CHAPTER 12

PubChem: Integrated Platform of Small Molecules and Biological Activities

Evan E. Bolton*, **Yanli Wang***, **Paul A. Thiessen***, and **Stephen H. Bryant***,[1]

Contents

* National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Department of Health and Human Services, 8600 Rockville Pike, Bethesda, MD 20894, USA

[1] Corresponding author. E-mail: bryant@ncbi.nlm.nih.gov

Annual Reports in Computational Chemistry, Vol. 4
ISSN 1574-1400, DOI: 10.1016/S1574-1400(08)00012-1

1. INTRODUCTION

PubChem [1], an open repository for experimental data identifying the biological activities of small molecules, is a part of the Molecular Libraries and Imaging (MLI) component of the National Institutes of Health (NIH) Roadmap for Medical Research initiative [2]. This program includes the Molecular Libraries Screening Center Network (MLSCN), grant-supported experimental laboratories, and a shared compound repository, referred to as the Molecular Libraries Small Molecular Repository (MLSMR) offering biomedical researchers access to chemical samples.

PubChem archives the molecular structure and bioassay data from the MLSCN and other contributors. PubChem provides search, retrieval, and data analysis tools to optimize the utility of these results. PubChem further enhances the research utility of the MLSCN output by including other public sources of chemical structure and bioactivity information and by integration of this data with other NIH biomedical knowledgebases. The primary aim of PubChem is to provide a public on-line resource of comprehensive information on the biological activities of small molecules accessible to molecular biologists as well as computational and medicinal chemists.

Initially launched September 2004, PubChem follows the GenBank [3] approach, whereby investigators make direct data submissions. PubChem depends on its contributors to help keep the database as comprehensive, current, and accurate as possible. The processing of PubChem is highly automated, as opposed to being manually curated, keeping the overall database cost low. The open repository nature of PubChem has a 25 year precedent in biology, for example, GenBank, SwissProt [4], PDB [5], etc., but there is less of a precedent for this model in chemistry.

The location of PubChem at the National Center for Biotechnology Information (NCBI) [6] provides the unique ability to integrate directly with a substantial wealth of biomedical information, over thirty databases with information ranging from scientific articles to genes, available within the NCBI Entrez search system [7]. By leveraging and integrating with these resources, PubChem provides a powerful, publicly accessible platform for mining biological information of small molecules.

2. DESCRIPTION

PubChem is organized as three distinct databases: PubChem Substance, PubChem Compound, and PubChem BioAssay. PubChem Substance contains descriptions of chemical samples, provided by data depositors, and links to information on their biological activities. The description includes PubChem Compound identifiers in cases where the chemical structures of compounds in the sample are known. Links providing information on biological activity include those to PubMed [8] citations, protein 3-D structures [9], links to contributor websites, and to biological testing results available in PubChem BioAssay.

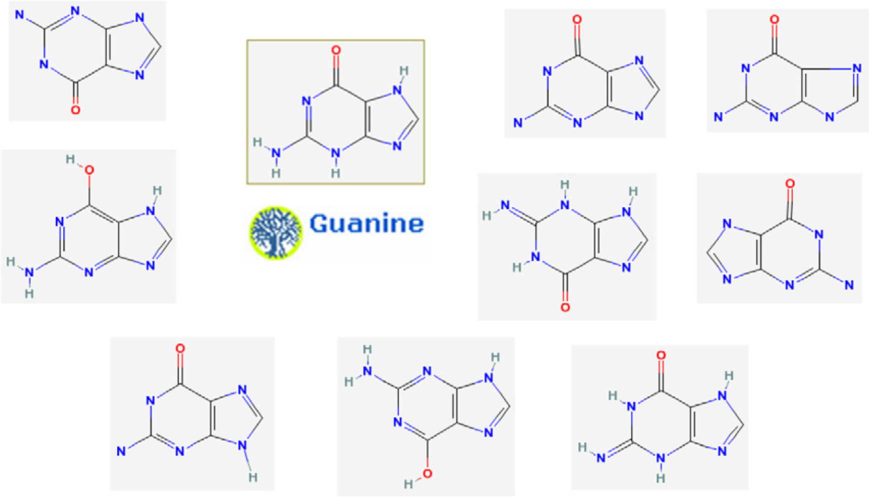


FIGURE 12.1 Different structural representations of guanine deposited in PubChem.

PubChem Compound contains the unique chemical structure content of PubChem Substance. Compounds may be searched by computed chemical properties and are pre-clustered by structure comparison into identity and similarity groups. Whenever possible, compounds are linked via PubChem Substance to information on their biological activities.

PubChem BioAssay contains the results of biological activity testing from a variety of sources. It provides searchable descriptions of each bioassay, including conditions and readouts specific to the screening procedure. PubChem BioAssay provides outcomes for the depositor's tested substances as links to PubChem Substance. Associations between biological testing results and the unique chemical structures are also generated to provide a comprehensive overview of the biological profile of tested compounds.

Abstracting the unique chemical structure content in PubChem Substance to create PubChem Compound is not always trivial. Widely adopted standards or rules for chemical structure representation do not exist, with various groups or individuals adopting preferences based on their organizational needs. Further complicating matters is that PubChem accepts chemical information from a multitude of depositors, each with the potential to represent identical chemical structures in a different way. For example, a molecule as simple as guanine (Figure 12.1) has a number of equivalent representations readily recognizable by a chemist as guanine. Programming a computer to recognize such chemical representations as being the same is nonetheless a challenge.

The normalization method used by PubChem to identify unique chemical content is referred to here as "standardization." This procedure involves a series of automated processing steps, outlined in Figure 12.2, to determine when a provided chemical structure description is well defined and chemically reasonable. The standardization processing steps involve: verification that each atom is a known

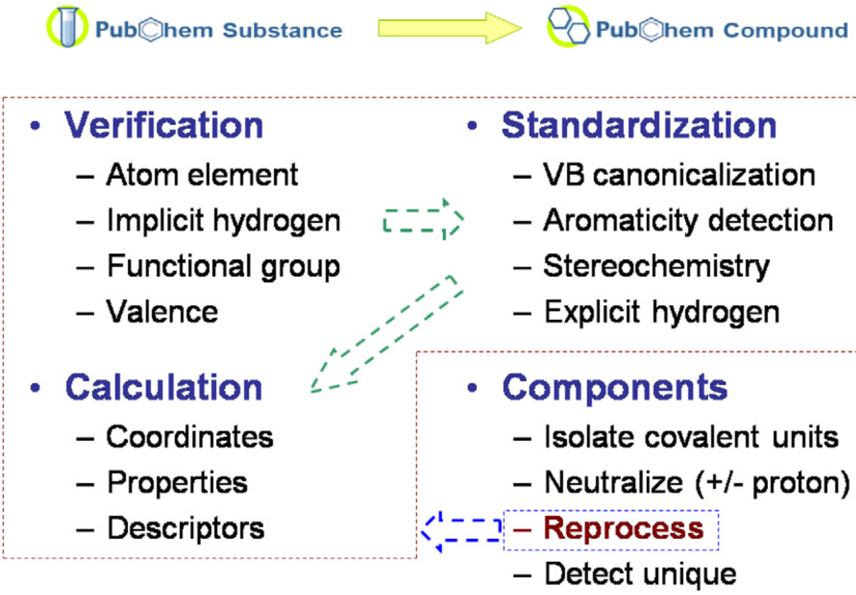


FIGURE 12.2 Overview of the processing performed on chemical structures deposited in PubChem.

element, assignment of implicit hydrogens to organic elements missing valences, normalization of functional group representations, validation that each atom valence and formal charge is reasonable, valence-bond canonicalization for tautomer and resonance invariance, extended aromaticity resonance detection and annotation, stereochemical center identification, and conversion of implicit to explicit hydrogens for unambiguous atom valences. Additional processing is performed to isolate unique covalent units within the chemical sample description of mixtures, which are acid/base neutralized when possible, and reprocessed using the above procedure. Subsequent processing of each standardized structure involves computation of 2-D depiction coordinates and calculation of basic chemical properties (e.g., molecular weight, molecular formula, etc.) and chemical descriptors (e.g., canonical SMILES [10], InChI [11], IUPAC name [12], etc.).

Contributed substance descriptions that do not include a chemical structure or that fail the PubChem chemical structure standardization procedure do not enter or have links to the PubChem Compound database. Prior to analysis or any modification of chemical structure input, care is taken to preserve the original structure description. The result of the normalization methodology employed is a uniform representation of the chemical structure content contained within the PubChem Substance database.

