Module 4 Assignment: Understanding Public Chemical Databases

PART I. Getting familiar with PubChem

1. Read these articles and answer the following questions.

- What is the difference between a substance and a compound in PubChem? (https://pubchemblog.ncbi.nlm.nih.gov/2014/06/19/what-is-the-difference-between-a-substance-and-a-compound-in-pubchem/)
- Compound Summary Page Redesigned (https://pubchemblog.ncbi.nlm.nih.gov/2014/10/20/compound-summary-page-redesigned/)
- Substance Record Page Released (https://pubchemblog.ncbi.nlm.nih.gov/2015/04/09/substance-record-page-released/)
- PubChem adds a “legacy” designation for outdated data (https://pubchemblog.ncbi.nlm.nih.gov/2015/11/16/pubchem-adds-a-legacy-designation-for-outdated-data/)
- “§2.4. Availability of compounds for subsequent experiments” in “Getting the most out of PubChem for virtual screening” (http://www.tandfonline.com/doi/full/10.1080/17460441.2016.1216967) [If you don’t have access to this article, Author’s original manuscript for this paper is available as an attachment at the end of Module 4.]

(a) Explain the difference between the PubChem Substance and Compound databases in two or three sentences.

(b) Explain what the Compound Summary page of a compound is.

(c) Explain what the Substance Record page of a substance is.

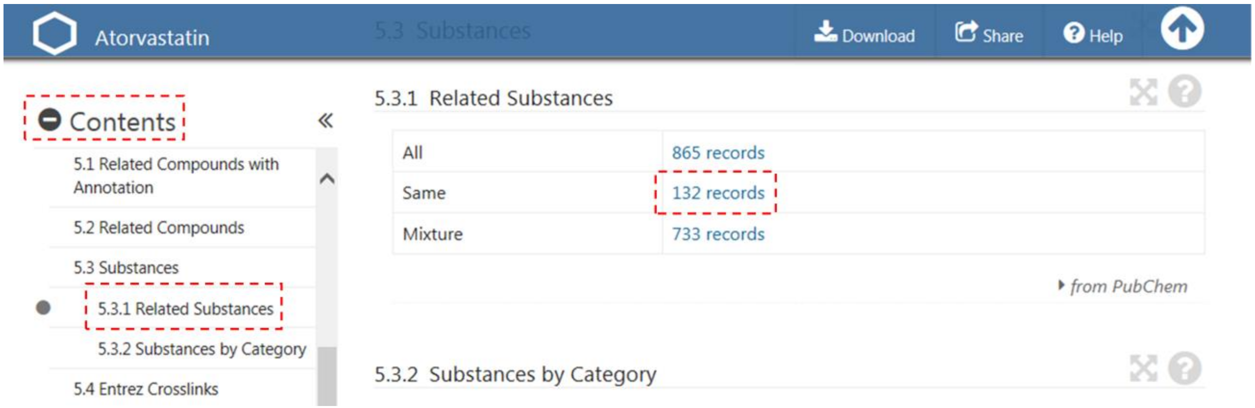
(d) Explain the reason why the “legacy” designation was introduced in PubChem in two or three sentences.

(e) Among the menus available on the top of the PubChem home page (https://pubchem.ncbi.nlm.nih.gov) is "Today's Statistics". The number of compounds/substances/assays shown under this menu does not include "non-live" records. What does "non-live" mean here?

2. While the PubChem Substance database is an archive in nature, data providers often want to update their substance information archived in PubChem. For this reason, PubChem keeps all different "versions" of a substance record and shows the most recent version on its Substance Record page by default (Click **here** to read about what the Substance Record page is). Go to the PubChem home page (https://pubchem.ncbi.nlm.nih.gov) and follow the steps described below.

