Published in final edited form as:

Methods Mol Biol. 2016 ; 1473: 153–159. doi:10.1007/978-1-4939-6346-1_16.

Accessing the High Throughput Screening Data Landscape

Daniel P. Russo¹ and Hao Zhu^{1,2,*}

¹Rutgers Center for Computational and Integrative Biology, Rutgers University-Camden, NJ 08102, USA

²Department of Chemistry, Rutgers University-Camden, NJ 08102, USA

Abstract

The progress of high throughput screening (HTS) techniques is changing the chemical data landscape by producing massive biological data from tested compounds. Public data repositories (e.g., PubChem) receive HTS data provided by various institutes and this data pool is being updated on a daily basis. The goal of these data sharing efforts is to let users quickly obtain the biological data of target compounds. Without a universal chemical identifier, the repositories (e.g. PubChem) provide users various methods to query and retrieve chemical properties and biological data by several different chemical identifiers (e.g., SMILES, InChIKey, IUPAC name, etc.). The major challenge for most users, especially computational modelers, is obtaining the biological data for a large dataset of compounds (e.g. thousands of drug molecules) instead of a single compound. This chapter aims to introduce the steps to access the public data repositories for target compounds with specific emphasis on the automatic data downloading for large datasets.

Keywords

Compounds; Chemical identifier; Biological data; PubChem

1. Introduction

In the past decade, the use of high throughput screening (HTS) in drug discovery and chemical toxicity evaluations has greatly facilitated the progress of cheminformatics studies. The massive data generated from HTS studies provides scientists a new vision of the biological effects induced by the compounds being tested. As a consequence, data sharing efforts to make these data easily available to communities have been undertaken within the same period. Data repositories such as PubChem, ChEMBL, BindingDB, and the Comparative Toxicogenomics Database (CTD), have become daily-used scientific tools for many scientists. With a chemical identifier (e.g. InChIKey), the biological data for a target compound can be obtained by clicking a button on the web based search portals of these repositories.

*Corresponding author: Hao Zhu, 315 Penn St., Department of Chemistry, Rutgers Center for Computational and Integrative Biology, Rutgers University-Camden, NJ 08102, hao.zhu99@rutgers.edu.

The PubChem project (<https://pubchem.ncbi.nlm.nih.gov>), initiated and hosted by the National Center for Biotechnology Information (NCBI) of the National Institutes of Health (NIH), is the largest public chemical data source. The goal of PubChem is to provide information for the biological activities of small molecules, and as of September 2015, it has received over 60 million unique chemical structures and 1 million biological assays from over 350 contributors [1]. The PubChem data repository consists of three primary databases: Substance, Compound, and BioAssay. The PubChem Substance database, indexed by the PubChem substance identifier (SID), contains chemical structures, synonyms, registration IDs, descriptions, related links, database cross-reference links to PubMed, protein 3D structures, and biological screening results. The PubChem BioAssay Database contains experimental testing results of the chemical substances described within the PubChem Substance database. Each HTS assay has a unique assay identifier (AID) in this database. The PubChem Compound database contains validated chemical depiction information that is provided to describe substances in PubChem Substance database. The chemical identifier, PubChem compound ID (CID), records unique structural information and, in turn, allows target compounds to be queried within the PubChem Compound database. Compound records are supplemented with textual structural identifiers such as Simplified Molecular-Input Line-Entry Systems (SMILES), the IUPAC International Chemical Identifier (InChI), and a more compact version, the InChI key[2–5].

