
Subject: Count # scaffolds by plate ID
Posted by [chemtv](#) on Mon, 03 May 2021 14:35:28 GMT
[View Forum Message](#) <> [Reply to Message](#)

Hi folks,
I am looking at some plated cpds where I have individual IDs and the assay plate IDs that I am trying to analyze. I want to calculate Murcko scaffolds then count the number of different scaffolds on each plate, (rough estimate of diversity). For datawarrior I suspect this is where macros can be your friend but I'm not familiar with writing macros. Has anyone else tried something like this? Even in Knime I'm not sure how to get to counts of a previous "aggregation". Any thoughts?
Thanks,
Greg

Subject: Re: Count # scaffolds by plate ID
Posted by [nbehrnd](#) on Mon, 03 May 2021 16:02:16 GMT
[View Forum Message](#) <> [Reply to Message](#)

Once the structures are read by DW, you may launch via chemistry -> analyze scaffolds the assignment of the Murcko scaffold. If you like, DW may write a new .dwar file with a frequency count of these common denominators, too:

In these files, the idcode of the identified scaffold is separated by a tabulator from the integer, as documented in the archive attached below. Is this the direction of analysis you would like to automate?

If so, it suffices to identify the pathway you would go to process the data manually, and to document this by recording a macro (``macro -> start recording`` to initiate, ``macro -> stop recording`` to complete DW's «training»). You may export then export the macro as a file as an individual .dwam file, a plain ASCII file.

The one attached below as example requires some adjustment in line #5 about the path and file name for the new .dwar (about the frequency count) to be written by you. This edit has to happen outside DW. Then, DW already working, load this macro adjusted .dwam file (`macro -> import macro`) and let it run (`macro -> run macro`, then select `export_scaffold`). As set up by now, it will write `murcko_scaffolds.dwar` as permanent record.

Note, regardless of your .dwar processed, above macro will yield a file `murcko_scaffolds.dwar` with the frequency count. Because of the static file name of the output it is possible you accidental overwrite the results of a precedent analysis.

Norwid

File Attachments

- 1) [murcko_scaffold.png](#), downloaded 803 times
- 2) [scaffold.zip](#), downloaded 310 times

3) [export_scaffold.dwam](#), downloaded 364 times

Subject: Re: Count # scaffolds by plate ID
Posted by [chemtv](#) on Tue, 04 May 2021 12:47:36 GMT
[View Forum Message](#) <> [Reply to Message](#)

Thanks for the quick reply. Yes I use the frequency file to plot the scaffold distributions, but I need the cpd and plate IDs recorded with the scaffolds found. This list would be great to see in a cell. I could export and extract the data externally if it were there. The issue is I don't see any options to get at this data with the scaffolds.

Thanks again,
Greg

Subject: Re: Count # scaffolds by plate ID
Posted by [chemtv](#) on Tue, 04 May 2021 12:49:35 GMT
[View Forum Message](#) <> [Reply to Message](#)

hmm, maybe I could merge the freq file with the regular murcko outfile by the scaffold and get to the plate IDs?

Subject: Re: Count # scaffolds by plate ID
Posted by [nbehrnd](#) on Tue, 04 May 2021 17:06:16 GMT
[View Forum Message](#) <> [Reply to Message](#)

I think I grasp your intent. By running the scaffold analysis, you aim for a table which either separates the entries by category like

```
| scaffold | frequency | plate addresses | individual label |
|-----+-----+-----+-----|
| isoxazolone | 2 | 12, 91 | compd18, compd13 |
| antipyrine | 4 | 18, 19, 20, 21 | compd23, compd24, compd25, compd26 |
| ... | ... | ... | ... |
```

or one where the number in the plate (plate address) and the compound label are next to each other like

```
| scaffold | frequency | plate address / individual label |
|-----+-----+-----|
| isoxazolone | 2 | 12, compd18; 91, compd13 |
| antipyrine | 4 | 18, compd23; 19, compd24; 20, compd25; 21, compd26 |
```

| ... | ... | ... |

Here, «plate address» is used just as alias for the number of the well in the plate == entry line in the initial .dwar file to process.

Norwid

Subject: Re: Count # scaffolds by plate ID
Posted by [amorrison](#) on Wed, 05 May 2021 04:23:35 GMT
[View Forum Message](#) <> [Reply to Message](#)

Hi Greg,
Would either of these achieve what you are aiming to do.

a) Generate Murcko scaffolds then calculate 'Frequency of same value in same category' - `frequencyInCategory(category-column, value-column)`. Where the category-column is the plateID and the value-column is the Murcko.

b) Generate Murcko scaffolds then merge rows on both Murcko and PlateID. This would give you the list of compoundIDs with the same Murcko in one cell which you could then do a `valueCount` to get the number of compounds.

Best,

Angus

Subject: Re: Count # scaffolds by plate ID
Posted by [chemtv](#) on Wed, 05 May 2021 21:23:11 GMT
[View Forum Message](#) <> [Reply to Message](#)

Starting with: 500K smiles, cpdID, plateID
Here's what I'm ultimately trying to get to...

plateID	Num Scaffolds	Scaffolds (smiles of scaffold)
abc01	5	phenyl, pyridine, indole, thiazole, naphthyl
abc02	3	phenyl, pyrimidine, naphthyl
abc03	2	indole, oxazole

Actually the 1st two columns are what I need.

Greg

Subject: Re: Count # scaffolds by plate ID
Posted by [nbehrnd](#) on Thu, 06 May 2021 12:24:18 GMT
[View Forum Message](#) <> [Reply to Message](#)

It is possible to use the structures of one .dwar file for a comparison against structures in a second .dwar file. Because I did not identify (yet) how to point from the frequency table generated specifically to the SMILES strings of the Murcko scaffolds of the original file, I wrote a little Python script to address the initial question. The script depends on standard CPython and the RDKit library only. For a small library (say, .ie. 3k entries), I perceive the rate of computation as fast enough; of course your mileage may vary when submitting 500k.

The archive, in addition to the script, equally documents a test run with said script.

Norwid

File Attachments

- 1) [scaffold_seeker.py](#), downloaded 1755 times
 - 2) [scaffold_seeker.zip](#), downloaded 350 times
-

Subject: Re: Count # scaffolds by plate ID
Posted by [chemtv](#) on Fri, 07 May 2021 18:11:53 GMT
[View Forum Message](#) <> [Reply to Message](#)

Thanks Norwid, I will take a look at what you sent.
