Subject: Importing compounds with assay data from Pubchem into Datawarrior
Posted by JonW on Fri, 23 Jul 2021 11:12:44 GMT
View Forum Message <> Reply to Message

Many thanks Thomas and colleagues
Data warrior is a "life saver" for non coding people like me - thank you so much for you and your team for developing it.

Its also one of (I think it is THE) easiest software to use for a variety of informatics / prediction modeling etc.

Importing / uploading to Datwarrior is usually really straightforward compared to other programs.

For example, it's possible to download compounds from Pubchem and via a text / csv file import into Datawarrior with no editing. This works really well and is easy to do.

However when trying to download, say 30,000 compounds that are also P450 inhibitors an error comes up, maybe because the request has too many data points or maybe it cant be done via a web page (Pubchem talks about Restful / ftp but for non coding people this seems very difficult to do - 2 hours trying to figure this out did not prove fruitful!

My questions/requests/comments are:

1. A request - Can Datawarrior create a built in search for Pubchem along similar lines to the Chembl facility? Pubchem seems to have so many more compounds with accompanying biodata in it compared to Chembl.

2. In the meantime - how can you import compounds with specific biodata from Pubchem into datawarrior?

Can you use Datawarriors URL import? I tried this for example with a list of P450 enzyme inhibition active compounds from Pubchem (by copying and pasting the url, but this did not work:
https://pubchem.ncbi.nlm.nih.gov/protein/P10633#section=Chem icals-and-Bioactivities

Any ideas regarding the URL or any other "non coding" ways to get data form Pubchem into Datawarrior? As stated above, you cannot simply download a csv file from Pubchem (at least for obtaining biodata) from the above link as an error occurs, even though Pubchem provides a downlaod "button" for it.

Many thanks in advance
Jon

Subject: Re: Importing compounds with assay data from Pubchem into Datawarrior
Posted by thomas on Fri, 23 Jul 2021 13:17:45 GMT
View Forum Message <> Reply to Message

Hi Jon,

unfortunately, I am afraid, I cannot help you much. I never fully explored what is needed to keep a fully updated copy of PubChem's bioactivity on a structure searchable server, but it always seemed to me that it would be beyond the effort, I would be willing to invest. I just tried to download a few data files from the PubChem website (pubchem.ncbi.nlm.nih.gov). Some download button scripts simply didn't do anything (Firefox on Ubuntu), but with most of them I could download either a CSV, SDF or gzipped SDF. All of those I could open successfully in DataWarrior ('.sdf.gz' need the newest dev update). Unless somebody else would volunteer to write the code to retrieve the PubChem bioassay data and merge it with the structures where needed and keep everything updated on a searchable server engine, I won't have the time to provide easy DataWarrior search access.

If you manage to download PubChem data files (csv or sdf) and have trouble reading them into DataWarrior because of limited csv/sdf capabilities, then please let me know. I would do my best to fix these issues on the DataWarrior side.

Thomas

Subject: Re: Importing compounds with assay data from Pubchem into Datawarrior
Posted by JonW on Fri, 23 Jul 2021 16:39:12 GMT
View Forum Message <> Reply to Message

Hi Thomas
Thanks for your efforts. Are you saying that you could download the bioassay csv files or just the substance / chemical files? The latter I can download and open in DW. The former - they dont download at all.

Can you confirm its biodata that you could download (e.g section 3 Chemicals and Bioactivities 3.1 Tested Compounds https://pubchem.ncbi.nlm.nih.gov/protein/P10633#section=Chemicals-and-Bioactivities)?
See attached file for exact location of the csv file on Pubchem (circled in red) -= this "download Button" doesnt work for me.

Aso the url import in DW - are you saying this cant be used either for Pubchem urls?

Perhaps other members of the forum have found ways to specifically download Pubchem compounds/substance WITH bioassay data In the same csv file) and could share how they did that and uploaded it into Data warrior?
Many thanks in advance
jon

File Attachments

1) Pubchem download error.docx, downloaded 29 times

Subject: Re: Importing compounds with assay data from Pubchem into Datawarrior
Posted by JonW on Sun, 25 Jul 2021 15:18:10 GMT
View Forum Message <> Reply to Message

Apologies for the "winking smileys" - I did not want to include them in my posts - they just "appeared" when I used "underline".

Also, I have noticed 43 people have looked at this thread in the space of 24 hours. Please folks add some help or some suggestions here. I can't believe I am the only one wanting to add / upload PubChem data to Datawarrior, this fantastic piece of software.

Or PM me if you prefer with any helpful tips or suggestions.
Many thanks in advance

Subject: Re: Importing compounds with assay data from Pubchem into Datawarrior
Posted by nbehrnd on Mon, 26 Jul 2021 19:55:23 GMT
View Forum Message <> Reply to Message

Dear Jon,

there was indeed an problem to access and subsequently keep a local copy of the .csv files. By now, yhe support by PubChem was able to correct the underlying error. The downloaded file contains a (late) column labelled «cmpdname»

Open the file in a text editor to copy all content into the working memory of the computer. In DataWarrior, paste these information via Edit -> Paste Special -> New From Data With Header Row. Establish the structures via Chemistry -> Add Structures From Name which opens a new interface. To guide DW's action, select the column of interest, «cmpdname».

Norwid

File Attachments
1) name2structure.png, downloaded 96 times
2) Data_From_Clipboard.dwar, downloaded 33 times

Subject: Re: Importing compounds with assay data from Pubchem into Datawarrior
Posted by thomas on Thu, 19 Aug 2021 18:56:03 GMT
View Forum Message <> Reply to Message

Hi Jon and Norwid,

I can confirm that Jon's URL now works, but I remember I had trouble before.
Structure retrieval based on names is the second best solution, although the openmolecules server knows most of the PubChem names unless they are very long.

If you can get an SD-File with SIDs or CIDs, the result will be more complete and potentially may have less errors.

Thomas
