



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 Chemoinformatics area

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

All of them are implemented in RDKit!! ;)

RETURN TO ISSUE | < PREV APPLICATION NOTE NEXT >

rdScaffoldNetwork: The Scaffold Network Implementation in RDKit

Franziska Kruger, Nikolaus Stiefl, and Gregory A. Landrum*

Cite this: *J. Chem. Inf. Model.* 2020, 60, 7, 3331–3335

Publication Date: June 25, 2020

<https://doi.org/10.1021/acs.jcim.0c00296>

Copyright © 2020 American Chemical Society

[RIGHTS & PERMISSIONS](#)

Article Views

701

Altmetric

7

Citations

-

[LEARN ABOUT THESE METRICS](#)

Share Add to Export



Read Online

PDF (870 KB)

Supporting Info (1) »

SUBJECTS: Fragmentation, Chemoinformatics, Isotopes, ▾

<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 Stief*

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

Publication Date: May 6, 2020

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

Copyright © 2020 American Chemical Society

[RIGHTS & PERMISSIONS](#)

Article Views

965

Altmetric

9

Citations

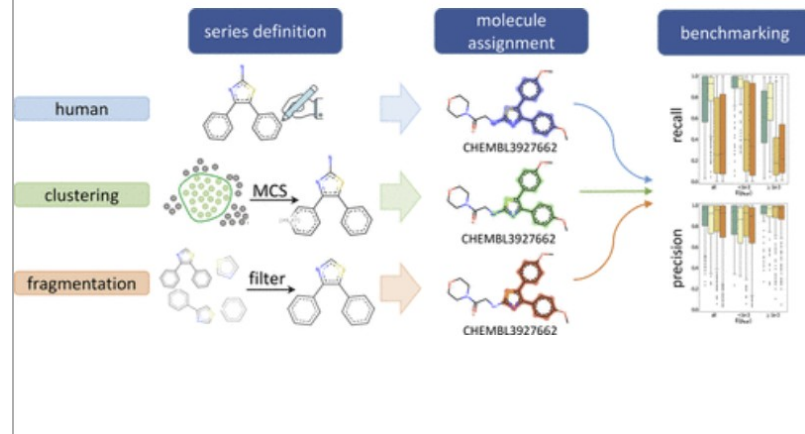
1

[LEARN ABOUT THESE METRICS](#)

Share Add to Export



<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 **author** which is developed by NextMove.
Author is High-Performance Chemical Database Searching tool.

UPGMA/ Butina clustering
and MCS calculation

series definition
(MCS or fragments)

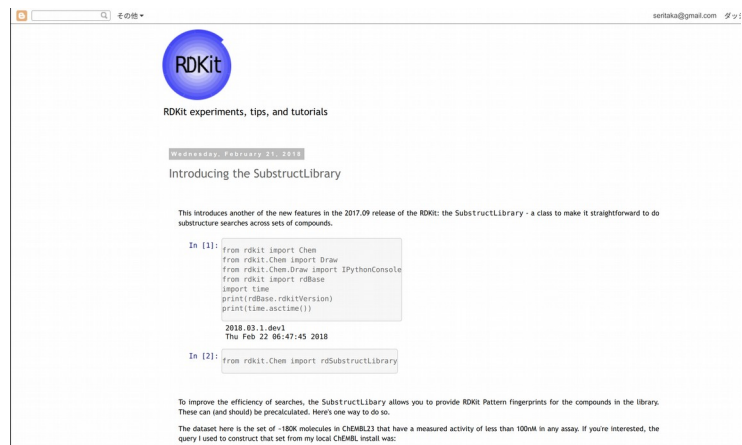
assigning molecules
(substructure matching)

automatically-identified
series

```
utilsStructureEval.py x
utilsStructureEval.py
1  #!/usr/bin/env python3
2  # -*- coding: utf-8 -*-
3  """
4  Created on Mon Jan 27 16:29:23 2020
5
6  @author: krueger1
7  """
8  from rdkit import Chem
9  from rdkit.Chem import rdFMCS
10 import author
11
12
13 def MCSFromMolList(molList, chemdb, Nchembl):
14     MCSsmarts2=rdFMCS.FindMCS(molList,atomCompare=rdFMCS.AtomCompare.CompareAny,bondCompare=rdFMCS.BondCompare.CompareOrderExact,ringMatchesRingOn
15     MCSsmarts2=rdFMCS.FindMCS(molList,atomCompare=rdFMCS.AtomCompare.CompareElements,bondCompare=rdFMCS.BondCompare.CompareOrder,ringMatchesRingOn
16     if MCSsmarts2=='': fChembl2=1
17     else: fChembl2=getFChembl(MCSsmarts2,chemdb,Nchembl)
18     if MCSsmarts=='': fChembl=1
19     else: fChembl=getFChembl(MCSsmarts,chemdb,Nchembl)
20     if fChembl2<fChembl:
21         fChembl=fChembl2
22     MCSsmarts=MCSsmarts2
23     return fChembl,MCSsmarts
24
25 def getFChembl(qry,chemdb,Ntot,qryformat='Smarts'):
26     if qryformat=='Smarts':
27         results=chemdb.search(qry)
28     elif qryformat=='MDL':
29         with open(qry) as f:
30             qryarhor=author.Query(f.read(),"Mdl")
31             results=chemdb.search(str(qryarhor))
32     fChembl=(len(results)+1)/(Ntot+2)
33     return fChembl
```