3. DATA RELATIONSHIPS

The fundamental relationships between the three PubChem databases are straightforward. PubChem Substance identifiers (SIDs) relate to PubChem Compound identifiers (CIDs) through chemical structure standardization. Each substance, if it standardizes, will have a corresponding CID that is the main "standardized" form of that substance, representing the whole structure. There may also be "component form" CIDs that include unique covalently bonded units, when the substance is a mixture, or an acid/base charge-neutralized form, when the substance is ionized. A parent compound is assigned to each CID, when possible, to identify the primary organic component. PubChem Assay identifiers (AIDs) contain activity data for SIDs. If a substance is associated with a compound, the assay outcome for the SID can be associated implicitly with a CID, as well.

A critical concept for the advanced PubChem user is that of combining and transforming sets of identifiers between the three PubChem databases, based on the above identifier relationships. For instance, there is a many-to-one relationship between SIDs and "standardized" CID, as more than one Substance depositor may have supplied the chemical structure that standardizes to a given CID. (In fact, even within a particular depositor's records, there may be redundant structures because of different sample origins, tautomeric forms, etc.). Also, the perceptive reader will notice there is not a direct relationship between BioAssay (AID) and Compound (CID) identifiers. To discover assays linked to a CID, there is an expansion of that CID to all SIDs for which that CID is the standardized form; AIDs can be associated with CIDs linked to any of these SIDs.

Many of the PubChem tools perform such transformations of the ID space implicitly, such as assay tools that work with sets of CIDs, or Entrez searches of CID chemical property indices in PubChem Substance, like IUPAC name, that actually come from standardized compounds. It can be important to understand these implicit relationships when navigating through PubChem, especially when searching and analyzing records across multiple databases.

As of March 2008, PubChem contains more than: 1,000 bioassays, 28 million bioassay test outcomes, 40 million substance contributed descriptions, and 19 million unique compound structures contributed from over 70 depositing organizations. While the majority of screening data were contributed by NIH funded screening centers under the MLSCN network, PubChem BioAssay database also contains test outcomes from a number of other organizations, including the sixty tumor cell line assays from DTP/NCI [13], toxicity data from the DSSTox program at EPA [14], and bioactivity data extracted from literature by the BindingDB project [15].

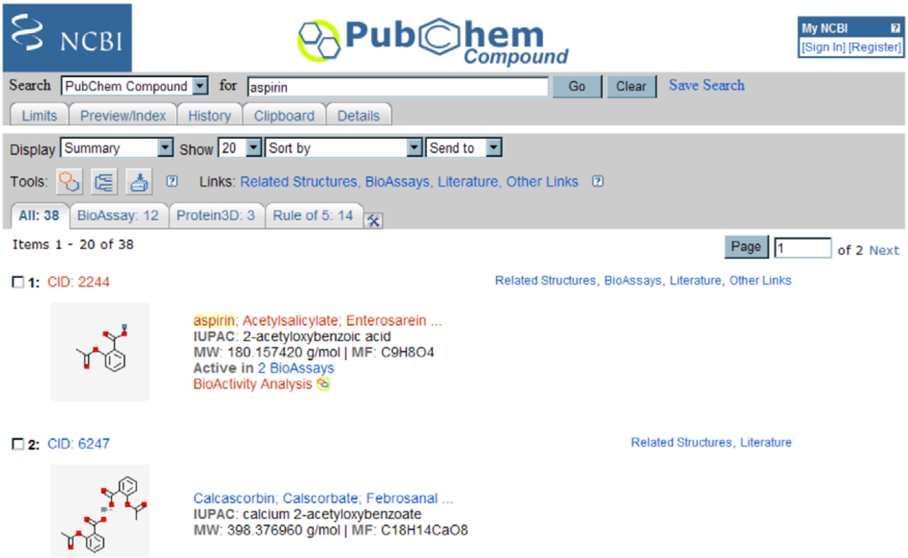
4. INTERFACE

The primary interface to PubChem data is through the NCBI search engine, Entrez. This web-based interface is simple, yet powerful, with many features not immediately apparent to those unfamiliar with the Entrez system. This section is intended as both an introduction and a guide to the more advanced Entrez features, and the types of Entrez PubChem queries that can be performed.

4.1 Entrez

There are a number of entry points to Entrez. The simplest is to go to the NCBI home page (http://www.ncbi.nlm.nih.gov/) from which one can input a search term (or terms) and initiate a search by activating the 'Go' button. By default, if a specific database is not selected in the search menu, the search is performed across all +30 databases available within Entrez, of which PubChem is a part. This "global query" result lists the count of records for the query in each of the Entrez databases. To see the PubChem query results, simply select one of the three PubChem databases (Substance, Compound, or BioAssay), and a detailed report for records matching the query is displayed for that database. One can also begin at the PubChem home page (http://pubchem.ncbi.nlm.nih.gov/) where an equivalent search of one of the three PubChem databases may be initiated through the input form at the top.

Figure 12.3 shows the result of searching for the word "aspirin" in Entrez's PubChem Compound database. This default display of multiple records in Entrez is referred to as a document summary (DocSum) report and is common to all Entrez databases. At the top are the common Entrez controls (database selection and search input box) and tabs for other Entrez tools (e.g., Limits, History, etc.) some of which are described in more detail below. Note that the format of this page evolves over time, but the basic controls remain the same. Moving down the DocSum page, the next section contains controls to change the display type; the default is "Summary" (as shown). Each Entrez database has report styles that vary in type and detail of information shown, the overall format is the same—a list of records,



Search PubChem Compound for aspirin

Display Summary Show 20 Sort by Send to

Items 1 - 20 of 38

□ 1: CID: 2244 Related Structures, BioAssays, Literature, Other Links

aspirin; Acetylsalicylate; Enterosarein ...
IUPAC: 2-acetyloxybenzoic acid
MW: 180.157420 g/mol | MF: C9H8O4
Active in 2 BioAssays
BioActivity Analysis

□ 2: CID: 6247 Related Structures, Literature

Calcascorbin; Calscorbate; Febrosanal ...
IUPAC: calcium 2-acetyloxybenzoate
MW: 398.376960 g/mol | MF: C18H14CaO8

FIGURE 12.3 Partial view of an Entrez document summary (DocSum) report page for the PubChem Compound query "aspirin".

each with report-specific information displayed. Also, controls exist to enable one to sort the results by various means or to export the DocSum to a file or printer.

PubChem databases have a number of additional controls that operate on a query result list, such as icon buttons (provided after "Tools") for assay data analysis and chemical structure download. There are pop-up link menus (provided after "Links" on the same line as "Tools") that provide powerful query result list operations. Also, the pop-up link menus exist for each record, but they function only on the individual record. The meaning of these links is detailed further in the following sections.

The last set of tabs shown in Figure 12.3 (e.g., All, BioAssay, Protein3D, etc.) before the actual record summaries are filters that apply to the current result list. For example, in Figure 12.3, the "BioAssay: 12" tab indicates 12 of the 38 results have associated bioassay data. This tab, if selected, will indicate which 12 records have associated BioAssay data and will allow the result list to be refocused to consider only the 12 records by clicking the "push pin" icon on the tab that appears. These filter tabs are in fact customizable through the "MyNCBI" system. One can create and store new filters in MyNCBI that can be applied to any search in a given database. As depicted in Figure 12.3, the MyNCBI tool is accessed via the box in the upper right corner of the page.

The remainder of the Entrez DocSum page contains the paginated list of results of the current search. While the details will vary according to the specific database and report style, information is shown for each record matching the query result. In the case of the example in Figure 12.3, this includes links to the detailed summary page of each item and other Entrez database records associated with those in the current database.

4.2 Advanced features

Entrez is basically a multi-database search engine. Under the surface are a vast number of details on which fields are available to be searched, what the many types of links mean, how the core Entrez controls function, and so on. All of this may be a bit daunting to the casual user, but understanding these details unlocks the true power of Entrez. This section serves as a guide to PubChem's Entrez databases, including what indices, links, and filters are available, and how these combine together to create an advanced query refinement system.

4.2.1 Entrez indices

An index is a piece of information tied to individual records and matched directly to a user's query in an Entrez search. Each index consists of text, numeric, or date values. Each Entrez database has its own set of indices. These indices are named according to the type of information they contain, for example, the indices "IUPACName" or "MolecularWeight" in PubChem Substance and Compound. Some indices may have multiple values for each record. For example, the index "Synonym" corresponds to chemical or common names of a substance, any number of which may be supplied by the depositor.

