3.4: Data Organization in PubChem as a Data Aggregator

PubChem Aggregator Overview

PubChem (https://pubchem.ncbi.nlm.nih.gov) is a data aggregator, meaning that it collects data from other data sources. As of February 2017, PubChem's data are from more than 500 organizations, including government agencies, university labs, pharmaceutical companies, substance vendors, and other databases. An up-to-date list of PubChem's data sources is available at the PubChem Sources page (https://pubchem.ncbi.nlm.nih.gov/sources). To better understand the features of this page, read this article on PubChem Blog:

Data Sources Page

(http://go.usa.gov/xk7xU)

PubChem organizes its data into three inter-linked databases: Substance, Compound, and BioAssay

(See **Table 1**), which can be searched from either the PubChem home page (https://pubchem.ncbi.nlm.nih.gov) or the web page of one of the three PubChem databases.

Table 1. Three inter-linked databases in PubChem.

Database	URL	Identifier
Substance	https://www.ncbi.nlm.nih.gov/pcsubstance	SID
Compound	https://www.ncbi.nlm.nih.gov/pccompound	CID
BioAssay	https://www.ncbi.nlm.nih.gov/pcassay	AID

Individual data contributors deposit information on chemical substances to the Substance database (https://www.ncbi.nlm.nih.gov/pcsubstance). Different data contributors may provide information on the same molecule, hence the same chemical structure may appear multiple times in the Substance database. To provide a non-redundant view, chemical structures in the Substance database are normalized through a process called "standardization" and the unique chemical structures are identified and stored in the Compound database (https://www.ncbi.nlm.nih.gov/pccompound). The difference between the Substance and Compound databases is explained in more detail in this blog post.

Compounds and Substances

What is the difference between a substance and a compound in PubChem?
(http://1.usa.gov/1nl9ePL)

Descriptions of biological experiments on chemical substances are stored in the BioAssay database (https://www.ncbi.nlm.nih.gov/pcassay). The unique identifiers used to locate records in these three databases are called SID (Substance ID), CID (Compound ID), and AID (Assay ID) for the Substance, Compound, and BioAssay databases, respectively.

All information in the Substance database is submitted by individual data depositors. However, the Compound database does contain information that are not submitted by data depositors, but annotated by the PubChem team. [In the context of scientific databases, annotation refers to the process of adding extra information to a database entry (for example, a compound in the Compound database and an assay in the BioAssay database)]. The annotated information is always presented with its provenance information (that is, the source of the information). The list of all the annotation sources used in PubChem is available at the PubChem Sources page (https://pubchem.ncbi.nlm.nih.gov/sources). From this page, one may download all the annotations from a particular source.