Scaffold analysis in Python with RDKit and pandas

Samo Turk[1], Sameh Eid[1], Andrea Volkamer[1], Friedrich Rippmann[2] and Simone Fulle[1]

[1] BioMed X Innovation Center, Im Neuenheimer Feld 583, 69120 Heidelberg, Germany

[2] Merck KGaA, Merck Serono, Global Computational Chemistry, Frankfurter Str. 250, 64293 Darmstadt, Germany

Dealing with big data analysis can be a significant challenge for a computational chemist. While there are some established off the shelf tools, one often finds them cumbersome and prefers the flexibility provided by programming languages. In this line, Python [1] is one of the most popular programming languages used in science. This is mainly due to its simple but elegant syntax as well as a large number of freely available modules and libraries. Python is getting popular even in the domain of data analysis which was traditionally reserved for R [2].

The most mature data analysis library for Python is pandas [3] and the most mature chemistry toolkit with Python bindings is RDKit [4]. With the help of RDKit, pandas can be chemistry aware, enabling rapid analysis of large amounts of chemical data. In addition, interactive programming environments such as Ipython [5] make chemical data analysis even more approachable.

Here, we will present some basic chemoinformatics operations provided by RDKit and pandas as well as new functions contributed by us to the RDKit community. For example, we will demonstrate how to read in chemical data, calculate descriptors, perform filtering based on these descriptors, and how to visualize these results. In particular, we will demonstrate how our contributions to the RDKit code allow efficient scaffold analysis for a set of compounds.

[1] Python Programming Language – Official Website - http://www.python.org/
[2] The R Project for Statistical Computing - http://www.r-project.org/
[3] Python Data Analysis Library — pandas: Python Data Analysis Library - http://pandas.pydata.org/
[4] RDKit - http://www.rdkit.org/
[5] IPython - http://ipython.org/