By default, when one enters a simple query in the Entrez search interface, that query is matched against all indices in that database. For example, if one searches "aspirin" in PubChem Compound, Entrez will report back any records with an index that contain "aspirin" as (any word in) a synonym, a depositor comment, etc. This is why a text search for "aspirin" also currently brings up the structure of acetaminophen, considering one of the names supplied by a depositor for acetaminophen is "Aspirin-Free Anacin," and so an unrestricted search for "aspirin" will match this record, as well.

It is possible to narrow the search to a particular index by adding the index name in brackets after the term itself. For example, "aspirin[CompleteSynonym]" returns only a single record, the actual structure of aspirin, because only that record has a synonym that matches that query exactly. Also, this shows that some Entrez indices are configured to require an exact match to the entire index, while others allow matches to any individual word in the longer text.

For numeric indices, one can perform a search for a range of values by using minimum and maximum values separated by a colon and followed by the index name in brackets. For example, to find all chemical structures in PubChem Compound with a count of hydrogen bond donors between 0 and 5, the range query would be "0:5[HydrogenBondDonorCount]." In the case of floating point range queries, such as finding all chemical structures with a molecular weight between 214.31456 and 215.0 g/mol, one would use the query "214.31456:215[MolecularWeight]."

Multiple indices may be searched simultaneously using Entrez's Boolean operators. For example, a query in PubChem Compound of "Br[Element] AND 1[CovalentUnitCount]" will find all chemical structures containing the element bromine and that are not part of a mixture. Please note that Entrez Boolean operators are capitalized (e.g., "AND," "OR," and "NOT").

By default, Entrez removes whitespace, some punctuation, and other special characters from the query string. To make sure Entrez treats the query as a single word or phrase, despite special characters, simply enclose the query in quotation marks. For example, to search the PubChem Compound database using the InChI string of aspirin, one would use ""InChI=1/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)/f/h11H"[InChI]" as the query.

Knowing what indices are available in a database is the key to maximizing the power of an Entrez search. The indices may be listed by going to the "Preview/Index" tab in Entrez, and opening the menu on the bottom left. Also, this page provides an interface for constructing index-specific queries. A complete list and description of the Entrez indices available for the three PubChem databases are detailed in the "Indices and Filters in Entrez" section of the help documentation: (http://pubchem.ncbi.nlm.nih.gov/help.html).

4.2.2 Entrez links

In addition to index searching, Entrez provides cross links or associations between records in different Entrez databases, or within the same database. These links may be applied to an entire search result list, via the links pop-up menus at the top of a DocSum page (see Figure 12.3), or to an individual record, via link menus on the right side of each entry in the DocSum.

Links provide a way to discover relevant information in other Entrez databases based on a user's specific interests. Equivalently, one may think of this as a way to transform an identifier list from one database to another based on a particular criteria. From PubChem Substance, for example, the link "PubChem BioAssays, Active" provides all assays where that particular substance (or any substance within a multi-record list) was found to be active, where the meaning of "active" is specific to and defined by a particular assay depositor. In a similar cross-database fashion, activating the "PubChem Same Substances" link from a PubChem Compound record will lead to all deposited substances exactly matching that compound, providing a method to see which depositors deposited a particular compound. Some links operate within the same database, going to records that are related in some way. For example, the "Similar Compounds" link from a structure in PubChem Compound will take the user to a DocSum display of all compounds that have a 2-D Tanimoto-based similarity of at least 90% to the structure.

4.2.3 Entrez filters

Filters are essentially Boolean bits (true or false) for all records in a database that indicate whether or not a given record has a particular property. Filters may be used to subset other Entrez searches according to this property, by adding the filter to the query string. For example, the "pcsubstance_pcassay" inter-database filter has a "true" bit for every substance that has associated PubChem BioAssay data, such that a search for "100:200[MolecularWeight] AND pcsubstance_pcassay[Filter]" in PubChem Substance will return a list of all substances with molecular weight from 100.0 to 200.0 g/mol and that have associated PubChem BioAssay data.

Filters are related to links in that the majority of filters in the PubChem databases are generated automatically based on the presence of links. In the above example the "pcsubstance_pcassay" filter has a "true" bit for every substance for which a PubChem BioAssay link is present (e.g., in the pop-up menus of the Entrez DocSum for that substance).

There are some special filters that are not link-based. The query "all[Filter]" simply returns every record in a given Entrez database. A database may have other special filters defined, such as the "has_pharm" filter in PubChem Compound that indicates whether a given chemical structure has a known pharmacological action.

The filters for each Entrez database may be listed by going to the "Preview/Index" tab in Entrez, opening the menu on the bottom left, selecting "Filters," and pressing the "Index" button. Also, this page provides an interface for adding filters to Entrez queries. A complete list and description of the custom Entrez filters available for the three PubChem databases are detailed in the "Indices and Filters in Entrez" section of the help documentation (http://pubchem.ncbi.nlm.nih.gov/help.html).

4.2.4 Entrez history

Entrez is a query refinement engine. In addition to enabling complex searches across databases, as described above, Entrez has a history mechanism (Entrez history) that automatically keeps track of a user's searches, temporarily caches them (for eight hours), and allows one to combine search result sets with Boolean logic. For example, say a structure search (described elsewhere in this document) has been completed, resulting in a list of 10,000 compounds. One may wish to narrow this search by other means, such as to find all compounds in the original search result that satisfy the "Lipinski Rule of 5" [16]. To do this, one would go to the "History" tab in Entrez, where all recent searches are listed, and find the history number in the leftmost column corresponding to the structure search in question (e.g., something looking like "#5 : 10,000 document(s)"). Then, in the search form, at the top of the page, one would use this history number to formulate a query such as "#5 AND lipinski_rule_of_5[Filter]," to narrow the original result to only those records that satisfy both the original query and the "Lipinski Rule of 5."

Entrez history is used heavily by PubChem tools (which are not a part of Entrez) so results of user searches can be used as a subset for further manipulation. For example, the chemical structure download service (described below) reads Entrez history items, so one can generate an SDF file containing just those compounds found in a PubChem Compound Entrez result set. For example, the BioAssay tools (also described below) make frequent use of Entrez history, so that structure queries can be used to subset assay results in a chemical structure analog series.

It is important to note that Entrez history is database-specific. One cannot use it to combine search results between databases (e.g., to 'AND' together a CID list with an AID list). Cross-database links must first be used as set transformation operators, so all ID lists are in the same database. For example, following the "PubChem BioAssays" link from a set of CIDs will create a new set of AIDs that have any test results for the set of CIDs (again with the implicit understanding that CID is first expanded to SID, which is built into the CID-AID links). From there, one may combine this set of AIDs with other search results in the BioAssay database using the Entrez Boolean logic.

Understanding which ID space transformations are implicit and which may be performed explicitly through links or other tools, is crucial to successful use of the advanced PubChem tools. With Entrez history, the user has complete control over the set logic used in sophisticated query refinement. Both of these concepts become even more important when dealing with the PubChem programmatic tools (described below).

5. TOOLS

We have described how PubChem databases are integrated into Entrez, enabling detailed and flexible searches across the PubChem data; however, Entrez is essentially a text search engine and is not amenable to more detailed chemical and bioassay data analysis. Such analysis must be handled by specialized applications.

As the PubChem data content grows, there is an ever increasing need for facile methods of efficient large-scale data management and analysis.

Researchers require the ability to obtain comprehensive summaries of the biological activities of small molecules. In addition, scientists are interested in other chemicals which share structural or physical property similarities to known bioactive entities, or have similar biological activity profiles. To this end, the PubChem BioAssay system provides additional data analysis tools for utilizing and analyzing the biological activity data. These include tools for comparison of test results across multiple experiments, visualizing and exploring structure-activity relationships, and summarizing bioactivity information.

There are two general categories of specialized applications provided by PubChem: those that deal with chemical structure information and those that deal with bioassay data. These categories are not totally distinct; however, as several of the PubChem tools, such as structure-activity analysis and structure clustering, directly bridge the two. These particular tools are closely integrated with Entrez, as searches in one may be used as starting points in the other, but they are conceptually and operationally separate from Entrez. The goal of this section is to describe available tools and how they combine together to form a unified platform for mining PubChem chemical and biological data.

