2D Quantitative Structure Property Relationship Study of Mycotoxins by Multiple Linear Regression and Support Vector Machine

A. Jabbari[1], J. B. Ghasemi[2] and F. Shiri[3]

[1,2]Chemistry Department, Faculty of Sciences, K.N.Toosi University of Technology, Tehran, Iran

[3]Faculty of Chemistry, Razi University, Kermanshah, Iran

Abstract

In the present work, support vector machines (SVMs) and multiple linear regression (MLR) were used for quantitative structure –property relationship (QSPR) study of retention time (t_R) in standardized liquid chromatography–UV–mass spectrometry of 67 mycotoxins (aflatoxins, trichothecenes, roquefortines and ochratoxins) based on molecular descriptors calculated from the optimized 3D structures. By applying missing value, zero and multicollinearity tests with a cutoff value of 0.95 and genetic algorithm method of variable selection the most relevant descriptors selected to build QSPR models. Multiple linear regression and support vector machines methods were employed to build QSPR models. The applicability domain of the model was investigated using William's plot. The effects of different descriptor on the retention times are described.

Introduction

Fungi are major plant and insect pathogens, but they are not nearly as important as agents of disease in vertebrates, i.e., the number of medically important fungi is relatively low [1].Studies have shown that a number of mycotoxins have carcinogenic properties. Some of them are clearly DNA-reactive and for others DNA reactivity may not be the mode of action. When the endpoint is cancer, in vitro or in vivo studies may need to be designed to elucidate possible molecular events related to gene expression, modifications of relevant proto-oncogenes or tumor suppressor genes, and genomic instability, as this will help in gaining an understanding of the mode of action underlying the carcinogenic process and in the characterization of hazard. quantitative structure- property relationship (QSPR) is a useful tool to predict the retention time avoiding long and tedious separation optimization. The QSPR study can also tell us which of the structural factors may play an important role in the determination of retention time.

Materials & Methods

The data set for this investigation was extracted from a work reported by K.F. Nielsen et al. [2]. This data set was randomly divided into two groups: training (50 compound) and prediction (17 compounds) sets. The molecular structures of data set were sketched using ChemDraw Ultra module of the CS ChemOffice 2005 molecular modeling software ver. 9, supplied by Cambridge Software Company. Each molecule was "cleaned up" and energy minimization was performed using Allinger's MM2 force filed and further geometry optimization was done using semiempirical AM1 (Austin Model) Hamiltonian and PM3 methods by default on the 3D-structure of molecules. A total of 54 molecular descriptors of differing types based on 3D structures were calculated to describe compound structural diversity. By applying missing value, zero and multicollinearity tests with a cutoff value of 0.95 and genetic algorithm method of variable selection the most relevant descriptors selected to build QSPR models. The GA is implemented in MATLAB (version 7.1, MathWorks, Inc.). After the descriptor was selected, multiple linear regression was used to develop the linear model of the property of interest, The SPSS software, (SPSS Ver. 11.5, SPSS Inc.), performed

multiple linear regression (MLR) analysis and variable selection by using stepwise method for the variable selection and modeling.

Results

In this paper new QSPR models have been developed for predicting the t_R of a diverse set of mycotoxins from the molecular structure alone. The best linear model contained 4 molecular descriptors (C log p, ElcE, DPLL and LUMO). Positive value in the regression coefficient for C log p demonstrates that with the increase of C log p, the value of t_R increases as well.
In reversed-phase chromatography, compounds with higher hydrophobicities would make stronger 11interactions with mobile phase which lead to having larger t_R within the compounds. The other descriptors (LUMO, DPLL and ElcE)are electronic and their regression coefficient is negative, it means as they increase, t_R decreases. We have compared tow linear models MLR and SVM with the data set. The linear kernel function was used for the SVR model in our study.

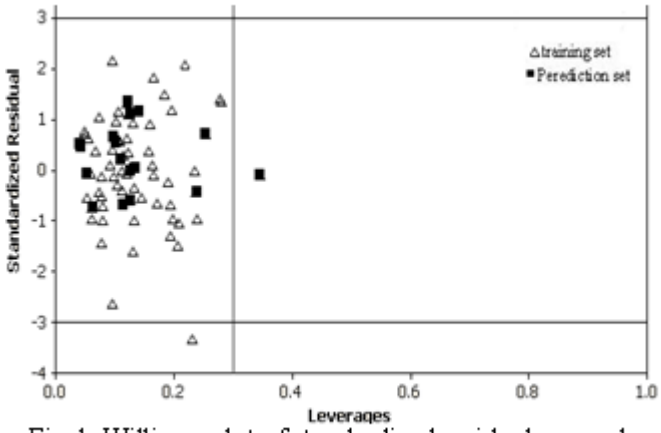

Fig 1. Williams plot of standardized residual versus leverage

The correlation and predictability measure by r^2 and q^2 are for SVM 0.931 and 0.932 and MLR 0.923 and 0.915 respectively. The obtained results show that both MLR and SVM methods could model the relationship between t_R and their electronic and thermodynamic descriptors, while model using SVM based on these same sets of descriptors produced even better model with a good predictive ability than the MLR model. SVM exhibit the better whole performance due to embodying the structural risk minimization principle and some advantages over the other techniques of converging to the global optimum and not to a local optimum. The Williams plot for the presented SVM model was shown in Figure 1. From this plot, the applicability domain[3] is established inside a squared area within ±3 standard deviations and a leverage threshold h^* of 0.3. For making predictions, predicted t_R data must be considered reliable only for those compounds that fall within this AD on which the model was constructed. It can be seen from Figure 1 that the majority of compounds in the dataset are inside this area.

Conclusions

In recent years, QSAR/QSPR methods have been paid attention to as an interesting complement, or even as an expensive, time consuming alternative laboratory data. By performing the model validation, it can be concluded that the presented model is a valid model and can be effectively used to predict the t_R of mycotoxins with an accuracy approximating the accuracy of experimental t_R determination. Moreover, the mechanism of the model was interpreted, and the applicability domain of the model was defined.

References

1) J.W. Bennett, M. Klich, Mycotoxins, *Clin. Microbiol. Rev.*, 16(3), (2003) 497-516
2) K.F. Nielsen, J. Smedsgaard, Fungal metabolite screening: database of 474 mycotoxins and fungal metabolites for dereplication by standardised liquid chromatography-UV-mass spectrometry methodology, *J. Chromatogr. A,* 1002(1-2), (2003) 111-136
3) P. Gramatica, Principles of QSAR models validation: internal and external, *QSAR Comb. Sci.*, 26(5), (2007) 694-701