Subject: suggest: adjustment .sdf export
Posted by nbehrnd on Fri, 15 May 2020 15:59:41 GMT
View Forum Message <> Reply to Message

Prior to further analysis of a library,[1] its entries were deduplicated by Data -> merge equivalent rows, using content of the structure column as sole criterion. The work with the .sdf subsequently generated by DataWarrior worked fine if the compound name column used the row number.

Yet, retaining the information of the molecules' name -- here, a PubChem identifier -- may be useful as a structure may be attributed more than one.[2] The corresponding choice of compound name column to equate automatic may then yield a .sdf which is not understood, e.g. by openbabel (version 3.0.0, April 2020).

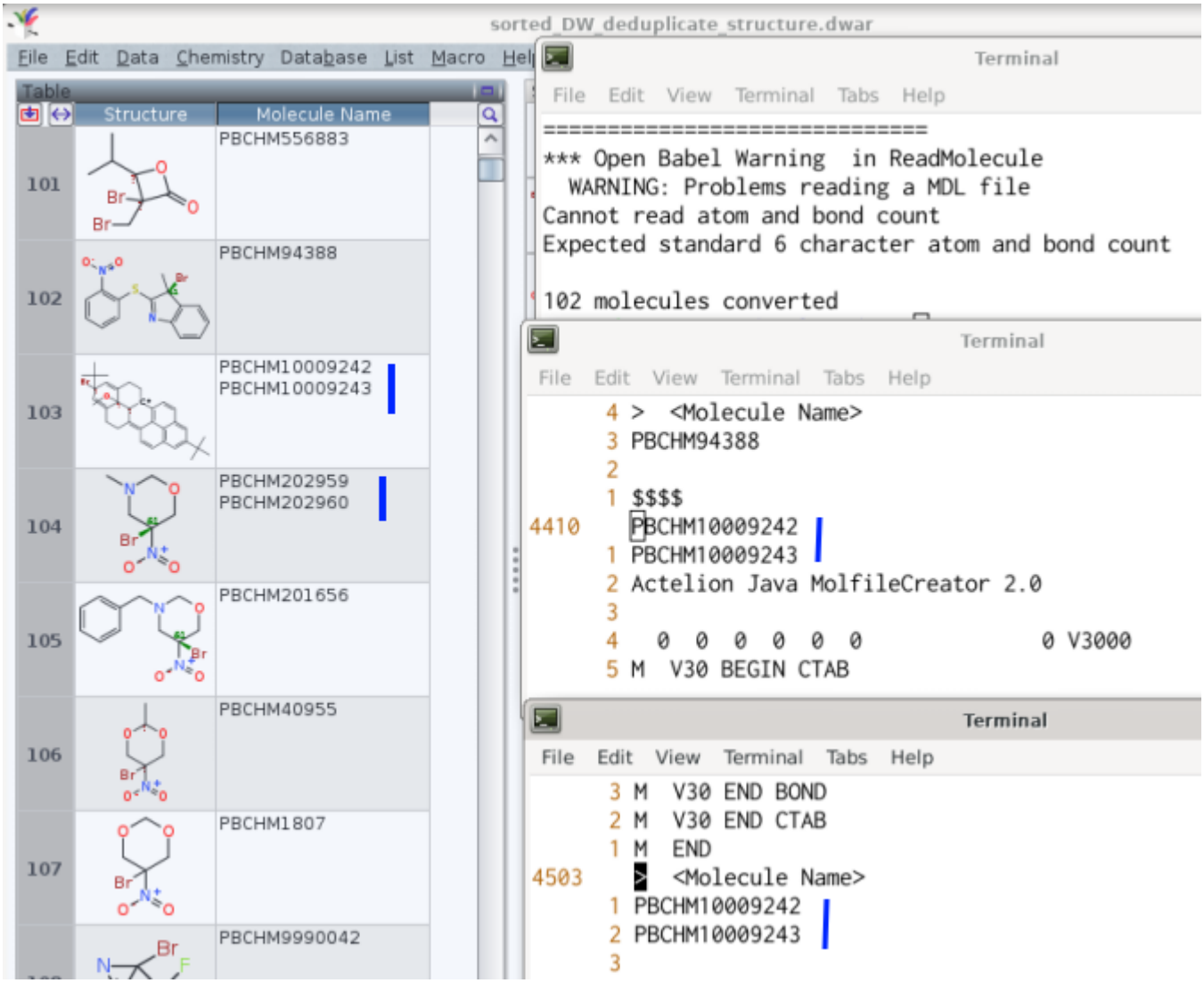
The suggestion for this type of .sdf export by DW is to report the molecules names in the data's header / footer on one line, separated only by a blank space.

The archived .dwar equally contains cells with more then one multiple occurrence of the same PubChem number (e.g. cell #46 about PBCHM2982, PBCHM47354, and PBCHM40585).
The desideratum for cases like this one is to retain only one occurence of each PubChem number per cell.

[1] https://github.com/IanAWatson/Lilly-Medchem-Rules/blob/master/test/example_molecules.smi, revision Apr 26, 2020
[2] E.g., https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702940/

File Attachments

1) format_suggest.png, downloaded 1089 times



2) testinput.zip, downloaded 658 times
3) sorted_DW_deduplicate_structure.dwar.zip, downloaded 690 times

---

Subject: Re: suggest: adjustment .sdf export
Posted by thomas on Sat, 23 May 2020 10:48:35 GMT
View Forum Message <> Reply to Message

Thank you very much for the detailed description of the problem. I have fixed the issue and in the next update it will be included. The behaviour is now to replace any NEWLINE characters by a '; ' string when writing the content of an associated compound name column into the first line of the molfile. When DataWarrior reads an SD-File with these entries again, it recognizes the names as separate ones again, because for DataWarrior the '; ' is a natural separator.