(a) After selecting the "Compound" tab above the search box, type "60823" in the search box and click the "Go" button. This will direct you to the Compound Summary page for CID 60823 (atorvastatin). (Click **here** to read about what the Compound Summary page is.) [You will learn how to search PubChem in much more detail for next two modules (Modules 5 and 6).]

(b) Scroll down until you see "Contents" on the left column. Expand this table of contents by clicking the "+" sign before "Contents". Locate the "Related Substances" section and click the record count for the "Same" item under that section.



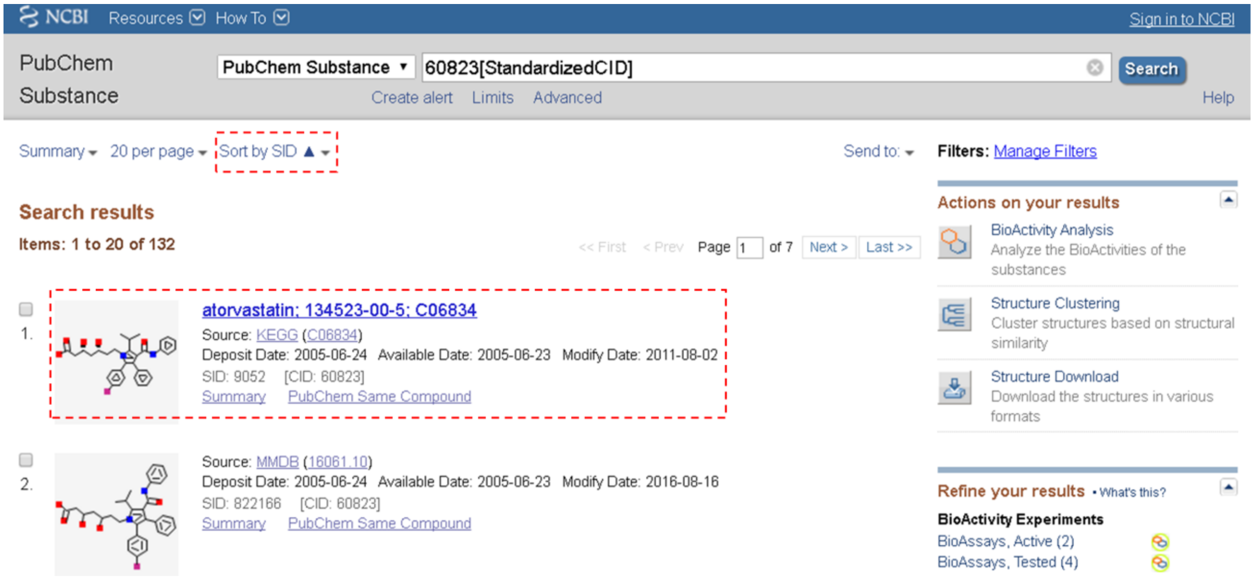
5.3.1 Related Substances

All	865 records
Same	132 records
Mixture	733 records

from PubChem

5.3.2 Substances by Category

(c) The previous step will lead you to the web page that presents 132 substance records associated with CID 60823 (atorvastatin). This page is called a Document Summary (DocSum) page, because it presents a (very brief) summary of retrieved records. Sort the list by SID (in ascending order) and click the substance that appear on the top of the DocSum page (that is, SID 9052).



60823[StandardizedCID]

Search results

Items: 1 to 20 of 132

1. atorvastatin; 134523-00-5; C06834
Source: KEGG (C06834)
Deposit Date: 2005-06-24 Available Date: 2005-06-23 Modify Date: 2011-08-02
SID: 9052 [CID: 60823]
Summary PubChem Same Compound

2. Source: MMDB (16061.10)
Deposit Date: 2005-06-24 Available Date: 2005-06-23 Modify Date: 2016-08-16
SID: 822166 [CID: 60823]
Summary PubChem Same Compound

