A MACHINE LEARNING MODEL FOR VIRTUAL SCREENING OF POTENTIAL *TC*CPY51 INHIBITORS

Flores Junior, L.A.P.[1*]; Peçanha, B.R.B.[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*
**lapfjunior@id.uff.br*.

Introduction

The protozoan *Trypanosoma cruzi* (*T. cruzi*) is the etiological agent of Chagas disease. This disease represents a serious public health problem due to around 8 million people being infected worldwide, mainly in Latin America, where this disease is endemic [1]. The treatment of this disease is based on drugs nifurtimox and benznidazole, which are most effective in the early stages of infection and retains several side effects [2,3]. In the search for new drug candidates, the sterol 14α-demethylase cytochrome P450 (CYP51) has been evaluated as a potential therapeutic target for *T. cruzi* due to its role in the ergosterol biosynthesis, which is essential for the parasite's survival [4]. Among the drug search strategies for new treatments, those focused on biological targets of the etiologic agent, such as the CYP51 enzyme, seem promising [5]. This work focused on constructing a machine learning (ML) model for virtual screening to identify potential hit candidates as *Tc*CYP51 inhibitors.

Material and Methods

We obtained a *Tc*CYP51 inhibitors library from the PubChem server, which compounds have the same biological assay protocols and chemical structure diversity (PubChem AID:1159558) [8]. The 2D structures of selected compounds were generated by DataWarrior software, from which we discarded the duplicate structures, inorganic salts, and undefined chirality [6]. After that, the dataset obtained was classified by pIC_{50} values ($-\log IC_{50}$) and divided into active and inactive sets. The descriptors were obtained by the RDKit Fingerprint in KNIME 4.1.3 software [7], considering the following types: Morgan, Feat Morgan, RDKit, Avalon, AtomPair, Layered, Torsion, and MACCS. The dataset built by active and inactive compounds was divided into the train and test sets (70:30 ratio) using the linear, random, and stratified partition modes.

The ML model was developed using the training set, considering the Xgboost and random forest algorithms, which were validated by internal cross-validation ten times. Then, ML model validation was made by the test set using the fallowed statistical metrics: AUC-ROC, sensitivity (Se), specificity (Sp), precision (Ac), F-score (F1), and Matthew's correlation coefficient (MCC).

Results and Discussion

We identified 584 *Tc*CYP51 inhibitor compounds from PubChem. Among them, 139 compounds were classified as the active set ($pIC_{50} \geq 5.0$ potent inhibitors) since they have a diversity of chemical structures and IC_{50} values from the same biological assay [8]. The inactive set was generated with 164 compounds ($pIC_{50} \leq 4.5$) that were submitted to the same biological assay (PubChem AID:1159558).

The Gradient Boost algorithm provided better statistics parameters for training and test sets than the Random Forest algorithm. For the test set, the AtomPair model with stratified partition mode retrieved the best metrics of AUC-ROC (0.87), MCC (0.69), and F-score (0.833), with a recall of 83.3% of actives (Se), 85.7% of inactive (Sp) and Ac of 84.6%. AUC-ROC > 0.8 indicate a high

capacity for prediction. Still, other parameters need to be observed, as the MCC value can show the performance of the model as a total prediction (+1), a random prediction (0), or a complete disagreement between prediction and observation (-1) [9].
To improve the algorithm's performance and generate a model with a better discriminatory power model, we used a consensus between the fingerprints by the average of the predictive activity for each fingerprint, followed by the evaluation of statistical metrics [10]. The consensus model generated with the 4 types of fingerprints (AtomPair, Morgan, Feat Morgan, and Torsion) showed a better performance of statistical parameters: AUC - ROC (0.99), MCC (0.93) and F-score (0.96), with the recall of 96.8% active (Se), 96.6% inactive (Sp) and 96.7% Ac.

Conclusion

In this study, we generated a high-performance ML model to discriminate inactive and active compounds for *Tc*CYP51 enzyme inhibition. The model obtained from the fingerprint consensus showed improved performance, increasing its statistical parameters (AUC-ROC: 0.99, MCC: 0.93). The developed ML model can be used in the virtual screening of compounds to identify potential *Tc*CYP51 inhibitors based on the structure of ligands, with a good discrimination between active and inactive compounds.

.

Acknowledgments
The authors thank for the support of the sponsors FAPERJ (E-26/210.068/2021) and CAPES (Finance Code 001).

Bibliographic References

[1] WHO. 2018. Available in: https://www.who.int/docs/default-source/ntds/chagas-disease/chagas-2018-cases.pdf?sfvrsn=f4e94b3b2 Accessed Oct. 1, 2022.
[2] ALDASORO, E. et al. What to expect and when: benznidazole toxicity in chronic Chagas' disease treatment. **J. Antimicrob. Chemother**.2018, 73 (4), 1060-1067.
[3] MIRANDA, M.R.; SAYÉ, M.M. Chagas disease treatment: From new therapeutic targets to drug discovery and repositioning. **Curr. Med. Chem**. 2019, 26 (36), 6517-6518.
[4] LEPESHEVA, G.I.; FRIGGERI, L.; WATERMAN, M.R. CYP51 as drug targets for fungi and protozoan parasites: past, present and future. **Parasitology**. 2018, 145 (14), 1820-1836.
[5] SANTOS-JÚNIOR, P.F.S. et al. Sterol 14α-demethylase from trypanosomatidae parasites as a promising target for designing new antiparasitic agents. **Curr. Top. Med. Chem**. 2021(21), 1900-1921.
[6] SANDER, T. et al. DataWarrior: An open-source program for chemistry aware data visualization and analysis. **J. Chem. Inf. Model.** 2015, 55 (2), 460-473.
[7] Knime software, Copyright by KNIME AG, Zurich, Switzerland, version 4.1.3, 2020.
[8] NHI. PubChem Bioassay Record for AID 1159558, TcCYP51 enzymatic inhibition, Source: GlaxoSmithKline (GSK). https://pubchem.ncbi.nlm.nih.gov/bioassay/1159558. Accessed Oct. 3, 2022.
[9] Kurczab, R.; Smusz, S.; Bojarski, A.J. The influence of negative training set size on machine learning-based virtual screening. **J. Cheminformatics.** 2014, 6 (1), 1-9.
[10] Rácz, A., & Keserű, G. M. Large-scale evaluation of cytochrome P450 2C9 mediated drug interaction potential with machine learning-based consensus modeling., **J. Comput. Aided Mol. Des.** 2020, 34(8), 831–839.