Int. J. Mol. Sci. **2010**, *11*, 3052-3068; doi:10.3390/ijms11093052

OPEN ACCESS

International Journal of
Molecular Sciences
ISSN 1422-0067
www.mdpi.com/journal/ijms

Article

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

Roya Khosrokhavar [1], Jahan Bakhsh Ghasemi [2],* and Fereshteh Shiri [3]

[1] Food and Drug Laboratory Research Center, MOH & ME, Tehran, Iran;
E-Mail: khosrokhavar_r@yahoo.com

[2] Department of Chemistry, Faculty of Sciences, K.N., Toosi University of Technology,
Tehran 16617, Iran

[3] Faculty of Chemistry, Razi University, Kermanshah, Iran; E-Mail: fereshteh.shiri@gmail.com

* Author to whom correspondence should be addressed; E-Mail: Jahan.ghasemi@gmail.com;
Tel.: +98-21-228-502-66; Fax: +98-21-228-536-50.

*Received: 6 July 2010; in revised form: 15 August 2010 / Accepted: 16 August 2010 /
Published: 31 August 2010*

Abstract: In the present work, support vector machines (SVMs) and multiple linear regression (MLR) techniques were used for quantitative structure–property relationship (QSPR) studies 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 were selected to build QSPR models. MLRand SVMs methods were employed to build QSPR models. The robustness of the QSPR models was characterized by the statistical validation and applicability domain (AD). The prediction results from the MLR and SVM models are in good agreement with the experimental values. The correlation and predictability measure by r^2 and q^2 are 0.931 and 0.932, repectively, for SVM and 0.923 and 0.915, respectively, for MLR. The applicability domain of the model was investigated using William's plot. The effects of different descriptors on the retention times are described.

Keywords: QSPR; mycotoxins; SVM; MLR; genetic algorithm; William's Plot

1. 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. Growth of fungi on animal hosts produces diseases collectively known as mycoses, while dietary, respiratory, dermal, and other exposures to toxic fungal metabolites produce diseases collectively called mycotoxicoses. Mycotoxicoses are examples of "poisoning by natural means" and thus are analogous to the pathologies caused by exposure to pesticides or heavy metal residues. The symptoms of mycotoxicosis depend on the type of mycotoxin; the amount and duration of the exposure; the age, health, and sex of the exposed individual; and many poorly understood synergistic effects involving genetics, dietary status, and interactions with other toxic insults. Thus, the severity of mycotoxin poisoning can be compounded by factors such as vitamin deficiency, caloric deprivation, alcohol abuse, and infectious disease status. In turn, mycotoxicoses can heighten vulnerability to microbial diseases, worsen the effects of malnutrition, and interact synergistically with other toxins [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. Mycotoxins may also cause developmental effects including birth defects, affect the reproductive system, affect the immune system, exhibit hormonal activity, affect specific target organs and may be neurotoxic. In addition to these diverse organ or site-specific actions, mycotoxins may affect the gastrointestinal system, cause skin irritation, have hematological effects and reduce growth [2–4].

Mycotoxins usually enter the body via ingestion of contaminated foods, but inhalation of toxigenic spores and direct dermal contact are also important routes. Mycotoxins occurring in food commodities are secondary metabolites of a range of filamentous fungi, which can contaminate food or food crops throughout the food chain. Although many hundreds of fungal toxins are known, a more limited number are generally considered to play an important part in food safety and for these a range of analytical methods have been developed [5].

Microfungi are a rich source of chemical diversity [6–8], and together with the actinomycetes they are the source of more than 50% of metabolites utilized by the pharmaceutical industry in either the native form or as derivatives [9–12].As only a small part of mycota is known and most fungi produce several unknown metabolites, fungi are still one of the most promising microbiotic sources for new lead compounds. Therefore, developing theoretical models to predict the property (e.g., retention time) of mycotoxins is necessary as they toxicity is very important for humans and animals.

Since the chemical diversity is very high within the micro-fungi almost all types of chemical structure can be expected in an extract, e.g., small acids, alcohols, ketones, alkaloids, antraquinones and cyclic peptides. To cope with this broad range of chemical structures, most methods are based on reversed- phase liquid chromatography combined with diode array detection (DAD) and atmospheric ionization [electrospray ionization (ESI) and atmospheric pressure chemical ionization (APCI)] mass spectrometry (MS). Nearly all methods use water–acetonitrile gradient elution on reversed-phase C_{18}

and C_8 columns, although methods for very polar and highly ionized components, using perfusion chromatography and hydrophilic interaction chromatography have been described [13].

However, only a few reports have investigated the quantitative correlation between the molecular parameters and the property of retention time of mycotoxins [14]. The computational methods used to calculate/predict retention time can be classified into two categories. One approach is to use a mathematical equation to correlate retention time with the molecular parameters. The other methods are more empirically based on QSPR approaches using multiple linear regression (MLR) and support vector machine (SVM) techniques. Of those previous studies that aimed to predict the retention time, the most promising method has been to use the QSPR approach: QSPR methods have been successfully used to predict many physicochemical properties. The advantage of this approach over other methods lies in the fact that the descriptors used can be calculated from the structure alone and are not dependent on any experimental properties. Once the structure of a compound is known, any descriptor can be calculated, no matter whether it is found or not. This means that once a reliable model is established, we can use this method to predict properties of compounds. Therefore, quantitative structure- property relationship (QSPR) is a useful tool to predict the retention time, avoiding long and tedious separation optimization. QSPR studies can also tell us which of the structural factors may play an important role in the determination of retention time.

