Questions

1. This question is designed to check if you have a clear understanding of how a text search works in PubChem.

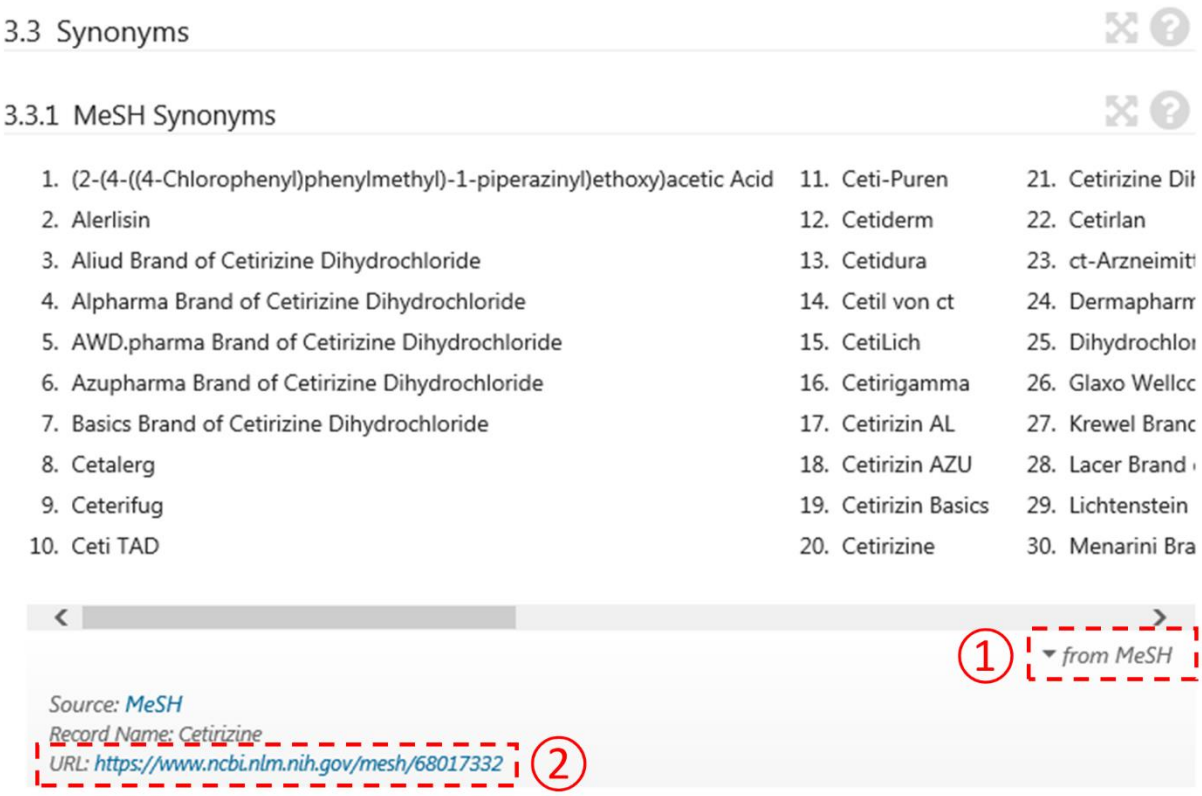
   (a) Go to the PubChem homepage (https://pubchem.ncbi.nlm.nih.gov) and select the "Compound" tab above the search box. Perform three searches using the queries listed on the table below, and record the number of returned compounds and the CIDs of hits.

Query	Number of hits	Returned CIDs
zyrtec[completesynonym]		
zyrtec[synonym]		
zyrtec		

   (b) Find the most frequently occurring **covalently-bonded unit** in the compounds in the table above. What is the CID of that covalent-bonded unit? [A covalently-bound unit (or simply called covalent unit) consists of a group of covalently-bonded atoms in a compound record. Some compounds in PubChem are mixtures of two or more covalently-bonded units. While the number of components in a mixture is conceptually similar to the number of covalently-bonded units, the term "component" often leads to some ambiguity. For example, is NaCl a single component or a mixture of two components? The use of covalently bonded units (instead of components) removes this ambiguity because it is well accepted that NaCl is bonded through an ionic bond.)]

   (c) What is the CID that is returned from the query "zyrtec[synonym]" but not from "zyrtec[completesynonym]"? Explain why this CID was returned from "zyrtec[synonym]" but not from "zyrtec[completesynonym]"

   (d) What is the CID that does not have the common covalently-bonded unit in (b)? Explain why it was returned from the query "zyrtec". [HINT: you will need to compare the depositor-provided synonyms for this CID with the MeSH term "cetirizine" (and its entry terms), which can be accessed through the link below the "MeSH Synonyms" section.]



