Subject: Sorting, counting and deleting different elements (e.g., Iodine) in a dataset
Posted by Jo W on Sun, 11 Sep 2022 22:54:09 GMT
View Forum Message <> Reply to Message

How do you find out the distribution and numbers of different elements that might occur in a dataset of drug-like molecules and delete those compounds that contain these specific elements?

For example if you download 5000 HIV active organic compounds from say PubChem containing a diverse set of different structures, there will be some compounds that for example contain selenium atoms or iodine.
These type of elements are not common in many datasets for biological screening and can distort and/or cause poor model predictions to occur.
So, how can you collate these compounds in Datawarrior and quickly analyse their frequency and also selectively remove them?

I know you can set up a filter for example "molecular formula" or "smiles" and then type in "Se" and then the filter "hides" all the selenium containing compounds (if you reverse the filter) and you can also tell how many compounds in the dataset were "hidden" and therefore get a figure of the selenium-containing compounds in the dataset.

However it's very laborious to do this for all other elements (accepting that you want for example C,H,O,N elements to remain), and also this peace-meal approach does not let you visualise the number of compounds in the dataset that contain for example, selenium, iodine, chlorine, phosphorous, etc.
For example, it would be good to see the following as a table in DW:
C,H,N,F - 90 compounds
C,H,O,P - 200 compounds
C,H,O,Se - 10 compounds etc
and maybe to visualise them as a histogram. Then for example removing the selenium-containing compounds from the dataset, to see what effects on the model they have.
So how can this be achieved in DW?

Subject: Re: Sorting, counting and deleting different elements (e.g., Iodine) in a dataset
Posted by nbehrnd on Tue, 13 Sep 2022 06:55:24 GMT
View Forum Message <> Reply to Message

Dear Jon,

this is indeed an interesting question to think about further. As an early concept, equally based on the previous assignment of the Hill formula, I wrote a DW macro which subsequently uses a regular expression (regex) in a if-clause to test if the entry in question contains each of the elements (CHNO) at least once. (See the attachement below.)

Though using a macro likely eases the task (as in to offer reproducible action regardless the size of the data set, and rate of processing), there might be some obstacles ahead to extend the approach, i.e. to use multiple «filters» / «detectors» at once. To check for (CHNF), or

(CHOP), or (CHOSe) as you intend is going to generate categories. This is not a problem for drawing a histogram with DW, but the syntax to probe, e.g. currently for (CHNO)

if(matchregex(MolecularFormula, "^C.*H.*N.*O.*"), "CHNO", "")

basically states

«check the regex expression on the Hill formula; if evaluated .True. return CHNO (which later may counted by DW plotting the histogram) -- else (equivalent to .False. / there is no match) return nothing».

Normally, I would try using the now empty return (above "there is no match") to nest a second test, e.g., «now test for CHOP». However, contrasting to «binning the data» as in «entries with a molecular mass, and user defined thresholds to establish categories based on this property in common»*), this approach doesn't work well enough here, because a molecule belonging to the category of (CHNO) simultaneously may belong to the category of (CHNF). So here, the discern neither is by one category in common (molecular mass), nor are the categories to probe in a relationship like (partial or complete) sub/super sets of each other.

*) DW allows to bin continuous data in preparation e.g., of a histogram; then, the bin size (e.g., interval of the molecular masses per class) applied however is uniform all across the data.

Norwid

File Attachments

1) Random_Molecules.dwar, downloaded 573 times
2) probe_CHNO.dwam, downloaded 569 times

Subject: Re: Sorting, counting and deleting different elements (e.g., Iodine) in a dataset
Posted by Jo W on Tue, 13 Sep 2022 20:13:33 GMT
View Forum Message <> Reply to Message

Many thanks Norwid
I will get back to you asap
Best regards
Jon

Subject: Re: Sorting, counting and deleting different elements (e.g., Iodine) in a dataset
Posted by thomas on Tue, 13 Sep 2022 20:44:36 GMT
View Forum Message <> Reply to Message

Dear Jon,

the following macro would be another example of how you could achieve this:
The macro adds the chemical formula and removes all numbers from it. Then it counts how often any atom combination exists, removes duplicate atom combinations, defines a custom order on the column that contains the atom combination string based on its frequency, and then creates a view with a bar chart showing all combinations with frequencies sorted by the frequencies.

To try out the macro on any data set with chemical structures copy the following macro text. Then in DataWarrior select "Macro->Paste Macro". The run the macro with "Macro->Run Macro->CountAtomCombinations"

```
<macro name="CountAtomCombinations">
<task name="addMolecularFormula">
structureColumn=Structure
</task>
<task name="findAndReplace">
isStructure=false
column=Molecular Formula
isRegex=true
what=[0-9]
with=
</task>
<task name="addCalculatedValues">
columnName=Atom Combination Frequency
isOverwrite=false
formula=frequency(MolecularFormula,"Molecular Formula")
</task>
<task name="deleteDuplicateRows">
caseSensitive=true
columnList=Molecular Formula
addCount=false
</task>
<task name="selectView">
viewName=Table
</task>
<task name="sortRows">
column=Atom Combination Frequency
selectedFirst=false
descending=true
</task>
<task name="setCategoryCustomOrder">
sortMode=mean
isStructure=false
column=Molecular Formula
isAscending=true
sortColumn=Atom Combination Frequency
</task>
<task name="new2DView">
```

where=center
whereView=Table
newView=2D View
</task>
<task name="setPreferredChartType">
type=bars
column=Atom Combination Frequency
viewName=2D View
mode=mean
</task>
<task name="assignOrZoomAxes">
high1=0.0
low1=1.0
column1=Molecular Formula
millis=1000
viewName=2D View
</task>
</macro>

Subject: Re: Sorting, counting and deleting different elements (e.g., Iodine) in a dataset
Posted by Paul on Wed, 28 Sep 2022 19:02:47 GMT
View Forum Message <> Reply to Message

Many years ago, this problem was presented as an Excel challenge that could not use macros/VBA.

I've kept and still use the spreadsheet because it's so simple and addresses the need.

Formulas from DataWarrior are pasted into the sheet and elemental composition data can then be pasted into a text file and merged back into the DWAR file. If Data Warrior has comparable CHAR functions perhaps a similar approach is possible. Granted, it does require moving data to then from Excel, but it's straight forward.

The attached spreadsheet handles only 10 formulas/rows at a time to keep the attachment size small. Expanding the formula rows in Excel is trivial, but the spreadsheet can get rather slow for thousands of rows at a time.

File Attachments

1) ATOMCALCS.xlsx, downloaded 516 times
