Clustering of large amount of molecules

RDKit UGM 2018
Takayuki Serizawa
pen@iwatobipen

There are many RDKitters in Japan!

rdkit-users-jp.github.io

rdkit-users-jp WEB pages

View My GitHub Profile

rdkit-users-jp

rdkit-users.jp は RDKit の日本ユーザー会です。
RDKit のユーザーであれば、どなたでも参加することが可能です。

公式ドキュメントの翻訳

TransifexにてRDKitのドキュメントを翻訳しています。

- https://www.transifex.com/rdkit-users-jp/document-translation/

コミュニケーションツール

Slack（おすすめ）

- Join Slack

Maling list

- https://groups.google.com/forum/#!forum/rdkit-users-jp/

Twitter

- hash_tag: #rdkitjp

Hosted on GitHub Pages — Theme by orderedlist

http://rdkit-users.jp/

Motivation

I’d like to cluster large set of molecules!

I am a member of *J-CLIC.

J-CLIC: Japan Compound Library Consortium

- **> 16 companies participate**
- **collaborative collection > 150,000 cpds over 5 years**

Why need fast clustering ?

- **To compare in-house / external compound libraries.**
- **Pairwise similarity calculation is time consuming.**

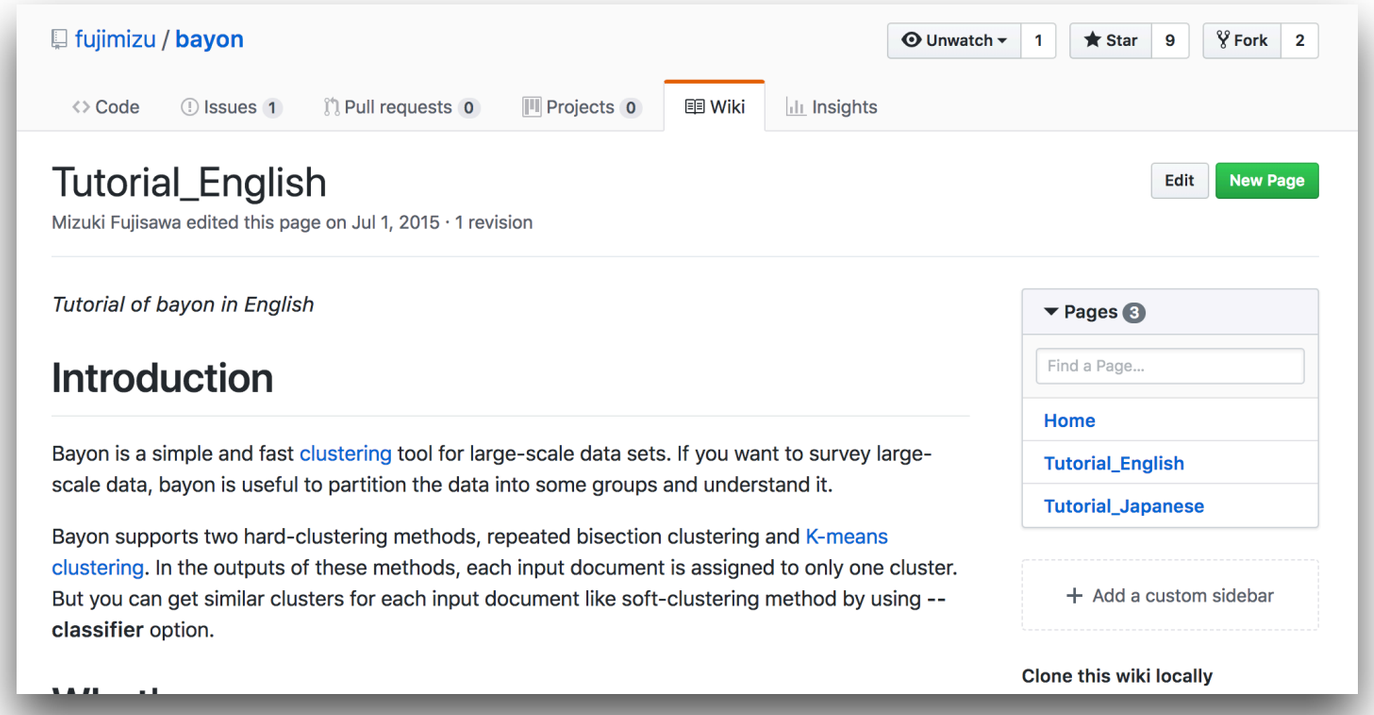
Contrib/Fastcluster

- Conducts molecular fingerprint calculation and clustering molecules.
  - Calls Bayon in subprocess.

There is bug in the script.

