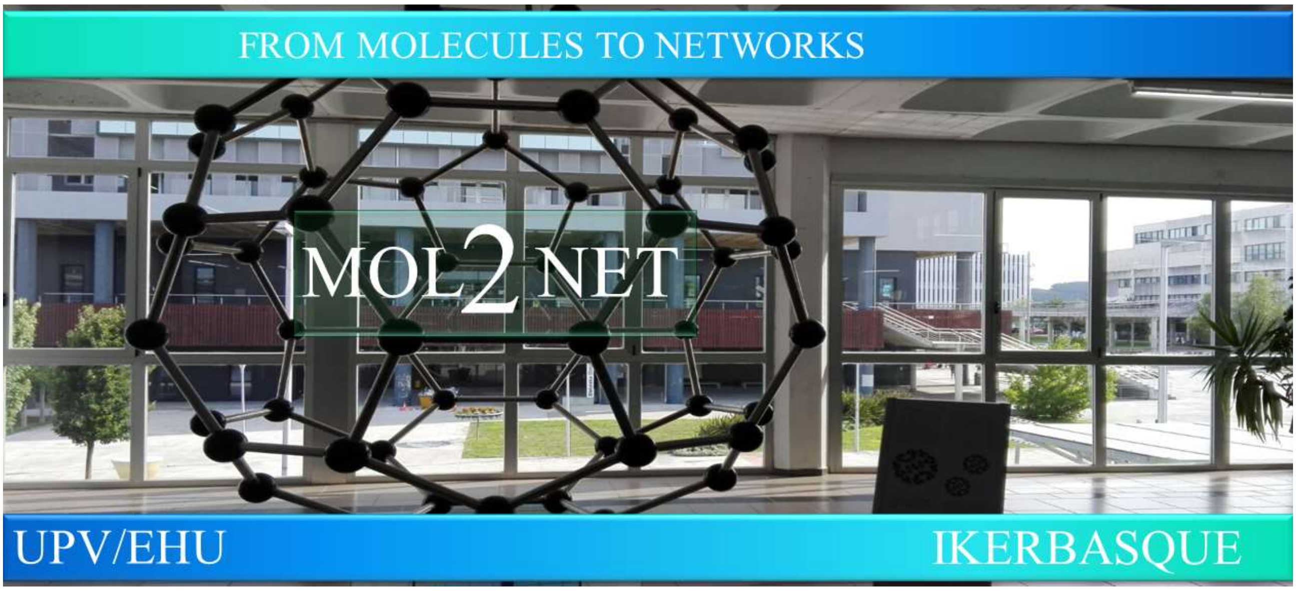
MOL2NET, International Conference Series on Multidisciplinary Sciences



Machine Learning Analysis of α-amylase Inhibitors

Karel Diéguez-Santana[a]* and Bakhtiyor Rasulev[b]

[a]Department of Organic and inorganic Chemistry, University of Basque Country UPVEHU, Leioa, Spain.

[b]Department of Coatings and Polymeric Materials, North Dakota State University, Fargo, ND 58102, USA

Abstract: In this work we report a Machine Learning study of a dataset involving the α-amylase inhibitors. The prediction of α-amylase inhibitory activity as anti-diabetic is carried out using LDA and classification trees (CT). A large data set of 640 compounds for α-amylase was selected to developing the ML models. In the case of CT-J48 have the better classification model performances with values above 80- 90% for the training and prediction sets, correspondingly. The best model shows an accuracy higher than 95% for training set; the model was also validated using 10-fold cross-validation procedure and through a test set achieving accuracies values of 85.32% and 86.80%, correspondingly.

Keywords: Anti-diabetic Agents; Decision trees; QSAR; Linear Discriminant Analysis.

E-mail: karel.dieguez.santana@gmail.com

1. INTRODUCTION

1.1 Diabetes mellitus.

According to World Health Organization estimations, about >1.0 million people decease yearly as a direct consequence of *Diabetes Mellitus* (DM) (Organization 2014). Safe and appropriate glucose lowering is important, especially in older people with DM. There are some conventional therapies at hand, such as: consumption of oral hypoglycemic agents, improvement of insulin effect on target cells, stimulation of endogenous insulin secretion and inhibition of dietary starch degradation by glycosidases such as α-amylase (DM type-2). Since α-amylase is a key enzymes in insulin adjustment, their inhibition is a therapeutic target for retarding glucose absorption and suppressing postprandial hyperglycemia (Bhandari, Jong-Anurakkun et al. 2008). Taking the abovementioned issues into account, the main goal of this research was to build the largest diverse datasets for the α-amylase inhibitory activity. Also, we used Dragon's molecular descriptors to classify the compounds according to their inhibitory activity using ML techniques.

