Bioinformatics, 2020, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month 2020
Applications Note

Application Note: Data and Text Mining

ScaffoldGraph: an open-source library for the generation and analysis of molecular scaffold networks and scaffold trees

Oliver B. Scott and A.W. Edith Chan*

Wolfson Institute of Biomedical Research, University College London, Gower Street, London WC1E 6BT, UK

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: ScaffoldGraph (SG) is an open-source Python library and command-line tool for the generation and analysis of molecular scaffold networks and trees, with the capability of processing large sets of input molecules. With the increase in high-throughput screening (HTS) data, scaffold graphs have proven useful for the navigation and analysis of chemical space, being used for visualisation, clustering, scaffold-diversity analysis and active-series identification. Built on RDKit and NetworkX, SG integrates scaffold graph analysis into the growing scientific/cheminformatics Python stack, increasing the flexibility and extendibility of the tool compared to existing software.

Availability and implementation: SG is freely available and released under the MIT license at https://github.com/UCLCheminformatics/ScaffoldGraph

Contact: edith.chan@ucl.ac.uk

1 Introduction

Molecular scaffolds are a core concept in medicinal chemistry. Representing the core structure of a molecule, the composition of a molecular scaffold plays a large role in determining global molecular properties, as well as determining basic shape and geometric orientation of substituents (Hu et al., 2016). As such, scaffolds form an important part of modern drug discovery, especially in techniques such as parallel synthesis where combinatorial compound libraries are constructed around various central ring-systems.

Despite the ubiquity of the scaffold concept, defining a scaffold *in silico* can be subjective when there is ambiguity over which portion of the molecule is considered the core. The popular Bemis and Murcko definition (Bemis and Murcko, 1996) ignores these ambiguities, defining a scaffold as the union of rings plus linker atoms connecting them. To reduce ambiguity, it has been proposed that a molecule can be represented by a tree or network of interrelated scaffolds, defined by the iterative removal of peripheral rings from the Bemis and Murcko structure. The scaffold tree (Schuffenhauer et al., 2006) is one of these proposed methods in which a molecule is represented through a linear arrangement of sub-scaffolds, where during ring dissection only the 'most peripheral' ring is removed, determined by a series of chemically intuitive rules. Aggregation of multiple linear graphs forms a scaffold tree. The scaffold network approach (Varin et al., 2011) deals with the ambiguity of determining the most peripheral ring by an exhaustive enumeration of all possible decompositions. Networks constructed from multiple molecules can be aggregated to form a large multipartite directed acyclic graph. The authors used constructed networks and trees for active series identification, identifying scaffolds statistically enriched for bioactivity using a method called compound set enrichment. With the increase in the availability of HTS data through platforms such as PubChem, this analysis is useful in identifying scaffolds which may be further optimized into drug-like molecules. Scaffold network and tree representations constructed for amodiaquine are shown in Figure 1.

2 ScaffoldGraph

ScaffoldGraph (SG) is a Python library and command-line interface (CLI) tool for the generation and analysis of scaffold networks, trees and various permutations of these concepts. SG focuses on being flexible and extensible, integrating into the open-source Python scientific stack, enabling efficient data analysis and the ability to integrate into drug discovery pipelines.

Whilst there is existing open-source software capable of scaffold network/tree construction, including Scaffold Network Generator (SNG) (Matlock et al., 2013) and Scaffold Hunter (Wetzel et al., 2009), SG aims to extend previous implementations through the implementation of additional features and through the provision of a library of methods for the construction and analysis of scaffold graphs. Some key features are highlighted below.

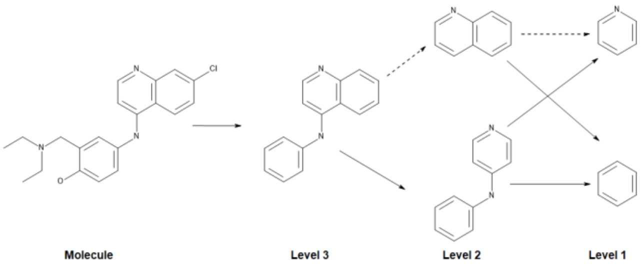

Fig. 1. Scaffold network and tree representations of amodiaquine. The scaffold network encompasses the whole diagram while the scaffold tree only constitutes the dotted path where the removed ring has been determined by a set of chemically intuitive rules. The top level scaffold, in this case level 3, will always represent the Bemis and Murcko scaffold while the subsequent levels represent the removal of a peripheral ring and the resultant side-chains