PubChem HTS data were obtained from various sources including university, industry or government laboratories. One of the initial missions of PubChem is to function as the repository to host the data generated by the HTS projects supported by the NIH's Molecular Libraries Program. A full listing of data sources can be found at <http://pubchem.ncbi.nlm.nih.gov/sources#assay>. The types of HTS data from these sources include results obtained from binding assays, functional cell assays, and Absorption, Distribution, Metabolism, Exclusion and Toxicity (ADMET) assays. The HTS data points can be qualitative (e.g., active or inactive), quantitative (e.g., the half-maximal effective concentration of a drug), or both. PubChem stores the HTS data of a compound in two fields: the *activity outcome* and the *active concentration*. The *activity outcome* either identifies the relevant compound as a chemical probe (i.e., a positive control of a HTS assay) or qualitatively transforms the experimental data into one of the following categories: active, inactive, unspecified/inconclusive, or untested. On the other hand, the *active concentration* stores the HTS data quantitatively as a concentration value in μM unit as well-defined biological endpoints, such as the half-maximal activity response (e.g., IC₅₀, EC₅₀, etc.).

There are two methods to obtain HTS data by accessing the data sharing repositories, such as PubChem. The data can be obtained by querying manually with individual compounds' textual chemical identifiers. However, if the goal is to download all relevant HTS data for a large set of compounds, automatic data extraction is needed. This chapter will use PubChem as an example to show how to obtain HTS data for target compounds, especially for a large set of compounds.

2. Materials

To extract the HTS data for target compounds via public data repositories (i.e. PubChem), the following software needs to be downloaded/installed on the computer:

- A web browser (e.g., Mozilla FireFox, Google Chrome, Microsoft Internet Explorer, Apple Safari)
- Microsoft Excel® or other spreadsheet program
- A programming package (e.g., Java, Python, Perl, C#)
- A file archiver that supports .gz decompression such as WinZip or 7-zip (Windows users only)

3. Methods

3.1 Accessing HTS data manually through the PubChem portal

Similar to popular internet search portals (e.g. Google®), PubChem provides users a manual search function by which queries can be made using various chemical identifiers. Each unique compound in the PubChem Compound database has an individual page listing standardized chemical information and properties, including a list of all submitted biological testing results. For a target compound existing in PubChem database, its biological testing data can be exported and downloaded as a comma-separated values (CSV) file and managed using Microsoft Excel®. Figure 1 shows the screenshot of the plain text file for the biological data for aspirin, with PubChem CID 2244 (downloaded from PubChem on February 15, 2016). The biological data of a compound is summarized by including not only the bioassay identifier (AID) and the associated testing results, but also detailed information of the bioassays and the definitions of the activities. This file can be obtained by inputting various identifiers of aspirin to their appropriate categories. Figure 2 shows a screenshot of the homepage to the PubChem search function. More information on the appropriate search option for a given identifier can be found in the Notes section. The biological data of a single target compound can be accessed by the following steps:

- Step 1** Open a web browser and visit the PubChem Compound search tool at: <https://pubchem.ncbi.nlm.nih.gov/search/search.cgi>.
- Step 2** Select the appropriate search tab.
- Step 3** Enter the correct information (e.g. chemical name as shown in Figure 2) and click “Search”. Using a unique identifier (e.g., PubChem CID) will result in the desired compound. Otherwise, manually analyzing the search results (i.e. a list of compounds containing the input information) is required.
- Step 4** From the compound summary page, scroll down to “BioAssay Results”. Click “Refine/Analyze” and select “Go To Bioactivity Analysis Tool” from the pull-down menu.
- Step 5** On the Bioactivity Analysis Tool page, click “Download Table”.

The resulting bioassay information for that compound will be automatically retrieved as a plain text file.

3.2 Retrieving PubChem HTS data through Web Services

If the goal is to download the HTS data for a large dataset (e.g. consisting of more than 1,000 compounds), automatic querying is needed by executing a coding script. To this end, PubChem offers specialized data retrieval services through a programmatic interface: PubChem Power User Gateway (PUG). The PUG provides quick access to PubChem data retrieval functions. Information on all the available PUG services can be found in the reference [6] as well as within the PubChem portal (<https://pubchem.ncbi.nlm.nih.gov/pug/pughelp.html>). The most broadly applicable function to retrieve HTS data for large chemical dataset is PUG-REST. PUG-REST, which uses a Representational State Transfer (REST)-style interface, allows users to construct Uniform Resource Identifiers (URLs) to retrieve data from PubChem. Through this way, PUG-REST is easily integrated with all programming/scripting languages that can post URLs (e.g., Java, Python, Perl, C#). Using PUG-REST, multiple records can be accessed automatically to fulfill the request of retrieving HTS data for large datasets. More information on all the features available through PUG-REST can also be found on the PubChem portal (https://pubchem.ncbi.nlm.nih.gov/pug_rest/PUG_REST.html).

