CMA4CH 2010, Mediterraneum Meeting

Application of Multivariate Analysis and Chemometry to Cultural Heritage and Environment
3rd ed., Taormina, Sicily island, Italy, Europe, 26-29 September 2010

Hybrid Strategies in Variable Selection for PLS Regression in a QSAR Study on Indulines Derivatives as 5-HT_{2c} Receptor Antagonists

H. Kaviani, J.B. Ghasemi*
Chemistry Department, Faculty of sciences, K. N. Toosi University of Technology, Tehran, Iran

Abstract

Hybridization methods are able to combine some beneficial features of a number of chemometrics methods. On the other hand, variable selection is the most important part in regression modeling. In this work, a hybrid approach that combines genetic algorithm (GA) and variable importance in projection (VIP) is proposed to achieve a variable selection method in PLS analysis. This method was applied for QSAR studies of a set of 1–(3–Pyridylcarbamoyl) induline derivatives. Thanks to hybridization method, not only lowering of the variable selected numbers was possible, but also a low dimensional PLS model with interpretable variables was obtained. The squared regression coefficient of prediction for training and test sets obtained by PLS model were 0.911 and 0.843 respectively. The effect of each class of descriptors in the final model, on the binding affinity of indulines derivatives, thoroughly explained and discussed in a descriptive manner.

Introduction

QSAR studies discover relationship between physiochemical properties of molecules and their structures. Selecting of most relevant and high correlated descriptors is a vital step in QSAR studies and it improves the performance of learning models, provides faster and more cost-effective predictors. However, evaluation of all possible feature subsets is computationally intractable and so is feasible only for very limited feature set. Thus practical feature selection algorithms are exclusively heuristic and necessary prior to building PLS models. With a larger dataset, variable prediction will be more problematic. Some efforts have done to solve this problem, for instance by combining of GA and PLS for variable selection in QSAR studies [1, 2].

The goal of this paper is to develop a methodology for variable selection using GA and recently reported method VIP Scores for a set of 85 derivatives of 1-(3-pyridylcarbamoyl) induline, which are agonists and antagonists that bind to the 5-HT_{2c} receptor and exhibit anxiolytic and antidepressant activity.

Materials & Methods

The structures of the molecules were drawn in HyperChem. After molecular optimizing, the resulted geometry was transferred into the program Dragon in order to obtain all of the descriptors [3]. The calculated descriptors were first analyzed to decrease the redundancy and correlation existing in the descriptor data matrix. Then the following steps applied into data matrix; (1) Removing variables consisting of many zero value and nearly constant variable by standard deviation index, (2) Running genetic algorithm [4] , (3) Performing calibration through PLS and OSC preprocessing on the training set and obtaining descriptors VIP. After drawing descriptors' VIP vector, those with VIPs that look like as noises were eliminated, (4) Finally performing the PLS model by remaining descriptors with 4 latent variables and examining on the test set.The calibration and prediction qualities were quantified with R^2 (training set) and Q^2 (test set), Root mean square errors also were obtained to evaluate the model quality.

Results

Fig. 1 demonstrates the plot of predicted activity by regression model against the experimental activity. 11The closeness of the data to the hypothetical perfect fit of the data to a linear model. Moreover, the low values of root - mean square error of prediction confirm the prediction ability and the accuracy of the resulted models for external test set.

Some parts of the present data set has been used by other research groups to develop QSAR based molecular models, MIA - QSAR and CoMFA. The comparison of the present model to the previous models reveals that our model has distinct priority according to the statistical parameters over MIA and CoMFA models. The presented model is more interpretable than MIA - PLS and is simpler than CoMFA model [5].

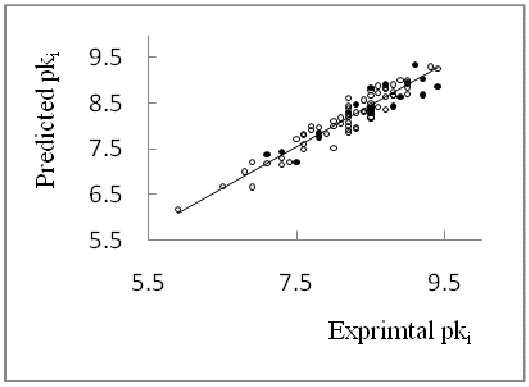

Fig. 1; Predicted vs. the Experimental activity (•) Test set (○) Training set

Conclusions

In the present study, the combination of GA, VIP and PLS is used to develop a variable selection method and regression technique, as a powerful optimization tool and PLS as a robust statistical method are applied to variable selection and regressing. GA-VIP-PLS can build multiple models simultaneously, which enables users to inspect and select different models. GA-VIP-PLS was applied to QSAR studies of 1-(3-pyridylcarbamoyl) induline derivatives, and many better model were built.

References

1) B.S. Dayal, J.F. MacGregor, Improved PLS Algorithms, *J. Chemometr.*, 11(1), (1997) 73-85.
2) H. Kubinyi, Evolutionary Variable Selection in Regression and PLS Analyses, *J. Chemometr.*, 10(2), (1996) 119-133
3) Package S.R.L. Talete, DRAGON for Windows (software for molecular descriptors calculation), Version 3.0 (2003)
4) R. Leardi, Genetic Algorithms in Chemometrics and Chemistry: A review, *J. Chemometr.*, 15(7), (2001), 559-569
5) M.P. Freitas, Multivariate image analysis applied to QSAR: Evaluation to a series of pontial anxiolytic agents, *Chemometrics Intell. Lab. Syst*, 91(2), (*2008)* 173-176