Identification of Molecular Descriptors for Toxicity Prediction of Small Molecules

Döring K, Grüning BA, Lucas X, Günther S
kersten.doering@pharmazie.uni-freiburg.de
Department of Pharmaceutical Bioinformatics, Institute for Pharmaceutical Sciences,
University of Freiburg, Germany

Introduction

According to the REACH legislation (Registration, Evaluation, Authorisation and Restriction of Chemicals), it has to be shown that environmental chemicals or potential drugs are nontoxic before placing them on the market [1,2]. An alternative approach to the common way of animal tests is the use of *in silico* methods for detecting toxicological effects [3]. We use machine learning (artificial neural networks) and combine it with the generation of molecular descriptors for the prediction of a compound's toxicity.

"... in silico methods as an alternative to animal testing."

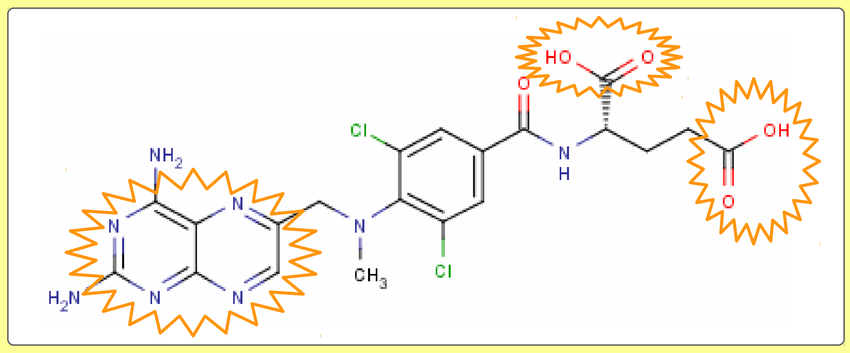
Dataset

1000 very toxic compounds: LD_{50}: 0.0-4.5 mg/kg bw

1000 nontoxic compounds: LD_{50}: 1,000-10,000 mg/kg bw

First results are based on a subset of an "in-house" database of around 20,000 chemical compounds, tested on mice intravenously (LD_{50}).

Dichloromethotrexate

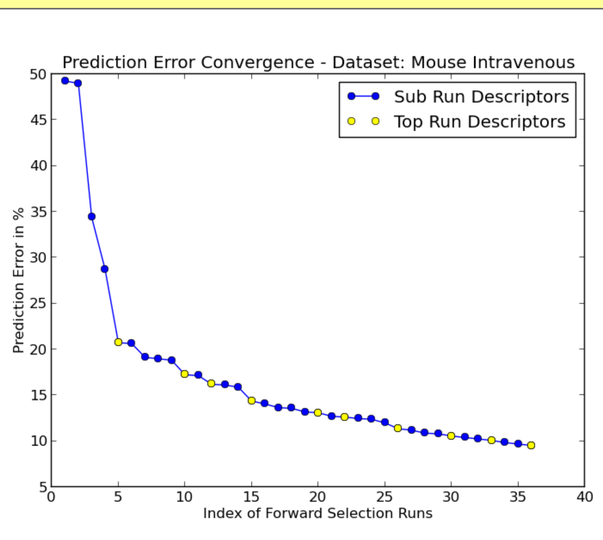


Toxicity:

1,021 mg/kg bw (LD_{50}: nontoxic)

This small molecule contains two times the descriptor "Carboxylic_acid" and one annelated ring structure.

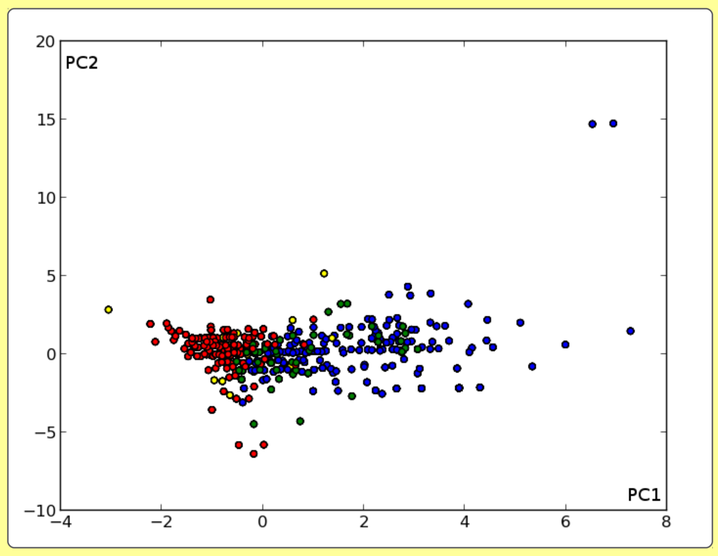
Identification of Molecular Descriptors



Order of descriptors:
Carboxylic_acid
Primary_alcohol
Annelated_rings
Salt
Halogen_on_hetero
Alkene
Hetero_N_basic_no_H
Tertiary_carbon
Sulfonic_acid
CH-acidic_strong

The classifier has selected the 10 best descriptors from a set of almost 600 descriptors by a forward selection (trained on around one third of the 2,000 compounds). The machine learning algorithm finds the best start descriptor (first yellow dot) and then tries to find the best subset until convergence (last yellow dot).

Toxicity Prediction



Legend:

Red:	*toxic*
Green:	*FN (toxic, but prediction nontoxic)*
Blue:	*nontoxic*
Yellow:	*FP (nontoxic, but prediction toxic)*

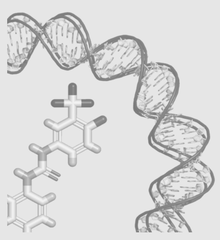
With the current descriptor set, the classifier is able to separate the toxic and nontoxic molecules. The prediction accuracy on this dataset has been 90.31% with 89.68% sensitivity and 90.92% specificity. The figure shows the classification of 1,263 test compounds with the first two principal components of the 10D descriptor space.

References
[1] Krauth *et al.*, 2013. Instruments for Assessing Risk of Bias and Other Methodological Criteria of Published Animal Studies: A Systematic Review. Environ Health Perspect. [Epub ahead of print]
[2] http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm
[3] E. Mombelli, 2008. An evaluation of the predictive ability of the QSAR software packages, DEREK, HAZARDEXPERT and TOPKAT, to describe chemically-induced skin irritation. Altern Lab Anim 36:15-24.

"... what makes a molecule toxic?"

Future Prospects

Descriptor sets will be further evaluated on different datasets including cross-validation of already used as well as unseen data to improve the classifier's generalisation error.



The working group of Pharmaceutical Bioinformatics at the Institute for Pharmaceutical Sciences develops algorithms and software for pharmaceutical research. Our fields of research include the modeling of molecular interactions, prediction of biological effects of molecules, identification of potential new drug agents, analysis of gene expression and methylation data as well as text and data mining. The working group is part of the University of Freiburg's Research Group Program of the Excellence Initiative of the federal and state governments.

http://www.pharmaceutical-bioinformatics.com/

This work is part of the CoRS project, which is funded by the German National Research Foundation (DFG, Lis45).

DFG Deutsche Forschungsgemeinschaft

UNI FREIBURG