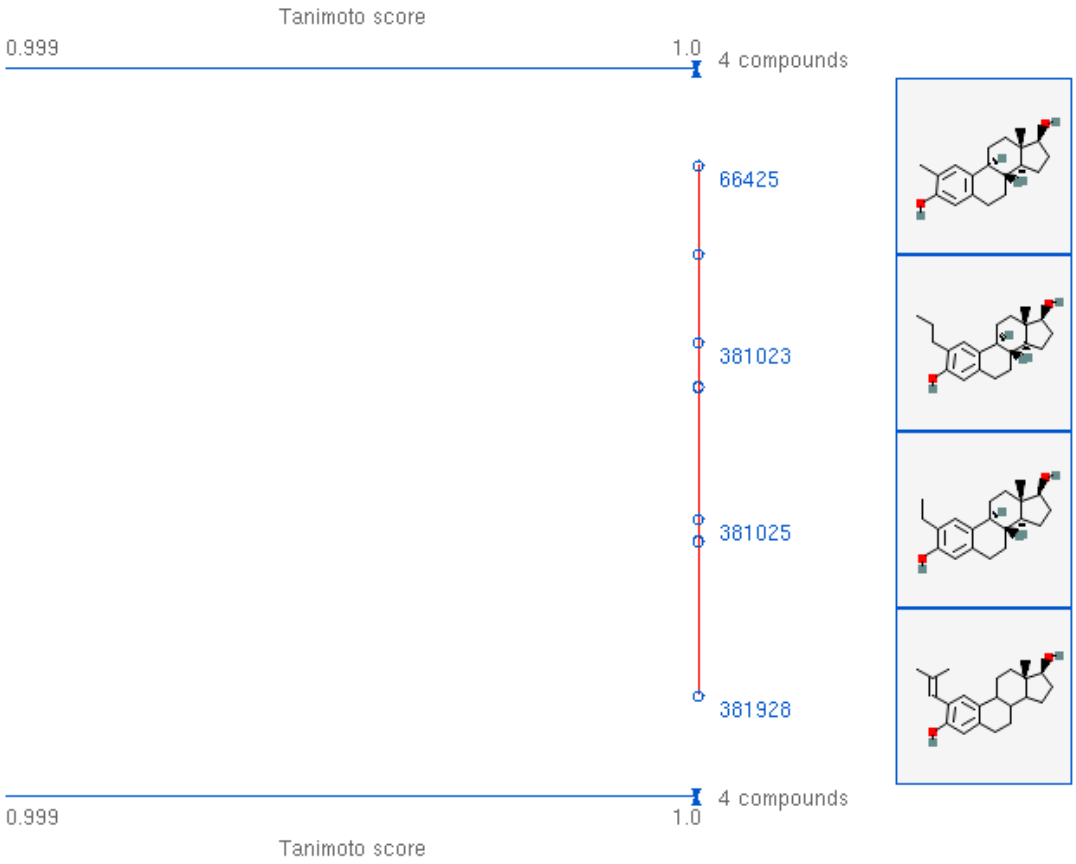



This list can be downloaded (select PubChem Download from display combo box) as SDF file which contains all the identification data as well as some chemical properties of the structures. It is also possible to take a quick look at all assays in which all of these structures were tested by selecting PubChemBioActivity Summary (see below).