2. This question tests whether you can search for compounds using molecular property values.

   (a) Read this wikipedia article (https://en.wikipedia.org/wiki/Lipinski's_rule_of_five) and summarize what Lipinski's rule of 5 is.

   (b) Search PubChem for compounds that satisfy each requirement of Lipinski's rule of five as well as all the requirements and record the number of compounds in the table below. The queries necessary for these tasks are also given in the table. Note that XLogP is used instead of LogP. (XLogP is a theoretical LogP value predicted by a computer algorithm.) Also note that XLogP has no lower-bound value, while the lowest possible value for the three properties is zero. Currently, the lowest XlogP value in PubChem is -107.5 (for CID 59172357). Therefore, the lower-bound for the XLogP query is set to a sufficiently low value (-1000).

	Criteria	Entrez Query	Number of CIDs
#1	HBD $\leq$ 5	0:5[HydrogenBondDonorCount]	
#2	HBA $\leq$ 10	0:10[HydrogenBondAcceptorCount]	
#3	MW $\leq$ 500	0:500[MolecularWeight]	
#4	LogP $\leq$ 5	-1000:5[XLogP]	
	Compounds satisfying all requirements.	#1 AND #2 AND #3 AND #4	

(c) Read the paper by Congreve et al. (Drug. Discov. Today, 2003, 8(19):876; http://dx.doi.org/10.1016/S1359-6446(03)02831-9) and summarize what Congreve's rule of 3 is and why it was introduced?

(d) Based on the table in (b) as a template, make a table that summarizes the number of compounds that satisfy Congreve's rule of 3. Perform Entrez searches for them and record the number of hits returned.

	Criteria	Entrez Query	Number of CIDs
#1			
#2			
#3			
#4			
	Compounds satisfying all requirements.		

(e) What is the percentage of compounds satisfying Congreve's rule of 3, relative to **all compounds in PubChem**?

3. Some compounds in PubChem have information on experimentally determined three-dimensional (3-D) molecular structures (presented in the "Protein Bound 3-D Structures" section of the Compound summary page). These structures are provided by the Molecular Modeling Database (MMDB), which curates 3-D structures from Protein Data Bank (PDB). For example, the protein-bound 3-D structure of penicillin V can be accessed via the following URL:

   https://pubchem.ncbi.nlm.nih.gov/compound/6869#section=Biomolecular-Interactions-and-Pathways

   On the other hand, some compounds have links to experimental 3-D structures archived in the Cambridge Structural Database (CSD) hosted by the Cambridge Crystallographic Data Centre (CCDC). For example, the 3-D structure of penicillin V archived at CSD-CCDC can be accessed via this URL:

   https://pubchem.ncbi.nlm.nih.gov/compound/6869#section=Crystal-Structures

Use PubChem's classification browser and advanced search builder to find the answer to the following questions.

(a) How many compounds in PubChem have protein-bound 3-D structures?

(b) How many compounds in PubChem have 3-D structures archived in CSD-CCDC?

(c) How many compounds in PubChem have *both* PDB structures *and* CSD 3-D structures?

(d) How many compounds in PubChem have *any* experimental 3-D structures (either from PDB or CSD)?

(e) What is the ratio of the compounds with both PDB and CSD structures to the compounds with any experimental 3-D structures?