Actions on your results

BioActivity Analysis
Analyze the BioActivities of the substances

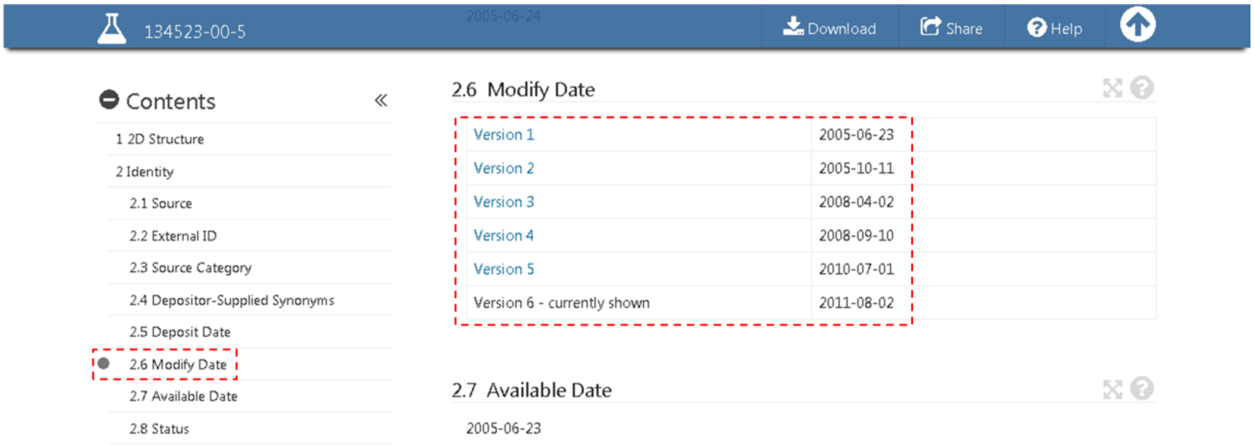
Structure Clustering
Cluster structures based on structural similarity

Structure Download
Download the structures in various formats

Refine your results • What's this?

BioActivity Experiments
BioAssays, Active (2)
BioAssays, Tested (4)

(d) Clicking SID 9052 directs you to the Substance Record page for SID 9052. From the Table of Contents, locate the "Modify Date" section. Now you see a table that shows a list of the dates when this substance record was modified.



Contents

1 2D Structure
2 Identity
2.1 Source
2.2 External ID
2.3 Source Category
2.4 Depositor-Supplied Synonyms
2.5 Deposit Date
2.6 Modify Date
2.7 Available Date
2.8 Status

2.6 Modify Date

Version 1	2005-06-23
Version 2	2005-10-11
Version 3	2008-04-02
Version 4	2008-09-10
Version 5	2010-07-01
Version 6 - currently shown	2011-08-02

2.7 Available Date

2005-06-23

(e) Click each version of this record and fill in the table below with the information under the "Depositor Comments" and "Cross-References" sections.

Version 1	Comments	None
	RegID	C06834
	RN	134523-00-5
	DBURL	http://www.genome.ad.jp/kegg/kegg2.html
	SBURL	http://www.genome.ad.jp/dbget-bin/www_bget?cpd:C06834
Version 2	Comments	
	RegID	
	RN	
	DBURL	
	SBURL	
Version 3	Comments	
	RegID	
	RN	
	DBURL	
	SBURL	
Version 4	Comments	
	RegID	
	RN	
	DBURL	
	SBURL	
Version 5	Comments	
	RegID	
	RN	
	DBURL	
	SBURL	

Version 6	Comments	Same as: D07474
	RegID	None
	RN	None
	DBURL	None
	SBURL	None

(f) The “Depositor Comments” section of version 6 of this record has a link to another substance record in PubChem. What is the SID of this substance record? What is the CID of the compound associated with this substance record? Is this CID the same as the one used in (a)?

(g) Briefly summarize (in three or four sentences) how depositor-provided information on SID 9052 has evolved over time.

