Subject: Re: Feature query - molecular fingerprints and library diversity
Posted by thomas on Mon, 15 Jun 2020 19:26:59 GMT
View Forum Message <> Reply to Message

I am not sure, whether I correctly understand your questions. Is it the following?

1) How many distinct descriptors do we have within a given compound library?

The topological descriptors in DataWarrior are comprised of 512 or even 1024 fragments or hash codes derived from fragments. Therefore, different structures usually have different combinations of contained fragments and, therefore, different descriptors. Except for very rare cases the number of different compounds is equal to the number of different descriptors. By creating a list of duplicate structures or removing duplicate structures you can easily determine the number of distinct structures, which usually is the number of distinct descriptors. You may use a trick make the encoded descriptor visible and to directly remove rows with duplicate descriptors: open the dwar file in a text editor and remove the four column property lines of a descriptor column, which look like this:
<columnName="FragFp">
<columnProperty="parent Structure">
<columnProperty="specialType FragFp">
<columnProperty="version 1.2.1">

After removing these lines, DataWarrior does not know anymore that the column contains descriptors that are associated to the chemical structure. Instead it shows the text encoded descriptors in a visible column, which you can use to remove duplicates.
It depends on the descriptor, whether some very similar structures actually end up with the same descriptor. The simple descriptors do not use stereo-chemistry. Therefore, different stereo isomers indeed have the same descriptor. The SkeletonSpheres descriptor does not have these issues and you will have a hard time finding two different molecules that have an identical SkeletonSpheres.

Or do you ask this?

1) How many of the 227787 fragments listed in the paper (or any other fragment collection) are a substructure of at least one compound of a given compound collection?

2) Create a list of fragments found in a given molecule including the count?

Both cannot be done with DataWarrior. But with a little Java programming using the open-source framework OpenChemLib that wouldn't be a big undertaking.

Please le tme know, if I misinterpreted your question.

Thomas