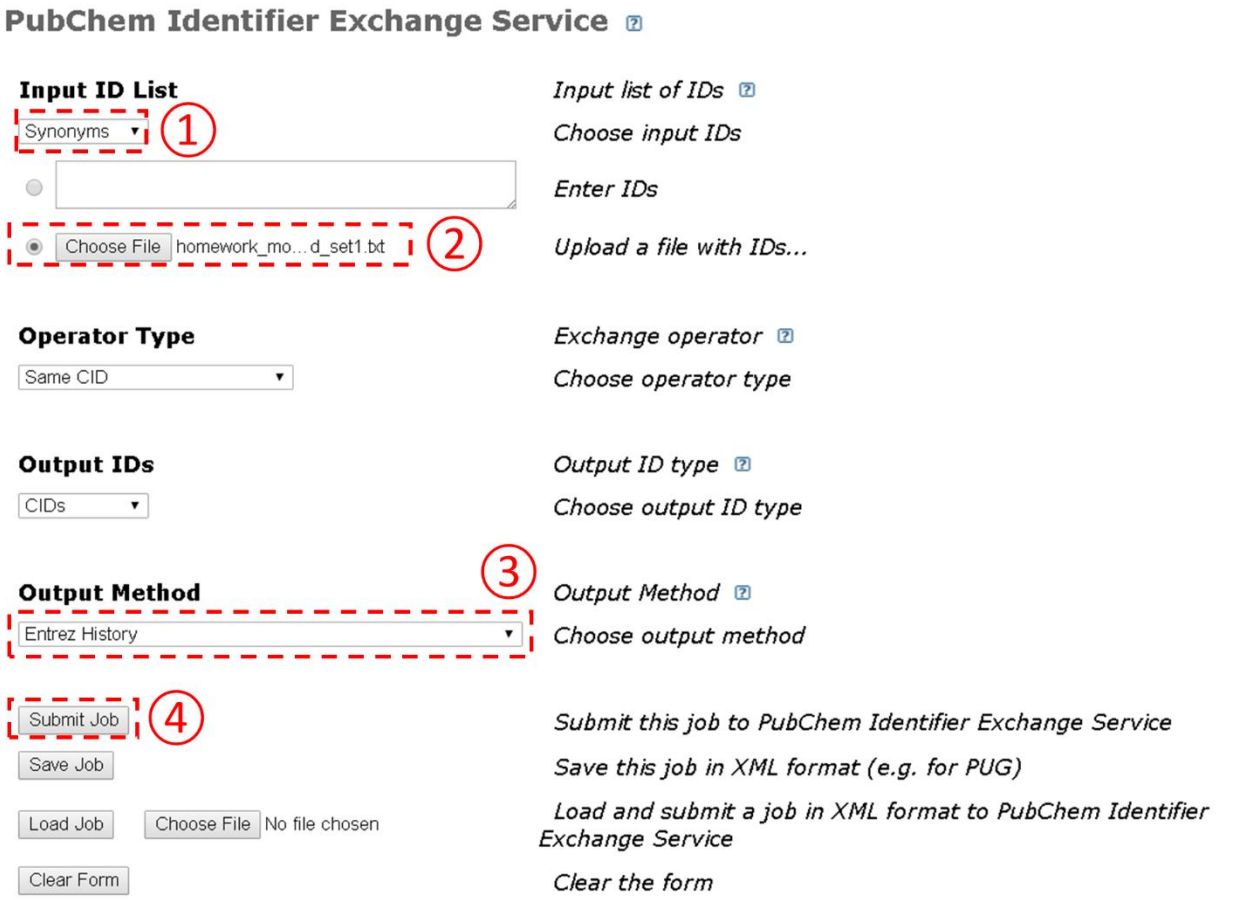
(f) Explain the difference between PDB (http://www.wwpdb.org/) and CSD (https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/).

(g) Suggest a reason why there are not so many compounds with both PDB and CSD structures.

4. This question involves the use of PubChem's Identifier exchange service (https://pubchem.ncbi.nlm.nih.gov/idexchange/idexchange.cgi) to search the Compound database using multiple chemical names as queries.

(a) Make a text file that contains the following 10 chemical names:
1-(1-Phenylcyclohexyl)pyrrolidine
3,4-Methylenedioxymethamphetamine
Allylprodine
Barbital
Cocaine
Lorazepam
Methadone
Normorphine
oxycodone
Phenylacetone

(b) Go to the Identifier Exchange Service, and follow the steps illustrated in the Figure below to search the Compound database using the text file generated from the previous step as an input to the Identifier Exchange Service. How many compounds do you get on the DocSum page?



(c) Considering that the input file had only ten chemical names, some of them must have resulted in "multiple" hits. To check what chemical name(s) resulted in multiple hits, repeat the search in (b) again, but with the "Output Method" option set to "Two column file showing each input-output correspondence". What chemical names result in multiple CIDs and what CIDs are associated with them? [In this question, it is not difficult to manually check what they are (because we have only 16 compounds returned from 10 compounds), but it wouldn't be feasible if you are dealing with hundreds of compounds.]

(d) Collect the canonical and isomeric SMILES for the CIDs returned in step (c).

(e) For each chemical name returned in (c), discuss the difference among the multiple compounds (CIDs) returned from the chemical name search.

5. This question tests whether you know how to obtain desired information from the PubChem Data Sources page (https://pubchem.ncbi.nlm.nih.gov/sources).

(a) What is the total number of data sources of PubChem information?

(b) How many data sources does PubChem collect annotations from?

(c) What kind of annotations does PubChem collect from NCI Investigational Drugs?

(d) Download the UV data from NCI Investigational Drugs in JSON, and fill in the following table with the information for the compound that appears first in the downloaded file

Compound Name	
CID	
UV data	