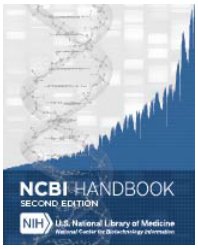
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: Morris R. Small Molecules and Biological Activities. 2013 Dec 9. In: The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013-.
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



Small Molecules and Biological Activities

Rana Morris[1]

Created: December 9, 2013.

Scope

The chemical and bioactivity resources at the National Center for Biotechnology Information (NCBI) are coordinated as part of the PubChem Project with data accessible as part of the PubChem Compound and Substance databases and PubChem BioAssay database, respectively. Currently the PubChem set of resources provides access to information about over 36 million chemical compounds and experimental studies with results for over 700 thousand biological activity assays.

History

The PubChem databases were originally developed to serve as a central data repository for the NIH-sponsored Molecular Libraries Roadmap Project (Pilot phase 2004–2008). The purpose of this project was to identify novel research reagents (protein function modulators and molecular probes) and discover candidate compounds for drug development. A large set of substances were tested for biological activity in high-throughput assays. The substances were cataloged in the PubChem Substance database and the description, supporting information, and results of the studies were stored in the PubChem BioAssay database. To increase the utility of these databases, a derivative database, PubChem Compound, was created by the PubChem group to consolidate and provide a cross-referencing platform for common chemical components of the substances.

Since its inception, the scope of these resources has been dramatically expanded for use by an expanding and diverse variety of researchers, from organic chemists to molecular biologists to drug design specialists. In addition to the data from the original NIH Molecular Libraries Initiative, substance information and biological activity assays are now submitted by various organizations:

- NIH Molecular Libraries participants and other laboratories with Bioassay Screening Results
- The NIH Substance Repository
- Organizations specializing in these areas:
  - Biological Properties
  - Chemical Reactions
  - Database Vendors
  - Imaging Agents
  - Journal Publishers
  - Metabolic Pathways
  - Patents

Author Affiliation: 1 NCBI; Email: morrisrc@ncbi.nlm.nih.gov.

Corresponding author.

- Physical Properties
- Protein 3D Structures
- siRNA Reagent Providers
- Substance Vendors
- Theoretical Properties
- Toxicology Properties

Dataflow

Data Submissions to PubChem

The PubChem Upload portal enables researchers, laboratories, and organizations to submit small molecule and bioassay data to the PubChem Substance and PubChem BioAssay databases. Data can be submitted to the PubChem Substance or BioAssay databases in any of the following ways:

- PubChem Submission Wizards—for small numbers of submissions, a series of guided forms assist novice submitters in entering substance and/or assay data without requiring knowledge of detailed data specifications. After data is typed or imported, the Upload system will prepare a properly formatted file that conforms to the data specifications.
- Pre-formatted File Uploads—for small or large numbers of records, Upload accepts pre-formatted files in several formats which may be GZIP-compressed. For PubChem Substances, formats include: SDF – Chemical Structure Data File, CSV – Comma-separated Variables, and for BioAssay formats include: ASN.1 – Abstract Syntax Notation 1, XML – Extensible Markup Language, CSV – Comma-separated Variables).
- FTP Depositions of Pre-formatted Files—for large and/or frequent data uploads, depositions of pre-formatted files by FTP are recommended. Once the data are in the FTP directory, submitters can review and edit them, and commit them to PubChem using the Upload system.

While the PubChem group and the user community-at-large recommend that submitters supply as much information about their substances and bioactivity assays as possible (including such things as chemical structures of tested substances and related PubMed records), the following are the minimal data that must be submitted:

To PubChem Substance:

- Submitter information
  - Name
  - Organization
  - Address
  - Contact Information
  - Source Classification
  - Signed Data Transfer Agreement
- Substance “name”
- Source’s identifier

To PubChem BioAssay:

- Submitter information (same as PubChem Substance Submitter information)
- Assay Information
  - Descriptive Assay title
  - Source’s identifier
  - Assay Information

- Assay type
- List of Substances tested
- Description
- Protocol
- Data Table with Defined Outcomes
- Definitions

The PubChem Upload system can also be used for updating existing PubChem records. Existing PubChem Substance or BioAssay records can be retrieved and loaded into the Web interface for direct editing and review of the revised records. Once corrections have been made and verified, they will be committed and stored in the system.