To construct a URL for PUG-REST data retrieval, a text URL string needs to be created and it contains four parts: base, input, operation, and output. The construction of this kind of URL is shown in Figure 3. The input section of the URL describes the target database (BioAssay, Compound, or Substance) to be queried, the category of the identifier, along with the identifier information. The operation section designates the information to be retrieved (“assaysummary” in this case to retrieve HTS data). The output section specifies the format of the output file. Figure 3 also shows several examples of URLs used to retrieve HTS data through PUG-REST.

Inputting a constructed URL into a web browser will result in the display of bioassay information for this target compound in the desired format. Within a programming script, individual URLs can be constructed for each compound in a large dataset in an automated fashion. In this case, the HTS data for all compounds in this dataset will be retrieved in the preferred output format.

3.3 Downloading master HTS database from PubChem

While the PubChem database is exceptionally large, sometimes it is necessary to transfer all the HTS data from PubChem to a local server for further analysis. File Transfer Protocols (FTP), which are common tools to share/transfer files over the internet, can be used to realize this goal. PubChem offers the download of all three databases through an FTP site. Most operating systems (e.g., Windows, Mac, Linux) support access to FTP sites, allowing for easy file transfers from the data server (e.g., PubChem) to the user’s local computer. Using an FTP, the entire PubChem BioAssay database can be accessed and downloaded in four formats: Abstract Syntax Notation (ASN), CSV, JavaScript Object Notation (JSON), and Extensible Markup Language (XML). The overall BioAssay database is large and

requires all entries to be distributed into compressed folders (in .zip format) with each folder containing a maximum of 1,000 assays and their associated assay data. Folders are named after the set of AIDs it contains, where an individual file in these folders is named after the AID information it contains. For example, within the folder “0000001_0001000” all the HTS data corresponding to AIDs 1–1000 can be found. Likewise, the file in this folder, “1.xml”, contains data from assay with AID 1. For storage efficiency, these files are compressed using the GZip algorithm (.gz format). While both Linux and Mac users will find support for GZip decompression, Windows users are required to download a file archiver program that supports GZip decompression (e.g., WinZip or 7zip). The following steps are needed to access these data:

Step 1 Connect to the PubChem FTP site containing the BioAssay database files.

For Windows users: open a Windows Explorer window by clicking the “Start” button and then clicking “Computer”. In the address bar, type <ftp://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay/>.

For Mac users: Click “Go” then click “Connect to Server...”. In the “Server Address” field type: <ftp://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay/>. When asked to enter name and password, choose to connect as “Guest”.

For Linux users: Open any file manager window. Click “File” and then click “Connect to Server...”. In the “Server” field type <ftp://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay/>.

Step 2 Select the folder containing the desired format. For example, the “XML” folder contains the bioassay database in XML format.

Step 3 Copy the compressed folders from to the desired directory on the local computer.

It is important to realize that PubChem updates its databases frequently.

3. Notes

The PubChem search portal accepts a variety of search options. For example, exact queries (under the “Name/Text” tab) can be performed based on general identifiers, such as CID, IUPAC name, etc.. Alternatively, queries can also be performed based chemical structure information. Under the “Identity/Similarity” search option, structural identifiers (i.e. SMILES or InChI) can be used. Commonly used chemical structure file formats, such as Structure Data File (SDF) format, can also be accepted under this search tab. If no identifier is available for a compound, there is also an option to draw the chemical structure using the drawing tool.

