Improving Drug Discovery Decision Making using Machine Learning and Graph Theory in QSAR Modeling

ERNST AHLBERG HELGEE

UNIVERSITY OF GOTHENBURG

Department of Chemistry
University of Gothenburg
Gothenburg, Sweden
2010

AKADEMISK AVHANDLING

för avläggande av doktorsexamen i naturvetenskap som, med medgivande från Institutionen för Kemi Göteborgs Universitet, kommer att försvaras offentligt fredagen den 5 mars 2010 kl 10.00 i sal HA4, Hörsalsvägen 4 Göteborg. Avhandlingen kommer att försvaras på engelska. Fakultetsopponent är Prof. Jean-Loup Faulon, Evry University, Evry, Frankrike

Abstract

During the last decade non-linear machine-learning methods have gained popularity among QSAR modelers. The machine-learning algorithms generate highly accurate models at a cost of increased model complexity where simple interpretations, valid in the entire model domain, are rare.

This thesis focuses on maximizing the amount of extracted knowledge from predictive QSAR models and data. This has been achieved by the development of a descriptor importance measure, a method for automated local optimization of compounds and a method for automated extraction of substructural alerts. Furthermore different QSAR modeling strategies have been evaluated with respect to predictivity, risks and information content.

To test hypotheses and theories large scale simulations of known relations between activities and descriptors have been conducted. With the simulations it has been possible to study properties of methods, risks, implementations and errors in a controlled manner since the correct answer has been known. Simulation studies have been used in the development of the generally applicable descriptor importance measure and in the analysis of QSAR modeling strategies. The use of simulations is spread in many areas, but not that common in the computational chemistry community. The descriptor importance measure developed can be applied to any machine-learning method and validations using both real data and simulated data show that the descriptor importance measure is very accurate for non-linear methods.

An automated method for local optimization of compounds was developed to partly replace manual searches made to optimize compounds. The local optimization of compounds make use of the information in available data and deterministically enumerates new compounds in a space spanned close to the compound of interest. This can be used as a starting point for further compound optimization and aids the chemist in finding new compounds. An other approach to guide chemists in the process of optimizing compounds is through substructural warnings. A fast method for significant substructure extraction has been developed that extracts significant substructures from data with respect to the activity of the compound. The method is at least on par with existing methods in terms of accuracy but is significantly less time consuming.

Non-linear machine-learning methods have opened up new possibilities for QSAR modeling that changes the way chemical data can be handled by model algorithms. Therefore properties of *Local* and *Global* QSAR modeling strategies have been studied. The results show that *Local* models come with high risks and are less accurate compared to *Global* models.

In summary this thesis shows that *Global* QSAR modeling strategies should be applied preferably using methods that are able to handle non-linear relationships. The developed methods can be interpreted easily and an extensive amount of information can be retrieved. For the methods to become easily available to a broader group of users packaging with an open-source chemical platform is needed.

Keywords: machine-learning, QSAR, descriptor importance, local and global models, method of manufactured solutions, automated compound optimization, drug design

ISBN: 978-91-628-8018-7