2.1 Features

- The ability to construct both scaffold networks and scaffold trees. An option to construct HierS type scaffold networks (J. Wilkens et al., 2005) in which fused ring-systems are not dissected during the pruning procedure.
- A utility to easily build custom rules and rule-sets for scaffold prioritization during scaffold tree construction.
- A utility for performing compound set enrichment (Varin et al., 2011, 2010) using scaffold trees or networks coupled with a binomial or Kolmogorov-Smirnov non-parametric hypothesis test.
- Graph structures are built using NetworkX, thus providing the user with a familiar API and multiple graph analysis algorithms out-of-the-box.
- A CLI, analogous to that of SNG (Matlock et al., 2013), with the addition of a flexible and accessible Python library for ease of integration into data analysis pipelines for drug discovery.

2.2 Performance

SG offers substantial performance benefits compared to existing software. Benchmarks were performed against SNG, the current best performing software, using a random selection of 150,000 molecules from the PubChem assay 1063. All tests were performed on an Intel Core i7-6700 3.4 GHz with 32GB of RAM.

SG took 15m 25s to generate a scaffold network from the 150,000 molecule dataset, while SNG took 27m 6s. SG can be run in parallel using multiple input sources, aggregating graphs after construction. When parallelized over 4 processes, SG took just 4m 29s to generate a scaffold network from the same 150,000 molecule dataset.

2.3 Implementation

SG is implemented in the Python programming language, built using the open-source cheminformatics package RDKit (Landrum, 2016) for parsing and manipulating molecular structures, and the network analysis package NetworkX (Hagberg et al., 2008) providing graph structures which scaffold graphs are built upon. For the implementation of compound set enrichment, the SciPy library (Virtanen et al., 2019) is employed for non-parametric hypothesis testing. Further implementation and usage details can be accessed at the GitHub repository.

3 Conclusion

SG provides an accessible Python library and command-line tool for the generation and analysis of scaffold networks and trees. Intended for use within drug discovery pipelines, SG extends existing implementations such as SNG, increasing the flexibility and accessibility of this type of analysis whilst improving performance.

Funding

Oliver B. Scott is supported by the BBSRC grant BB/R506229/1.

Conflict of Interest: none declared.

References

Bemis, G.W. and Murcko, M.A. (1996) The Properties of Known Drugs. 1. Molecular Frameworks. J. Med. Chem., 39, 2887–2893.

Hagberg, A.A. et al. (2008) Exploring Network Structure, Dynamics, and Function using NetworkX. In, Varoquaux,G. et al. (eds), Proceedings of the 7th Python in Science Conference. Pasadena, CA USA, pp. 11–15.

Hu, Y. et al. (2016) Computational Exploration of Molecular Scaffolds in Medicinal Chemistry. J. Med. Chem., 59, 4062–4076.

Wilkens, S.J. et al. (2005) HierS: Hierarchical Scaffold Clustering Using Topological Chemical Graphs. J. Med. Chem., 48, 3182–3193.

Landrum, G. (2016) RDKit: Open-Source Cheminformatics Software.

Matlock, M.K. et al. (2013) Scaffold network generator: a tool for mining molecular structures. Bioinformatics, 29, 2655–2656.

Schuffenhauer, A. et al. (2006) The Scaffold Tree − Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. J. Chem. Inf. Model., 47, 47–58.

Varin, T. et al. (2010) Compound Set Enrichment: A Novel Approach to Analysis of Primary HTS Data. J. Chem. Inf. Model., 50, 2067–2078.

Varin, T. et al. (2011) Mining for bioactive scaffolds with scaffold networks: Improved compound set enrichment from primary screening data. J. Chem. Inf. Model., 51, 1528–1538.

Virtanen, P. et al. (2019) SciPy 1.0--Fundamental Algorithms for Scientific Computing in Python.

Wetzel, S. et al. (2009) Interactive exploration of chemical space with Scaffold Hunter. Nat. Chem. Biol., 5, 581–583.