5.1 Summary pages

The Entrez DocSum reports serve a limited quantity of data to help navigate and subset records. Detailed information is provided by PubChem summary pages. Each record in an Entrez DocSum contains a link that leads to the more detailed information on a specific record. Typically these pages are reached through Entrez, but one can also navigate to them directly. For PubChem Substance SIDs, the summary page URL is of the form:

http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?sid=1234

where the SID (substance identifier) is provided as an argument. Similarly, for a PubChem Compound the URL is like:

http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=2244.

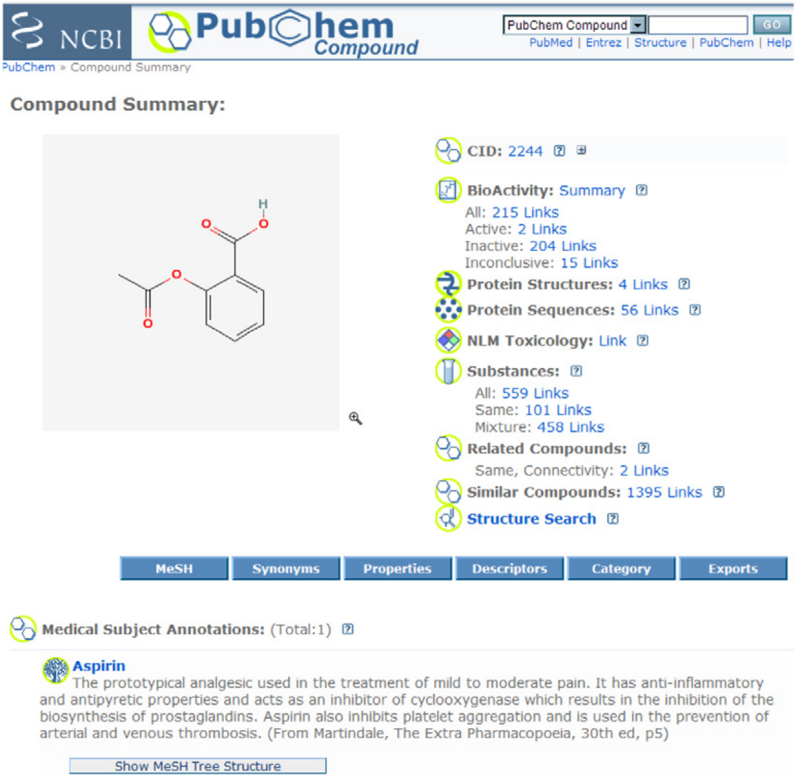
For a PubChem BioAssay summary page, the URL has the form:

http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=910.

In general, summary pages contain the detailed information necessary to understand how individual PubChem records combine information into a comprehensive system of interconnected data.

5.1.1 Compound/substance summary

The layout of the substance and compound summary pages are very similar. The content of this summary is heavily dependent on the information provided by depositors and our ability to integrate contributed information with biomedical



Compound Summary:

CID: 2244

BioActivity: Summary
All: 215 Links
Active: 2 Links
Inactive: 204 Links
Inconclusive: 15 Links

Protein Structures: 4 Links

Protein Sequences: 56 Links

NLM Toxicology: Link

Substances:
All: 559 Links
Same: 101 Links
Mixture: 458 Links

Related Compounds:
Same, Connectivity: 2 Links

Similar Compounds: 1395 Links

Structure Search

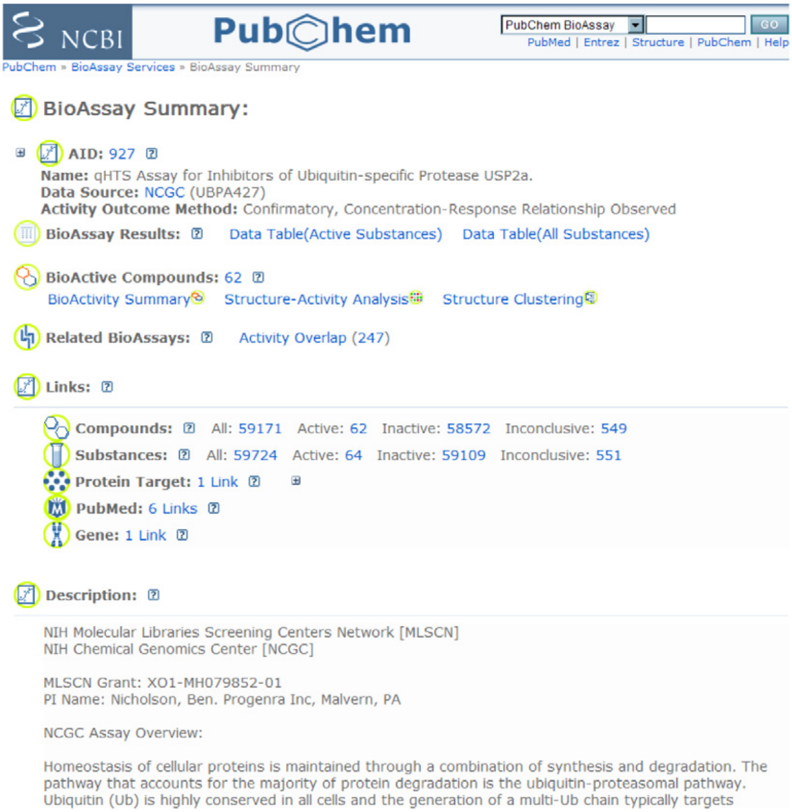
Medical Subject Annotations: (Total:1)

Aspirin
The prototypical analgesic used in the treatment of mild to moderate pain. It has anti-inflammatory and antipyretic properties and acts as an inhibitor of cyclooxygenase which results in the inhibition of the biosynthesis of prostaglandins. Aspirin also inhibits platelet aggregation and is used in the prevention of arterial and venous thrombosis. (From Martindale, The Extra Pharmacopoeia, 30th ed, p5)

FIGURE 12.4 Partial view of a compound summary for aspirin (CID 2244).

resources at NCBI. Summary pages, despite being continually refined as content is added or usability improved, provide an overall summary of what is known about a particular substance or compound. In general, a compound or substance summary will contain these basic aspects: a depiction of a chemical structure; indication of where or how the record originated (e.g., who contributed the record); links to a set of related inter-database Entrez resources, such as a protein 3-D structures or literature articles; links to known biomedical information (e.g., pharmacological actions of a drug); a list of synonyms or names associated with the record; computed chemical structure properties and descriptors; and record download controls. Figure 12.4 depicts an example of a compound summary for aspirin.

Substance summary pages are distinctly different from compound summary pages in two important ways. First, a substance summary provides access to the depositor's original structure information as well as the standardized form of the substance (when applicable), with the standardized form always shown by default. Second, a substance summary only provides information provided by a single depositor, whereas a compound summary page aggregates information across all depositors providing substances that standardize to that compound.


NCBI
PubChem
PubChem BioAssay
GO
PubMed | Entrez | Structure | PubChem | Help
PubChem » BioAssay Services » BioAssay Summary

BioAssay Summary:

AID: 927

Name: qHTS Assay for Inhibitors of Ubiquitin-specific Protease USP2a.
Data Source: NCGC (UBPA427)
Activity Outcome Method: Confirmatory, Concentration-Response Relationship Observed

BioAssay Results: Data Table(Active Substances) Data Table(All Substances)

BioActive Compounds: 62
BioActivity Summary Structure-Activity Analysis Structure Clustering

Related BioAssays: Activity Overlap (247)

Links:

Compounds: All: 59171 Active: 62 Inactive: 58572 Inconclusive: 549
Substances: All: 59724 Active: 64 Inactive: 59109 Inconclusive: 551
Protein Target: 1 Link
PubMed: 6 Links
Gene: 1 Link

Description:

NIH Molecular Libraries Screening Centers Network [MLSCN]
NIH Chemical Genomics Center [NCGC]

MLSCN Grant: XO1-MH079852-01
PI Name: Nicholson, Ben. Progenra Inc, Malvern, PA

NCGC Assay Overview:

Homeostasis of cellular proteins is maintained through a combination of synthesis and degradation. The pathway that accounts for the majority of protein degradation is the ubiquitin-proteasomal pathway. Ubiquitin (Ub) is highly conserved in all cells and the generation of a multi-Ub chain typically targets

FIGURE 12.5 Partial view of a bioassay summary for a confirmatory (secondary screen) assay for ubiquitin-specific protease USP2a (AID 927).

