STRUCTURE-BASED VIRTUAL SCREENING AND MACHINE LEARNING MODELS FOR VALIDATION OF ACTIVITY PREDICTION OF THROMBOXANE SYNTHASE INHIBITORS (CYP5A1)

Peçanha, B. R. B.[1*]; Flores Jr., L. A. P.[1]; Lima, C. H. S.[2]; Dias, L. R. S.[1**]
[1]*Universidade Federal Fluminense/Faculdade de Farmácia, Laboratório de Química Medicinal, R. Mario Viana 523, Niterói, RJ, Brasil*
[2]*Universidade Federal do Rio de Janeiro/Instituto de Química, Av. Athos da Silveira Ramos 149, Rio de Janeiro, RJ, Brasil*
[*]*brunapecanha@id.uff.br* [**]*lrsdias@id.uff.br*

Introduction

Molecular docking is a useful tool for structure-based virtual screening (SBVS), allowing the identification of potential inhibitors for several targets, using a scoring function (SF) to predict the ligand-protein binding affinity (Berishvili et al., 2018). This tool has some disadvantages concerning the accuracy in biological prediction when compared to experimental data (Nogueira; Koch, 2019). In this context, the post-processing of docking SFs using machine learning algorithms (MLA) made it possible to obtain classification models comparable to biological experiments (Yasuo; Sekijima, 2019). An attractive target for the development of classification models using MLA is Thromboxane synthase (CYP5A1 or TXS), an enzyme involved in platelet aggregation that plays an important role in thrombotic events (Mesitskaya et al., 2018). Thus, we developed a classification model using MLA from the docking SF to validate activity prediction of TXS inhibitors.

Method

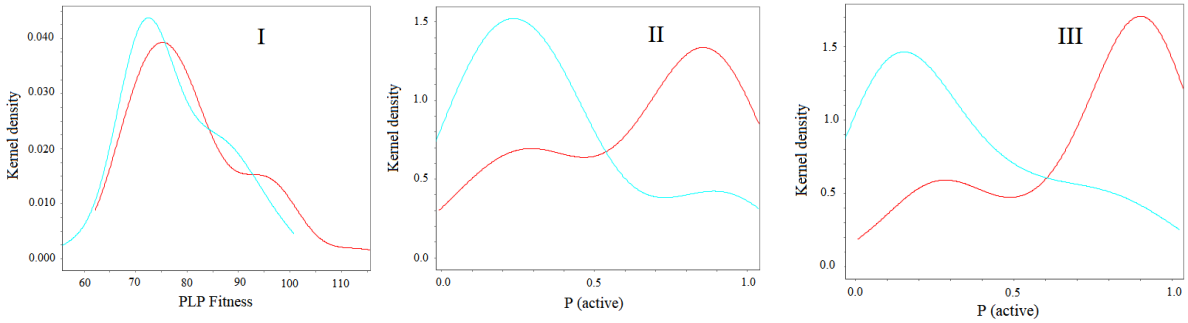
The 3D-structure of TXS was obtained according to the description in the available literature (Yang et al., 2017). TXS inhibitors (TXSI) were searched in ChemBL databases with IC_{50} values in different biological assays. The duplicated structures, inorganic salts, and undefined chirality were removed and the remaining 333 TXSI were classified using pIC_{50} values ($-\log IC_{50}$) in a range from 4.0 to 9.0, from which 144 strong inhibitors ($pIC_{50} \geq 7.3$) considered active compounds and 144 weak inhibitors ($pIC_{50} \leq 7.0$) considered inactive compounds were selected. 3D TXSI were generated using Open Babel (pH=7.8) (O'Boyle et al., 2011). Molecular docking studies were performed using GOLD 2020.1 (Jonnes et al., 1997), considering the SFs ChemPLP, ChemScore, and GoldScore. The molecular descriptors were calculated in the DataWarrior program 5.2.1 (Sander et al., 2015). The classification models were developed using KNIME software 4.1.3. The datasets were normalized (Z-score), filtered by linear correlation, and partitioned in the training sets (70%) and test sets (30%) considering the linear, random, and stratified partition modes. The tested MLA were XGBoost (Extreme Gradient Boosting), Random Forest, Support Vector Machine (SVM), Naive Bayes, and Linear Regression. The tenfold internal cross-validation was carried out with the training sets. Models containing better metrics of the following: AUC-ROC, sensitivity (Se), specificity (Sp), accuracy (Ac), F-score (F1) and Matthews's correlation coefficient (MCC), were chosen to validate activity prediction. Kernel density plot was used to illustrate discrimination between active and inactive compounds. The applicability domain (APD) of training sets was calculated for the test sets using Enalos node on KNIME.