3. Some records in PubChem are “non-live”, meaning that they are “not searchable”, although they do exist in the database. This exercise is designed to help students better understand what non-live records are.

(a) In the previous exercise, you searched the PubChem Compound database for CID “60823”. Now search the Compound database using the query “73336290”. What message do you get?

(b) Repeat the search using the query “60823” (which we already know returns a hit) to go to the Compound Summary page for CID 60823. What is the URL of this page (i.e., the web address)?

(c) Replace the string “60823” in the URL with “73336290” and press the “Enter” key on the keyboard. This will get you to the Compound Summary page of CID 73336290, which you were not able to find in (a). Indeed, this page presents the message “NOTE: NON-LIVE RECORD. See the related substances for more information.” How many substances are associated with CID 73336290? What are their SIDs?

(d) Go to the Substance Record page for the substance associated with CID 73336290 (by clicking the SID listed under the “Related Substances” section of its Compound Summary page). What is the version number of this substance record?

(e) Go to the Substance Record page for the most recent version of the substance record in (d). This page shows the message: “NOTE: REVOKED RECORD. See the revoke reason and the revision history for more information.” What is listed as the revoke reason for this record?

PART II. Finding information on caffeine

4. Go to the PubChem home page (https://pubchem.ncbi.nlm.nih.gov). After selecting the “Compound” tab above the search box, type “caffeine[CompleteSynonym]” in the search box and click the “Go” button. It will lead you to the Compound Summary page for caffeine (CID 2519), which presents all information available in PubChem for this chemical. [The suffix “[CompleteSynonym]” means that you want to search for compounds whose name exactly matches the query string (caffeine). You will learn how to search PubChem in much more detail for next two modules (Modules 5 and 6).]

(a) Scroll down until you see “Contents” on the left column. Expand this table of contents by clicking the “+” sign before “Contents”. Locate the “boiling point” and “melting point” sections (under “Experimental Properties” of “Chemical and Physical Properties”). Fill the table below using the information presented in these sections. Because the data in these sections are given in either ºC or ºF, you need to convert from one unit to the other to fill the table.

Primary Data source	Boiling points in ºC	Boiling points in ºF
HSDB		
ILO-ICSC		
CAMEO Chemicals	177.8°C	352° F at 760 mm Hg (sublimes) (NTP, 1992)

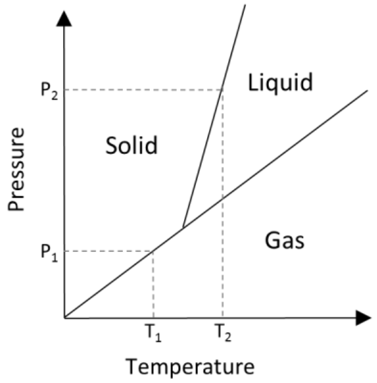
Data source	Melting points in ºC	Melting points in ºF
HSDB		
DrugBank (Phys Prop)		
ILO-ICSC		
CAMEO Chemicals	237.8 °C	460° F (NTP, 1992)
Human Metabolome Database		

(b) The reported boiling points of caffeine are actually the sublimation temperature, as indicated in parentheses. Explain the difference between boiling and sublimation.

(c) Suggest a reason why PubChem reports the sublimation temperature under the boiling point section?

(d) The sublimation temperature and melting point of caffeine summarized in the table does *not* seem reasonable if you compare them with each other. Explain why they are not reasonable.

(e) Below is the schematic phase diagram of caffeine. Based on this diagram, what kind of information do you need to resolve the data accuracy issue identified in (d).



(f) Among the data sources in the table, how many sources explicitly reported the atmospheric pressure at which the phase transition temperatures were measured? What are they?

(g) Whenever possible, PubChem presents data with its provenance (source). For each boiling and melting point value, click the source name to find more detailed information about the data provenance. In most case, PubChem provides a link to the data reported on the original data source. Click this link to check the melting and boiling point data on the data sources' websites. Does any data source provide additional information about the boiling and melting points of caffeine that may help you figure out the issue with the phase transition temperatures of caffeine?