The PubChem databases have a rolling update schedule. As new data is submitted to PubChem, it is processed for addition to both the relevant primary databases as well as to PubChem Compound (when applicable). Data is released to the website and updated on the FTP site as soon as it is ready, generally within 48 hours.

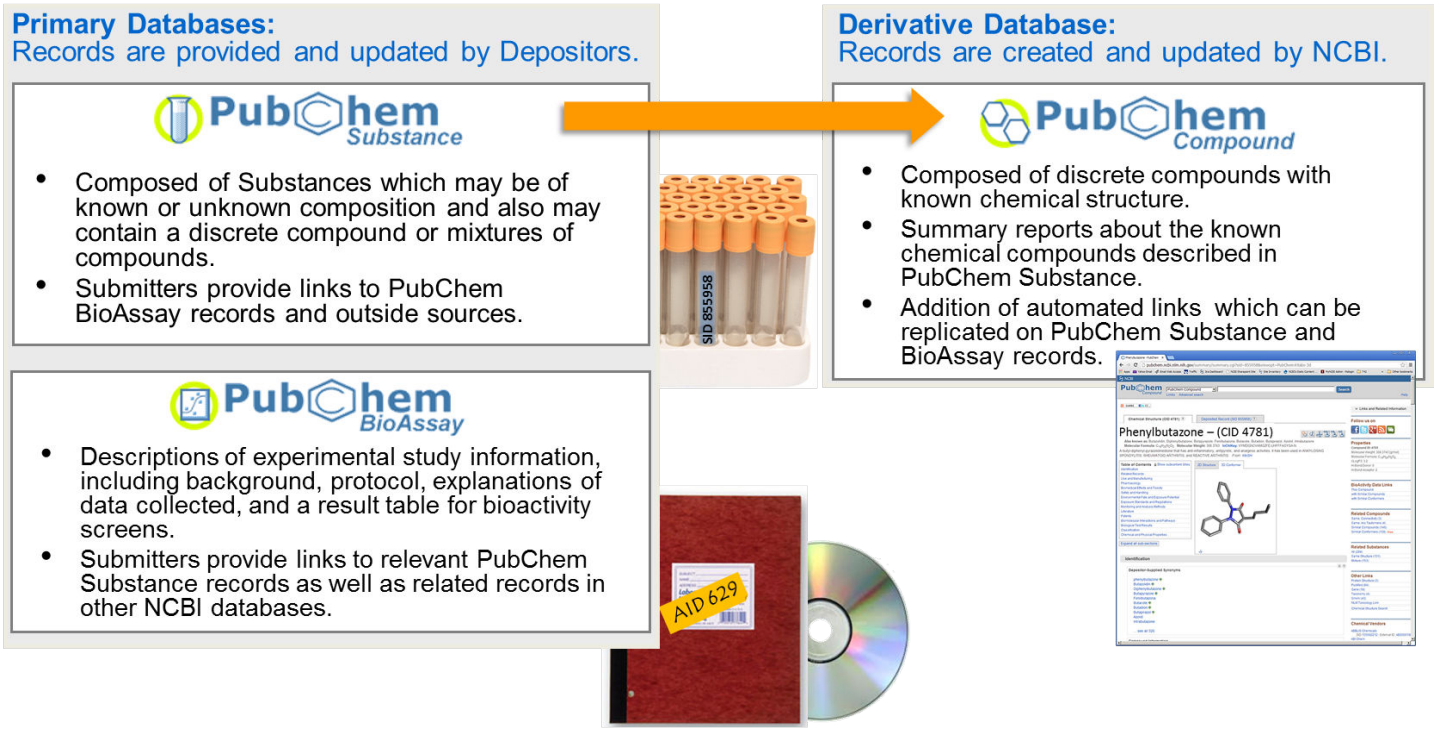
Primary and Derivative Databases

PubChem's primary databases contain records of data as submitted to the PubChem group:

- PubChem Substance: Information provided to PubChem by submitters. Substances may be of known or unknown chemical composition, and may contain a discrete compound or mixtures of compounds. Links and information in the record are provided and maintained by submitters. Please note that multiple records representing the same substances may be submitted by different laboratories or organizations.
- PubChem BioAssay: Information provided to PubChem by submitters including study background, protocol, and results for experimental bioactivity screens of chemical substances that have been submitted to PubChem Substance.