5.1.2 BioAssay summary

A BioAssay summary displays descriptive information and a summary of the assay results. This includes an overview and background of what the assay attempts to achieve, the assay protocol utilized, references, definition of all reported assay outcomes, indication of the primary result fields, and explanation of the criteria used when considering samples as active or inactive. One can use the “Related BioAssay, Depositor” link to find additional screening performed for a particular assay project. An example bioassay summary is depicted in Figure 12.5.

5.2 Structure search

The PubChem structure search tool enables one to query and subset PubChem Compound by a variety of chemical structure search types and optional filters. The chemical structure search service may be directly accessed using the URL:

http://pubchem.ncbi.nlm.nih.gov/search/.

The supported query input formats for the structure search tool are SMILES, SMARTS [17], InChI, CID (PubChem Compound identifier), molecular formula, and SDF [18]. There is also an online JavaScript-based chemical structure sketcher through which a query may be manually drawn, edited, or imported. The sketcher is compatible with modern web browsers and does not require special software to be downloaded or installed.

Multiple chemical structure search types are available. Identity search enables one to find identical PubChem records at different levels of "sameness" through consideration of structural connectivity and either the presence or absence of isotopic and stereochemical information. Similarity searches locate chemical structures similar to a query, using a percent similarity measure employing the Tanimoto equation [19] and a dictionary-based fingerprint, analogous to the MACSS structure-based keys [20], that are described on the PubChem FTP site: (ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt). Molecular formula searches employ a flexible query containing the count of particular elements in a chemical structure. Substructure searches locate records that contain all atoms in a particular chemical structure query pattern. Superstructure searches locate records that comprise a subset of atoms in a particular chemical structure query pattern.

While the input query and search type are all that are necessary to perform a structure search in PubChem, there are numerous choices by which one may narrow the search to smaller subsets of PubChem. For example, one may search only within a previous Entrez search result, or even a previous structure search result, or upload a file of CIDs against which the search is to be performed. One may filter based on a wide variety of properties, such as molecular weight, heavy atom count, presence or absence of stereochemistry, assay activity, elemental composition, depositor name or category, etc. Most of these subset operations could be accomplished through appropriate Entrez index queries followed by Boolean operations on structure search results; however, the structure search tool provides a convenient one-step interface for chemical search refinement.

All compound structure searches are queued on a set of devoted NCBI servers. The user is taken to a search status page after submitting a query, with a meter showing the relative progress of the requested task. By default, a structure-based query is allowed to take as much time as necessary to complete, but may be limited in the total count of result structures; however, query time and result limits are customizable. Another key feature is the ability to import and export PubChem structure queries to an XML file, which allows one to repeat a particular compound query without filling out the search form again, to share a complex query with a colleague, or to serve as an example for constructing queries for the PubChem Power User Gateway (PUG) interface (described later).

5.3 Structure standardization

Given that PubChem modifies chemical structure information to normalize its representation, it is important for contributors and users of PubChem to explore or understand these changes (e.g., when attempting to integrate external resources with PubChem). With this aim in mind and in the spirit of structure modification transparency, the PubChem chemical structure standardization tool was created. Chemical structure standardization may be directly accessed using the URL:

http://pubchem.ncbi.nlm.nih.gov/standardize/.

This service takes as input a chemical structure and (if standardization is possible) outputs a chemical structure. Allowed structural input and output formats include SMILES, InChI, or SDF file; however, the input and output formats need not be the same. As with structure search, the standardization service is queued on PubChem servers, meaning a request may not start right away or may not complete immediately. One may also import and export standardization requests to a local XML file to serve as an example for constructing queries for the PUG interface (described in detail later).

5.4 Structure downloads

After working with PubChem to achieve a particular subset for a query of interest, it is often important for a user to export resulting substance or compound records from PubChem for further local analysis. The structure download tool prepares PubChem Substance or PubChem Compound records as an export from Entrez in a number of formats. While all PubChem data is available on the PubChem FTP site (via the URL ftp://ftp.ncbi.nlm.nih.gov/pubchem/), being able to interact with a user-selected subset is substantially more convenient. The structure download tool may be directly accessed using the URL:

http://pubchem.ncbi.nlm.nih.gov/pc_fetch/.

Using the download service is straightforward. The user need only perform a search using any combination of Entrez and PubChem-specific search tools, then go to the download page from the PubChem Substance or Compound Entrez DocSum using the download link as indicated by a button with a disk icon. After the user selects an export format, a file containing the exported records will be prepared (on queued PubChem servers, meaning a download may not start or finish immediately) and then served to the user as an URL specifying the download location. It is important to understand that records retrieved from PubChem Substance contain the original deposited information, whereas those from PubChem Compound are standardized forms of the deposited structural information.

A number of formats are available for data export. These formats include SDF, image, small image, SMILES, InChI, XML, and either text or binary ASN.1. The PubChem native archive data format is ASN.1; all other formats are converted from the original ASN.1. The XML formatted data is exactly equivalent to the ASN.1 in content. SDF format is the industry standard for conveyance of chemical structure information and is readily imported into a large number of chemistry programs. Unfortunately, the SDF format is unable to handle all aspects of the ASN.1 data and may not contain all archived information. The PubChem ASN.1 specification, XML schema, and a description of PubChem SDF structure data (SD) tags are all found on the PubChem FTP site in the "specifications" directory.

ASN.1 is a binary format. NCBI utilizes a textual description of ASN.1 that is both computer and human readable (to some extent), but that is not a standard type of ASN.1 data format. This means ASN.1 parsing libraries other than NCBI's may be unable to read it. The PubChem ASN.1 text format does provide a relatively facile means for users to find pertinent information stored in the archive format by simple inspection.

The PubChem download service exports chemical structure images. The images are either 300 × 300 or 100 × 100 pixels in size. The image format is PNG and images are packaged as SID or CID-numbered files in a zip (.zip) archive.

Exports of structural descriptors, SMILES and InChI, provide chemical structure information in a simple tab-delimited text file containing CID or SID and either the isomeric SMILES or InChI strings. Given the very nature of the formats of SMILES and InChI, not all chemical structure information can be identically represented. For example, SMILES encodes only covalent bonds, while PubChem supports the additional concepts of ionic, complex, and dative bonds. Most small molecules in PubChem can be reproducibly interconverted between InChI, SMILES, and PubChem ASN.1 formats without loss of chemical structure information.

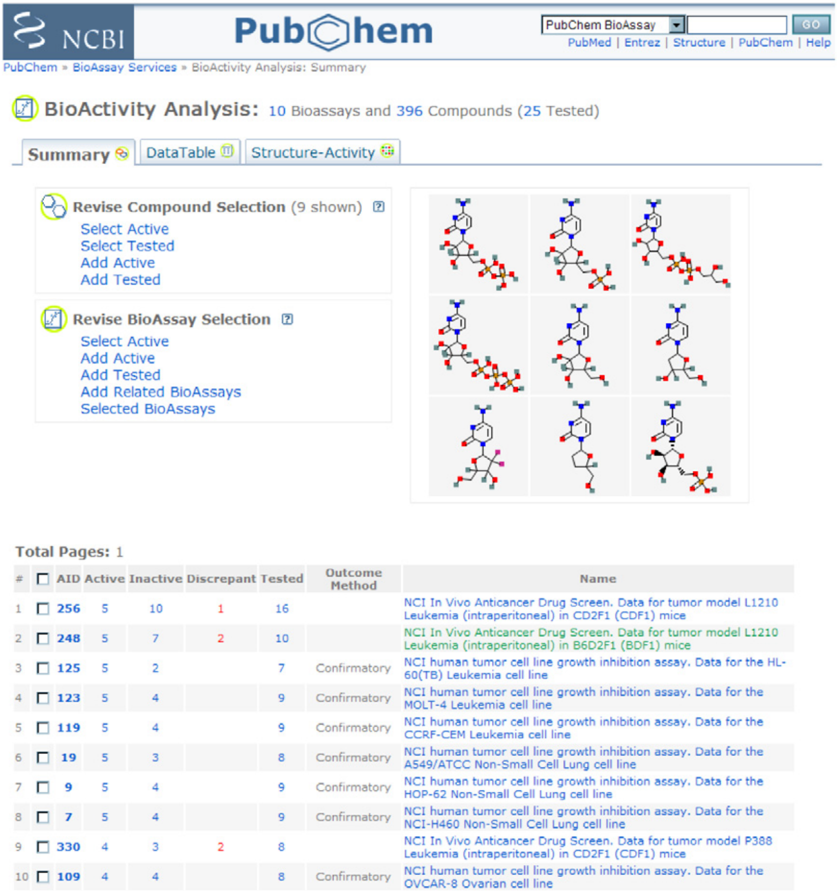
Files may optionally be compressed in standard gzip (.gz) or bzip2 (.bz2) formats. Downloads through the structure download tool are limited to a maximum of 250,000 records per request. Image downloads are limited to 50,000 per request due to the inherent limitations of the zip (.zip) format. As with the other structure tools, the structure download service is accessible using the PubChem Power User Gateway (PUG).

