4.2: Text Search in PubChem

Basics

Text search allows one to find chemical structures using one or more textual keywords, which may be chemical names (e.g., “aspirin”) or any word or phrase that describe molecules of interest (e.g., “cyclooxygenase inhibitors”). One can perform a text search from the PubChem homepage, by providing a text query in the search box. If the query is **a phrase or a name with non-alphanumeric characters**, **double quotes should be used around the query**. Various indices can be individually searched by suffixing a text query with an appropriate index enclosed by square brackets (for example, the query ***“N-(4-hydroxyphenyl)acetamide”[iupacname]***). Numeric range searches of appropriate index fields can be performed using a “:” delimiter (for example, the query ***100.5:200[molecularweight]*** for a molecular weight range search between 100.5 and 200.0 g/mol). One can see what search indices are available in PubChem from the drop-down menu on the “PubChem Compound Advanced Search Builder”, which can be accessed by clicking the “advanced” link (next to the “Go” button) on the PubChem homepage. Queries may be combined using the Boolean operators “**AND**”, “**OR**”, and “**NOT**”. These Boolean operators must be capitalized.

Depositor-supplied synonyms

Conceptually, data in a database are stored in the same way as we would record them in a table or excel spreadsheet. The rows in the table correspond to compounds, and the columns correspond to properties or descriptions for those compounds (e.g., melting and boiling points, chemical names, toxicity, bioactivity, target proteins, and so on). These columns are commonly called “data fields”. You may want to perform a search against all data fields or only a particular field. To search the (depositor-provided) chemical name field of the records in the PubChem Compound database, a chemical name query needs to be suffixed with either of the “[synonym]” or “[completesynonym]” index. The “[synonym]” index invokes search for molecules whose names contain the query chemical name as a part (that is, **partial matching**), and the “[completesynonym]” index invokes search for those whose names completely match the query (that is, **exact matching**). If no index is given after the query, PubChem will search all data fields. Compare the following searches for “aspirin” against the PubChem Compound database.

- **aspirin[completesynonym] (1 hit, as of Feb. 26, 2017)**
  https://www.ncbi.nlm.nih.gov/pccompound/?term=aspirin%5Bcompletesynonym%5D
- **aspirin[synonym] (98 hits)**
  https://www.ncbi.nlm.nih.gov/pccompound/?term=aspirin%5Bsynonym%5D
- **aspirin (103 hits)**
  https://www.ncbi.nlm.nih.gov/pccompound/?term=aspirin

Note that the URLs for these searches contain the query strings (following the string “?term=”), and that the square brackets enclosing the Entrez indices “completesynonym” and “synonym” are replaced with the strings “%5B” and “%5D”. Because the first query resulted in only one hit, the user is directed to the Compound Summary page for the hit compound (CID 2244). On the other hand, because the other two queries result in multiple hits, the results are presented on the DocSum pages.

When either “[completesynonym]” or “[synonym]” is used, it is the **“depositor-provided synonyms”** fields of the compound records in PubChem that is searched for the query string. The depositor-provided synonyms field for a compound contains a filtered list of chemical names (synonyms) provided by individual data providers for the substances associated with that compound. These synonyms are presented in the “Depositor-provided synonyms” section on a Compound Summary page. To see the variety of synonyms for a compound, check the following link [to the Depositor-provided synonyms” section of the Compound Summary page for CID 2244 (aspirin)]:

https://pubchem.ncbi.nlm.nih.gov/compound/2244#section=Depositor-Supplied-Synonyms

For CID 2244, there are more than 700 depositor-supplied synonyms. These synonyms include not only those commonly used in chemistry class (e.g., common names, IUPAC names, CAS registry numbers) but also those used in many other places (e.g., database identifiers, chemical vendor catalogues, the name of products that contains the chemical, code numbers internally used in a company, and so on).

As mentioned above, the search for aspirin with the “[completesynonym]” index specified returns only one compound (CID 2244). It means that one of many names of this compound exactly matches the query string “aspirin”. On the other hand, the search for aspirin with the “[synonym]” index returns additional 97 compounds. It means that at least one of the names of each these

compound partially match the query string (that is, the compound contains the string “aspirin” in one of its names). Interestingly, the results from the last two queries include acetaminophen (CID 1983), which is the active ingredient of Tylenol. Check the following link to the depositor-provided synonyms section of CID 1983 to see what synonyms of Tylenol contains the string “aspirin”:

https://pubchem.ncbi.nlm.nih.gov/compound/1983#section=Depositor-Supplied-Synonyms

Some of the synonyms of Tylenol contains the phrase “aspirin-free” or “non-aspirin”. Note that Tylenol was returned from a search for “aspirin” (through partial matching using the [synonym] index).

MeSH Synonyms

The National Library of Medicine (NLM)’s Medical Subject Headings (MeSH)[1,2]

is a controlled vocabulary thesaurus of medical terms arranged in a hierarchical structure. It is used for indexing scientific articles from biomedical journals for PubMed and cataloging medical books, documents, and audiovisual materials, in order to facilitate retrieval of medical information at various levels of specificity.

Many of MeSH terms are chemical names (e.g., for drugs, nutrients, metabolites, toxic chemicals, and so on). PubChem performs an automated annotations of PubChem records with MeSH terms (by means of chemical name matching), creating associations between PubChem records and PubMed articles that share the same MeSH annotation. The MeSH term that match a (depositor-provided) synonym of a compound in PubChem is presented with its entry terms under the “MeSH Synonyms” section of the Compound Summary page of that compound.

Go to the Compound Summary page for CID 171511 via the following link to check the MeSH synonyms and Depositor-supplied Synonyms sections.

References

(1) Medical Subject Headings (MeSH) (https://www.nlm.nih.gov/mesh/)

(2) Medical Subject Headings (MeSH®) Fact Sheet (https://www.nlm.nih.gov/pubs/factsheets/mesh.html)