The PubChem Compound database is a derivative database created by a consolidation of PubChem Substance Records with additional curated information created and maintained by the PubChem group.

- PubChem Compound: Composed of discrete chemicals with known molecular structure. Summary reports are generated from the known chemical compounds described in PubChem Substance and are supplemented with additional information. Links back to related PubChem Substances, PubChem BioAssays, and other NCBI and external sites are provided by the PubChem group.



For example, a Molecular Libraries Screening Center Laboratory at Emory University submitted an Estrogen Receptor-alpha coactivator high-throughput chemical binding screen. This was given the PubChem BioAssay Identifier AID 629. Around the same time, a collaborating group, the Molecular Libraries Screening Center's Small Molecule Repository, submitted information about the substances in their system including 86,096 that were tested in this particular assay. One of those tested and found to be active in AID 629 was given the PubChem Substance Identifier SID 855958, which was one of the many linked to the PubChem BioAssay record.

In addition to AID 629, this same substance (SID 855958) was tested in over 2,900 assays by various groups. Further information about this chemical was uploaded to the PubChem group by several other organizations including biological property databases, journal publishers, and chemical vendors as submissions to create their own PubChem Substance records. As of December 1, 2013, 284 PubChem Substance records were submitted for this same chemical. The PubChem group then consolidated all information and submitted links for this chemical and included additional calculated chemical properties and information provided by several other key sources (NLM's Medical Subject Headings, MeSH, DailyMed, and Hazardous Substances Data Bank [HSDB]) and created a reference record in the PubChem Compound database with a PubChem Compound identifier CID 4781. This record contains links back to all corresponding PubChem Substance and BioAssay records.

Access

Searching PubChem Databases Using the Web Interface

NCBI's chemical and bioactivities data can be accessed through the NCBI website and the PubChem project homepage. Searching for PubChem data can be performed through the common Web Search (Entrez) mechanism. For the PubChem Compound, PubChem Substance and PubChem BioAssay databases, there are highly specialized and useful Limits pages to assist in specifying key types of data. These databases also maintain Entrez Advanced pages with Advanced Search Builders and Search History tables.

For users who would like to retrieve records from a list of PubChem-related record identifiers (PubChem Compound Identifiers – CIDs, PubChem Substance Identifiers – SIDs, and BioAssay Identifiers – AIDs), NCBI's Batch Entrez permits the upload of a text file with subsequent retrieval from the selected database.

In addition to text-based searches using the Web Search (Entrez) system, the PubChem group has created a Chemical Structure Search mechanism in which two dimensional drawings or textural representations of chemical structures (SMILES – Simplified molecular-input line-entry system, SMARTS – Smiles arbitrary target specification, InChIs – International Chemical Identifiers) can be used to query the PubChem Compound or Substance databases to find records for chemicals with matching or similar structures.

Downloading PubChem Data from the Web

After performing a search, groups of retrieved records can be downloaded using the "Send to" function common to all Web Search (Entrez) displays. Individual PubChem records can be downloaded directly from the Web interface with links at the top-right-hand side of the screen. PubChem records display record download buttons for available file formats (SDF, CSV, ASN.1, XML).

The PubChem group has also created Download Facility tools which enable the quick upload of a list of PubChem record identifiers (CIDs, SIDs, or AIDs) and download of PubChem Compound or Substance information (PubChem Structure Download Facility) or PubChem BioAssay information (BioAssay Download Facility) in a number of file formats.

For users who would like to download the data in bulk, files are available on the PubChem FTP site, which is mirrored on the Aspera-plugin version of the NCBI FTP site.

