SciForum
Mol2Net

Development of QSAR Models for Identification of CYP3A4 Substrates and Inhibitors

Flavia C. Silva, Ekaterina V. Varlamova, Rodolpho C. Braga, and Carolina H. Andrade*
Labmol – Laboratory for Molecular Modeling and Drug Design, Faculty of Pharmacy, Federal University of Goias, Goiania, Goiás, 74605-170, Brazil.
* Author to whom correspondence should be addressed; E-Mail: carolina@ufg.br. Tel: + 55 62 3209-6451; Fax: + 55 62 3209-6037.

Published: 4 December 2015

Abstract: The pharmacokinetic properties of absorption, distribution, metabolism and excretion (ADME) play a crucial role in drug discovery and development, since many drug candidates fail due to an inappropriate pharmacokinetic profile. Cytochrome P450 enzymes are predominantly involved in Phase 1 metabolism of xenobiotics. Thus, it is important to better understand and prognosticate substrate binding and inhibition of CYP450. The goal of this study was to obtain QSAR (Quantitative Structure-Activity Relationship) models to identify substrates and inhibitors of CYP3A4. The data sets were collected and curated from online available databases and literature. Several QSAR models were obtained and validated according to the recommendations of the Organization for Economic Co-operation Development (OECD). The combination of different descriptors and machine learning methods led to robust and predictive QSAR models with high coverage. The interpretation of developed models was performed using the predicted probability maps (PPMs). These maps help to encode major structural fragments to classify compounds as inhibitors or not inhibitors of CYP3A4. In conclusion, the obtained models can reliably identify substrates and non-substrates, and inhibitors and non-inhibitors of CYP3A4, which is very important in the early stages of the development of new drugs.

Keywords:QSAR; *in silico*; drug metabolism; substrate; inhibitor; CYP3A4.

1. Introduction

Many drug candidates fail during the drug development process in clinical trials due to an inappropriate pharmacokinetic profile. For this reason, the study of the pharmacokinetic properties absorption, distribution, metabolism, excretion, and toxicity (ADME/Tox) of a drug candidate is important to reduce time and

increase the chances of success during drug discovery and development[1].

ADME/Tox properties are the major contributors to the failures of new drugs in the development pipeline and often the underlying biological mechanism of toxicity is related to metabolism. Metabolic liability can lead to a number of diverse issues, including drug−drug interactions, in particular enzyme inhibition and induction, which in turn may cause therapeutic failure toxicity, and adverse effects[2].

Cytochrome P450 (CYP) enzymes are predominantly involved in Phase 1 metabolism of xenobiotics. CYP3A4 is the most abundant cytochrome isoenzyme present in liver and is responsible for the metabolism of more than 50% of the marketed drugs[3]. The main goal of this study was to develop robust and predictive models that can be used to classify compound as inhibitor/non-inhibitor or substrate/non-substrate of CYP3A4 for identifying and discarding drug candidates with potential metabolism issues.

2. Results and Discussion

The statistical results of QSAR models generated for substrates of CYP3A4 (dataset I), using the test set compounds, are summarized in Figure 1.

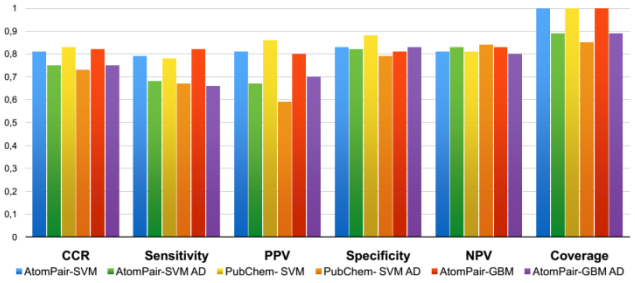

Figure 1. Statistical results of predictions of QSAR models for CYP3A4 substrates evaluated by 5-fold external cross-validation.

The combination of different descriptors and machine learning (ML) methods led to robust and predictive QSAR models for substrates of CYP3A4, with correct classification rate (CCR) values ranging between 0.65-0.83 and coverage of 0.69-0.89. However, among the best three selected models (Atom Pair-SVM; PubChem-SVM; Atom Pair-GBM), the model generated by combining Atom Pair-GBM without considering DA showed a higher sensitivity and lower difference between the values of sensitivity and specificity obtained the best ability to classify correctly both substrates as non-substrates of CYP3A4.

The statistical results of binary and multiclass QSAR models for CYP3A4 inhibitors (data set II) are illustrated on Figure 2.

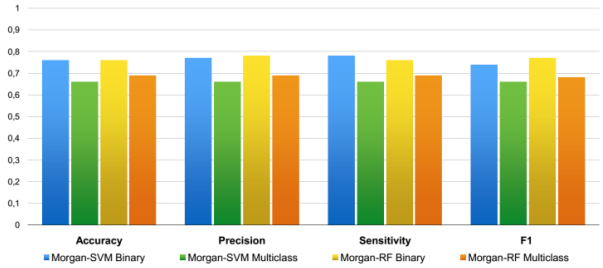

Figure 2. Statistical results of predictions for the best binary and multiclass QSAR models for CYP3A4 inhibitors evaluated by 5-fold external cross-validation.