After the calculation of molecular descriptors, many different chemometrics methods, such as multiple linear regression (MLR), partial least squares regression (PLS), different types of artificial neural networks (ANN), genetic algorithms (GAs), and support vector machine (SVM) can be employed to derive correlation models between the molecular structures and properties. As a new and powerful modeling tool, support vector machine (SVM) has gained much interest in pattern recognition and function approximation applications recently. In bioinformatics, SVMs have been successfully used to solve classification and correlation problems. SVMs have also been applied in chemistry, for example, the prediction of retention index of protein [15], and other QSAR studies. Compared with traditional regression and neural networks methods, SVMs have some advantages, including global optimum, good generalization ability, simple implementation, few free parameters, and dimensional independence [16]. The flexibility in classification and ability to approximate continuous function make SVMs very suitable for QSAR and QSPR studies. In the present paper, we introduce the applications of support vector regression (SVR) for correlation problems in QSAR and compare its performance with MLR method.

2. Results and Discussion

54 descriptors were calculated by the ChemOffice software. By applying missing value, zeroand multicollinearity tests with a cutoff value of 0.95 and variable selection by genetic algorithm, the number of descriptors was reduced to 22.The stepwise regression routine was used to develop the linear model for the prediction of the retention time of mycotoxins using calculated structural descriptors. The best linear model contained four molecular descriptors. The regression coefficients of the descriptors, Mean effect and variable inflation factors (VIF) are listed in Table 1.

Positive values in the regression coefficients show that the indicated descriptors contribute positively to the value of t_R, whereas negative values indicate that the greater the value of the descriptor, the lower the value of t_R. In other words, increasing the electronic energy (ElcE), dipole length (DPLL)and Lowest Unoccupied Molecular Orbital energy (LUMO) will decrease t_R, and the increase in the C log*P* increases the extent of t_R of the compounds.

Table 1. Details of the constructed QSPR model.

Descriptor	**Coefficient**	**Mean effect**	**VIF**[e]
C logP[a]	2.6951(±0.2248)	5	1.006
ElcE[b]	-0.0002(±0.0001)	8	1.246
DPLL[c]	-1.091(±0.2981)	-3.875	1.556
LUMO[d]	-1.6922(±0.5521)	0.594	1.287
Constant	3.1912(±1.7569)	_	_

a = The octanol/water partition coefficient

b = Electronic energy

c = Dipole length

d = Lowest Unoccupied Molecular Orbital energy

e = Variable inflation factors

With comparison of the mean effects of the descriptors appearing in MLR model, it is observed that the ElcE of the molecules has the largest effect on the t_R of the compound. The mean effect of a descriptor is the product of its mean and the regression coefficient in the MLR model [17].

Based on the variable inflation factor (VIF) values of the four descriptors shown in Table 1, it has been found that the descriptors used in the model have very low inter-correlation. Correlation between these descriptors and property as correlation matrix of measured data are given in Table 2. Correlation coefficients measure how closely two values (descriptor and property) are related to each other by a linear relationship. If a descriptor has a correlation coefficient of 1, it describes the property exactly. A correlation coefficient of zero means the descriptor has no relevance. It is seen that C logP is positivelycorrelated to the property with a correlation coefficient equal to 0.82126.

Table 2. Correlation matrix for MLR model.

	t_R	**C logP**	**ElcE**	**DPLL**	**LUMO**
t_R	1				
C logP	0.821263	1			
ElcE	-0.21234	0.05977	1		
DPLL	-0.07144	0.004813	-0.32903	*1*	
LUMO	-0.12041	-0.05044	0.000773	-0.45025	1

After establishing models by MLR, the support vector machines were used to compare the performance of MLR based on the same subset of descriptors. Similar to other multivariate statistical models, the performances of SVM for regression depend on the combination of several parameters. They are capacity parameter *C*, *ε* of *ε*-insensitive loss function, the kernel type *K*, and its corresponding parameters. *C* is a regularization parameter that controls the tradeoff between maximizing the margin and minimizing the training error. If *C* is too small, then insufficient stress will be placed on fitting the training data. If *C* is too large, then the algorithm will overfit the training data. The linear kernel function was used for the SVR model in our study for investigation of the linear relationship between the theoretical molecular descriptors and the retention time. The optimal value for *ε* depends on the type of noise present in the data, which is usually unknown. Even if enough knowledge of the noise is available to select an optimal value for *ε*, there is the practical consideration of the number of resulting support vectors. *ε*-insensitivity prevents the entire training set meeting boundary conditions and so allows for the possibility of sparsity in the dual formulation's solution. So, choosing the appropriate value of *ε* is critical from theory. To find an optimal ε, the root mean squares error (RMSE) on LOO cross-validation on different *ε* was calculated. The curve of RMSE *versus* the epsilon (*ε*) is shown in Figure 1. The optimal *ε* was found to be 0.014. The other important parameter is regularization parameter *C*, whose effect on the RMSEis shown in Figure 2. The optimal *C* was found to be 4.

Figure 1. The selection of the optimal epsilon for SVM (C = 4).

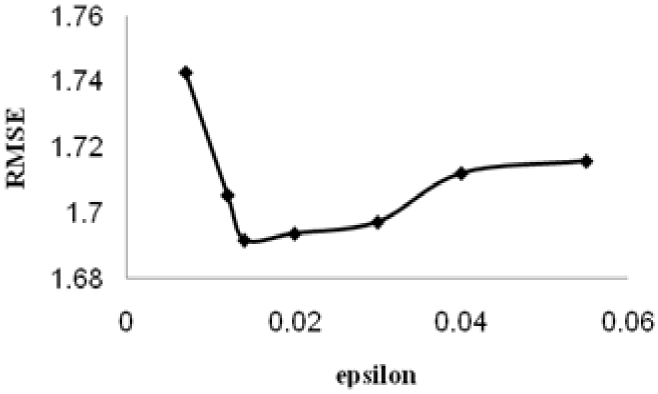
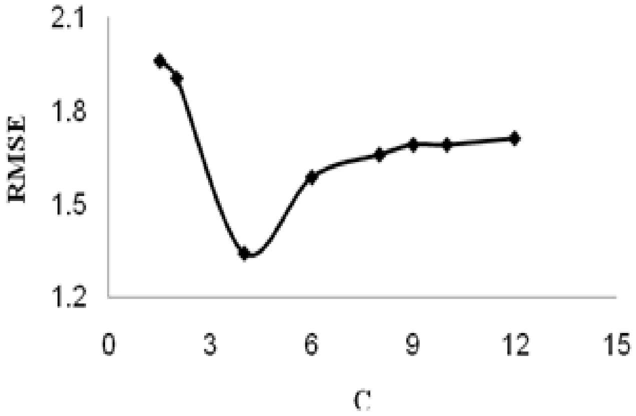


Figure 2. The selection of the optimal capacity factors for SVM (ε = 0.01).



