WWTF Project VRG19-009

Algorithmic Data Science for Computational Drug Discovery

Abstract

Drug discovery is a lengthy and expensive process that suffers from a small and decreasing success rate. Finding a drug in the search space of possible molecules, which is estimated to contain up to 10^60 structures, is often compared to "finding a needle in a haystack". Although it is possible to perform millions of pharmacological tests in a reasonable time using automated high-throughput screening devices, only a negligible fraction of the chemical space can be synthesized and tested experimentally. Data science is of major importance to exploit the available information in order to direct the search for a new drug candidate to the relevant chemical subspace. The amount of available data in the life sciences grows rapidly, including the outcome of high-throughput experiments as well as data on target proteins, their structure and interaction. On the one hand, this poses new algorithmic challenges regarding the scalability of data mining and machine learning methods to very large molecular data sets. On the other hand, there is well justified hope to obtain more realistic models from the analysis of very large quantities of data with methods tailored to this domain. If this is the case, such methods will allow for a greater automation of the drug discovery process reducing the required time and costs drastically. The goal of the project is to achieve this through the following work plan. First, we will develop efficient methods for the graph based similarity search in large molecular databases using index data structures. Our methods will support graphs with complex vertex and edge annotations to incorporate conformational features regarding the shape and flexibility. On this basis we will develop algorithms for mining relevant substructures, either preserving a certain biological property or having an extraordinarily strong effect on the property. These substructures will be of interest for direct inspection by domain specialist, but also for the machine learning approaches we will develop. For the former we will provide visual analysis modules extending the KNIME analytics platform. We will analyze existing machine learning approaches for graphs regarding their ability to learn successfully in scenarios with equivalent substructures and will identify their weaknesses. Based on this analysis we will develop new machine learning approaches, which will overcome these issues, e.g., by providing this information explicitly as part of the input or by learning structural equivalences in an end-to-end fashion. Finally, we will study generative models for molecular graphs, which allow to construct new compounds with desired properties from structural building blocks starting with their core structure. The developed methods will be released as KNIME extensions for use in chemical data mining workflows and evaluated in real-world pharmacoinformatics tasks.

Scientific disciplines:
102033 - Data mining (40%) | 301207 - Pharmaceutical chemistry (30%) | 102019 - Machine learning (20%) | 102031 - Theoretical computer science (10%)

Keywords:
machine learning; data mining; graph algorithms; cheminformatics; pharmacoinformatics

VRG leader:	Nils Kriege
Institution:	University of Vienna
Proponent:	Wilfried Gansterer
Institution:	University of Vienna

Status: Ongoing (01.05.2020 - 30.11.2028) 103 months

Funding volume: EUR 1,466,230

Further links about the involved persons and regarding the project you can find at
https://archiv.wwtf.at/programmes/vienna_research_groups/VRG19-009