Machine Learning models for the identification of new molecules active against Dengue Virus

Rodolpho C. Braga[1]* (PQ), Ekaterina V. Varlamova (PQ)[1], Sean Ekins[2] (PQ), Carolina H. Andrade[1] (PQ)**

***rodolphobraga@yahoo.com **carolina@ufg.br**

[1]***LabMol – Laboratory for Molecular Modeling and Drug Design**, Faculty of Pharmacy, Federal University of Goiás, Goiania – GO, Brazil.*
[2]***Collaborations Pharmaceuticals Inc.**, 5616 Hilltop Needmore Road, Fuquay Varina, NC 27526, USA.*

Keywords: *Dengue virus 2, QSAR, drug design, virtual screening*

Abstract

Machine learning (ML) and virtual screening were performed for search of new hits against Dengue Virus 2.

Introduction

The World Health Organization estimates that from 50 to 100 million people become infected with Dengue virus annually in more than 100 countries. About 550,000 patients require hospitalization and 20,000 die annually. There is no specific treatment and it is only possible to treat the symptoms of the disease. So there is an urgent need for a drug to treat Dengue virus. The main goal of this work was to develop ML models and to virtual screening of databases for prioritizing compounds to be tested against the four Dengue virus serotypes. Specific goals of were: *i)* data collection and curation; *ii)* ML model building and virtual screening; *iii)* experimental validation of selected hits.

Results and Discussion

Data with activity against Dengue virus 2 for modelling was obtained from online public database PubChem (PubChem AID: 651640). The data set consists of compounds of two classes: active and inactive. The data was curated as described by Fourches, Muratov and Tropsha[1]. Being unbalanced (2,650 active and 249,804 inactive compounds) the under-sampling technique was used for balancing the data set. Three different data sets were used for modelling with a ratio of active and inactive compounds as 1:1, 1:2 and 1:4. Two types of descriptors were calculated: SiRMS[2] (Simplex representation of molecular structure) and Morgan fingerprints[3]. Atoms in a simplex were differentiated on the base of different characteristics: atoms individuality, partial atom charge, lipophilicity, atomic refraction, possibility of atom to be hydrogen donor or acceptor in H-bond. Two ML algorithms (SVM and RF) were used for model generation. Five-fold external cross-validation was used for the estimation of predictive power of obtained models. Models with ratio 1:1 showed the best results with balanced accuracy more than 90%. These models were combined in consensus model (Figure 1). The 1:1 dataset was further explored with Bayesian, Recursive partitioning (forest and single tree, Discovery Studio, Biovia) and SVM (R) algorithms using FCFP_6 descriptors and 8 simple descriptors. These lead to models with Receiver Operator Characteristic (ROC) after 5 fold testing (0.94-0.96). The consensus model and Bayesian Model were used for virtual screening of the ChemBridge and Prestwick databases. Ten hits and FDA approved drugs were selected for experimental validation against the four Dengue virus serotypes. Results of the screen are expected to be available at the time of the conference.

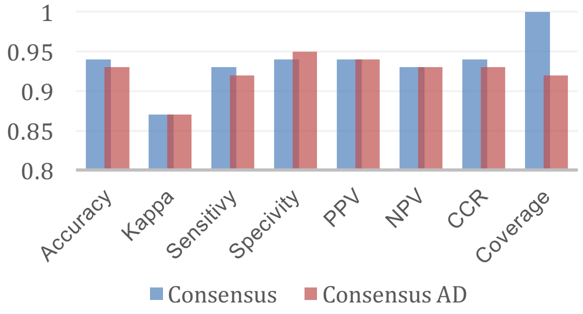

Figure 1. Statistical characteristics for consensus model for balanced data set.

Conclusions

Robust and predictive QSAR models were generated with a balanced Dengue dataset using different types of descriptors and ML methods and used for virtual screening of the ChemBridge and Prestwick databases. Selected hits may also be of interest for screening against related viruses like Zika Virus.

Acknowledgements

CNPq, CAPES and FAPEG.

[1] Fourches, D., Muratov E, Tropsha A. *J Chem Inf Model.* **2010,** *50,* 1189-204
[2] Kuz'min, V.E., Artemenko, A.G., Muratov, E.N. *J Comput Aided Mol Des.* **2008**, *22*, 403-421.
[3] Morgan, H. L. *J. Chem. Doc.*, **1965**, *5*, 107-113.