Satisfied with the robustness of the QSPR model developed using the training set, we applied the QSPR model to an external data set of 17 mycotoxins comprising the test set. The predicted results are given in Table 3.The squared correlation coefficient between experimental and predicted t_R values for the test set for both models is significant. Figure 3 shows the quality of the fit. Also the random distribution of residuals about zero mean in Figure 3 confirms the good predictive ability of the models.

Figure 3. t_R estimated by MLR (top panel) and SVM (bottom panel) modeling *versus* experimental values and residual *versus* experimental t_R.

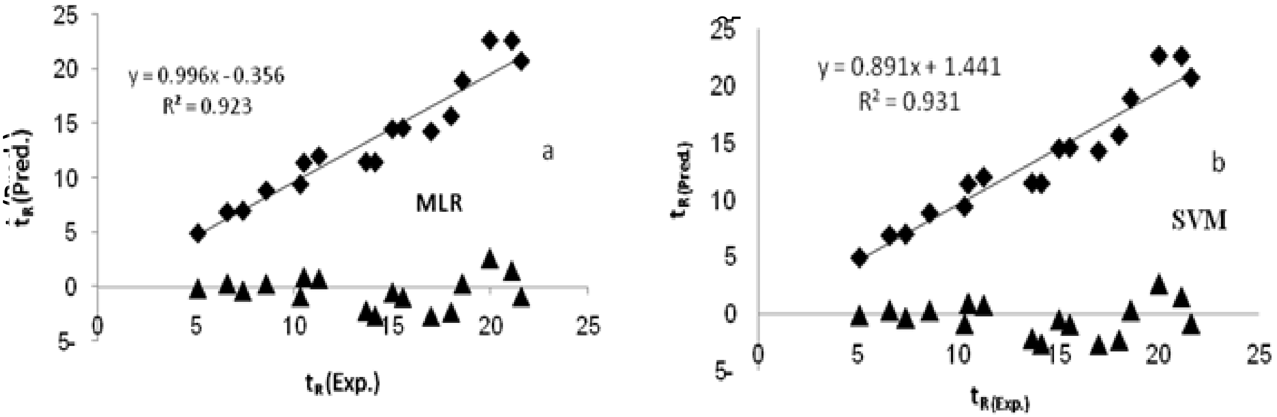


Table 3. Comparison of experimental and predicted values of t_R for prediction set by MLR and SVM models.

No.	Exp. (t_R)	MLR model		SVM model	
		Pred. (t_R)	RE (%)	Pred. (t_R)	RE (%)
21	5.1	4.97	2.55	5.03	1.37
4	6.6	6.91	-4.7	7.99	-21.06
23	7.4	7.03	5	8.35	-12.84
41	8.59	8.88	-3.38	10.08	-17.35
3	10.33	9.44	8.62	10.25	0.77
38	10.51	11.43	-8.75	12	-14.18
24	11.28	12.03	-6.65	12.37	-9.66
27	13.69	11.51	15.92	11.74	14.24
34	14.15	11.48	18.87	12.53	11.45
13	15.03	14.52	3.39	15.18	-1
25	15.56	14.61	6.11	14.79	4.95
37	17	14.29	15.94	15.08	11.29
11	18.02	15.7	12.87	16.37	9.16
46	18.6	18.91	-1.67	19.39	-4.25
65	20	22.66	-13.3	22.11	-10.55
29	21.12	22.61	-7.05	20.43	3.27
55	21.6	20.74	3.98	19.84	8.15

The statistical parameters calculated for the MLR and SVM models are represented in Table 4. In this table, statistical parameters root mean squared error of prediction (RMSEP),standard error of prediction (SEP),relative error of prediction (REP%) and the others parameters obtained by applying the MLR and SVM methods to the test set indicate a good external predictability of the QSPR models. The results also 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 an even better model with a better predictive ability than the MLR model.SVM performs better on the whole due to embodying the structural risk minimization principle and the advantage over other techniques of converging to the global optimum and not to a local optimum.

Table 4. The statistical parameters obtained by applying the MLR and SVM methods to the prediction set.

Parameters	MLR	SVM
RMSEP	1.504	1.341
REP[a] (%)	10.902	9.719
SEP[b]	1.551	1.382
q^2	0.915	0.932
R^2	0.923	0.931
$(R^2\text{-}R_0^2)/R^2$	0.001	0.0118
$(R^2\text{-}R'^2_0)/R^2$	0.0108	0.0011
r_m^2	0.894	0.833
k	0.996	0.891
k'	0.926	1.045
NDS[c]	4	4

a = Relative error of prediction.

b = Standard error of prediction.

c = Number of descriptors.

