3_MoleculeSetsAndSimilarityExample

October 23, 2019

1 Molecule sets and similarity example

Andrea Volkamer

1.1 Working with Tables and csv Files

[1]:
```
import pandas as pd
df = pd.read_csv('./data/EGFR-course.csv', delimiter=',', names=['Smiles',␣
↪'Name'], header=None)
df.head()
```

[1]:
```
                                              Smiles         Name
0     COc1cc2ncnc(Nc3ccc(F)c(Cl)c3)c2cc1OCCCN1CCOCC1    Gefitinib
1         C#Cc1cccc(Nc2ncnc3cc(OCCOC)c(OCCOC)cc23)c1    Erlotinib
2  CS(=O)(=O)CCNCc1ccc(-c2ccc3ncnc(Nc4ccc(OCc5ccc...    Lapatinib
3  CN(C)C/C=C/C(=O)Nc1cc2c(Nc3ccc(F)c(Cl)c3)ncnc2...     Afatinib
4  C=CC(=O)Nc1cc(Nc2nccc(-c3cn(C)c4ccccc34)n2)c(O...  Osimertinib
```

[2]:
```
df.shape
```

[2]:
```
(5, 2)
```

[3]:
```
df.columns
```

[3]:
```
Index(['Smiles', 'Name'], dtype='object')
```

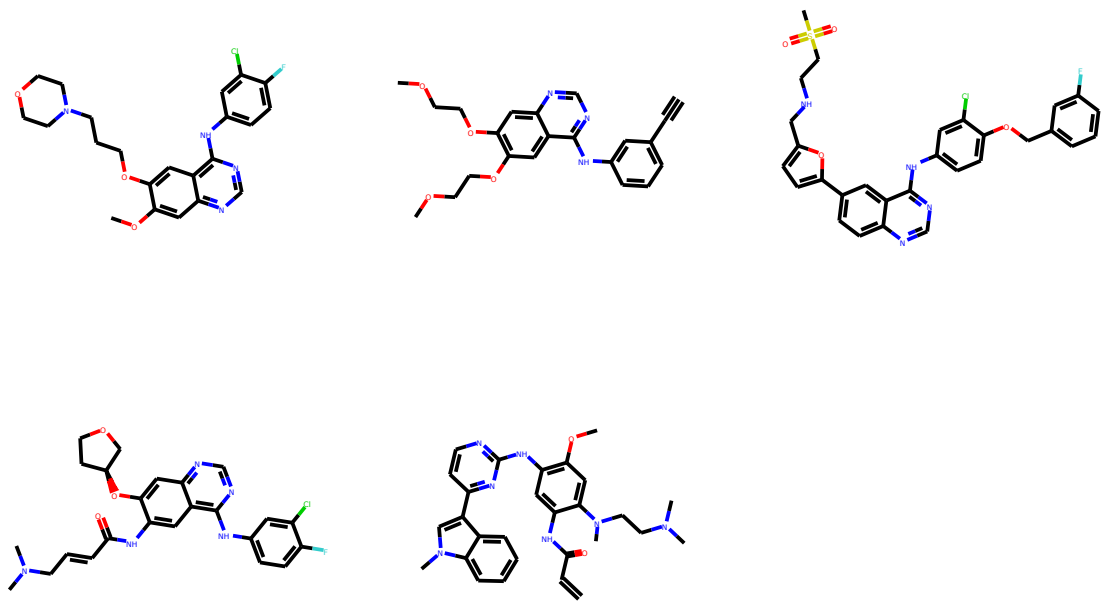
[4]:
```
from rdkit import Chem
from rdkit.Chem.Draw import IPythonConsole
from rdkit.Chem import Draw, PandasTools

PandasTools.AddMoleculeColumnToFrame(df, smilesCol='Smiles')

# Draw molecules
df[['Name', 'ROMol']]
```

[4]:
```
          Name                                                ROMol
0    Gefitinib  <img src="...
1    Erlotinib  <img src="...
2    Lapatinib  <img src="...
3     Afatinib  <img src="...
```

```
4  Osimertinib  <img src="...
```

[5]:
```
# Or
Draw.MolsToGridImage(list(df.ROMol), useSVG=True)
```

[5]:



Add descriptor column to dataframe.

