Automated identification of chemical series with RDKit

pen@iwatobipen

https://iwatobipen.wordpress.com/

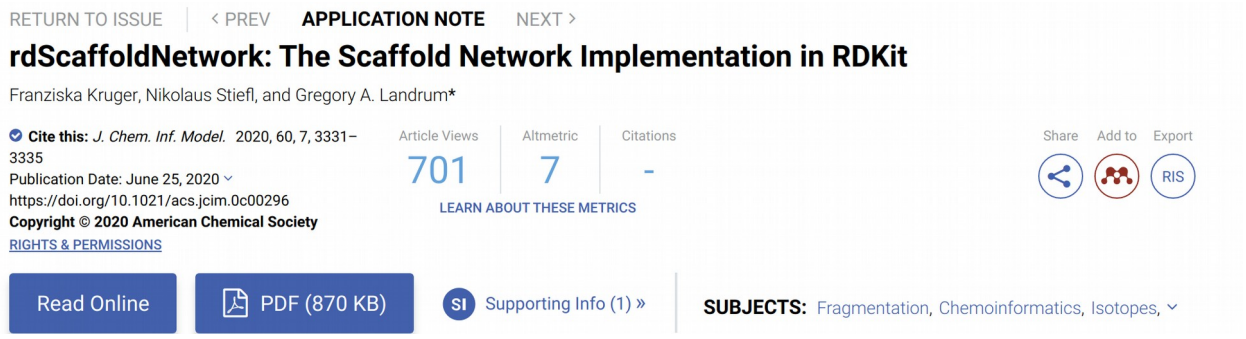
https://github.com/iwatobipen

Chemical series are very important basis for SAR. How do you define them?

Fragmentation of molecules is the common strategy in Cemoinformatics area

1) Bermis Murcko, the RECAP and BRICS
2) Scaffold trees / Scaffold networks
3) etc..

All of them are implemented in RDKit!! ;)



https://pubs.acs.org/doi/10.1021/acs.jcim.0c00296

Recent publication about new approach from Nikolaus et. al.

The authors shared their code in SI!

RETURN TO ISSUE | < PREV CHEMICAL INFORMATION NEXT >

Automated Identification of Chemical Series: Classifying like a Medicinal Chemist

Franziska Kruger, Nikolas Fechner, and Nikolaus Stiefl*

Cite this: *J. Chem. Inf. Model.* 2020, 60, 6, 2888–2902

Publication Date: May 6, 2020

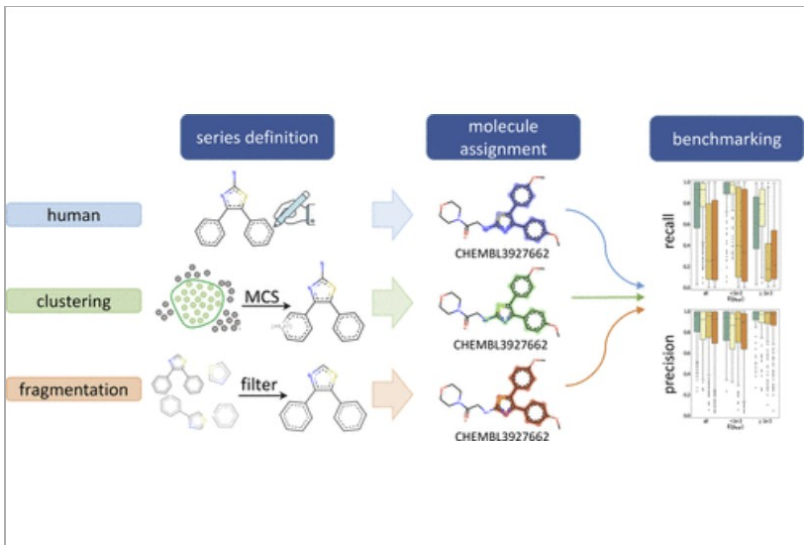
https://doi.org/10.1021/acs.jcim.0c00204