2.1. Definition of the Applicability Domain of the Model

Once a QSPR model is obtained, another crucial problem is the definition of its applicability domain (AD). For any QSPR model, only the predictions for chemicals falling within its AD can be considered reliable and not model extrapolations. There are several methods for defining the AD of QSPR models [18], but the most common one is determining the leverage values for each compound [19]. To visualize the AD of a QSPR model, the plot of standardized residuals *versus* leverage values (*h*)(the William’s plot) was exploited in this study, which played a double role. Firstly, it described the impacts of the objects on models by the values of their leverages. Leverage indicates a compound’s distance from the centroid of *X*. The leverage of a compound in the original variable space is defined as [20]:

$$\hbar_i = x_i^T(X^TX)^{-1}x_i \tag{1}$$

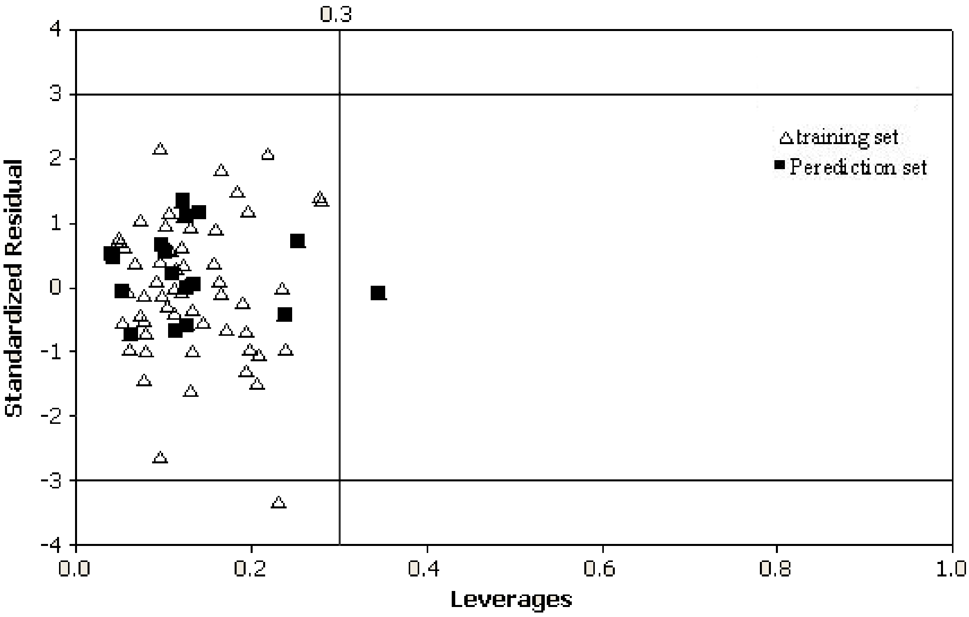
where *xi* is the descriptor vector of the considered compound and *X* is the descriptor matrix derived from the training set descriptor values. The warning leverage (*h**) is defined as [18]:

$$\hbar^* = \frac{3P}{n} \tag{2}$$

where *n* is the number of training compounds, *p* is the number of model variables plus one. The leverage (*h*) greater than the warning leverage (h^*) suggested that the compound was very influential on the model. Secondly, it presented the Euclidean distances of the compounds to the model measured by the cross-validated standardized residuals. The cross-validated standardized residuals greater than three standard deviation (*s*) units classified the compound as a response outlier.

The Williams plot for the presented SVM model is shown in Figure 4.From this plot, the applicability domain 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 4 that the majority of compounds in the data set are inside this area. However, only one compound in prediction set(squares at 0.33 h) slightly exceeds the critical hat value that the developed SVM model has good generalizability and predictivity for the compound with descriptor values significantly far from the centroid of the descriptor space. Also, compound **2** in the training set is wrongly predicted (>3 s), but with lower leverage values ($h < h^*$).These erroneous predictions could probably be attributed to wrong experimental data rather than to molecular structures [19].

Figure 4. Williams plot of standardized residual *versus* leverage.



2.2. Interpretation of Descriptors

By interpreting the descriptors in the regression model, it is possible to gain some insight into factors that are likely to govern the retention time of mycotoxins. In regard to this point that all the descriptors in the final model together attributethe same property or activity, each one of the descriptors or their related coefficient takes into account a definitive amount of variance within property. However it can be concluded that the interpretation of a combination set of the descriptors would be much better than considering the result of the single descriptors. Of the four descriptors, C logP is thermodynamic and LUMO, DPLL and ElcE are electronic descriptors.

The octanol/water partition coefficient (C logP) characterizes the effectiveness of hydrophobicity of the compounds. C logP values can be calculated from molecular structure by summation of fragment values, which captures the nature of the hydrophobic regions of the molecule separately from hydrophilic regions. In the other words, it can be estimated from hydrophobic contributions of the chemical groups present in complex molecules [21,22]. The fact that similar descriptors have been reported to correlate with partition coefficients of different compounds suggests that this correlation model has wider applications [23]. A positive value in the regression coefficient for C logp demonstrates that with the increase of C logp, the value of t_R increases as well. In reversed-phase chromatography, compounds with higher hydrophobicities would make stronger interactions 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 that as they increase, t_R decreases. In particular, electronic parameters are considered important in the establishment of QSAR models and are helpful to quantify different types of intermolecular and intramolecular interactions, as these interactions are usually responsible for properties of chemical and biological systems [24]. Dipole length is the electric dipole moment divided by the elementary charge. Electric dipole is a vector quantity, which encodes displacement with respect to the centre of gravity of positive and negative charges in a molecule. Dipole length encodes information about the charge distribution in molecules and is important for modeling polar interactions. Large substituents decrease the DPLL valuem which is not desirable [25,26]. The ElcE descriptor has the largest effect on the t_R of the compounds. The ElcE is the total electronic energy given in electron volt at 0 ℃ [27]. Involvement of electronic factors suggests the occurrence of either charge transfer or dipolar interactions. The transfer of a pair of electrons from the HOMO to the LUMO is, by definition, a reaction between a Lewis acid and a Lewis base. Thus, the parameter LUMO is a measure of the ability of a molecule to interact with the π and *n*-electron pairs of the other molecules. The reduction in energy in molecular orbital is the driving force for chemical bond formation [28].The negative sign of the corresponding regression coefficient between t_R and LUMO indicates that, t_R increase with decrease in the magnitude of LUMO index. The present results reinforce previous findings [29,30].