5.5 BioActivity analysis

Beyond a summary description, one would like to view, analyze, and display the actual bioassay data. PubChem provides an integrated suite of tools, each presented as an individual tab, for this purpose. One would use the bioactivity summary tool to, at a glance, be able to examine an overview of the bioassays tested for a list of substances or compounds. To be able to subset and analyze substances or compounds tested in a set of bioassays, one would use the structure-activity analysis tool. To view the actual bioassay outcomes, one would use the data table tool.

5.5.1 BioActivity summary

The BioActivity summary tool is a powerful data analysis tool that provides a comprehensive view of biological activity information available for one or more small molecules. It allows one to compare and examine biological outcome counts across multiple assays, enabling common groups of compounds tested in different assays to be rapidly located (e.g., for Structure–Activity Relationship (SAR) analysis). Furthermore, it allows one to select specific test results to view via the 'Data Table' tab and to perform exploratory data analysis via the 'Structure–Activity' tab. Figure 12.6 depicts an example bioactivity summary.



BioActivity Analysis: 10 Bioassays and 396 Compounds (25 Tested)

Total Pages: 1

#	□	AID	Active	Inactive	Discrepant	Tested	Outcome Method	Name
1	□	256	5	10	1	16		NCI In Vivo Anticancer Drug Screen. Data for tumor model L1210 Leukemia (intraperitoneal) in CD2F1 (CDF1) mice
2	□	248	5	7	2	10		NCI In Vivo Anticancer Drug Screen. Data for tumor model L1210 Leukemia (intraperitoneal) in B6D2F1 (BDF1) mice
3	□	125	5	2		7	Confirmatory	NCI human tumor cell line growth inhibition assay. Data for the HL-60(TB) Leukemia cell line
4	□	123	5	4		9	Confirmatory	NCI human tumor cell line growth inhibition assay. Data for the MOLT-4 Leukemia cell line
5	□	119	5	4		9	Confirmatory	NCI human tumor cell line growth inhibition assay. Data for the CCRF-CEM Leukemia cell line
6	□	19	5	3		8	Confirmatory	NCI human tumor cell line growth inhibition assay. Data for the A549/ATCC Non-Small Cell Lung cell line
7	□	9	5	4		9	Confirmatory	NCI human tumor cell line growth inhibition assay. Data for the HOP-62 Non-Small Cell Lung cell line
8	□	7	5	4		9	Confirmatory	NCI human tumor cell line growth inhibition assay. Data for the NCI-H460 Non-Small Cell Lung cell line
9	□	330	4	3	2	8		NCI In Vivo Anticancer Drug Screen. Data for tumor model P388 Leukemia (intraperitoneal) in CD2F1 (CDF1) mice
10	□	109	4	4		8	Confirmatory	NCI human tumor cell line growth inhibition assay. Data for the OVCAR-8 Ovarian cell line

FIGURE 12.6 Partial view of a BioActivity Summary for cytidine (CID 596) and its 2-D similarity neighbors for all bioassays tested within PubChem.

BioActivity summary provides a set of functions that allows one to revise the substance/compound and assay sets. For example, one may focus only on a subset of compounds that are active in one or more of the selected assays using the 'Compound | Select Active' link, or explore additional screen sets where the given compounds were considered active using the 'Assay | Add Active' link. PubChem provides multiple access points for this service. For compounds or substances tested found in Entrez, one can launch this service for each individual record using the direct "BioActivity Analysis" link, or, for all of the records from an Entrez search, through the launching point at the "Tool" area.

5.5.2 Structure–Activity Analysis

Structure–Activity Analysis is an exploratory tool which performs single linkage clustering analysis for small molecules and their biological screening information in a "heatmap" style display. With this web based tool, a list of assays may be provided and clustered based on activity profile of tested compounds or based on protein target sequence similarity. A set of compounds entered can be clustered either by activity spectrum or 2-D chemical structure similarity. Facilities are provided for navigating between various PubChem web tools and Entrez, and can be accessed throughout the heatmap display. For example, one may identify a compound cluster, click the blue circle near the node of a compound cluster, and select "Compound in Entrez" from the pop-up menu to send the compounds back to Entrez for display. One may also "zoom in" on a sub-region in the heatmap display and request test results generated by multiple screenings for a cluster of compounds, using the embedded tool menu. The service provides various "Revise" functions allowing one to change the selection of compounds or assays. Using the "Revise" function, one can continue the analysis by combining additional screening results. With these versatile functions, the Structure–Activity analysis tool provides a powerful service for iterative analysis of the complex screening data and associated chemical and biological information in PubChem and NCBI resources. An example structure activity analysis is provided in Figure 12.7.

The structure clustering aspect of the Structure–Activity Analysis tool is also available as a separate standalone service for the examination of the similarity of a list CIDs. The tool is called Structure Clustering. Considering the functionality is a subset of the Structure–Activity Analysis tool, it is not described in further detail.

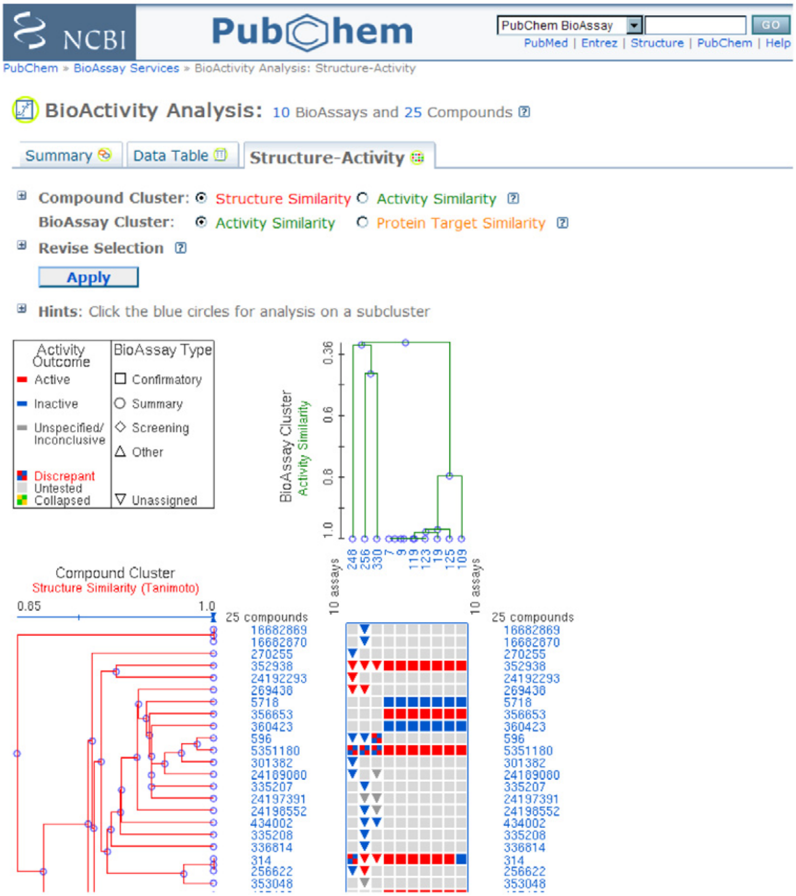
5.5.3 Data Table

To obtain the actual test results in a tabular format, with a single compound or substance per row, one uses the Data Table tool. Able to handle multiple assays and multiple compounds or substances, the Data Table provides various means to "collapse" the data view, including compound (as opposed to substance) specific operations, ignoring or including stereochemistry, and grouping by parent compound (for compound salt form invariance). One may also choose how to handle duplicate and conflicting outcomes resulting from the various methods. Pagination and per column sorting controls are available and all data may be exported in different ways.

The Data Table tool is multifunctional with separate tabs for views of concise (only primary results) or all data. Additional controls for plotting bioassay data columns and for subsetting the displayed data using particular data values or data ranges are provided by the "Plot" and "Select" tabs, respectively. Figure 12.8 depicts an example data table view.