3.RESULTS AND DISCUSSIONS

3.1 Anti-diabeticLDAmodels development and validation

After applying the IMMAN filtering and FSW-FS techniques over each 0D, 1D, and 2D descriptor family, the total number of descriptors was reduced up to 21. Further application of the BS method resulted in subsets of 10 and 14 descriptorsforα-amylase and or α-glucosidase, respectively. The canonical discriminant functions for classifying the anti-diabetic inhibitory activity of each enzyme are given by**Eqs.1** and **2**.

$$\textbf{\textit{Class (}}\boldsymbol{\alpha}\textbf{\textit{-Amylase)}} = + 0.637 - 3.185*BEHp4 + 45.770*JGI3 - 0.559*BELe5 + 0.549*MATS6v + 1.621*BELe6 - 0.294*ESpm11u + 0.665*BELv3 + 1.992*BEHv7 + 2.980*BELp5 - 1.794*MATS4v \qquad \textbf{\textit{(1)}}$$

$N = 640$

Wilks's lambda values are less than the unity (λ < 1) for discriminant functions (**1**) and (**2**). These values are significant sincetheir corresponding ji-squared statistics are larger than the critical values (*p < 0.0001*). Also, Mahalanobis' distance values are associated lo highly significant *F*-ratio scores (*p < 0.0001*). These results indicate that LDA functions (**1**) is able to effectively separate the group of inhibitory compounds from the group of non-inhibitory compounds. We use the term "non-inhibitory compounds" to refer to the fact that this compounds are not confirmed inactive (CI) in screening assays. On the other hand, results are an indirect evidence that automatically selected MDs have a higher discriminatory ability.

It can also be observed thatthe α-amylase QSAR model (**1**) yields adequate classification performancesof 75.31% and 77.57% for the training and test set, respectively. The MCC calculated is larger than 50%, which represent the balance and adequate level of the true positive and negative rates. We took 0.5 as an acceptable threshold for the MCC since the class prevalence is unknown a priori. Considering the previous analyses, we can state that the LDA-QSAR models (**1**) and (**2**), developed from 10 and 14 automatically selected descriptors, are statistically robust and can be applied as virtual screening tools to predict the anti-diabetic activity of new molecular entities.

Anti-diabetic decision-tree-based models development and validation

The nonlinear decision tree algorithm J48 showed good performances over the training, cross-validation, and test sets. Specifically, over the training data set j48 models exhibited an accuracy score (MCC) of 88.13% (0.77) for α-amylase target. For the case of the test repository, tree-based QSAR models yielded accuracies (MCC) of 83.18% (0.67). Moreover, this virtual high-throughput screening tool is able to classify with relatively low *fpr*. It means that they are able to effectively diminishing the wrong assessment of potential active (positive) compounds. The *fpr* values behaved below the 12% threshold [LDA(α-Amylase): 3.67 % training, 9.17% cross-validation, and 11.82% for the test sets]. Although this values are higher than the previous case, they keep below 15%, which is also recommended (Castillo-Garit, Casanola-Martin et al. 2017).

REFERENCES

Bhandari, M. R., N. Jong-Anurakkun, G. Hong and J. Kawabata (2008). "α-Glucosidase and α-amylase inhibitory activities of Nepalese medicinal herb Pakhanbhed (Bergenia ciliata, Haw.)." Food Chemistry **106**(1): 247-252.

Castillo-Garit, J. A., G. M. Casanola-Martin, H. Le-Thi-Thu, H. Pham-The and S. J. Barigye (2017). "A Simple Method to Predict Blood-Brain Barrier Permeability of Drug-Like Compounds Using Classification Trees." Medicinal Chemistry **13**(7): 664-669.

Organization, W. H. (2014). "Global Health Estimates: Deaths by Cause, Age, Sex and Country, 2000-2012." Geneva, WHO.