3. Experimental Section

3.1. Data Set

The data set for this investigation was extracted from a work reported by Nielsen *et al.* [13]. These data are listed in Table 5. It can be seen from the table that the data set is diverse, consisting of aflatoxins, trichothecenes, roquefortines and ochratoxins. This data set was randomly divided into two groups: training (calibration) and prediction (test) sets. The training and prediction sets consisted of 50 and 17 molecules, respectively. The values of t_R were used as the dependent variables.

Table 5. Experimental retention time (t_R) of 67compounds.

NO.	Compound	t_R(min)	NO.	Compound	t_R(min)
Aflatoxins and their precursors					
1	Aflatoxicol I	12.45	9	Austocystin A	21.57
2	Aflatoxin B_1	11.50	10	Averufin	25.65
3	Aflatoxin B_2	10.33	11	5-Methoxysterigmatocystin	18.02
4	Aflatoxin B_2 α	6.60	12	Dihydroxysterigmatocystin	17.70
5	Aflatoxin G_1	10.16	13	Methoxysterigmatocystin	15.03
6	Aflatoxin G_2	8.97	14	Sterigmatocystin	18.91
7	Aflatoxin G_2α	5.00	15	Norsolorinic acid	31.08
8	Aflatoxin M_1	7.21	16	Parasiticol	10.73

Table 5. *Cont.*

NO.	Compound	t_R (min)	NO.	Compound	t_R (min)
Trichothecenes					
17	Nivalenol	1.27	27	HT-2 Toxin	13.69
18	Fusarenone X	2.35	28	T-2 Toxin	17.06
19	Deoxynivalenol	1.54	29	Acetyl-T-2 toxin	21.12
20	3-Acetyldeoxynivalenol	5.21	30	Trichodermin	16.13
21	15-*O*-Acetyl-4-deoxynivalenol	5.10	31	Trichodermol	9.69
22	Scirpentriol	1.82	32	7-α-Hydroxytrichodermol	2.59
23	15-Acetoxyscirpenol	7.40	33	Verrucarol	2.89
24	Diacetoxyscirpenol	11.28	34	4,15-Diacetylverrucarol	14.15
25	3α-Acetyldiacetoxyscirpenol	15.56	35	Trichothecin	16.29
26	Neosolaniol	3.19	36	Trichothecolone	3.63
37	Trichoverrol A	10.16			
Roquefortines ,ergot amines and related alkaloids					
38	Agroclavine-I	17.00	51	Ergotamin	19.60
39	Auranthine	10.51	52	Fumigaclavine C	21.40
40	Aurantiamine	10.49	53	Marcfortine A	19.59
41	Aurantioclavine	14.30	54	Marcfortine B	17.39
42	Chanoclavine-I	8.59	55	Meleagrin	18.90
43	Costaclavine	17.00	56	Oxalin	21.60
44	Cyclopenin	11.60	57	Pyroclavine	14.81
45	Cyclopenol	6.20	58	Roquefortine C	20.50
46	Cyclopeptin	12.05	59	Roquefortine D	6.09
47	Dihydroergotamin	18.60	60	Rugulovasine A and B	8.43
48	Elymoclavine	5.34	61	Secoclavine	20.40
49	Epoxyagroclavine-I	10.00	62	α-Ergocryptin	19.20
50	Ergocristine	25.10			
Ochratoxins					
63	Ochratoxin α	5.60	66	Ochratoxin B-ethyl ester	19.41
64	Ochratoxin A-methyl ester	22.49	67	Ochratoxin α-methyl ester	16.16
65	Ochratoxin B-methyl ester	20.00			

3.2. Descriptor Generation and Reduction

The molecular structures of data set were sketched using the ChemDraw Ultra module of the CS ChemOffice 2005 molecular modeling software version 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. The descriptors calculated accounts three important properties of the molecules:

(a) thermodynamic, (b) electronic and (c) steric, as they represent the possible molecular interactions which determined the retention time of the studied molecules.

After the calculation of molecular descriptors, any parameter which is not calculated (missing value) for any number of the compounds in the data set is rejected in the first step. Some of the descriptors were rejected because they contained a value of zero for all the compounds and have been removed (zero tests).In order to minimize the effect of colinearity and to avoid redundancy, we used amulticollinearity test with a cutoff value of 0.95, and subsequently discarded 10 parameters. Finally, a total set of 44 remaining descriptors were achieved and used to select the optimal subset of descriptors that have a significant contribution to the t_R property.

3.3. Descriptor Selection and Model Building

The basic strategy of QSPR analysis is to find optimum quantitative relationships between the molecular descriptors and desired property, which can then be used for the prediction of the property from only molecular structures. One of the most important problems involved in QSPR studies is to select optimal subset of descriptors that have significant contribution to the desired property. The well-known genetic algorithm is just a well-accepted method for solving this kind of problems.

After correlation analysis of the descriptors, we used MLR analysis on the molecular descriptors that resulted in genetic algorithm (GA) variable selection procedure. The GA-algorithm applied in this paper uses a binary representation as the coding technique for the given problem; the presence or absence of a descriptor in a chromosome is coded by 1 or 0. The GA performs its optimization by variation and selection via the evaluation of the fitness function (RMSECV). The algorithm used in this paper is an evolution of the algorithm described in Ref. [31], whose parameters are reported in Table 6. In our study, a genetic algorithm procedure was used for selection of descriptors using the PLS Toolbox (version 2.0, Eigenvector Company, USA). The GA is implemented in MATLAB (version 7.1, MathWorks, Inc.). By performing GA, 22 descriptors were retained for next analysis step.

Table 6. Parameters of genetic algorithm (GA).

