Subject: Assign cluster name based on cluster size
Posted by mcmc on Thu, 07 Apr 2022 11:46:09 GMT
View Forum Message <> Reply to Message

it looks as if the cluster numbers that are generated by the "cluster compounds" algorithm are rather arbitrarily assigned (I guess in chronological order).
I feel it could be useful to sort clusters by size. That is, cluster 1 would be the largest one, then 2, etc.
Probably an easy fix, yet very helpful?

Subject: Re: Assign cluster name based on cluster size
Posted by nbehrnd on Sat, 09 Apr 2022 18:29:25 GMT
View Forum Message <> Reply to Message

The following doesn't offer the automatic sort (and perhaps eventual cumulative distribution) you look for.

Yet because each cluster's molecule is labeled by the integer of the cluster, one may use it and request DataWarrior to plot a histogram (bin size 1, starting by 1 along the abscissa, number of molecules per bin along the ordinate) to identify the most populated bin/to check if the distribution of bins perhaps is bi/polymodal. For smaller numbers of bins, it might be helpful to add the number of molecules (view options for statistical graphs) to the drawing.

Norwid

File Attachments
1) example.png, downloaded 167 times

Subject: Re: Assign cluster name based on cluster size
Posted by mcmc on Mon, 11 Apr 2022 15:27:41 GMT
View Forum Message <> Reply to Message

Thanks Norwid. Meanwhile I realized that with the current cluster numbering, similar clusters tend to have adjacent numbers. That is, cluster 404 resembles 405. I guess that has some advantages too.

Also I observed that a SALI analysis provides "neighbor count" which seems to be the same as cluster size (minus 1). That in turn, gives a filter that can be used to zoom in on the most populated clusters.

Subject: Re: Assign cluster name based on cluster size
Posted by nbehrnd on Fri, 22 Apr 2022 20:46:45 GMT

View Forum Message <> Reply to Message

Hello mcmc,

I just completed a small Python script to process DataWarrior's results about structure similarity (Chemistry -> Cluster Compounds) exported as text file (File -> Save Special -> Textfile). It identifies the clusters, sorts these based on the number of molecules in each clusters, updates the molecules' cluster labels (1, 2, 3,...) accordingly and writes a new .txt file one may read with DW by (Ctrl + O). There are two sorts possible: a) «the more molecules in the cluster, the lesser the integer used as label of the cluster», a pattern possibly matching best your intent. Though with the optional flag -r you equally may reverse the sort for b) «the more molecules in the cluster, the greater the label».

The .zip archive attached below includes the .py script and describes early results when processing a small set of test data. It assumes the first column labeled «Cluster No» contains the cluster labels assigned by DataWarrior (which is the program's default header).

Norwid

File Attachments

1) 2022-04-26_datawarrior_clustersort.zip, downloaded 211 times

---

Subject: Re: Assign cluster name based on cluster size
Posted by DrCJM on Fri, 05 Aug 2022 05:46:53 GMT
View Forum Message <> Reply to Message

Late comment!

It's easy to make a new column with the number of compounds in each cluster - all cluster members will obviously have the same number.

After clustering, which creates the "Cluster No" column, I calculate a new column (usually called "Cluster Count") with the function:

frequency(ClusterNo, "Cluster No")

You can then sort on Cluster Count or resize markers on graphs by Cluster Count etc. Very useful if you just want to show the Representative compounds on a graph, but size them by the number of similar compounds they represent.

So doesn't rename the cluster but allows you to do the sorts of things you might want to do after renaming them.

Craig.