CLASSIFICATION MODELS BASED ON MACHINE LEARNING FOR THE PREDICTION OF mPGES-1 INHIBITORS

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*
[*]*brunapecanha@id.uff.br* [**]*lrsdias@id.uff.br*

Introduction

Microsomal prostaglandin E synthase-1 (mPGES-1) is an α-helical homotrimeric integral membrane inducible enzyme involved in the production of prostaglandin E2 (PGE2). The mPGES-1 inhibition is a therapeutic strategy for the treatment of pain, inflammation, and some cancers, besides a substitute for the use of cyclooxygenase-2 inhibitor anti-inflammatory drugs (coxibs) [1]. Molecular docking is a widely used tool for structure-based virtual screening for drug design, allowing the identification of potential inhibitors for several targets, using a scoring function (SF) to predict the ligand-protein binding pose in the active site. However, this tool has some disadvantages concerning the accuracy in biological prediction when compared to experimental data [2]. In this context, classification models based on machine learning are very useful to improve accuracy in activity prediction of docking scoring functions [3]. Thus, we developed a machine learning classification (MLC) model to rank mPGES-1 inhibitors from docking scores and physicochemical (PC) properties to validate activity prediction.

Methods

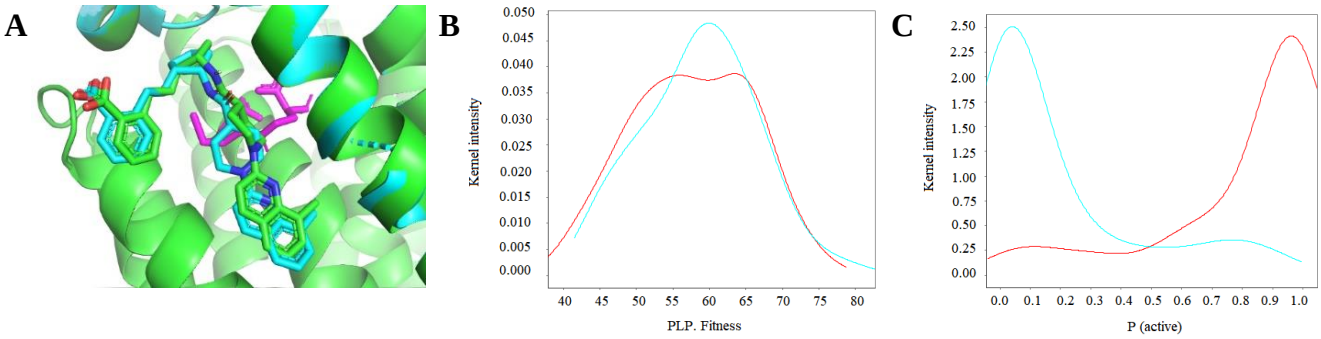
The homotrimer of human mPGES-1 (Uniprot: O14684) was modeled using templates (PDB: 3DWW and 4AL0) in SWISS-MODEL server [4]. The 3D ligand structure (PDB: 7DN) of the monomeric structure [1] (PDB: 5TL9, resolution: 1.20 Å) was inserted by overlap in the homotrimer model using PyMOL software [5]. Molecular docking was performed using GOLD 2020.1 [6], considering the SFs ChemPLP, ChemScore, and GoldScore. mPGES-1 inhibitors were selected in ChemBL databases [7] with IC_{50} values. The 3D ligand structures were built by Open Babel [8] and filtered to remove duplicates, inorganic salts, undefined chirality and compounds with molecular weight over 640. The inhibitors (847) were classified using pIC_{50} values ($-\log IC_{50}$) in a range from 4 to 10, of which 419 were selected as active ($pIC_{50} \geq 7.3$, strong inhibitors) and 398 as inactive ($pIC_{50} < 6.8$, weak inhibitors). PC properties were calculated by DataWarrior version 5.2.1 [9]. MLC model was developed using KNIME software 4.1.3 [10] and the data were normalized (Z-score), filtered by linear correlation (0.5), and partitioned into training set (70%) and test set (30%) for validation, considering the linear, random, and stratified partition modes. The algorithm used was Extreme Gradient Boosting (XGBoost) and a tenfold internal cross-validation was carried out to the training set. Statistical parameters were observed to evaluate the models as: AUC-ROC, sensitivity (Se), specificity (Sp), accuracy (Ac), F-score (F1) and Matthews's correlation coefficient (MCC). Kernel density plot was used to illustrate discrimination between active and inactive compounds. Enalos node for KNIME was used to calculate the applicability domain (APD).