(h) Provide a short paragraph (less than 100 words) that evaluates the accuracy of the phase transition temperatures in PubChem and other databases, based on your answers to questions (a) ~ (g).

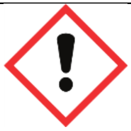
(i) The term "curation" refers to the process of critically reviewing the data as well as checking its accuracy in order to provide the data and related information for the scientific community. What you did while answering the questions (a)~(h) may be considered as a part of manual curation efforts, and the paragraph written in (h) is essentially a curated review of the phase transition temperature data of caffeine. Provide the total (estimated) time you spent for answering questions (a) ~ (h), and use it to estimate the amount of time necessary to manually curate the melting and boiling data for 10, 100, 1000, 10000 compounds, respectively.

(j) Curation can be done either manually or in an automated way. Discuss (in less than 50 words) the pros and cons of manual and automated curations and when each approach would be appropriate.

PART III. Understanding toxicity of caffeine

5. Go back to the PubChem Compound Summary page for caffeine (CID 2519). Locate the “GHS classification” section (under “Hazard Identification” of “Safety & Hazards”). This section shows the GHS (Globally Harmonized System of Classification and Labelling of Chemicals) information of caffeine collected from authoritative organizations. (Click **here** to read more about the GHS information.)

(a) By default, the “GHS classification” section shows the GHS information from a single organization. To view information from all organizations, click the “View GHS classification from all sources” at the end of the section. Based on the information on this page, fill in the table below.

	EU Regulation & Austrailia	Japan NITE	ECHA
Pictograms			
GHS hazard statement code	H302		

(b) What does each pictogram in the table mean? Refer to the PubChem GHS classification page (https://pubchem.ncbi.nlm.nih.gov/ghs/#_pict).

(c) What does each hazard statement code means? Refer to the PubChem GHS classification page (https://pubchem.ncbi.nlm.nih.gov/ghs/#_prec)

(d) According to the information collected in the table, caffeine has acute oral toxicity. However, from our experience, we know that the consumption of caffeine in a cup of coffee is usually okay. To better understand the GHS information on caffeine, it is necessary to review the oral toxicity data of caffeine. First fill in this table with the oral LD50 values for different animal species compiled in the "Non-Human Toxicity values" sections (under "Toxicological Information" of "Toxicity").

	Oral LD50 value
Rat	192 mg/kg
Mouse	
Dog	
Rabbit	
Guinea pig	
Hamster	

(e) Fill in the table with the caffeine content in typical food and drinks, which can be found in the "Food Survey Values" section (under "Ecological Information" of "Toxicity"). To compare these values with the oral LD50 values, divide the caffeine content by the average bodyweight of adult humans (assumed to 60 kg). Also calculate the per cent value of the caffeine content per kg. relative to the oral LD50 value for mice [that is, caffeine content per kg divided by LD50 (mice)].

	Caffeine content	Caffeine content per kg	Ratio
A single cup of espresso (30 mL)	64 mg	1.1 mg/kg	0.84 %
A 8-oz cup of automatic drip coffee			
A cup of tea (camellia sinensis)			
Soft drinks per 12 oz			
Energy drinks per 12 oz			

(f) Find information on human toxicity of caffeine from the "Human Toxicity Values" section (under "Toxicological Information" of "Toxicity"). Compare this information with the data collected in (d) and (e).

(g) Read the following articles about caffeine.

- http://www.fda.gov/downloads/UCM200805.pdf
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3839443/pdf/hpj4801-5.pdf
- http://www.fda.gov/food/recallsoutbreaksemergencies/safetyalertsadvisories/ucm405787.htm

(h) Based on the information collected from (a) through (g), write a short paragraph (less than 100 words) about the toxicity of caffeine and caffeine-containing products.