Accessing PubChem Data Using Programming Interfaces

The PubChem group provides scripting access for users in a number of formats. As with other NCBI databases, the PubChem resources are accessible using the NCBI Entrez Utilities API Interface (EUtilities). In addition, the Group has developed a PubChem-specific RESTful API Interface (PUG-REST) as well as an XML-intensive PubChem Power User Gateway (PUG).

Related Tools

Standardization Service mimics the initial steps of PubChem's data processing pipeline and, therefore, enables users to see how submitted structures would be handled. In addition, this service will enable the conversion of chemical structures in SMILES, InChI, or SDF formats to a different format. Standardization through this service is limited to a single structure at a time.

Identifier Exchange Service enables the upload and conversion of a list of identifiers (CIDs, SIDs, or Registry IDs), InChIs or synonyms with retrieval of identifiers for identical or similar PubChem Compound or Substance records. A file or correspondence table can be downloaded.

Data Dicer is an alternative to the Web Search (Entrez) Limits page, providing guided searches and tabular retrievals for information pertaining to bioactivity outcomes for gene/protein targets or screened small molecules in bioactivity assays listed in the PubChem BioAssay database.

Classification Browser searches PubChem records annotated with hierarchies/terms and displays the distribution of these in the context of the specific ontology. Please note that this only operates on the subset of PubChem records that have been annotated with terms from the available hierarchies.

Structure Search searches for identical or similar chemical structure records using CID or SIDs, Names (SMILES, SMARTS, InChI, Synonyms, MeSH terms), Molecular Formulas, Molecular Weights, SDF file, or even hand-drawn structures. Similarity thresholds can be customized and record retrieval filters can be set using an interface similar to that of the PubChem Web Search (Entrez) Limits page or using search histories from PubChem Compound or Substance databases. The 2D structure searching strategy uses a modified-Tanimoto scoring system, while the 3D structure searching strategy is based on comparison of calculated shapes of 3D

conformers with positions of the functional groups. Search results displayed sorted in decreasing order of the resulting calculated 2D- or 3D-similarity score are retrieved and displayed as a traditional PubChem Compound or Substance search results list.

Score Matrix Service enables the downloading of 2D- or 3D-similarity scoring matrix values calculated by the PubChem Structure Search.

Structure Clustering Tool assists in exploration of chemical-structure space and displays chemical structural similarity and population diversity relationships in a tree format using the Single Linkage algorithm. Calculations are based on 2D-similarity scores (modified-Tanimoto scoring) or 3D-similarity scores (conformer, functional group scoring).

BioActivity Summary/DataTable searches and retrieves biological screening results for individual or a set of chemical samples or displays the information content of PubChem BioAssay records. Tables are downloadable in a number of formats.

BioActivity Plot Service displays bioactivity assay data in customizable scatter plots or histograms.

Structure-Activity Relationship HeatMap enables interactive visualization for exploring PubChem data, displaying a chemical structure tree on a Y-axis with the bioassays clustered by relatedness on the X-axis. Activity data values are shown as colored squares in the resulting grid. The display of the grid is customizable with respect to color and zoom level and can be downloaded.

Webpage Widgets provides code for configurable PubChem data displays including record summaries as well as tables for placement in Web pages.

References

The PubChem Project

Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res. 2009 Jul 1;37(Web Server issue):W623-33. Epub 2009 Jun 4. doi: 10.1093/nar/gkp456.

PubChem Substance and PubChem Compound Databases

Bolton E, Wang Y, Thiessen PA, Bryant SH. PubChem: Integrated Platform of Small Molecules and Biological Activities. Chapter 12 IN Annual Reports in Computational Chemistry, Volume 4, American Chemical Society, Washington, DC, 2008 Apr.

PubChem BioAssay Database

Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Zhou Z, Han L, Karapetyan K, Dracheva S, Shoemaker BA, Bolton E, Gindulyte A, Bryant SH. PubChem's BioAssay Database. Nucleic Acids Res. 2012 Jan;40(1):D400-12. (Epub 2011 Dec 2) doi: 10.1093/nar/gkr1132.