Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Application note

OXFORD

Data and text mining

admetSAR 2.0: web-service for prediction and optimization of chemical ADMET properties

Hongbin Yang [1], Chaofeng Lou [1], Lixia Sun[1], Jie Li[1], Yingchun Cai[1], Zhuang Wang[1], Weihua Li[1], Guixia Liu[1] and Yun Tang [1,*]

[1] Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: admetSAR was developed as a comprehensive source and free tool for the prediction of chemical ADMET properties. Since its first release in 2012 containing 27 predictive models, admetSAR has been widely used in chemical and pharmaceutical fields. This update, admetSAR 2.0, focuses on extension and optimization of existing models with significant quantity and quality improvement on training data. Now 47 models are available for both drug discovery and environmental risk assessment. In addition, we added a new module named ADMEToptfor lead optimization based on predicted ADMET properties.
Availability: Free available on the web at http://lmmd.ecust.edu.cn/admetsar2/
Contact: ytang234@ecust.edu.cn
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Studies have shown that in early 1990s poor drug metabolism and pharmacokinetics (DMPK) was the major cause of clinical attrition in drug development. With the increasing appreciation of early DMPK investigation, the contribution of these factors has dramatically reduced from 40% in 1991 to 11% in 2000 (Khanna, 2012). This compelling data evidence that early evaluation of chemical absorption, distribution, metabolism, excretion and toxicity (ADMET) profiles plays a significant role in drug discovery. Although a variety of medium and high-throughput *in vitro* screening methods have been used widely, it is still a daunting task to keep up with the AMDET data of each compound (Hou and Wang, 2008). In response to this situation, *in silico* methods will be a good complement for the investigation of these properties and now, they have received considerable attention with various predictive models reported (Wang *et al.*, 2015; Yang *et al.*, 2018).

To accelerate the estimation of chemical ADMET profiles, our team released a webserver called admetSAR in 2012. Armed with over 210,000 experimental data for 96,000 compounds and 27 computational models, it can not only deliver free accessed chemical and biological information, but also provide prediction for most of the endpoints related to ADMET properties (Cheng *et al.*, 2012). With a user-friendly interface, admetSAR enables users to search for chemical information by its CASRN, common name or structure easily. To date, it has been widely applied for drug discovery and environmental risk assessment, and has been cited by 207 times (Web of Science, assessed on Jun. 20, 2018). It has also been embedded in DrugBank since version 4.0 in 2014 (Law *et al.*, 2014). However, the models in admetSAR were not so satisfactory, which were all built by LIBSVM combined with MACCS fingerprint without other options.

In response to the limitations mentioned above, we updated the server to version 2.0 by improving the existing models and adding many new ones with more training data. Notably, for each model, the final selection is an optimal one trained by various machine learning algorithms and molecular fingerprints. Additionally, we developed a module named ADMEToptto automatically optimize a query molecule by scaffold hopping in terms of their ADMET properties. We devote admetSAR 2.0 to be a platform for ADMET prediction and lead optimization in drug design (Figure S1). The user interface and architecture were also updated to facilitate the users.

2 Methods

The training data for building the predictive models were collected from databases such as DrugBank (Law *et al.*, 2014), ChEMBL (Gaulton *et al.*, 2016) and CPDB (Gold *et al.*, 2005), peer reviewed scientific papers, and

© The Author(s) (2018). Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from http://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty707/5085368
by Pius XII Memorial Library/Saint Louis University user
on 28 August 2018

high-throughput screening projects such as Tox21 (Attene-Ramos *et al.*, 2013) and CYP450 (Veith *et al.*, 2009). The number of training molecules and the sources of each model were summarized in Supplements. All the molecules were prepared by Pipeline Pilot (Version 2017 R2), including (1) removing salts; (2) standardization; (3) removing repetitions; (4) representing as canonical SMILES.

The molecules were represented by molecular fingerprints such as MACCS, Morgan and AtomParis implemented with RDKit. Machine learning algorithms including random forest (RF), support vector machine (SVM) and k-nearest neighbors (kNN) were used to build the models, which were implemented by scikit-learn package with python scripts. Five-fold cross-validation were employed to optimize the hyper-parameters such as the number of neighbors, k, in kNN, and the best model was selected according to their performance. Synthetic minority oversampling and random under sample techniques were employed to address some imbalance datasets such as blood brain barrier (BBB) and endocrine disrupting (ED). Regression models such as plasma protein binding (PPB) were built using graph convolutional neural network, which was implemented by DeepChem, an open source toolchain that uses deep learning in drug discovery (Altaetran *et al.*, 2017). Multi-label methods such as Binary Relevance, Classifier Chains, and Label Powerset were employed for ED, since most of the compounds in the training set have bioactivity results of more than one endpoint related to ED. The detailed descriptions are represented in Supplement.