The two best binary and multiclass models were generated using a combination of Morgan-SVM and Morgan-RF. These binary models showed equal values of accuracy 0.76, which corresponds to the percentage of molecules that are correctly classified by model. Furthermore, they showed sensitivity values of 0.74 and 0.77, respectively. The accuracy of these models was 0.77 and 0.78, respectively, whereas F1 was 0.76 and for both models. The multiclass models were also generated using the combination of Morgan-SVM and Morgan-RF. The Morgan-RF model presented precision value 0.69, while the Morgan-SVM was 0.66. The Morgan-RF model was also slightly higher in relation to F1 value,

with value of 0.69, compared to the value of 0.66 for the Morgan-SVM. However, multiclass and binary QSAR models showed similar statistical results. Therefore, both models were considered the best models to evaluate the inhibition of CYP3A4. In addition, predicted probability maps (PPMs) were generated by Morgan-RF models. The maps for drugs ketoconazole, tioconazole and miconazole are presented in Figure 3.

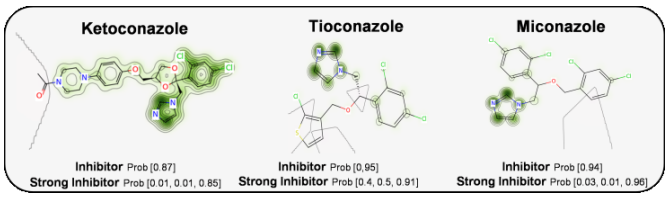

Figure 3. PPMs for selected antifungal drugs generated using Morgan-RF models. Green atoms/fragments have favorable contribution in the property (CYP3A4 inhibition); Gray: no contribution; Pink atoms/fragments have unfavorable contribution in the property (CYP3A4 non-inhibition). The bit vector size of Morgan was 1024 bits.

Miconazole, ketoconazole and tioconazole are antifungal drugs and CYP3A4 inhibitors. These three drugs were classified by the binary model as CYP3A4 inhibitors, and multiclass model considered the three drugs as strong inhibitors with high probability. The imidazole fragment in their structures outlined in green indicate that this fragment has favorable characteristics for the investigated property. These fragments have atoms which are capable of coordinating with heme group iron. The phenyl and thiophene rings are outlined in gray color, which features neutral contribution to the property. Gray isolines demarcate the separation of regions that have favorable and unfavorable contribution.

3. Materials and Methods

In this study, two large datasets were collected for profiling the CYP3A4 activity. The dataset I contained 8,214 compounds, in which 475 are substrates of CYP3A4 and 7,739 are non-substrates (inactive). The annotated dataset was gathered from the literature[4] and PubChem bioassay (Assay ID: 1851). The dataset II contained 9,186 compounds, in which 4,962 are inhibitors de CYP3A4 and 4,224 are non-inhibitors. The annotated dataset was gathered from ChEMBL340 assay. All the molecular modeling studies were performed using a workflow in KNIME platform developed in our laboratory. The dataset curation (removal of duplicates, structural conversion, normalization of specific chemotypes etc.) was performed using Indigo Open Source Standardizer following the workflow described by Fourches et al.[5] including the duplicate analysis. Binary and multiclass QSAR models were developed and validated according to the OECD principles. For generation of QSAR models we used the qsaR package fully integrated workflow KNIME 2.9[6]. The cross-validation procedure 5-fold was used to estimate the robustness of the model using the training set, while the test set was used to validate and estimate the predictive power of the generated models.

Because dataset I was highly unbalanced, it was not recommended to build binary QSAR models for the entire dataset. Therefore, a linear under-sampling strategy was used to investigate the more adequate dataset balancing. We generated five under-sampled datasets with substrates-to-non-substrates ratios of 1:1, 1:2, 1:3, 1:4, 1:8, and the unbalanced dataset. From the six different datasets splits generated, the balancing with proportion of 1:1 and the total unbalanced

dataset were selected because of the best statistical results and covering the largest chemical space. Thus, various QSAR models were generated using different types of descriptors and algorithms, in order to use more information from QSAR models. Four different types of molecular fingerprints were utilized in this study (Atom Pair[7], PubChem[8], MACCS[9] and FeatMorgan[10]), as well as four ML algorithms (SVM[11], GBM[12], PLSDA[13] and *k*NN[14]) were used to model generation, totaling in 16 different QSAR models.

For dataset II, the models for CYP3A4 inhibitors were generated using a 5-fold technique, *i.e.*, spliting the data set in modeling set and external validation set. We used only one type of molecular descriptor (Morgan) and two ML methods (SVM and RF[15]). For construction of multiclass models, the threshold activity was defined as follows: strong inhibitor ≤ 1 μM; weak-moderate inhibitor, property between 1 μM and 10 μM; non-inhibitor ≥ 10 μM[16].