Article Views 965 | Altmetric 9 | Citations 1

LEARN ABOUT THESE METRICS

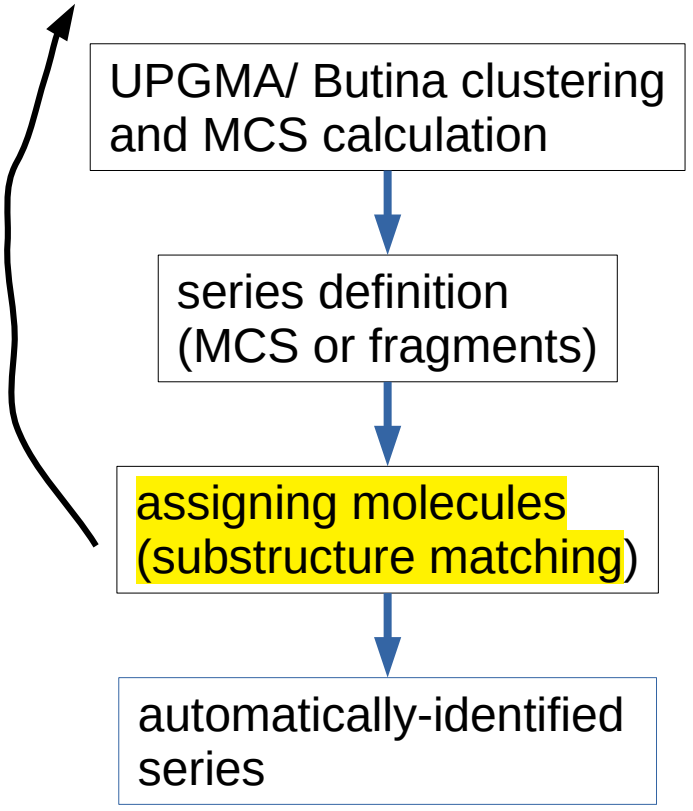
Share Add to Export RIS

https://pubs.acs.org/doi/abs/10.1021/acs.jcim.0c00204



Key point is MCS saearch and Fast SSS in ChEMBL!!!

Searching the frequency of query MCS in ChEMBL is required.
The original implementation used **arthor** which is developed by NextMove.
Arthor is High-Performance Chemical Database Searching tool.



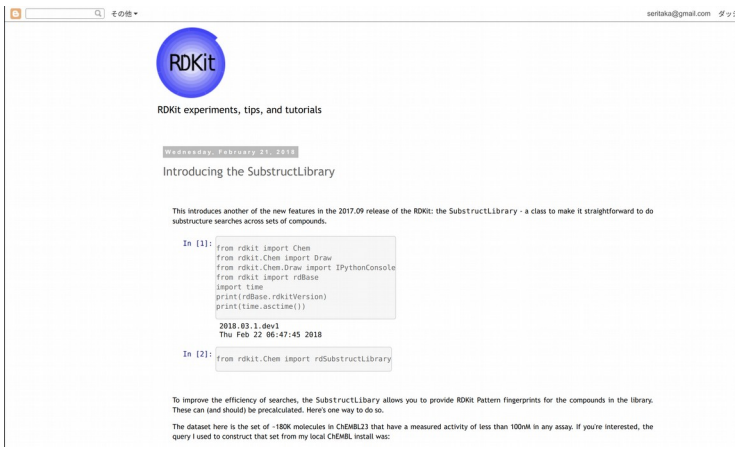
utilsStructureEval.py

```python
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Mon Jan 27 16:29:23 2020

@author: krugefr1
"""
from rdkit import Chem
from rdkit.Chem import rdFMCS
import arthor


def MCSFromMollist(mollist,chembldb,Nchembl):
    MCSSmarts2=rdFMCS.FindMCS(mollist,atomCompare=rdFMCS.AtomCompare.CompareAny,bondCompare=rdFMCS.BondCompare.CompareOrderExact,ringMatchesRingOn
    MCSSmarts=rdFMCS.FindMCS(mollist,atomCompare=rdFMCS.AtomCompare.CompareElements,bondCompare=rdFMCS.BondCompare.CompareOrder,ringMatchesRingOnl
    if MCSSmarts2=='': fChembl2=1
    else: fChembl2=getFChembl(MCSSmarts2,chembldb,Nchembl)
    if MCSSmarts=='': fChembl=1
    else:fChembl=getFChembl(MCSSmarts,chembldb,Nchembl)
    if fChembl2<fChembl:
        fChembl=fChembl2
        MCSSmarts=MCSSmarts2
    return fChembl,MCSSmarts

def getFChembl(qry,chembldb,Ntot,qryformat='Smarts'):
    if qryformat=='Smarts':
        results=chembldb.search(qry)
    elif qryformat=='MDL':
        with open(qry) as f:
            qryarthor=arthor.Query(f.read(),"Mdl")
        results=chembldb.search(str(qryarthor))
    fChembl=(len(results)+1)/(Ntot+2)
    return fChembl
```

J. Chem. Inf. Model. 2020, 60, 6, 2888–2902
Arthor https://www.nextmovesoftware.com/arthor.html

I would like implement the code with open source package...