4. Conclusions

HTS data can be accessed through public repositories for target compounds. Since there are several existing methods to obtain HTS data, the individual researcher can choose the suitable way to get the HTS data based on the actual needs (e.g. the size of the target dataset). Meanwhile, the HTS data retrieval procedure needs necessary computer skills and

background knowledge of the data sources, such as what have been described for PubChem data acquire in this chapter.

References

1. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, et al. PubChem Substance and Compound databases. *Nucleic Acids Res.* 2015 gkv951.
2. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci.* 1988; 28:31–36.
3. Weininger D, Weininger A, Weininger JL. SMILES. 2. Algorithm for generation of unique SMILES notation. *J Chem Inf Comput Sci.* 1989; 29:97–101.
4. Weininger D. SMILES. 3. DEPICT. Graphical depiction of chemical structures. *J Chem Inf Comput Sci.* 1990; 30:237–243.
5. Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I. InChI - the worldwide chemical structure identifier standard. *J Cheminformatics.* 2013; 5:7.
6. Kim S, Thiessen PA, Bolton EE, Bryant SH. PUG-SOAP and PUG-REST: web services for programmatic access to chemical information in PubChem. *Nucleic Acids Res.* 2015; 43:W605–W611. [PubMed: 25934803]

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	AID	Panel ID	SID	CID	Activity	AC Value (micromolar)	AC Name	BioAssay Name	GI	Target Name	Outcome Method	PubMed ID	GeneID	Target Count (active)	Target Count (tested)
2	157039	0	103164874	2244	Unspecified	7.76E-10	IC50	Antiaggregatory effect against arachidonic acid induced platelet aggregation in the human platelet rich plasma (PRP)			Confirmatory	11585451		12	541
3	89416	0	103164874	2244	Unspecified	7.00E-07	IC50	Inhibition of collagen induced platelet aggregation in platelet-rich human plasma (PRP)			Confirmatory	1172124		12	541
4	376764	0	103164874	2244	Unspecified	49700	IC50	Inhibition of sheep placental COX2 by liquid scintillation counter			Confirmatory	15679320		12	541
5	311317	0	103164874	2244	Unspecified	20000	Ki	Inhibition of cyclooxygenase			Confirmatory	17341061		12	541
6	376763	0	103164874	2244	Unspecified	18800	IC50	Inhibition of COX2 from ram seminal vesicles by liquid scintillation counter			Confirmatory	15679320		12	541
7	162666	0	103164874	2244	Unspecified	18000	IC50	Tested for inhibitory activity against Prostaglandin G/H synthase 2 from human			Confirmatory	7562922		12	541
8	349142	0	103164874	2244	Unspecified	7530	Ki	Inhibition of human carbonic anhydrase 1 esterase activity by noncompetitive Lineweaver-Burke plot	115449	Carbonic anhydrase 1	Confirmatory	18819808	759	12	541
9	625305	0	103164874	2244	Unspecified	6000	IC50	Antiinflammatory activity against TNF-alpha induced VCAM-1 protein expression in MAEC by ELISA			Confirmatory	21955479		12	541
10	399402	0	103164874	2244	Unspecified	5000	IC50	Inhibition of sheep placental cyclooxygenase COX2 assessed as PGE2 production	3914304	Prostaglandin G/H synthase 2	Confirmatory	9462646	443460	12	541

Figure 1.
Example of the 10 biological testing results for aspirin (PubChem CID 2244) downloaded in plain text format.

NCBI
PubChem
Structure Search

PubChem Compound Search

Limits Advanced search

Search By: Name/Text Identity/Similarity Substructure/Superstructure Molecular Formula 3D Conformer

Synonyms/Descriptors/MeSH term etc: Search

Clear

Figure 2.
The PubChem search tool interface as of February, 2016.

Figure 3.
Various queries to retrieve PubChem HTS data for aspirin using PUG-REST.