Subject: Re: molecular complexity descriptor
Posted by thomas on Sun, 04 Nov 2018 00:13:29 GMT
View Forum Message <> Reply to Message

Hi Tim,
the complexity calculation is conceptually very easy but computationally demanding. Its original version calculates the number of distinct structural fragments, which one can construct from a molecule by just cutting parts off. When doing this all delocalized bonds are retained, i.e. marked as such. Then the fragments is converted into a canonical code and added to the list if it is new. The fragment count grows in principal exponentially with the size of the molecule. Therefore we normalize the absolute fragment count by taking its logarithm and devide it by the molecule size. The more distinct fragments, the more complex is the molecule. Molecules with many symmetrical=equivalent atoms, substituents, or molecules with many re-occurring sub-structures are by this logic of low complexity.
For larger molecules the complete creation of all existing sub-structures is rather demanding in terms of memory and time. Therefore DataWarrior uses a fast and simplified version. We have found that if we limit the number of bonds that we allow a fragment to have, we have nevertheless a good estimator for the brute force method's result. DataWarrior limits the fragment generation to a maximum of 7 bonds and calculates the complexity as log(fragmentCount)/bondLimit with bondLimit=7 unless the molecule has less than 14 bonds. Then bondLimit is bondCount/2.

You can find the source code in FastMolecularComplexityCalculator.java as part of the DataWarrior source code.

More detailled info is here:

von Korff M., Sander T. (2013) About Complexity and Self-Similarity of Chemical Structures in Drug Discovery. In: Stavrinides S., Banerjee S., Caglar S., Ozer M. (eds) Chaos and Complex Systems. Springer, Berlin, Heidelberg