- I tried to replace MCS search code from arthor to RDKit postgresql cartridge at first but it is not enough.
- But Greg shared cool code to search chemical series, I would like to share the code in the UGM ;)
- RDKit has ultra fast substructure search module named **‘rdSubstructLibrary’**.



http://rdkit.blogspot.com/2018/02/introducing-substructlibrary.html

rdSubstructLibrary can search compounds in a short time!

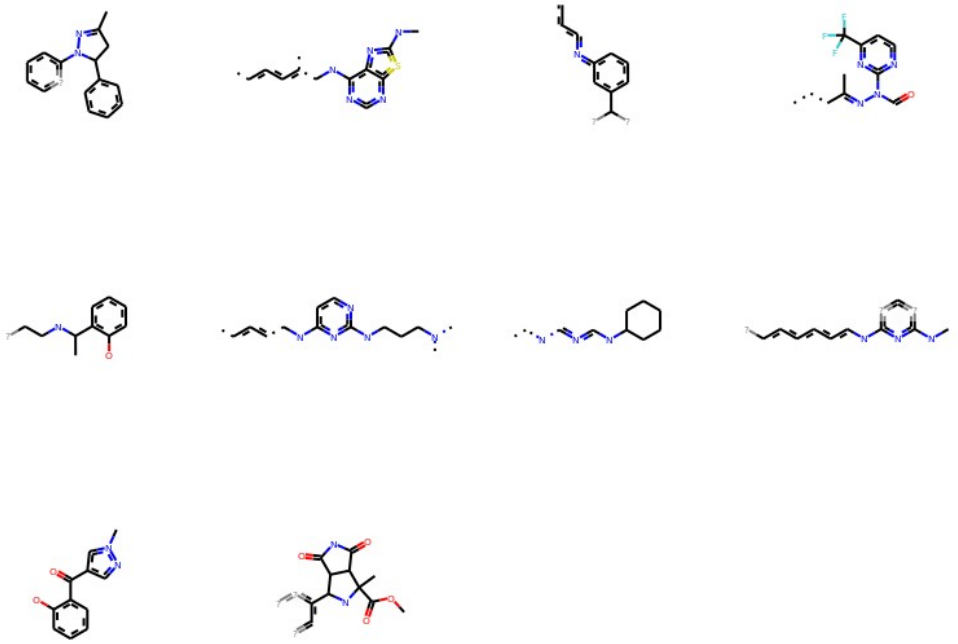
with rdkit-postgres

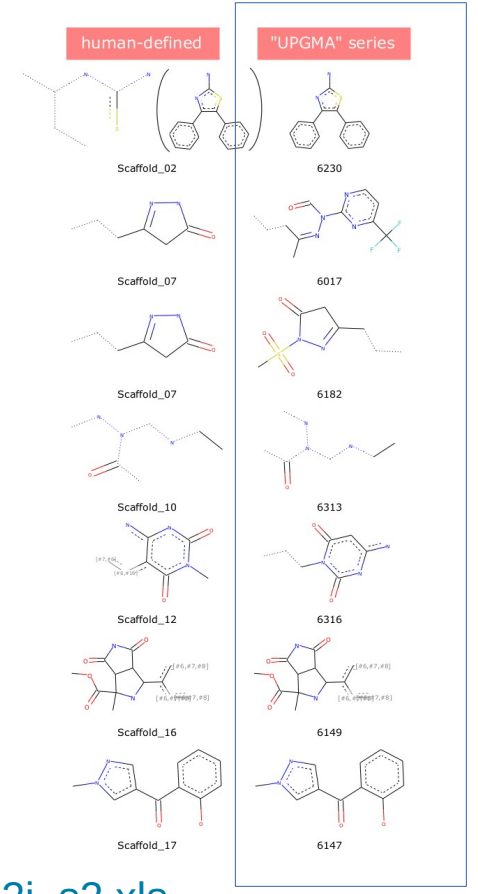
```
In [5]: q = chembldb.query(Mols.m)
        res = q.filter(Mols.m.hassubstruct(mol_from_smarts('a1aaaaa1'))).limit
        (10000)
        %time a=res.all()

CPU times: user 615 ms, sys: 8.61 ms, total: 624 ms
Wall time: 3.41 s

In [6]: # subsearch molecule with SMARTS and rdSubstructLibrary.

In [7]: %time a=library.GetMatches(Chem.MolFromSmarts('a1aaaaa1'), maxResults=1
        0000)

CPU times: user 910 ms, sys: 0 ns, total: 910 ms
Wall time: 342 ms
```