Results and Discussion

The molecular docking protocol was validated from redocking with 7DN ligand using the homotrimer model. The ChemPLP SF presented the lowest RMSD (1.27 Å) for the highest score pose of the protein-ligand complex (Fig. 1A). This validated protocol was used for the pose selection of mPGES-1 inhibitors. The docking scores and energies by protein residue (309) and PC descriptors (14) of mPGES-1 inhibitors were extracted, in a total of 323. The descriptors were selected through correlation filter of which 119 were used to generate the MLC model. A balanced database (1:1) and a tenfold cross-validation were used to reduce model overfitting (Berishvili et al., 2018). Compounds with

pIC_{50} values between 6.8 and 7.3 were excluded to minimize the border effect and improve the discriminant power of the models [11]. The stratified partitioning of MLC data showed the best results for the test set, based on the calculation of the probability (P) of the positive class (active): AUC-ROC (0.93), MCC (0,71), and F-score (0.858), with a recall of 86.5% of active (Se), 84.2% of inactive (Sp) and Ac of 85.4%. These values higher than 0.8 indicate a high capacity of the predictive model [11]. The MCC value can indicate a total prediction (+1), a random prediction (0), or a total disagreement between prediction and observation (−1). In Figure 1B, the kernel density plots show that only the PLP.Fitness score is not efficient to discriminate the compounds. But the kernel density plots of the MLC model (Fig 1C) showed excellent discrimination for active and inactive compounds [2]. The calculated APD value was 11.064 with the predictive activity considered reliable for 86.2% of test set compounds, those with domain values lower than APD, and considered unreliable for 34 compounds of the test set (13.8%) [11].

Figure 1. (A) Redocking of the ligand (PDB:7DN) and the protein (PDB: 5TL9), with RMSD 1.27 Å. Kernel density plots are showed: (B) only with the PLP.Fitness score, and (C) after generating the MLC model with docking scores and PC descriptors for compounds of test set. P (active) = Probability of positive class (active); Red line = actives; Cyan line = inactives.



Conclusion

In this study, we developed a machine learning classification (MLC) model using docking score and ligand descriptors to predict the activity of novel potential mPGES-1 inhibitors. The model showed excellent discrimination between active and inactive compounds, with a prediction accuracy of 85.4% and AUC-ROC 0.93. The reliability prediction for the test set (86%) indicates the MLC model developed to be used for virtual screening of potential inhibitors and for *in silico* drug repositioning studies based on mPGES-1 as a target for inflammation and related diseases.

Acknowledgments

The authors thank the Brazilian agencies FAPERJ (E -26/010.001318/2019), CAPES (Finance Code 001), and PQI-UFF 001/2018.

Bibliographic References

[1] Partridge, K., Antonysamy, S., Bhattachar, S., et al.: Discovery and characterization of [(cyclopentyl)ethyl]benzoic acid inhibitors of microsomal prostaglandin E synthase-1, Bioorg. Med. Chem. Lett., 2017, 27, pp. 1478–1483.

[2] Berishvili, V., Voronkov, A., Radchenko, E., et al.: Machine Learning Classification Models to Improve the Docking-based Screening: A Case of PI3K-Tankyrase Inhibitors, Mol. Inf., 2018, 37, 11, 1800030.

[3] Yasuo, N., Sekijima M.: Improved Method of Structure-Based Virtual Screening via Interaction-Energy-Based Learning, J. Chem. Inf. Model. 2019, 59, pp. 1050−1061.

[4] Waterhouse, A., Bertoni, M., Bienert, S., et al.: SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res., 2018, 46, W296-W303.

[5] Schrödinger, L. The PyMOL Molecular Graphics System, version 2.4.1., 2015.

[6] Jones, G., Willett, P., Glen, R., et al.: Development and validation of a genetic algorithm for flexible docking, J. Mol. Biol., 1997, 267, 3, pp. 727–748.

[7] Gaulton, A. Hersey A., Nowotka, M. et al.: The ChEMBL database in 2017. Nucleic Acids Res., 2017, 45, D945–D954.

[8] O'Boyle N., Banck, M., James, C., et al.: Open Babel: An open chemical toolbox. J. Cheminform., 2011, 3, 33.

[9] Sander, T., Freyss, J., von Korff, M., et al. DataWarrior: an open-source program for chemistry aware data visualization and analysis, J. Chem. Inf. Model., 2015, 55, 2, pp. 460–473.

[10] Knime software, Copyright by KNIME AG, Zurich, Switzerland, version 4.1.3, 2020.

[11] Jain, S., Grandits M., Richter, L., et al.: Structure based classification for bile salt export pump (BSEP) inhibitors using comparative structural modeling of human BSEP, J. Comp. Aided Mol. Des., 2017, 31, 6, pp. 507–521.