Subject: What do I do with the entries assigned to multiple categories?
Posted by Nastasia on Thu, 23 Jul 2015 10:31:25 GMT
View Forum Message <> Reply to Message

If I merge my structures by structure they are assigned ti multiple categories, it is important for me to have all the different descriptors separately, but at the same time it would be great to merge similar molecules. Is there any solution?

Subject: Re: What do I do with the entries assigned to multiple categories?
Posted by thomas on Thu, 23 Jul 2015 19:57:41 GMT
View Forum Message <> Reply to Message

I am not sure, whether I understand your question correctly, but it seems that grouping your structures by clustering or by shared scaffold would be a way of grouping them by similarity.

Subject: Re: What do I do with the entries assigned to multiple categories?
Posted by Nastasia on Fri, 24 Jul 2015 12:42:10 GMT
View Forum Message <> Reply to Message

I have a number of molecules, some of which are the same. However, they were retrieved from different sources (source is important, so it is a descriptor as well), when I merge them by structure, my descriptors are merged as well, so I have some sort of multiple categories case. And when I build polts I still have some individual cases and all the other entries are assigned to multiple category, and I need my descriptors to be treated separately. Same will happen if I try to merge common scaffolds.
What I need in the end: to be able to order my molecules by similarity, to plot them, and it will be super awesome if I could use common scaffold as markers' label, but not showing all the labels at the same time, more like one label for a group of similar markers.
All in all, I am trying to merge/group structures, but not to loose/merge any descriptors. How can I group them to have them in some categories like: all having the same scaffold1...scaffold100, but not merging them?
I hope I explained it) And thank you.

Subject: Re: What do I do with the entries assigned to multiple categories?
Posted by thomas on Sun, 26 Jul 2015 11:35:20 GMT
View Forum Message <> Reply to Message

I assume that you aim for visualizing the chemical space of your unique molecules and then use color to highlight, which regions of the space are covered by which source, possibly one plot per source. You certainly need to experiment a little with what works well for you data. What I can suggest is the following:
- make a copy of your file and (for simplicity) remove all columns except the chemical structure.
- remove rows with redundant structures and calculate the skeletonSpheres descriptor, which is the best for chemical similarity

- make a SOM (Data menu) with somewhat more reference vectors than compounds using the skeletonSpheres descriptor
- make a similarity analysis (Chemistry menu) with the skeletonSpheres descriptor
- make one or more scaffold analysis (with Murcko, most central ring system, ...)
- you may add labels to each scaffold with 'Add Row Numbers..." and choosing 'Use same number for same' Scaffold
- cluster dataset, if the number of compounds permit that
- save the file

Now you have a file with all compounds categorized in various aspects (scaffolds, cluster) and with x- and y-coordinates allowing to show the chemical space

- Now open your original file and select File->Merge, choose the previously created file, select the structure as merge key and add all other columns of that file to your original file.
- Create 2D-views to show the SOM space and Activity analysis space by using their x- and y-coordinates for the axes. You may quickly copy the views of the compound file with 'File->Open Special->Apply Template...' and choosing that file.
- Rows with the same structure appear in the views with the same coordinates, i.e. on top of each other. If you add a slight jittering, then you start to see multiple ones.
- To highlight all compounds of one category, e.g. all compounds from one source, you may use 'Set Focus...'. For that you either need to create a list for every category, or as a little trick, you set the focus to the selection. Then you create a new 2D-view with one axis unassigned and the other assigned to the category (e.g. source or scaffold number). In this view it is easy to select an entire category causing the selected rows to visually stick out in the other view that uses 'focus on selection'.
- You may use the 'category browser filter' and its dynamic feature to show the chemical space of the different sources one-by-one.

You could go a step further by taking the file with the unique compounds, make a copy, delete all rows except one of the scaffold columns, remove all redundant scaffolds and make a SOM or Activity analysis of that. This way you visualize scaffold space rather than compound space. If you save the file and merge it into your original using scaffold as key, you can do the same things with scaffold space that I have suggest above with compound space.

I hope you can derive something useful from these suggestions...

Thomas