with rdSubstructLibrary

https://iwatobipen.wordpress.com/2020/07/18/fast-substructure-search-module-of-rdkit-rdkit-memo/

The implementation worked as same as the original code

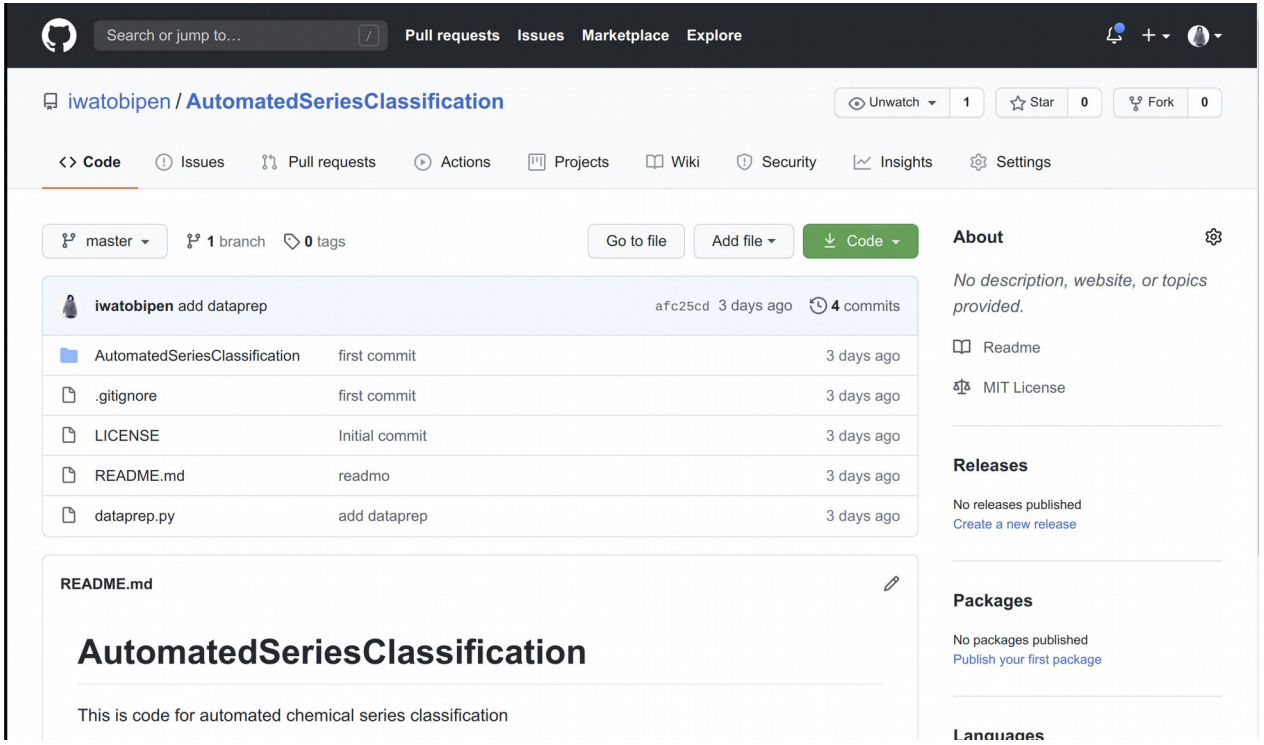




Data set came from
https://pubs.acs.org/doi/suppl/10.1021/jm020472j/suppl_file/jm020472j_s2.xls
ChEMBL27 was used in this work but original work used ChEMBL25 so, there are some differences between the article and the implementation.

Code Availability

I’ll add short documentation about how to run the code in a few days



https://github.com/iwatobipen/AutomatedSeriesClassification

Acknowledgements

- Greg Landrum
- RDKit community