Subject: Re: Similarity analysis using "find similar compounds..." - slow analysis of libraries
Posted by thomas on Sat, 28 Nov 2020 23:05:43 GMT
View Forum Message <> Reply to Message

I have changed the algorithm again. Now it just writes the highest similarity and the number of compounds with similarity above threshold into the open file. This accelerates again. Now a 16k by 16k comparison takes about 10 seconds on my computer. A million by a million would probably take around 12 hours.

Putting two sets into one file and use the procedure I suggested earlier would not work for your purpose, because it just uses the complete similarity matrix of all compounds without considering sets. But I hope, the current update works for you. It can be downloaded as development patch from the download page after clicking the 'read and understood' box. The links are in the small print.

This task actually does not need much memory. It basically needs to fit the first file into the memory, which should be possible with even a few million compounds, if the -Xmx setting is adapted. The second file's size doesn't matter much, because it is processed row by row.

Please let me know, if there are problems of any kind.