6. PROGRAMMATIC TOOLS

While giving access to all available PubChem data and functionality, interactive web-based interfaces are not particularly well suited to highly repetitive or automated tasks. Without programmatic tools, tasks such as performing specific data lookups for a large number of chemical structures would be tedious if not impossible to perform and a software tool that integrates with PubChem services and data would be difficult to create and maintain. With programmatic access to PubChem,



NCBI PubChem

PubChem BioAssay ▾ GO

PubMed | Entrez | Structure | PubChem | Help

PubChem » BioAssay Services » BioActivity Analysis: Structure-Activity

BioActivity Analysis: 10 BioAssays and 25 Compounds

Summary | Data Table | **Structure-Activity**

Compound Cluster: ● Structure Similarity ○ Activity Similarity

BioAssay Cluster: ● Activity Similarity ○ Protein Target Similarity

Revise Selection

Apply

Hints: Click the blue circles for analysis on a subcluster

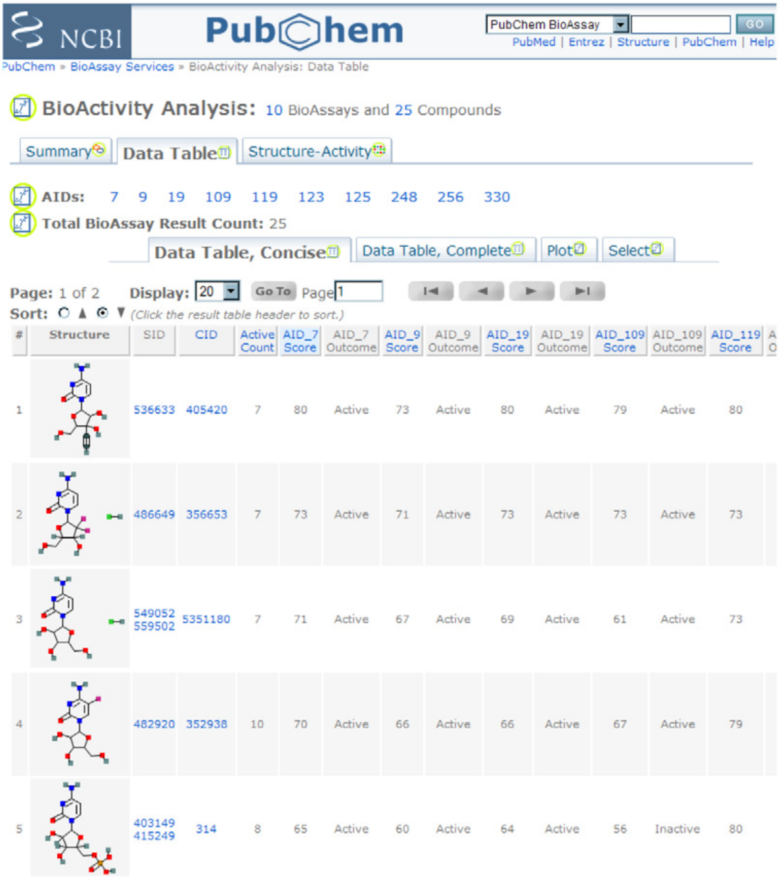
FIGURE 12.7 Partial view of a Structure–Activity Analysis for cytidine (CID 596) and it 2-D similarity neighbors for all bioassays tested within PubChem, with compounds clustered by 2-D structure similarity and assays clustered based on compound biological response.

data can be utilized in more imaginative and complex ways without the need to download the entire PubChem content or to duplicate PubChem functionality.

Two sets of synergistic tools are available for programmatic access to PubChem data, Entrez utilities [21] (eUtils) and the PubChem Power User Gateway (PUG). Entrez-based access is achieved through the use of eUtils. To provide access to the capabilities of PubChem tools, PUG is available. Together, these two facilities enable users to interact with PubChem using XML over HTTP.

6.1 Entrez utilities

Entrez has a suite of associated tools, collectively called eUtils. Together, these tools provide access to nearly all Entrez functionality, primarily through an XML



PubChem » BioAssay Services » BioActivity Analysis: Data Table

BioActivity Analysis: 10 BioAssays and 25 Compounds

AIDs: 7 9 19 109 119 123 125 248 256 330

Total BioAssay Result Count: 25

Page: 1 of 2 Display: 20 Go To Page 1

Sort: ○ ▲ ● ▼ *(Click the result table header to sort.)*

#	Structure	SID	CID	Active	AID_7	AID_7	AID_9	AID_9	AID_19	AID_19	AID_109	AID_109	AID_119	A
				Count	Score	Outcome	Score	Outcome	Score	Outcome	Score	Outcome	Score	O
1		536633	405420	7	80	Active	73	Active	80	Active	79	Active	80	
2		486649	356653	7	73	Active	71	Active	73	Active	73	Active	73	
3		549052 559502	5351180	7	71	Active	67	Active	69	Active	61	Active	73	
4		482920	352938	10	70	Active	66	Active	66	Active	67	Active	79	
5		403149 415249	314	8	65	Active	60	Active	64	Active	56	Inactive	80	

FIGURE 12.8 Partial view of a concise Data Table for cytidine (CID 596) and it 2-D similarity neighbors for all bioassays tested within PubChem.

over HTTP interface. These tools are described in detail elsewhere:

http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html.

The primary eUtil tools of most interest to PubChem users are eSearch, eFetch, ePost, eLink, eHistory, and eInfo. eSearch performs an Entrez search, with the same query syntax as web-based Entrez queries (e.g., to query PubChem Compound for the chemical name "aspirin"). eFetch returns an ID list from a prior search (e.g., the list of PubChem Compound identifiers (CIDs) from the aforementioned query of "aspirin"). ePost creates a new ID list by upload of a list of identifiers (e.g., substance identifiers (SIDs)). eLink follows a given link type to create a new ID list from an existing one (e.g., to find all PubChem BioAssay identifiers (AIDs) associated with a list of SIDs). eHistory returns information on current Entrez History entries. eInfo lists available Entrez indices and links for a given database.

Each of these eUtils applications can return data in XML format for automatic parsing by a script or application. Most eUtil tools have the option to use an Entrez history key, which include a web-environment (WebEnv) and Entrez history item (query_key) as arguments, as input or output. This enables Entrez history to store sets of identifiers temporarily, relieving the user's application of the burden of continually sending and receiving potentially large ID lists. The XML specification, in DTD form, for each eUtil tool can be found at the URL:

http://eutils.ncbi.nlm.nih.gov/entrez/query/DTD/index.html.

6.2 PUG

PubChem's Power User Gateway (PUG) is a single entry point to a vast array of PubChem functionality. It is not necessarily intended for the casual user, but rather for those who are seeking a low-level interface access to PubChem. Outlined here, PUG is documented with examples at:

http://pubchem.ncbi.nlm.nih.gov/pug/pughelp.html.

The basic design of PUG is simple, a central gateway to multiple PubChem functions. PUG does not take URL arguments. All communication with PUG is through XML over HTTP. To perform any request, one formulates input in XML and then sends it to PUG via an HTTP POST. PUG interprets the incoming request, initiates the appropriate action, and then returns results in XML format. With this design, PUG may be used with any scripting or programming language that has the ability to read and write XML, and to send and receive data via HTTP. The XML specification for the XML used by PUG may be found, in DTD and XML schema forms, respectively, at:

http://pubchem.ncbi.nlm.nih.gov/pug/pug.dtd,

http://pubchem.ncbi.nlm.nih.gov/pug/pug.xsd.

Either of these specifications may be used to guide the creation of valid input XML to send to PUG and to parse the returned results. Because PUG encompasses a wide variety of functions, its XML structure is necessarily complex. It may be easier to create input XML data with the help of a tool that can generate program code from or at least validate XML using a DTD or schema.

PubChem tools for structure search, standardization, and downloads are enabled via PUG, with more to be added. In each case, the options available through PUG are the same as those available through the interactive web pages, including all the advanced options and filters of the structure search service. In fact, most of the web tools can write out queries in PUG's XML format, which can be sent directly to PUG or used as templates for constructing new PUG requests.

As with the web-based tools, requests through PUG may be queued on PubChem servers. Thus, PUG may not deliver an answer directly in response to the initial request. Rather, for cases where execution may take some time, PUG will return a waiting message, along with a request identifier which is used to poll

PUG periodically for the status of that request. PUG responds with another waiting message if the operation is still in progress, an error message if it failed, or a success message with the final results, when the task is finished. It is up to the PUG user to add a periodic status check loop to handle these queued requests properly.

