Subject: Re: Similarity analysis using "find similar compounds..." - slow analysis of libraries
Posted by thomas on Thu, 26 Nov 2020 17:16:19 GMT
View Forum Message <> Reply to Message

I could confirm this with two files containing 16000 random structures each. Two reasons together cause the incredibly bad performance (many numbers and bad sorting):

A new column receives all individual similarity values to compounds of the second file, which are higher than the given threshold. With a 30% limit these are about 70% of the other file's molecules. Therefore, about 8000 to 12000 similarity values were put into every row of the first file. The individual similarity values are kept sorted by DataWarrior. Unfortunately, this was done in a very inefficient way by keeping the cell content as text, converting it to numbers to find the insert position for the new value. With just a few values this is not a problem, but with thousands of values, this was very expensive. I have updated the source code. The next development release early December will contain the fix. Now it takes about 2 minutes.

If your original idea was to get a distribution of all mutual similarities between any pair of molecules in a file, than there is a much faster way: Launch DataWarrior in development mode (with Java option '-Ddevelopment=true'). The you get a few undocumented additional items in the chemistry menu. 'Compare Descriptor Similarity Distribution' counts and shows a graph of the Gaussian like curve of all similarity values binned into 1-percent bins.