Subject: Molecules clustering, few questions
Posted by pocin on Wed, 04 Feb 2015 14:58:22 GMT
View Forum Message <> Reply to Message

Hi, I have three questions. I looked up the manual and also the publication without any luck.

I have clustered my library of 325 molecules into fixed number of clusters (50) without specifying the minimal co value using Flexophore descriptor.

In the manual, there is something about weighted mean similarities of cluster members. How do I find out this value? I would like to know how similar are molecules in my cluster and also how different are each clusters between each other.

Second - What is meant by "Is cluster representative"? Is it the most similar molecule to all others in the cluster?

Also is there any native way how to display cluster size or do I have to do it externally in a spreadsheet?

Thanks in advance and have nice rest of the day,
Robin

PS: Thank you for developing such an amazing software!

Subject: Re: Molecules clustering, few questions
Posted by thomas on Thu, 05 Feb 2015 22:26:57 GMT
View Forum Message <> Reply to Message

Hi Robin,

> ... weighted mean similarities of cluster members ...
while the clustering uses these values internally, they are not added to the result columns. I simply did not think that far. What you can do is a little more cumbersome but visual procedure:
- Create a 2D-view and put the (dynamic) Flexophore similarity on one axis
- Put cluster no, the structure or some activity value on the second axis
- set marker color to cluster no
- set marker shape to 'is represenative'
If you click on any compound now, then you see visually the similarities to all other cluster member and the rest of the compounds.

> What is meant by "Is cluster representative"? Is it the most similar molecule to all others in the cluster?
it is exactly that

> There is a way, which is admittedly not very intuitive, because it is not created directly and you must calculate it afterwards:
- Select 'Data->Add Calculated Values...'

- copy/paste this formula: frequency(ClusterNo,"Cluster No")
- You may set 'new column name' to 'Cluster Size' and press OK.
What the frequency function does within every row is: take the value from the 'Cluster No' column (e.g. 1), count how often this values is found in the entire dataset in the 'Cluster No' column and write the count value into the new column named 'Cluster Size' into the same row.

The clustering is very old functionality. It is reproducible and analytic, but requires the entire similarity matrix and plenty of resources if you have large files. I am aware that I should introduce something more efficient for large files. It is on the list among many other things...

Kind regards,

Thomas