[6]:
```
from rdkit.Chem import Descriptors
df['HeavyAtoms'] = df.apply(lambda x: Descriptors.HeavyAtomCount(Chem.
 ↪MolFromSmiles(x['Smiles'])), axis=1)
```

[7]:
```
df.head()
```

[7]:
```
                                                Smiles          Name  \
0     COc1cc2ncnc(Nc3ccc(F)c(Cl)c3)c2cc1OCCCN1CCOCC1    Gefitinib
1         C#Cc1cccc(Nc2ncnc3cc(OCCOC)c(OCCOC)cc23)c1    Erlotinib
2  CS(=O)(=O)CCNCc1ccc(-c2ccc3ncnc(Nc4ccc(OCc5ccc...    Lapatinib
3  CN(C)C/C=C/C(=O)Nc1cc2c(Nc3ccc(F)c(Cl)c3)ncnc2...     Afatinib
4  C=CC(=O)Nc1cc(Nc2nccc(-c3cn(C)c4ccccc34)n2)c(O...  Osimertinib

                                                ROMol  HeavyAtoms
0  <img src="...          31
1  <img src="...          29
2  <img src="...          40
3  <img src="...          34
4  <img src="...          37
```

1.2 Molecular similarity

Find most similar compound - which one is the most similar one to Gefitinib?

[8]:
```
from rdkit.Chem import AllChem
```

[9]:
```
from rdkit.Chem import rdFingerprintGenerator
from rdkit import DataStructs

# Generate fingerprints for the molecules
# By default the RDKit uses Morgan Fingerprints of radius 2 (MFP2)
fps = rdFingerprintGenerator.GetCountFPs(list(df.ROMol))

# Gefitinib as reference
fp1 = fps[0]

# Calculate similarity to reference
all_sim = DataStructs.BulkDiceSimilarity(fp1,fps)
df['sim2Gefitinib'] = all_sim
```

[10]:
```
df.head()
```

[10]:
```
                                                Smiles         Name  \
0     COc1cc2ncnc(Nc3ccc(F)c(Cl)c3)c2cc1OCCCN1CCOCC1    Gefitinib
1         C#Cc1cccc(Nc2ncnc3cc(OCCOC)c(OCCOC)cc23)c1    Erlotinib
2  CS(=O)(=O)CCNCc1ccc(-c2ccc3ncnc(Nc4ccc(OCc5ccc...    Lapatinib
3  CN(C)C/C=C/C(=O)Nc1cc2c(Nc3ccc(F)c(Cl)c3)ncnc2...     Afatinib
4  C=CC(=O)Nc1cc(Nc2nccc(-c3cn(C)c4ccccc34)n2)c(O...  Osimertinib

                                                 ROMol  HeavyAtoms  \
0  <img src="...          31
1  <img src="...          29
2  <img src="...          40
3  <img src="...          34
4  <img src="...          37

   sim2Gefitinib
0       1.000000
1       0.609195
2       0.517073
3       0.641711
4       0.371134
```

[11]:
```
df.sort_values(['sim2Gefitinib'], inplace=True, ascending=False)
```

[12]:
```
df.head()
```

[12]:
```
                                                Smiles         Name  \
0     COc1cc2ncnc(Nc3ccc(F)c(Cl)c3)c2cc1OCCCN1CCOCC1    Gefitinib
3  CN(C)C/C=C/C(=O)Nc1cc2c(Nc3ccc(F)c(Cl)c3)ncnc2...     Afatinib
1         C#Cc1cccc(Nc2ncnc3cc(OCCOC)c(OCCOC)cc23)c1    Erlotinib
```

```
2  CS(=O)(=O)CCNCc1ccc(-c2ccc3ncnc(Nc4ccc(OCc5ccc...    Lapatinib
4  C=CC(=O)Nc1cc(Nc2nccc(-c3cn(C)c4ccccc34)n2)c(O...  Osimertinib

                                                ROMol  HeavyAtoms  \
0  <img src="...          31
3  <img src="...          34
1  <img src="...          29
2  <img src="...          40
4  <img src="...          37

   sim2Gefitinib
0       1.000000
3       0.641711
1       0.609195
2       0.517073
4       0.371134
```

[13]:
```python
# Save file
df.drop('ROMol', axis=1).to_csv('./data/mytest_csvFile.csv')
```

[ ]:
```python

```