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'**.



rdSubstructLibrary can search compounds in a short time!

with rdkit-postgres

```
In [5]: q = chembl.db.query(Mols.m)
        res = q.filter(Mols.m.has_substruct(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

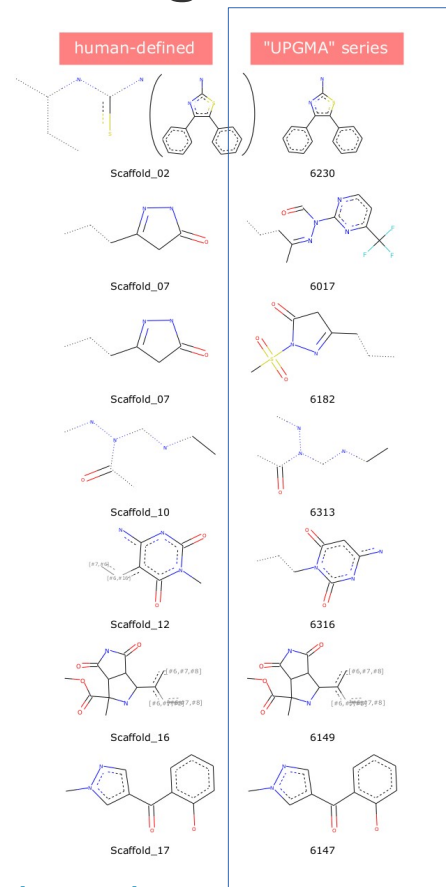
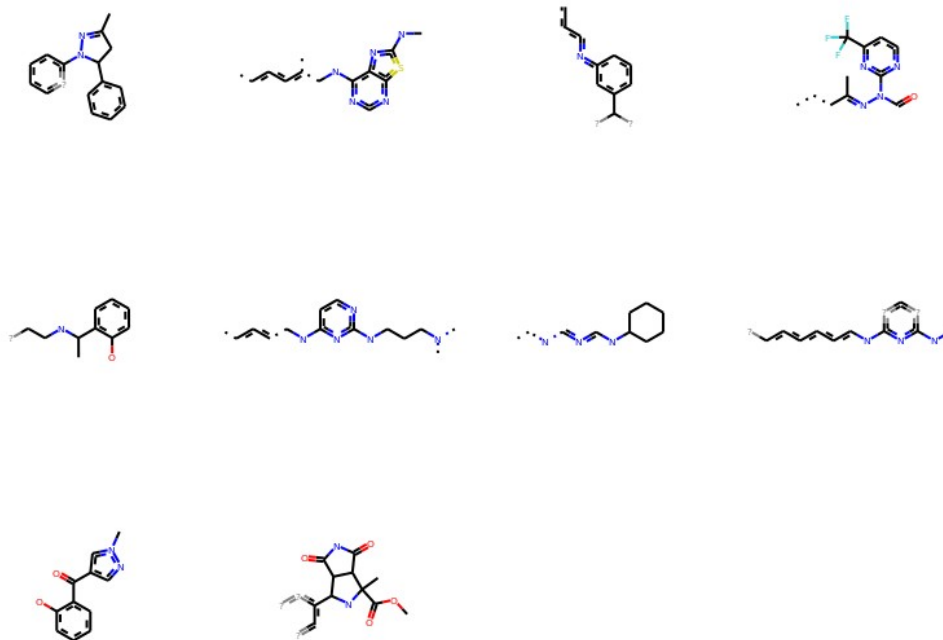
```
In [6]: # subsearch molecule with SMARTS and rdSubstructLibrary.
```

```
In [7]: %time a=library.GetMatches(Chem.MolFromSmarts('a1aaaaa1'), maxResults=10000)
```

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

with rdSubstructLibrary

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

The screenshot shows the GitHub repository page for `iwatobipen / AutomatedSeriesClassification`. The repository has 1 pull request, 0 stars, and 0 forks. The `Code` tab is selected, showing the file structure and commit history. The file structure includes `AutomatedSeriesClassification`, `.gitignore`, `LICENSE`, `README.md`, and `dataprep.py`. The commit history shows a single commit by `iwatobipen` with the message `add dataprep`. The `README.md` file is displayed, showing the title `AutomatedSeriesClassification` and the text `This is code for automated chemical series classification`. The right sidebar contains sections for `About`, `Releases`, `Packages`, and `Languages`.

Search or jump to... Pull requests Issues Marketplace Explore

Unwatch 1 Star 0 Fork 0

<> Code Issues Pull requests Actions Projects Wiki Security Insights Settings

master 1 branch 0 tags Go to file Add file Code

iwatobipen add dataprep afc25cd 3 days ago 4 commits

AutomatedSeriesClassification	first commit	3 days ago
.gitignore	first commit	3 days ago
LICENSE	Initial commit	3 days ago
README.md	readmo	3 days ago
dataprep.py	add dataprep	3 days ago

README.md

AutomatedSeriesClassification

This is code for automated chemical series classification

About No description, website, or topics provided. Readme MIT License

Releases No releases published Create a new release

Packages No packages published Publish your first package

Languages

<https://github.com/iwatobipen/AutomatedSeriesClassification>



Acknowledgements

- Greg Landrum
- RDKit community