JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

Prediction of Fungicidal Activities of Rice Blast Disease Based on Least-Squares Support Vector Machines and Project Pursuit Regression

HONGYING DU,[†] JIE WANG,[†,‡] ZHIDE HU,*[,†] XIAOJUN YAO,[†] AND XIAOYUN ZHANG[†]

Department of Chemistry, Lanzhou University, Lanzhou 730000, China, and Department of Biomedical Engineering, Yale University, New Haven, Connecticut 06511

Three machine learning methods, genetic algorithm-multilinear regression (GA-MLR), least-squares support vector machine (LS-SVM), and project pursuit regression (PPR), were used to investigate the relationship between thiazoline derivatives and their fungicidal activities against the rice blast disease. The GA-MLR method was used to select the most appropriate molecular descriptors from a large set of descriptors, which were only calculated from molecular structures, and develop a linear quantitative structure−activity relationship (QSAR) model at the same time. On the basis of the selected descriptors, the other two more accurate models (LS-SVM and PPR) were built. Both the linear and nonlinear modes gave good prediction results, but the nonlinear models afforded better prediction ability, which meant that the LS-SVM and PPR methods could simulate the relationship between the structural descriptors and fungicidal activities more accurately. The results show that the nonlinear methods (LS-SVM and PPR) could be used as good modeling tools for the study of rice blast. Moreover, this study provides a new and simple but efficient approach, which should facilitate the design and development of new compounds to resist the rice blast disease.

KEYWORDS: Quantitative structure−activity relationship (QSAR); fungicidal activity; genetic algorithm (GA); projection pursuit regression (PPR); least-squares support vector machine (LS-SVM)

INTRODUCTION

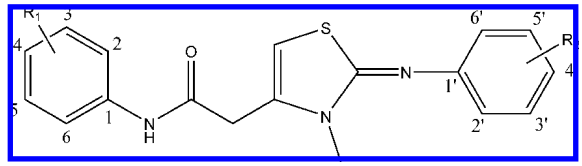
Rice blast disease is one of the most important and damaging diseases for rice, and it will cause substantial reduction in crop yields (*1*). This disease is caused by filamentous fungus, *Pyricularia oryzae* (teleomorph, *Magnaporthe grisea*). The pathogenic fungus directly penetrates into the rice plant from a cellular structure called an appressorium that is formed at the tip of the germ tube (*2*). The appressoria synthesize melanin, which is deposited between the plasma membrane and the appressorial cell well (*3*). The fungus mechanically punctures the hard epidermis of rice by using osmotic pressure and penetrates into rice (*4*). The melanization of the appressorium is essential for developing control agents against rice blast disease (*5*). With this aim, the fungus requires and uses melanin-derived pressure; thus, melanin biosynthesis inhibition has been shown to be a promising biochemical target for the discovery of new selective fungicides. To solve this question, many scientists have been attempting to find new chemicals that are effective for preventing rice blast disease.

Quantitative structure−activity relationship (QSAR) modeling (*6*−*9*) is a useful tool in activity assessments of many inhibitors. The advantage of the QSAR approach over the other methods lies in the fact that the descriptors used to build models can be calculated from the molecular structure alone and used computational algorithms to relate the key descriptors to the dependent property values of interest (*10*). Therefore, it is possible to explore these activities from the reliable models. However, the main problems encountered in this research are still the description of the molecular structure using appropriate molecular descriptors and the selection of suitable modeling methods. Many multivariate data analysis methods, such as multiple linear regression (MLR) (*11*) or partial least-squares (PLS) (*12*) and artificial neural network (ANN) (*13*), have been used in QSAR studies. However, the practical usefulness of MLR in QSAR studies is rather limited, because it provides relatively poor accuracy; ANN can offer satisfactory accuracy in most cases but tends to overfit the training data. In this study, two novel approaches least-square support vector machine (LS-SVM) and projection pursuit regression (PPR) were used to model the fungicidal activities of thiazoline against rice blast. The LS-SVM method, which is proposed by Suykens et al. (*14*), is a simplification of traditional support vector machine (SVM). It encompasses similar advantages with SVM and its own additional advantages. It only requires solving a set of linear

* To whom correspondence should be addressed. Telephone: +86-931-891-2540. Fax: +86-931-891-2582. E-mail: hu_zhide@yahoo.com.cn.
[†] Lanzhou University.
[‡] Yale University.