Results / Discussion

The TXSI database was composed by a balanced ratio (1:1) of active and inactive compounds (Berishvili et al., 2018). The compounds with pIC_{50} values between 7.0 and 7.3 were excluded to improve the discriminant power of the models (Jain et al., 2017). Docking SFs were used without

rescoring because the co-crystallized ligand of TXS was not available. The docking scores with zero variance and high Pearson correlation (0.7) were removed. Molecular descriptors were added with docking scores to evaluate the performance of the models with a filter correlation of 0.5. The training set was submitted to a tenfold cross-validation to reduce overfitting in MLA models. The XGBoost algorithm provided better statistics parameters for training sets for all SFs. For the test set, the ChemPLP model with stratified random partition mode retrieved better metrics of AUC-ROC (0.727), MCC (0.47) and F1 (0.716), with a recall of 67% of actives (Se), 80% of inactives (Sp) and Ac of 74%. AUC-ROC values above 0.8 indicate high capacity of prediction, but other parameters need to be observed as Ac, F1 and MCC (Jain et al., 2017). The MCC value can represent total prediction (+1), random prediction (0) or total disagreement between prediction and observation (−1). In Figure 1, kernel density plots of ChemPLP Fitness score (I) and ChemPLP model score (II) showed better discrimination through the model (Berishvili et al. 2018). The density plot of the ChemPLP model plus molecular descriptors with random partition mode (III) presented an improvement of discrimination of actives (76%), AUC-ROC (0.822), and F1 (0.736), with the same Ac and MCC. The APD values were 11.252 (ChemPLP model) and 61.475 (ChemPLP plus descriptors). The prediction for all compounds in the test sets was considered reliable with domain values lower than APD values (Jain et al., 2017).

Figure 1. Kernel density plots for test sets considering (I) the ChemPLP Fitness score, (II) ChemPLP with stratified Partitioning, (III) ChemPLP plus molecular descriptors with random Partitioning. P= Probability; Red line = active; Cyan line = inactive.



Conclusion

In this study, we developed structure-based classification models using machine learning algorithms and docking scores to validate activity prediction of TXSI. We observed better discrimination of actives with the ChemPLP model score than with the ChemPLP Fitness score, with a prediction accuracy of 74%. However, the model was improved by a combination of docking scores and molecular descriptors, increasing the AUC-ROC from 0.727 to 0.822, and discrimination to 76% of active compounds. The applicability domain showed higher reliability of predictions for the test set for both models.

Acknowledgments

The authors thank the Brazilian agencies FAPERJ (Emergency Support for Stricto Sensu Graduate Programs and Courses in the State of Rio de Janeiro, E -26/200.930/2017), CAPES (Finance Code 001), and the Institutional Qualification Program (PQI-UFF 001/2018).

Bibliographic References

Berishvili VP, et al. 2018. *Molecular Informatics*, 37 (11):1800030
Jain S, et al. 2017. *Journal of Computer-Aided Molecular Design*, 31(6):507–521.
Jonnes G, et. al. 1997. *Journal of Molecular Biology*, 267(3):727–748.
Mesitskaya DF, et al. 2018. *Cardiovascular & Hematological Agents in Medicinal Chemistry*, 16(2):81–87.
Nogueira MS; Koch O. 2019. *Journal of Chemical Information and Modeling*, 59(3):1238−1252.
O'Boyle NM, et al. 2011. *Journal of Cheminformatics*, 3:33.
Sander T, et al. 2015. *Journal of Chemical Information and Modeling*, 55(2):460–73.
Yang HC, et al. 2017. *Journal of Physical Chemistry B*, 121(50):11229−11240.
Yasuo N; Sekijima M. 2019. *Journal of Chemical Information and Modeling*, 59(3):1050−1061.