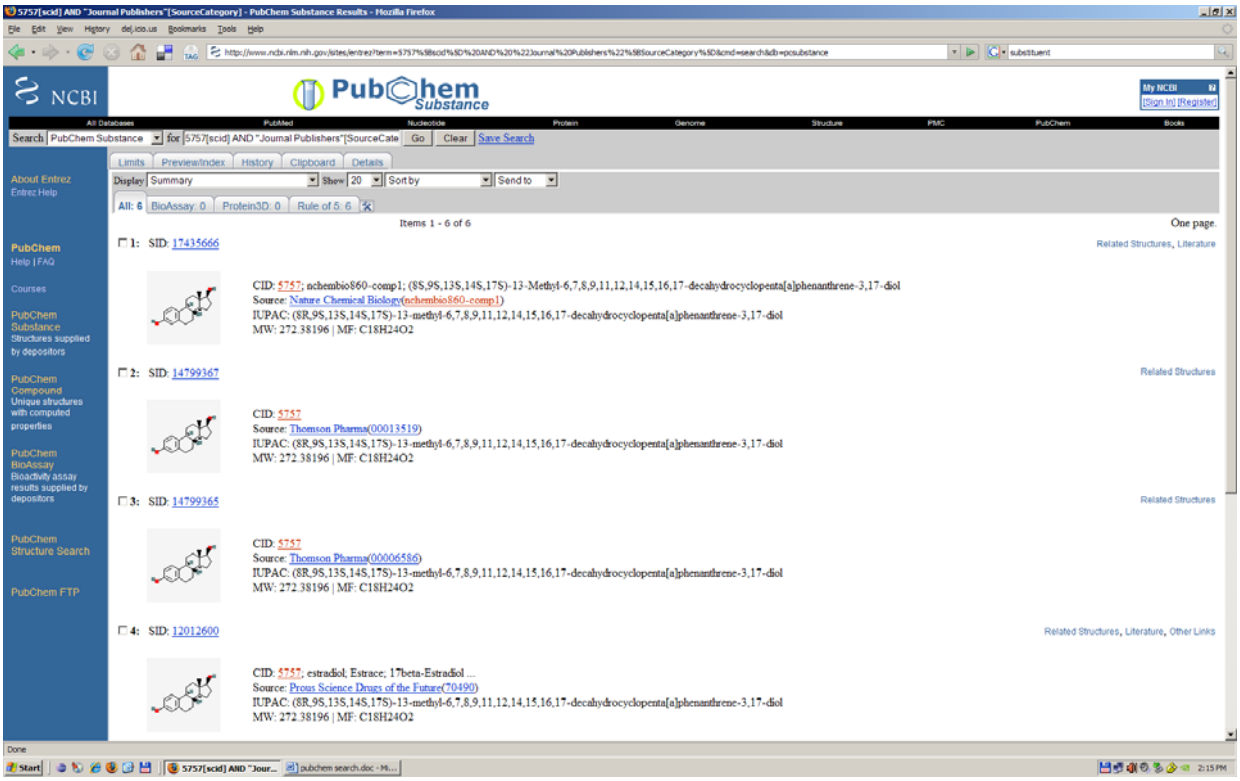

The page displays all assay in which chemical compounds similar to Estradiol were tested; assays in which any of the compounds were found to be active are listed first.

- Select only the assay in which at least one compound is active and then click Structure Activity Analysis. The page displayed contains summary information only for compounds and assays but one can easily notice that there are 4 compounds CIDs (66413, 5468270, 5468334, 388008) found active in multiple assays. The Compound Cluster tree displayed in the left area of the page is a useful tool for analyzing structure activity relationships. Click on the tree node of interest to display the compounds in a new web page.




All compounds in the cluster have a hydrophobic substituent at aromatic ring, 3 compounds are active in 2 assays, and 2 compounds are active in 4 assays.

- Return to the page were Estradiol is displayed (CID 5757) scroll to the bottom of the page and click Journal Publisher links.




- Click on the last link to open the paper published in Nature Chemical Biology.