Cross-Validation	**Random subset**
Number of subsets	4
Population size	64
Mutation rate	0.005
Window width	2
Initial term%	20%
Maximum generation	100
Convergence (%)	50
Cross-over	Double

Finally, descriptor-screening methods were used to select the most relevant descriptor to establish the models for prediction of the molecular property. Here, the stepwise regression method was used to choose the subset of the molecular descriptors.

After the descriptor was selected, multiple linear regression (MLR)[32] was used to develop the linear model of the property of interest, which takes the form below:

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n \quad (3)$$

In this equation, *y* is the property, that is, the dependent variable, x_1-x_n represent the specific descriptor, while b_1- b_n represent the coefficients of those descriptors, and b_0 is the intercept of the equation. The statistical evaluation of the data was obtained by the software SPSS. The SPSS software, (SPSS Ver. 11.5, SPSS Inc.), performed MLR analysis and variable selection by using stepwise method for the variable selection and modeling.

3.4. Theory of SVM

The foundation of support vector machines (SVM) has been developed by Vapnik, and they are gaining popularity due to many attractive features and promising empirical performance [33]. The formulation embodies the structural risk minimization (SRM) principle [32,33], which has been shown to be superior to the traditional empirical risk minimization (ERM) principle, employed by conventional neural networks. SRM minimizes an upper bound on VC dimension ("generalization error"), as opposed to ERM that minimizes the error on the training data. It is the difference that equips SVM with good generalization performance, which is the goal in statistical learning. Originally, SVM were developed for classification problems [34], and now, with the introduction of *ε*-insensitive loss function, SVM have been extended to solve nonlinear regression estimation [36].

Compared to other neural network regressors, there are three distinct characteristics when SVM are used to estimate the regression function. First of all, SVM estimate the regression using a set of linear functions that are defined in a high dimensional space. Second, SVM carry out the regression estimation by risk minimization where the risk is measured using Vapnik's *ε*-insensitive loss function. Third, SVM use a risk function consisting of the empirical error and a regularization term which is derived from the SRM principle.

In support vector regression (SVR), the basic idea is to map the data *x* into a higher-dimensional feature space *F* via a nonlinear mapping Φ, and then to do linear regression in this space. Therefore, regression approximation addresses the problem of estimating a function based on a given data set G = $\{(x_i,d_i)\}i^n$ (x_i is the input vector, d_i is the desired value, and *n* is the total number of data patterns). SVM approximate the function using the following

$$y = f(x) = w\Phi(x) + b \quad (4)$$

where Φ(*x*) denotes the element wise mapping from *x* into feature space. The coefficients *w* and *b* are estimated by minimizing

$$R_{SVMs}(C) = C\frac{1}{n}\sum_{i=1}^{n} L_e(d_i, y_i) + \frac{1}{2}||w||^2 \quad (5)$$

$$L_\varepsilon(d, y) = \{\begin{cases} |d-y| - \varepsilon|d-y| \geq \varepsilon \\ 0 \qquad otherwise \end{cases} \quad (6)$$

In Equation 5, R_{SVMs} is the regularized risk function, and the first term $C\frac{1}{n}\sum_{i=1}^{n} L_e(d_i, y_i)$ is the empirical error (risk). They are measured by the *ε*-insensitiveloss function (L_ε) given by Equation 6. This loss function provides the advantage of enabling one to use sparse data points to represent the decision function given by Equation 4. The second term $\frac{1}{2}||w||^2$, on the other hand, is the regularization term. *C* is referred to as the regularized constant, and it determines the tradeoff between the empirical risk and the regularization term. Increasing the value of *C* will result in the relative importance of the empirical risk with respect to the regularization term to grow.

ε is called the tube size, and it is equivalent to the approximation accuracy placed on the training data points. Both *C* and *ε* are user-prescribed parameters.

Finally, by introducing Lagrange multipliers (a_i, a_i^*) andexploiting the optimality constraints, the decision functiongiven by Equation 4 has the following explicit form:

$$f\left(x, a_i, a_i^*\right) = \sum(a_i - a_i^*)K(x, x_i) + b \tag{7}$$

Based on the Karush-Kuhn-Tucker (KKT) conditions of quadratic programming, only a number of coefficients (a_i, a_i^*) will assume nonzero values, and the data points associated with them could be referred to as support vectors. In Equation 7, the kernel function *K* corresponds to $K(x, x_i) = \Phi(x).\Phi(x_i)$. One has several possibilities for the choice of this kernel function, including linear, polynomial, splines, and radial basis function. The elegance of using the kernel function lies in the fact that one can deal with feature spaces of arbitrary dimensionality without having to compute the map $\Phi(x)$ explicitly.

The overall performances of SVM models were evaluated in terms of root mean square error (RMSE), which was defined as below:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n_s}(y_k - y^\wedge{}_k)^2}{n_s}} \tag{8}$$

where y_k is the desired output, $\hat{y}_k$ is the predicted value and *n*s is the number of samples in the analyzed set.

The predictive power of the models developed on the calculated statistical parameters standard error of prediction (SEP) and relative error of prediction (REP %) as follows:

$$SEP = \left[\frac{\sum_{i=1}^{n}(\tilde{y}_i - y_i)^2}{n-1}\right]^{0.5} \tag{9}$$

$$REP(\%) = \frac{100}{\bar{y}}\left[\frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - y_i)^2\right]^{0.5} \tag{10}$$

where $\tilde{y}_i$, y_i and $\bar{y}$ are the predicted, experimental and mean activity property, respectively.

All calculations in this work were carried out by using Matlab (V 7.1, The Mathworks, Inc.) and the SVM toolbox developed by Gunn [37].