10.1021/jf8022194 CCC: $40.75
Published on Web 10/25/2008

Table 1. Chemical Structures, Experimental Values, and Predicted Fungicidal Activities by MLR, LS-SVM, and PPR



number	R1	R2	experimental log *A*	GA-MLR	LS-SVM	PPR	number	R1	R2	experimental log *A*	GA-MLR	LS-SVM	PPR
1	4-CH_3	4′-F	1.96	2.07	2.00	1.98	51	4-OC_6H_5	3′-Br	0.90	1.08	1.02	0.92
2[a]	4-CH_3	4′-Br	1.94	2.01	2.00	2.14	52	4-OC_6H_5	3′-F	1.23	1.19	1.16	1.19
3	4-CH_2CH_3	4′-Cl	2.00	1.96	2.00	1.96	53	4-Cl	3′-Cl, 4′-F	1.98	2.11	1.97	1.98
4[a]	4-CH_2CH_3	4′-C_6H_5	1.99	1.87	1.91	2.00	54[a]	4-Cl	3′-Br	1.98	2.15	2.00	2.02
5	4-CH_2CH_3	4′-OCF_3	1.99	1.92	1.98	1.97	55	4-Cl	3′-F	1.96	2.10	1.95	1.97
6[a]	4-CH_2CH_3	4′-C_6H_{13}	2.00	1.92	1.93	1.97	56[a]	4-Cl	3′-OCH_3	1.76	2.03	1.98	1.94
7	4-CH_2CH_3	4′-C_4H_9	1.99	2.00	2.00	1.98	57	4-Br	3′-Br	1.99	2.01	2.00	1.97
8	4-CH_2CH_3	4′-OCF_3	1.93	1.95	1.99	1.98	58	4-Br	3′-F	1.90	1.85	1.89	1.92
9	4-CH_2CH_3	4′-$COCH_3$	2.00	2.01	1.99	1.98	59	4-OCF_3	3′-CF_3	1.70	1.30	1.40	1.57
10	4-CH_2CH_3	4′-OC_6H_5	1.98	2.11	2.00	2.01	60	4-OCF_3	3′-F	1.62	1.28	1.31	1.47
11	4-CH_2CH_3	4′-CH_2CN	2.00	1.99	1.99	2.02	61[a]	4-NO_2	3′-F	1.23	1.07	1.04	1.07
12	4-CH_2CH_3	4′-CN	1.92	1.91	1.96	1.97	62	4-CH_3	2′-F, 4′-F	1.99	2.10	2.01	2.02
13[a]	4-CH_2CH_3	4′-I	1.97	1.77	1.91	2.01	63[a]	4-CH_3	2′-F	1.94	2.08	2.00	1.99
14	4-CH_2CH_3	4′-OC_5H_{11}	1.96	1.85	1.91	1.94	64	4-CH_2CH_3	2′-F	1.99	1.85	1.94	1.96
15	4-$CH(CH_3)_2$	4′-CF_3	1.80	1.64	1.79	1.67	65	4-CH_2CH_3	2′-F, 4′-Br	2.00	1.72	1.89	2.01
16	4-OCH_3	4′-OCF_3	1.99	2.02	2.02	1.94	66	4-CH_2CH_3	2′-F, 4′-Cl	1.99	1.87	1.96	1.96
17	4-OCH_3	4′-OC_2H_5	1.98	1.89	1.94	1.96	67	4-CH_2CH_3	2′-Cl, 4′-F	1.98	1.90	1.97	1.96
18	4-OCH_3	4′-NO_2	2.00	1.99	1.96	1.98	68	4-CH_2CH_3	2′-CH_3, 3′-Cl	1.95	1.94	1.95	1.97
19	4-OCH_3	4′-Cl	1.97	2.07	2.01	1.97	69	4-CH_2CH_3	2′-CH_3, 4′-Br	1.98	1.88	1.95	1.89
20	4-OCH_3	4′-C_4H_9	1.98	2.09	2.02	2.00	70[a]	4-CH_2CH_3	2′-Br, 4′-CH_3	1.94	2.01	2.00	1.92
21	4-OCH_3	4′-SCH_3	1.96	1.99	1.96	1.93	71	4-CH_2CH_3	2′-OCH_3, 4′-NO_2	1.81	1.83	1.92	1.90
22	4-OC_4H_9	4′-CF_3	1.00	1.26	1.09	1.14	72	4-OCH_3	2′-F, 4′-F	1.97	1.99	1.98	1.97
23	4-OC_6H_5	4′-OC_6H_5	1.40	1.17	1.19	1.21	73	4-OCH_3	4′-Cl, 2′-F	1.97	1.99	1.98	1.97
24	4-OC_6H_5	4′-Br	1.23	1.14	1.12	1.09	74	4-OCH_3	2′-Cl	1.98	1.84	1.90	1.93