PPMs[17] were generated for visualization of favorable (positive) and unfavorable (negative) structural fragments for compound to be inhibitor or non-inhibitor of CYP3A4.

4. Conclusions

The largest publicly available data sets for substrates and inhibitors of CYP3A4 were collected, prepared and balanced. Robust and predictive QSAR models were generated for the identification of substrates (binary models) and inhibitors (binary and multiclass models). Obtained models can be used for identifying substrates and inhibitors of CYP3A4 in early stages of drug development. PPMs showed important contribution of some fragments probably responsible for interaction with the heme group of CYP3A4.

Acknowledgments

We are grateful to Coordination for the Improvement of Higher Education Personnel (CAPES), the National Counsel of Technological and Scientific Development (CNPq), and the State of Goiás Research Foundation (FAPEG) for their financial support and fellowships. We are also grateful to ChemAxon (Budapest, Hungary) for providing us with the academic license for their software.

Author Contributions

Conceived and designed the experiments: FCS, EV, RCB, CHA. Performed the experiments: FCS, EV, RCB. Analyzed the data: FCS, EV, RCB, CHA. Contributed analysis tools: FCS, EV, RCB, CHA. Wrote the paper: FCS, EV, RCB, CHA.

Conflicts of Interest

The authors declare no conflict of interest. The funders had no role in the study design, data collection, analysis, decision to publish, or preparation of this manuscript.

References and Notes

1. STEPAN, A. F.; MASCITTI, V.; BEAUMONT, K.; KALGUTKAR, A. S. Metabolism-guided drug design. **MedChemComm**, 2013, v, 4, p. 631-652.

2. LI, H.; SUN, J.; FAN, X.; SUI, X.; ZHANG, L.; WANG, Y.; HE, Z. Considerations and recent advances in QSAR models for cytochrome P450-mediated drug metabolism prediction. **Journal of computer-aided molecular design**, 2008, 11, p. 843–855.
3. KIRCHMAIR, J.; WILLIAMSON, M. J.; TYZACK, J. D.; TAN, L.; BOND, P. J.; BENDER, A.; GLEN, R. C. Computational prediction of metabolism: sites, products, SAR, P450 enzyme dynamics, and mechanisms. **Journal of chemical information and modeling**, 2012, 3, p. 617–648.
4. ZARETZKI, J.; RYDBERG, P.; BERGERON, C.; BENNETT, K. P.; OLSEN, L.; BRENEMAN, C. M. RS-Predictor models augmented with SMARTCy reactivities: robust metabolic regioselectivity predictions for nine CYP isozymes. **Journal of chemical information and modeling**, 2012, 6, p. 1637–1659.
5. FOURCHES, D.; MURATOV, E.; TROPSHA, A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. **Journal of chemical information and modeling**, 2010, 7, p. 1189–1204.
6. BRAGA, R. C.; ALVES, V. M.; SILVA, A. C.; LIAO, L. M.; ANDRADE, C. H. Virtual Screening Strategies in Medicinal Chemistry: The state of the art and current challenges. **Current topics in medicinal chemistry**, 2014, 16, p. 1899–1912.
7. CARHART, R. E.; SMITH, D. H.; VENKATARAGHAVAN, R. Atom Pairs as Molecular Features in Structure-Activity Studies : Definition and Applications. **Journal of chemistry information and computer sciences**, 1985, 2, p. 64-73.
8. STEINBECK, C.; HAN, Y.; KUHN, S.; HORLACHER, O.; LUTTMANN, E.; WILLIGHAGEN, E. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. **Journal of chemical information and computer sciences**, 2003, 2, p. 493–500.
9. TODESCHINI, R., CONSONNI, V. **Molecular Descriptors for Chemoinformatics**. 2 rd ed. MANNHOLD, R; KUBINYI, H; FOLKERS, G. Wiley-VCH: Weinheim, Germany, 2009, p-1-1257.
10. MORGAN, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. **Journal of Chemical Documentation**, 1965, 2, p. 107–113.
11. CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, 1995, 20 (3), p. 273-297.
12. FRIEDMAN, J. H. Greedy Function Approximation: A Gradient Boosting Machine. **Annals of Statistics,** 2001, 5, p. 1189–1232.
13. Barker, M.; Rayens, W. Partial Least Squares for Discrimination. **Journal Chemometrics**.**2003**, *3*, p.166–173.
14. ALTMAN, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. **The American Statistician,** 1992, 46 (3), p. 175–185.
15. BREIMAN, L. Random Forests. **Machine Learning**, 2001, 45 (1), p. 5–32
16. YAN, Z.; CALDWELL, G. W. Metabolism profiling, and cytochrome P450 inhibition & induction in drug discovery. **Current topics in medicinal chemistry**, 2001, 5, p. 403–425.

17. RINIKER, S.; LANDRUM, G. A. Similarity maps - a visualization strategy for molecular fingerprints and machine-learning methods. **Journal of cheminformatics**, 2013, 1, p. 43.