The combination of PUG and Entrez eUtils opens up a wide spectrum of programmatic tasks that can harness the true power of PubChem inside custom applications. An advantage to this approach, compared to having a local copy of PubChem data on the user's computer, is that the mass of PubChem data and complexity of the analysis functions are all maintained by PubChem, thus, the CPU cycles needed to perform the tasks are hosted by PubChem. The user needs only this basic interface to access PubChem infrastructure, at the relatively small investment of a little programming.

7. DEPOSITION SYSTEM

PubChem is an open repository. Organizations may contribute information about small molecules and integrate their public resource with PubChem, in part by providing URLs back to and from their website to PubChem. The types of PubChem depositors are greatly varied with contributors from government organizations, academic groups, chemical reagent and screening library suppliers, scientific journals, scientific data publishers, physical property databases, and more. To handle this quantity and diversity of data, PubChem created an on-line data deposition system for rapid contribution of substance and bioassay data. This system may be accessed via the URL:

http://pubchem.ncbi.nlm.nih.gov/deposit/.

Any organization may become a PubChem contributor. The deposition system allows potential depositors to obtain a test account quickly, to examine how their data will look in PubChem and to gain familiarity with the user interface. A test account is nearly identical to a deposition account except data cannot be added to PubChem when using a test account. To actually put data into PubChem, potential depositors must apply for a deposition account. Deposition accounts require a click-thru data transfer agreement that must be agreed upon prior to allowing data to be contributed. Essentially, this agreement enables the depositor to retain all rights to their information while allowing PubChem to display and distribute any provided information.

Deposition of substance information is performed using the industry standard SDF format, which may include using the SMILES or InChI formats as the chemical structure. Depositing properly formatted substance data into PubChem is as simple as uploading a file, via HTTP or FTP.

Deposition of assay information is performed in two parts. Creation of a new assay involves providing a description, protocol, target, readouts, and other associated information using a web-form or via an XML file. After the assay description is completed, assay test results can be readily provided by using the standard CSV (i.e., comma delimited) file format. Assays provide outcomes for substances.

As such, PubChem requires these substances to be available in PubChem prior to providing respective assay information.

Once data is put into PubChem, depositors may update their information at any time. Updates to existing PubChem records cause versioning to occur. PubChem is archival, in that retention of previous versions of records allows PubChem users to access a particular version of a substance or bioassay record, regardless of its revision history. It should be noted; however, that older version information is not presented by default.

Bioassays have two levels of versioning, being major and minor updates. Minor bioassay versions indicate changes to the bioassay textual description. Major bioassay versions indicate addition or reduction in the count of readouts. Major bioassay versioning requires all bioassay data to be completely restated by the depositor, considering the readouts changed in some way. Bioassay records also have substance-level outcome versioning. If a bioassay substance outcome is provided more than once by a depositor, previous reported results are versioned.

8. FUTURE DIRECTIONS

Expansion and enrichment of the bioassay data are ongoing, by adding annotations for small molecules and drugs using publicly available information, such as that provided at the National Library of Medicine (NLM) or the Food and Drug Administration (FDA). With efforts from the scientific community, bioassay data is becoming better annotated by linking target to protein classification resources or molecular pathway information. With further integration with NCBI resources such as PubMed and the Entrez search system, information contained within PubChem will become more discoverable and useful to a broader audience of scientists worldwide.

PubChem currently provides 2D-based data analysis and clustering tools. Small molecules are not flat. They have a rich diversity of 3-D shapes and 3-D orientation of features possible. Addition of a theoretical 3-D description of the PubChem Compound database may open new avenues in the understanding of bioassay outcomes by allowing combination of 2-D and 3-D data analysis and clustering techniques, thus enabling improved hypothesis generation and trend recognition implicit with a biological dataset. Neighboring 3-D descriptions of PubChem Compound, much like the 2-D similarity neighbors currently available, may help scientists identify and better understand interrelationships of the biological properties of small molecules. It is, to this end, that a 3-D description of the PubChem Compound database is in progress.

Programmatic access to PubChem using work-flow automation software (such as Taverna [22] and Pipeline Pilot [23]) and scripting languages (such as Python [24], Ruby [25] and PERL [26]), may enable researchers to make exciting new discoveries and to further leverage and integrate PubChem into their basic research. New interfaces using SOAP-based web services (via WSDL [27]) are in the making, to make access to PubChem easier and conceptually simpler to achieve. For those who would rather learn directly about the inner workings of PubChem data processing and analysis, a C++ API, based on the NCBI C++ toolkit [28], will be made available.

PubChem is a significant source of information on the biological properties of small molecules. The offering of tools and services associated with the access and mining of this data makes PubChem important to the work of scientists worldwide as an enabling resource. PubChem continues to grow and evolve as a function of time. New tools and services are in development and existing offerings are being refined. Feedback from the user community is an important and welcome part of this process to ensure the utility of PubChem to the community is maximized. The NCBI help desk (email: info@ncbi.nlm.nih.gov) is the primary locus for such input.

ACKNOWLEDGMENTS

This research was supported, in part, by the Intramural Research Program of the NIH, National Library of Medicine.

REFERENCES

1. http://pubchem.ncbi.nlm.nih.gov.
2. http://nihroadmap.nih.gov/molecularlibraries/.
3. Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., Wheeler, D. GenBank. Nucleic Acids Res. 2007, 35, D21–5.
4. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M. The Swiss-Prot Protein Knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 2003, 31, 365–70.
5. Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., Bourne, P. The Protein Data Bank. Nucleic Acids Res. 2000, 28, 235–42.
6. http://www.ncbi.nlm.nih.gov.
7. http://www.ncbi.nlm.nih.gov/sites/gquery.
8. http://pubmed.gov.
9. Chen, J., Anderson, J., DeWeese-Scott, C., Fedorova, N., Geer, L., He, S., Hurwitz, D., Jackson, J., Jacobs, A., Lanczycki, C., Liebert, C., Liu, C., Madej, T., Marchler-Bauer, A., Marchler, G., Mazumder, R., Nikolskaya, A., Rao, B., Panchenko, A., Shoemaker, B., Simonyan, V., Song, J., Thiessen, P., Vasudevan, S., Wang, Y., Yamashita, R., Yin, J., Bryant, S.H. MMDB: Entrez's 3D-structure database. Nucleic Acids Res. 2003, 31, 474–7.
10. OEChem, version 1.5.1, OpenEye Scientific Software, Inc., Santa Fe, NM, USA, http://www.eyesopen.com, 2007.
11. Stein, S., Heller, S., Tchekhovskoi, D. An Open Standard for Chemical Structure Representation: The IUPAC Chemical Identifier. Proceedings of the 2003 International Chemical Information Conference (Nimes), Infonortics 131–43. http://www.iupac.org/inchi/.
12. Lexichem, version 1.6, OpenEye Scientific Software, Inc., Santa Fe, NM, USA, http://www.eyesopen.com, 2007.
13. http://dtp.nci.nih.gov/webdata.html.
14. Richard, A., Williams, C. Distributed Structure-Searchable Toxicity (DSSTox) public database network: A proposal. Mutat Res. 2002, 499, 27–52. http://www.epa.gov/nheerl/dsstox/.
15. Liu, T., Lin, Y., Wen, X., Jorrisen, R., Gilson, M. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res. 2007, 35(Database Issue), D198–201. http://www.bindingdb.org/.

16. Lipinski, C., Lombardo, F., Dominy, B., Feeney, P. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Del. Rev. 2001, 46, 3–26.
17. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.
18. Dalby, A., Nourse, J., Hounshell, W., Gushurst, A., Grier, D., Leland, B., Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. J. Chem. Inf. Comput. Sci. 1992, 32, 244–55.
19. Tanimoto T. IBM Internal Report 17th Nov. 1957.
20. Durant, J., Leland, B., Henry, D., Nourse, J. Reoptimization of MDL keys for use in drug discovery. J. Chem. Inf. Comput. Sci. 2002, 42, 1273–80.
21. http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html.
22. http://taverna.sourceforge.net/.
23. http://accelrys.com/products/scitegic/.
24. http://www.python.org/.
25. http://www.ruby-lang.org.
26. http://www.perl.org/.
27. http://www.w3.org/TR/wsdl.
28. http://www.ncbi.nlm.nih.gov/IEB/ToolBox/CPP_DOC/.