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

Structural bioinformatics

eMolTox: prediction of molecular toxicity with confidence

Changge Ji[1,2,*], Fredrik Svensson[2,3], Azedine Zoufir[2] and Andreas Bender[2,*]

[1]Shanghai Engineering Research Center for Molecular Therapeutics and New Drug Development, School of Chemistry and Molecular Engineering, East China Normal University, Shanghai 200062 China, [2]Center for Molecular Informatics, Department of Chemistry, Lensfield Road, Cambridge CB2 1EW, UK and [3]IOTA Pharmaceuticals, St Johns Innovation Centre, Cowley Road, Cambridge CB4 0WS, UK.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: In this work we present eMolTox, a web server for the prediction of potential toxicity associated with a given molecule. 174 toxicology-related in *vitro/vivo* experimental datasets were used for model construction and Mondrian conformal prediction was used to estimate the confidence of the resulting predictions. Toxic substructure analysis is also implemented in eMolTox. eMolTox predicts and displays a wealth of information of potential molecular toxicities for safety analysis in drug development.

Availability: The eMolTox Server is freely available for use on the web at http://xundrug.cn/moltox.

Contact: chicago.ji@gmail.com or ab454@cam.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Drug-mediated toxicity is a heavy burden to the pharmaceutical industry, contributing to safety-related failures in development and the high cost of drug discovery (Cook, et al., 2014). In the past years and decades many biological methods were developed to test the potential toxicity of chemical compounds, including biochemical assays, cellular assays and model-organ systems. These data are now also publicly available on a large scale (Knudsen, et al., 2015), and while coverage of chemical space and consistency of measurements may not in every case be ideal we believe that still the time has come to now model such data on a large scale and to provide a public webservice to do so.

In this work, we present eMolTox, a publicly available web server for prediction of different kinds of toxic endpoints from toxicology related *in vitro*/*in vivo* experimental data and analysis of toxic substructure.

2 Methods

Various types of safety data are generated *in vitro* and *in vivo* (in animals and in humans), and this data can now be used to predict the toxicity potential of a drug candidate at an early stage (Blomme and Will, 2016). In our current work we have collected different types of toxicology data from public databases and literature, including off-target functional assays, cytotoxicity tests, mutagenicity tests, CYP450 inhibition assays, acute oral toxicity assays, transporter assays etc. 174 data sets were extracted in total, details of which are listed in Table S1. 2,048 bit ECFP_4 circular Morgan fingerprints and 196 different physiochemical descriptors (see Text S3) generated by RDKit (Landrum, 2006) were used for building each of the prediction model. Models were generated using Random Forests and Conformal Prediction.

QSAR models built from machine learning methods often meet with the problem of a poor understanding of the confidence of the prediction for the compound of interest (Norinder, et al., 2014). There is no guarantee that a QSAR model can predict all molecules in the chemical space with high confidence; however, a model that is aware what it can and what it cannot predict (with a given confidence value) is already of significant practical value. The conformal prediction framework is a recent development in machine learning (Shafer and Vovk, 2008) that can associate a reliable measure of confidence with a given prediction by using known data in the form of an additional calibration set.

Conformal prediction (CP) can be implemented as a simple wrapper to existing classifiers or regression algorithms. CP uses part of the training data set as a calibration set to calculate p-values for each possible class label through the ranking of nonconformity. The conformal

predictors output predictions together with an associated p-value, and a low p-value is interpreted as the label being unlikely (see Text S1) and it can be disregarded by the user.

For a given compound, a conformal predictor gives the p-value for active and inactive classes as p_1 and p_0 respectively. The output label under significance level ε can be defined as:

Active: $p_1 > \varepsilon$ and $p_0 \leq \varepsilon$

Inactive: $p_0 > \varepsilon$ and $p_1 \leq \varepsilon$

Uncertain: $p_1 > \varepsilon$ and $p_0 > \varepsilon$, $p_1 \leq \varepsilon$ and $p_0 \leq \varepsilon$

We use efficiency and validity to evaluate the performance of confidence estimation for the conformal predictor. Efficiency is defined as the single label prediction rate at a given significance value. The conformal confidence prediction was said to be valid if the frequency of errors was less than ε at a chosen confidence level 1- ε. There is a chance that a test molecule may have an undefined label under conformal prediction, which is termed 'inefficient prediction'. More detailed information about model construction and evaluation is provided in Supplementary Text S2.