3.5. Validation Test

The main goal in QSPR studies is to obtain a model with the highest predictive ability. In order to evaluate the predictive ability of our QSPR model, we used the method described by Golbraikh and Tropsha [38] and Roy and Roy [39]. The determination coefficient in prediction (q^2_{test}) was calculated using the following equation [39]:

$$q_{test}^2 = 1 - \frac{\sum(y_{pred\ test} - y_{test})^2}{\sum(y_{test} - \bar{y})^2} \tag{11}$$

where $y_{predtest}$ and y_{Test} are the predicted values based on the QSPR equation (model response) and experimental activity values, respectively, of the external test set compounds. $\bar{y}$ is the mean activity value of the training set compounds. Further evaluation of the predictive ability of the QSAR model for the external test set compounds was done by determining the value of r_m^2 using the following equation [39]:

$$r_m^2 = r^2 \left(1 - |\sqrt{r^2 - r_0^2}\ |\right) \tag{12}$$

where r^2 is the squared Pearson correlation coefficient for regression calculated using $Y = a + bx$; ''a'' is referred to as the y-intercept, ''b'' is the slope value of regression line, and r_0^2 is the squared correlation coefficient for regression without using y-intercept and the regression equation was $y = bx$. Both r^2 and r_0^2 between experimental and predicted values for the external test set compounds were calculated using the regression of analysis Toolpak option of Excel. If r_m^2 value for a give model is >0.5, it indicates the good external predictability of the developed model.

The values of k and k', slopes of the regression line of the predicted property *versus* actual property and *vice versa*, were calculated using the following equations [38]:

$$k = \frac{\sum y_i \tilde{y}_i}{\sum \tilde{y}_i^2} k' = \frac{\sum y_i \tilde{y}_i}{\sum y_i^2} \tag{13}$$

where $\tilde{y}_i$ and y_i are the predicted and experimental property, respectively. The values of k and k' are within the specified range of 0.85 and 1.15 [36]. The value of $[\mathrm{r}^2 - \mathrm{r}_0^2/\mathrm{r}^2]$ and $\left[\mathrm{r}^2 - \mathrm{r}_0^{2'}/\mathrm{r}^2\right]$ are less than 0.1 (stipulated value)[38]. r_0^2 and $r_0^{2'}$ are correlation coefficient of regression between the predicted and experimental property of compounds in the test set and *vice versa* without using y-intercept.

To further check the inter-correlation of descriptors variance inflation factor (VIF) analysis was performed. The VIF value is calculated from $1/1 - \mathrm{r}^2$, where r^2 is the multiplecorrelation coefficient of one descriptor's effect regressed on the remaining molecular descriptors. If the VIF value is larger than 10, information of the descriptor could be hidden by correlation of descriptors [40].

4. Conclusions

In recent years, attention has been paid to QSAR/QSPR methods as an interesting complement, or even as an expensive, time consuming alternative, to laboratory data. 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. We have compared two linear models, MLR and SVM, with the data set. The obtained results show that both MLR and SVM methods could model the relationship between t_R and their electronic and thermodynamic descriptors; on the same sets of descriptors, using SVM based produced a better model with a better predictive ability than the MLR model.SVM exhibit the better overall 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. By performing 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. It can be reasonably concluded that the proposed model would be expected to predict t_R for new organic compounds or for other organic compounds for which experimental values are unknown. Additionally, the presented method could also identify and provide some insight into what structural features are related to the t_R property of organic compounds.

References

1. Bennett, J.W.; Klich, M. Mycotoxins. *Clin. Microbiol. Rev.* **2003**, *16*, 497–516.
2. Magan, N.; Olsen, M. *Mycotoxins in Food Detection and Control*; Woodhead Publishing Limited:Cambridge, UK, 2000; p. 7.
3. Baggiani, C.; Giraudi, G.; Vanni, A. A molecular imprinted polymer with recognition properties towards the carcinogenic mycotoxin ochratoxin A. *Bioseparation* **2001**, *10*, 389–394.
4. El-Nezami, H.; Kankaanpaa, P.; Salminen, S.; Ahokas, J. Ability of dairy strains of lactic acid bacteria to bind a common food carcinogen, aflatoxin B_1. *Food Chem. Toxicol.* **1998**, *36*, 321–326.
5. Shephar, G.S. Determination of mycotoxins in human foods. *Chem. Soc. Rev*. **2008**, *37*, 2468–2477.
6. Frisvad, J.C.; Thrane, U.; Filtenborg, O. *Chemical Fungal Taxonomy*; Marcel Dekker: New York, NY, USA, 1998; pp. 289–301.
7. Mantle, P.G. *Secondary Metabolites of Penicillium and Acremonium*; Plenum Press: New York, NY, USA, 1987; pp.135–151.
8. Gloer, J.B. The chemistry of fungal antagonism and defense. *Can. J. Bot.***1995**, *73*, S1265–S1274.
9. Bull, A.T.; Ward, A.C.; Goodfellow, M. Search and discovery strategies for biotechnology: The paradigm shift. *Microbiol. Mol. Biol.* **2002**, *R64*, 573–606.
10. Bentley, R. Mycophenolic Acid: A One hundred year odyssey from antibiotic to immunosuppressant. *Chem. Rev*. **2000**, *100*, 3801–3826.
11. Constant, H.L.; Beecher, C.W.W. A method for the dereplication of natural product extracts using electrospray HPLC/MS. *Nat. Prod. Lett*. **1995**, *6*, 193–196.
12. Corley, D.G.; Durley, R.C. Strategies for database dereplication of natural products. *J. Nat. Prod.* **1994**, *57*, 1484–1490.
13. Nielsen, K.F.; Smedsgaard, J. Fungal metabolite screening: Database of 474 mycotoxins and fungal metabolites for dereplication by standardised liquid chromatography–UV–mass spectrometry methodology. *J. Chromatog. A* **2003**, *1002*, 111–136.
14. Steinmetz, W.E.; Rodarte, C.B.; Lin, A. 3D QSAR study of the toxicity of trichothecene mycotoxins. *Eur. J. Med. Chem*. **2009**, *44*, 4485–4489.
15. Song, M.; Breneman, C.M.; Bi, J.; Sukumar, N.; Bennett, K.P.;Cramer, S.; Tugcu, N. Prediction of protein retention times in anion exchange chromatography systems using support vector regression.*J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1347–1357.
16. Yao, X.J.; Panaye, A.; Doucet, J.P.; Zhang, R.S.; Chen, H.F.; Liu, M.C.; Hu, Z.D.; Fan, B.T. Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1257–1266.
17. Jalali Heravi, M.; Konuze, E. Use of quantitative structure property relationships in predicting the Kraft point of anionic surfactants. *Int. Electron. J. Mol. Des.* **2002**, *1*, 410–417.
18. Eriksson, L.; Jaworska, J.; Worth, A.P.; Cronin, M.T.D.; McDowell, R.M.; Gramatica, P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification and regression-based QSARs. *Environ. Health Perspect*. **2003**, *111*, 1361–1375.

19. Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **2007**, *26*, 694–701.
20. Netzeva, T.I.; Worth, A.P.; Aldenberg, T.; Benigni, R.; Cronin, M.T.D.; Gramatica, P.; Jaworska, J.S.; Kahn, S.; Klopman, G.; Marchant, C.A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G.Y.; Perkins, R.; Roberts, D.W.; Schultz, T.W.; Stanton, D.T.; vande Sandt, J.J.M.; Tong, W.; Veith, G.; Yang, C. Current status of methods for defining the applicability domain of (quantitative) structure–activity relationships. The report and recommendations of ECVAM Workshop 52. *Altern. Lab. Anim*. **2005**, *33*, 155–173.
21. Roberts, D.W. Application of octanol/water partition coefficients in surfactant science: A quantitative structure-property relationship for micellization of anionic surfactants. *Langmuir* **2002**, *18*, 345–352.
22. Leo, A.J. Calculating log P_{oct} from structures. *Chem. Rev.* **1993**, *93*, 1281–1306.
23. Katritzky, A.R.; Lobanov, V.S.; Karelson, M. QSPR: The correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.* **1995**, *24*, 279–287.
24. Khan, M.S.; Khan, Z.H. Molecular modeling for generation of structural and molecular electronic descriptors for QSAR using quantum mechanical semiemprical and ab initio methods. *Gen. Inf.* **2003**, *14*, 486–487.
25. Ghasemi, J.; Abdolmaleki, A.; Asadpour, S.; Shiri, F. Prediction of solubility of nonionic solutes in anionic micelle (SDS) using a QSPR model. *QSAR Comb. Sci.* **2008**, *27*, 338–346.
26. Todeschini, R.; Consonni, V.; Mannhold, R.; Kubinyi, H.; Timmerman, H. *Handbook of Molecular Descriptors*; Wiley-VCH in Weinheim: New York, NY, USA, 2000; pp. 324–345.
27. Melagraki, G.;Afantitis, A. A novel QSPR model for predicting *θ* (lower critical solution temperature) in polymer solutions using molecular descriptors. *J. Mol. Model* **2007**, *13*, 55–64.
28. Afantitis, A.; Melagraki, G.; Sarimveis, H.; Koutentis, P.A.; Markopoulos, J.; Igglessi-Markopoulou, O. A novel QSAR model for predicting induction of apoptosis by 4-aryl-4H-chromens. *Bioorg. Med. Chem.* **2006**, *14*, 6686–6694.
29. Altomare, C.; Cellamare, S.; Carotti, A.; Barreca, M.L.; Chimirri, A.; Monforte, A.M.; Gasparrini, F.; Villani, C.; Cirilli, M.; Mazza, F. Substituent effects on the enantioselective retention of anti-HIV 5-aryl-Δ2-1,2,4-oxadiazolines on R, R-DACH-DNB chiral stationary phase. *Chirality* **1996**, *8*, 556–566.
30. Altomare, C.; Carotti, A.; Cellamare, S.; Fanelli, F.; Gaspar-rini, F.; Villani, C.; Carrupt, P.A.; Testa, B. Enantiomeric resolution of sulfoxides on a DACH-DNB chiral stationary phase: A quantitative structure-enantioselective retention relationship (QSERR) study. *Chirality* **1993**, *5*, 527–537.
31. Holland, J.H. Genetic algorithms. *Sci. Am.* **1992**, *267*, 66–72.
32. Ghasemi, J.; Saaidpour, S.; Brown, S.D. QSPR study for estimation of acidity constants of some aromatic acids derivatives using multiple linear regression (MLR) analysis. *J. Mol. Struct.* **2007**, *805*, 27–32.
33. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 1995; pp. 217–231.
34. Burges, C.J.C. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Disc*. **1998**, *2*, 1–47.

35. Vapnik, V. *Estimation of Dependences Based on Empirical Data*; Springer: Berlin, Germany, 1982; pp. 137–145.
36. Vapnik, V.; Golowich, S.; Smola, A. Support vector method for function approximation, regression estimation, and signal processing. In *Neural InformationProcessing Systems*; MIT Press: Cambridge, MA, USA, 1997; Volume 9, pp. 281–287.
37. Image Speech and Intelligent Systems Research Group. Applying SVM Toolbox. University of Southampton: Southampton, UK. Available at: http://www.isis.ecs.soton.ac.uk/isystems/kernel/svm.zip (accessed on 16 August 2010).
38. Golbraikh, A.; Tropsha, A. Beware of q^2. *J. Mol. Graph. Model*. **2002**, *20*, 269–276.
39. Roy, P.P.; Roy, K. On some aspects of variable selection for partial least squares regression models. *QSAR Comb. Sci.* **2008**, *27*, 302–313.
40. Shapiro, S.; Guggenheim, B. Inhibition of oral bacteria by phenolic compounds. Part 1 QSAR analysis using molecular connectivity. *Quant. Struct. Act. Relat.* **1998**, *17*, 327–337.