Subject: Question about R-group occurences estimation
Posted by jvolvr on Tue, 21 Jan 2020 15:02:21 GMT
View Forum Message <> Reply to Message

Is there a way to calculate the number of occurrences of a given group (e.g. hydroxyls, methyls) in my data set through the SMILES of my compounds?

Subject: Re: Question about R-group occurences estimation
Posted by thomas on Tue, 21 Jan 2020 20:14:34 GMT
View Forum Message <> Reply to Message

not yet, but it absolutely makes sense to introduce a custom substructure count feature. The substructure search itself is able to count occurrences, thus this is more matter of where to place the functionality in the user interface. I will think about it... Thomas

Subject: Re: Question about R-group occurences estimation
Posted by nbehrnd on Tue, 11 Feb 2020 13:53:18 GMT
View Forum Message <> Reply to Message

There are two possible difficulties to retrieve methyl groups as-such in a list of SMILES. For one, the characteristic of them is the (single) carbon atom and searching just for this is less identifying than the string of c1ccccc1 about benzene, for example. Second, probably there would be a need to add explicit hydrogens on all SMILES before these would be easier to identify (which may be done, e.g., with babel).

If you have access to Python, then the additional module by rdkit (http://rdkit.org/) may be quite helpful to querry your SMILES with SMARTS. With the test file of smiles_list.smi attached below, the identification of methyl groups (in SMART's convention, expressed as [CH3]) works fine both locally -- per SMILES entry -- as well as in counting the globally:

```
from rdkit import Chem
smiles_source = "smiles_list.smi"
grand_total = 0

# example pattern to identify and count:
functional_group = Chem.MolFromSmarts('[CH3]')  #  methyl group

# alternative examples:
#functional_group = Chem.MolFromSmarts('c1ccccn1')  # for a pyridine
#functional_group = Chem.MolFromSmarts('C1CCCCC1')  # for cyclohexane

with open(smiles_source, mode="r") as source_file:
    for index, line in enumerate(source_file, start=1):
        molecule = Chem.MolFromSmiles(line.strip())
        match = molecule.GetSubstructMatches(functional_group)

        print("{:3} matches in entry {:2}: {}.".format(
            len(match), index, line.strip()))

        grand_total += len(match)

print("\nIn total {} instances were identified.".format(grand_total))
```

As DataWarrior relies on java, the implementation of the «ErtlFunctionalGroupsFinder» described by Fritsch et al. (https://jcheminf.biomedcentral.com/articles/10.1186/s13321-019-0361-8, open access) equally may be of interest for Thomas.

File Attachments

1) example.png, downloaded 1282 times
2) listing.png, downloaded 1201 times
3) smiles_list.smi, downloaded 631 times
4) example.py, downloaded 607 times

Subject: Re: Question about R-group occurences estimation
Posted by thomas on Wed, 12 Feb 2020 18:40:51 GMT

View Forum Message <> Reply to Message

This is now built in: "Chemistry->From Chemical Structure->Add Substructure Count..."
This opens a dialog to
-select the structure column
-to draw a fragment with optional query features
-and to define whether counted fragments may overlap on some atoms

It is in the current beta and will be in the next official version that is due in a few days