Apart from data driven predictors, eMolTox also includes a toxic substructure analysis. Structural alerts (also known as toxicophors/toxic fragments) are chemical substructures that indicate or associate to specific toxic endpoints. Structural alerts are widely accepted in chemical toxicology and regulatory decision making. We collected different kinds of public available structural alerts (Table S2) and analyzed whether a query compound contain a specific toxic substructure.

3 Results

The efficiencies and validities of selected data sets (see Table S3 for complete list) are listed in Table 1 and show that Mondrian conformal prediction gives similar prediction accuracy for both active and inactive sets where validity values are around 0.90 (which is the expected value, 1- ε) at a significance level of 0.1. One parameter in CP is that efficiency and validity are dependent on the significance value chosen, where high confidence may lead to high rate of undefined labels in the predictions. We investigated model performances at different significance levels (see Table S3). The average effiency at a significance level of 0.05, 0.10 and 0.20 is 0.65, 0.76 and 0.85 respectively. Hence, for the balance of efficiency and validity, we use 0.1 as the default significance level in eMolTox webserver. The users can also input the significance level manually which they prefer to use for their analyses.

Table 1. Performance of Conformal Predictor for the selected 8 Data Sets at a significance level ε = 0.10.

Biological Action	Efficiency	Validity (Positive set)	Validity (Negative set)
Modulator of Beta-1 AR	0.92	0.89	0.91
Modulator of HERG	0.83	0.94	0.90
Disrupt mitochondria	0.82	0.94	0.90
Block BSEP Pump	0.86	0.92	0.88
High Acute Rat Oral tox	0.86	0.83	0.91
Mutagenicity	0.72	0.90	0.90
Carcinogenic Potency	0.55	0.98	0.88
Inhibitors/Substrates of CYP450 3A4	0.87	0.92	0.89

4 Web Server

4.1 Interface features

The eMolTox web server offers the user many ways to submit query molecules, namely SMILES strings, IUPAC name, commercial name, CAS ID, InChI key or drawing a molecule.The results page of eMolTox (Supplementary Figs. S1 and S2) is composed of two main sections: Firstly, a table of all potential active endpoints the query compound might have, together with the confidence of each prediction. The structure of the most similar active compound in the database is also shown to rationalize predictions. Secondly, eMolTox provides a table showing all matched toxic or reactive substructures highlited on the molecule supplied together with the potential toxicity label. In addition, the full prediction result is provided as a compressed csv file for download, including a full result table with both positive and negative predictions. Molecules predicted to have undefined label are labeled as 'Inconclusive'. All training data set used for model building are furthermore available for download from the web server.

4.2 Implemention

The web server uses node.js code to run the interface functionality and python code to perform prediction and analysis. Models were developed using Python, Scikit-learn version 0.18 (Pedregosa, et al., 2011) and the nonconformist package (Linusson, 2017). RDKit (Landrum, 2006) is used for generating molecular fingerprints, matching structural alerts and drawing molecule picture. Chemical Identifier Resolver is used for converting molecular name and CAS id to SMILES strings.

Acknowledgements

We thank Dr. Dezsö Modos for helpful discussions.

Funding

We thank the National Key Research and Development Plan (2016YFA0501700), National Natural Science Foundation of China (Grants 21003048 and 21433004), Shanghai Natural Science Foundation (Grant 14ZR1411900) and China Scholarship Council for financial support.

Conflict of Interest: none declared.

References

Blomme, E.A.G. and Will, Y. Toxicology Strategies for Drug Discovery: Present and Future. *Chemical Research in Toxicology* 2016;29(4):473-504.

Cook, D., *et al.* Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat Rev Drug Discov* 2014;13(6):419-431.

Knudsen, T.B., *et al.* FutureTox II: In vitro Data and In Silico Models for Predictive Toxicology. *Toxicological Sciences* 2015;143(2):256-267.

Landrum, G. RDKit: Open-source cheminformatics. 2006.

Linusson, H. Nonconformist. In.; 2017.

Norinder, U., *et al.* Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination. *Journal of Chemical Information and Modeling* 2014;54(6):1596-1603.

Pedregosa, F., *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 2011;12:2825-2830.

Shafer, G. and Vovk, V. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research* 2008;9:371-421.