Bayon is powerful software for clustering

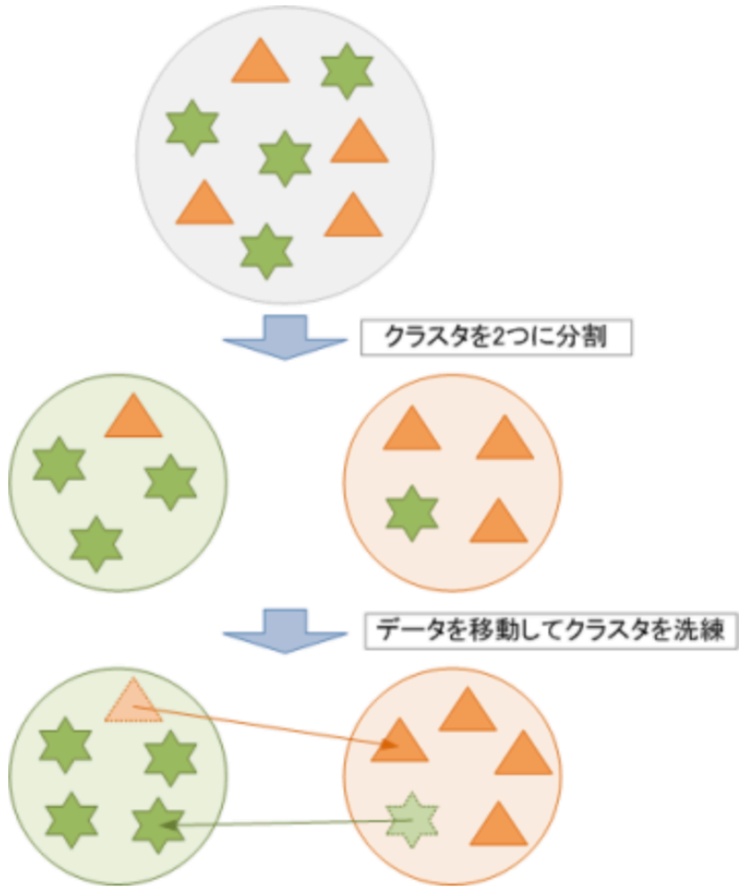
- **Supports Repeated Bisection and K-means**
- **Works very fast!**



https://github.com/fujimizu/bayon
https://github.com/fujimizu/bayon/wiki/Tutorial_English

Repeated Bisection

- Repetitively divide the into two
  - Applying K-means (k=2) repetitively



Benchmark

***Used ChEMBL24 dataset**
(1,820,030 smiles)

Number of output clusters

$ python fastcluster.py chembl_24_1.smi 10000

=>>> 17 minutes for fingerptint calculation
=>>> 18 minutes for clustering of 1,820,030 molecules

MacBook Pro (Mid 2014)
Processor 2.5 GHz Intel Core i7
Memory 16GB RAM

Thank you for your attention

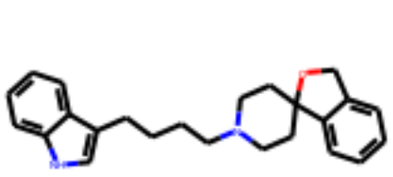
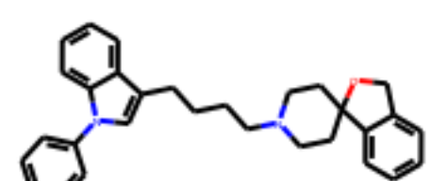
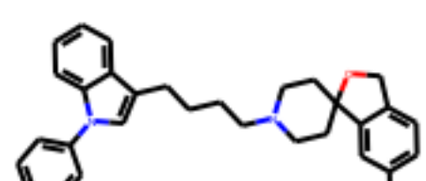
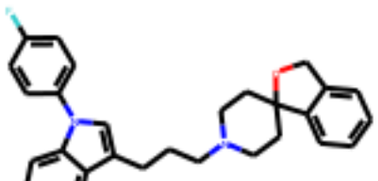
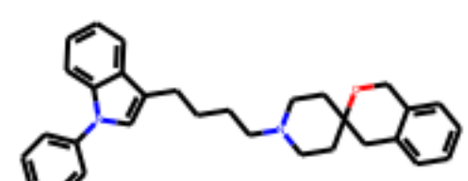
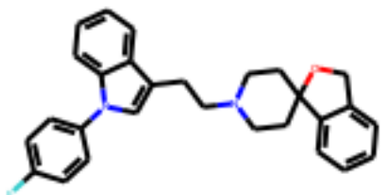
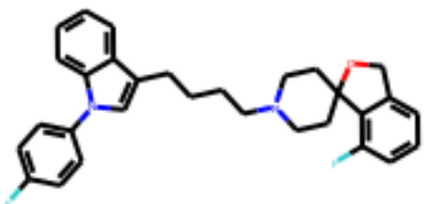
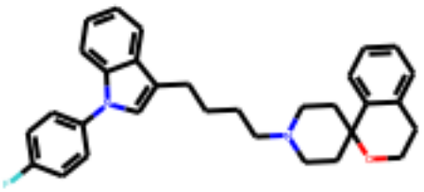
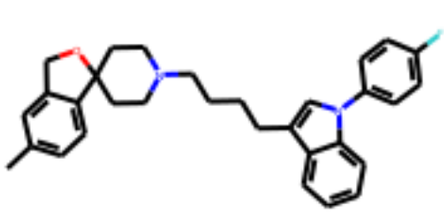
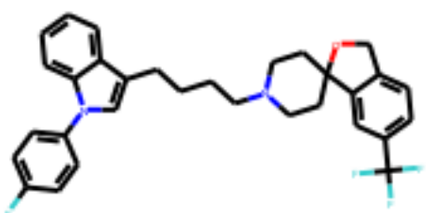
Any advice and suggestions are gratefully appreciated!

You can find me at:
https://iwatobipen.wordpress.com
https://github.com/iwatobipen

Appendix

Added clustered data and Notebook

CLS_ID_9900

CLS_ID_900