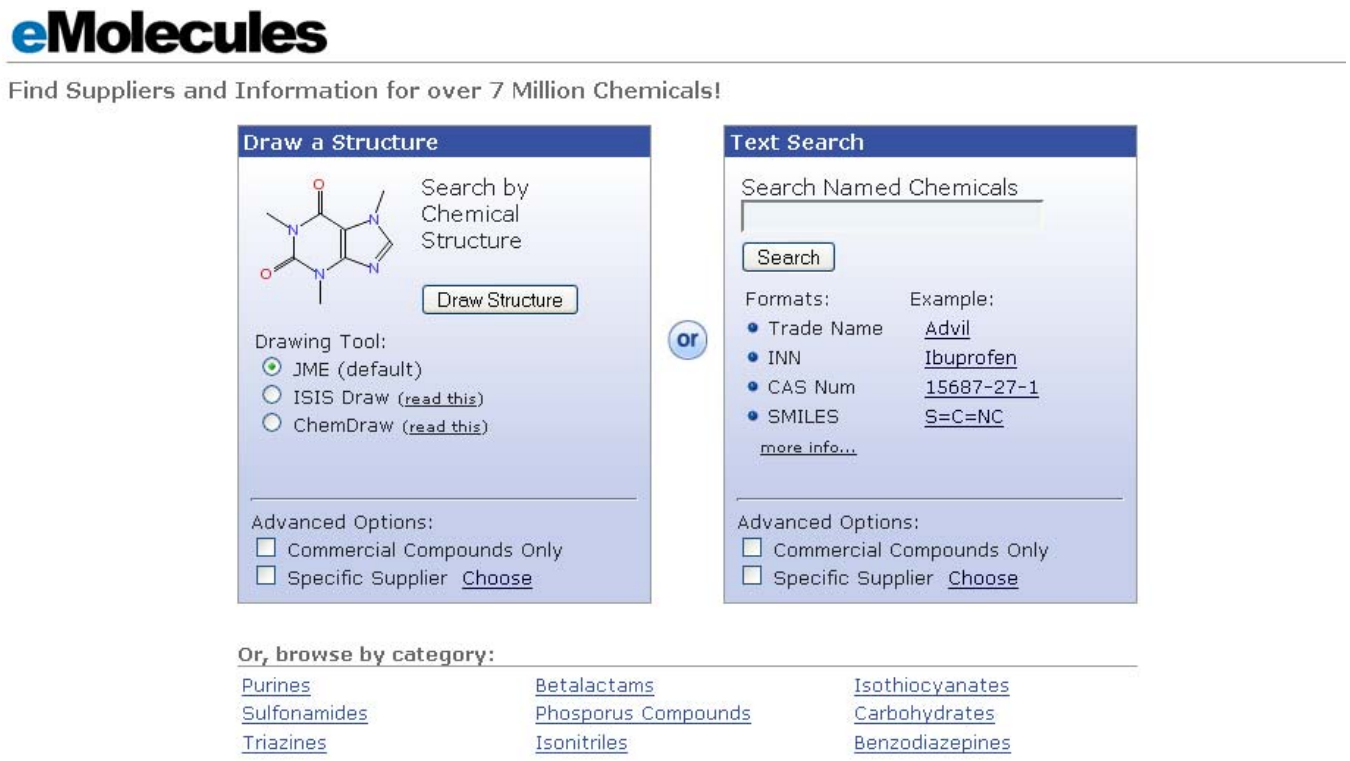

The paper describes how virtual screening can be applied to discover new compounds that bind selectively to GPR30.

**2. eMolecules**

eMolecules is one of the world's most comprehensive openly accessible search engine for chemical structures.

- The database contains around 8 million unique chemical structures from 19 million sources
- It allows unparalleled search speed—typically just seconds.
- The search can be done by drawing chemical structures or substructures using common industry tools — ISIS/Draw, ChemDraw, ChemSketch and JME.
- It has more than 4 million commercially available screening compounds and hundreds of thousands of building blocks and intermediates.
- It includes reference links to many prominent sources of public data for spectra, physical properties and biological data, including NIST WebBook, National Cancer Institute, DrugBank and PubChem.

The basic search can be done by (i) drawing chemical structures, and (ii) text, as can be seen below.




- Search the eMolecules database for beta estradiol
- Click on the PubChem Supplier's ID (10468968)

Are you re-directed to the same compound as before (CID 5757)? What is the difference between these two compounds?