An optimization module named ADMEToptwas embedded in this webserver, which can automatically replace scaffolds with others in our scaffold library. New molecules will be filtered by their ADMET properties to remove the unsuitable scaffolds. The candidate scaffolds are selected according to the scaffold similarity represented by the Tanimoto coefficient of their scaffold descriptors (Rabal *et al.*, 2015).

3 Results

The webserver uses SMILES as input structure data for the molecule to be predicted. InDraw (http://in.indraw.integle.com/) is embedded as a molecular editor to generate SMILES. At most 20 molecules can be submitted in a batch and the results can be downloaded as CSV files.

In total 40 binary endpoints (including a multi-label classification model consisting of 6 endpoints related to ED), 3 multi-class models, and 4 regression models were built. For classification models, five-fold cross validation was employed to evaluate the performance except for ED and BBB, which were evaluated with external validation. The AUC (area under the receiver operating characteristic curve) ranged from 0.625 to 0.992 with an average value of 0.842. The regression models were evaluated by random sampling, extracting 20% from the data dropping out of training and optimization of hyper-parameters. The R^2 were 0.668 for PPB, 0.810 for water solubility, 0.822 for *Pyriformis* toxicity, and 0.522 for acute oral toxicity. The detailed performances are shown in Supplement. The applicability domain (AD) is defined upon the physicochemical and topological properties including molecular weight, AlogP, numbers of H-bond donors and acceptors, and numbers of atoms and rings. The detailed description of AD can be found in the Supplement. In ADMEToptopt, more than 50 thousand unique scaffolds were extracted from chemicals in ChEMBL and Enamine. During scaffold hopping, up to 14 ADMET properties can be used as restrictive factors for lead optimization.

4 Conclusion

The webserver implements the state-of-the-art machine learning methods to build predictive models covering major ADMET properties for drug discovery. Therefore, admetSAR will facilitate medicinal chemists to design and optimize lead compounds with better ADMET properties. In addition, several eco-toxicity models are included that may be helpful for environmental risk assessment of industrial and agricultural chemicals. All these models fulfill four in five of the OECD principals (http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf, accessed on June 20, 2018) except the interpretability considering the fact that the used algorithms are complex. The algorithms and descriptors used for model building were implemented by open-source cheminformatics and machine learning libraries, so that they are transparent and reproducible, which are also important in evaluation of the quality of the models (Patel *et al.*, 2018). We will continue to improve and extend the computational models for prediction of chemical ADMET properties to make admetSAR a practical platform for drug discovery and other chemical research.

Funding

This work was supported by the National Key Research and Development Program of China (Grant 2016YFA0502304) and the National Natural Science Foundation of China (Grants 81373329 and 81673356).

References

Altaetran, H. *et al.* (2017). Low data drug discovery with one-shot learning. *Acs. Cent. Sci.*, **3**(4), 283–293.

Attene-Ramos, M. S. *et al.* (2013). The tox21 robotic platform for the assessment of environmental chemicals–from vision to reality. *Drug Discov. Today*, **18**(15-16), 716–723.

Cheng, F. *et al.* (2012). admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties. *J. Chem. Inf. Model.*, **52**(11), 3099–3105.

Gaulton, A. *et al.* (2016). The ChEMBL database in 2017. *Nucleic Acids Res.*, **45**(Database issue), D945–D954.

Gold, L. S. *et al.* (2005). Supplement to the carcinogenic potency database (cpdb): Results of animal bioassays published in the general literature through 1997 and by the national toxicology program in 1997-1998. *Toxicological Sciences*, **85**(2), 747–808.

Hou, T. and Wang, J. (2008). Structure-adme relationship: still a long way to go? *Expert Opinion on Drug Metabolism and Toxicology*, **4**(6), 759–770.

Khanna, I. (2012). Drug discovery in pharmaceutical industry: productivity challenges and trends. *Drug Discov. Today*, **17**(19-20), 1088–1102.

Law, V. *et al.* (2014). DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, **42**(1), D1091–D1097.

Patel, M. *et al.* (2018). Assessment and reproducibility of quantitative structure-activity relationsh by the nonexpert. *J. Chem. Inf. Model.*, **58**(3), 673–682.

Rabal, O. *et al.* (2015). Novel scaffold fingerprint (SFP): applications in scaffold hopping and scaffold-based selection of diverse compounds. *J. Chem. Inf. Model.*, **55**(1), 1–18.

Veith, H. *et al.* (2009). Comprehensive characterization of cytochrome P450 isozyme selectivity across chemical libraries. *Nat. Biotechnol.*, **27**(11), 1050–1055.

Wang, Y. *et al.* (2015). In silico ADME/T modelling for rational drug design. *Quarterly Reviews of Biophysics*, **48**(4), 488–515.

Yang, H. *et al.* (2018). In silico prediction of chemical toxicity for drug design using machine learning methods and structural alerts. *Front. Chem.*, **6**, 30.

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty707/5085368
by Pius XII Memorial Library/Saint Louis University user
on 28 August 2018