25[a]	4-OCF_3	4′-Cl	1.52	1.24	1.26	1.36	75[a]	4-OCH_3	2′-Br	1.92	1.82	1.88	1.92
26[a]	4-OCF_3	4′-OCH_3	1.40	1.05	0.98	1.06	76	4-OCH_3	2′-$CH(CH_3)_2$	1.90	1.94	1.94	1.97
27	4-NO_2	4′-C_4H_9	1.11	0.97	1.03	0.98	77	4-OCH_3	2′,4′-$(CH_3)_2$	1.96	2.01	1.97	1.91
28	4-NO_2	4′-F	0.90	1.05	0.97	1.06	78	4-OC_2H_5	2′-F, 4′-F	1.90	1.96	1.97	1.94
29	4-NO_2	4′-CN	0.70	0.95	0.73	0.74	79	4-OC_6H_5	2′-F, 4′-F	1.23	1.48	1.52	1.36
30	4-CHF_2	4′-$OCH(CH_3)_2$	1.70	1.51	1.65	1.57	80[a]	4-NO_2	2′-F, 4′-F	1.23	1.09	0.99	1.11
31	4-CHF_2	4′-Cl	1.52	1.50	1.60	1.52	81	4-OCF_3	2′,4′-$(CH_3)_2$	1.23	1.21	1.26	1.21
32	4-CHF_2	4′-OCH_3	1.40	1.39	1.51	1.55	82	4-OCF_3	2′-OCH_3	0.70	1.01	0.92	0.84
33	4-Cl	4′-C_6H_5	1.98	1.88	1.92	1.96	83	4-CH_2CH_3	4′-OCH_3	1.99	1.94	1.97	1.98
34[a]	4-Cl	4′-C_4H_9	1.97	1.90	1.95	1.92	84	4-CH_2CH_3	4′-Br	1.98	1.83	1.93	1.99
35[a]	4-CH_2CH_3	3′-NO_2	1.99	1.98	2.00	1.97	85[a]	4-OCH_3	4′-CF_3	1.99	1.95	1.96	1.99
36	4-CH_2CH_3	3′-Cl, 4′-F	1.98	1.96	1.98	1.98	86	4-OCH_3	4′-CH_2CH_3	1.98	2.12	2.01	2.02
37	4-CH_2CH_3	3′-NO_2, 4′-F	1.98	2.01	2.00	2.02	87	4-Cl	4′-OC_6H_5	1.92	1.82	1.89	1.90
38	4-CH_2CH_3	3′-CH_3, 4′-Br	1.98	1.98	1.99	1.96	88	4-Cl	4′-OCF_3	1.82	1.90	1.91	1.86
39[a]	4-CH_2CH_3	3′-Cl, 4′-CN	1.98	1.90	1.96	1.92	89	4-CHF_2	4′-Br	1.40	1.37	1.47	1.35
40[a]	4-C_4H_9	3′-Cl, 4′-F	1.30	1.24	1.11	1.06	90	4-OCF_3	4′-F	0.90	1.20	1.20	1.27
41	4-OCH_3	3′-CF_3, 4′-Cl	2.00	2.09	1.97	2.03	91	4-OCH_3	3′-NO_2	1.99	2.06	1.98	2.02
42	4-OCH_3	3′-CF_3	2.00	2.01	1.98	2.02	92	4-OCH_3	3′-F	1.97	2.03	2.00	1.94
43	4-OCH_3	3′-Br	1.99	2.02	2.02	2.01	93	4-OC_2H_5	3′-Cl, 4′-F	1.94	1.92	1.93	1.97
44	4-OCH_3	3′-Cl, 4′-F	1.98	2.05	2.01	1.95	94	4-OCF_3	3′-Br	1.70	1.40	1.55	1.62
45	4-OCH_3	3′-Cl, 4′-OCH_3	1.98	2.06	1.95	1.91	95	4-C_4H_9	3′-F	1.00	1.19	1.01	0.99
46[a]	4-OCH_3	3′,4′-Cl_2	1.98	2.05	2.00	1.95	96	4-CH_2CH_3	2′-Cl, 4′-CH_3	2.00	1.91	1.94	1.97
47[a]	4-OCH_3	3′-F, 4′-CH_3	1.98	2.04	1.96	1.95	97	4-CH_2CH_3	2′-CH_3, 4′-OCH_3	2.00	1.83	1.85	1.95
48	4-OCH_3	3′-CH_3	1.98	2.08	2.01	2.01	98	4-CH_2CH_3	2′,4′-Cl_2	1.96	1.89	1.96	2.00
49	4-OC_2H_5	3′-Br	1.91	1.92	1.92	1.95	99	4-OCH_3	2′-F	1.99	1.92	1.95	1.97
50	4-OC_4H_9	3′-F	1.00	1.22	1.04	1.07	100	4-CN	2′-F, 4′-F	1.23	1.24	1.24	1.20

[a] The test set.

equations (linear programming), which is much easier and computationally simpler than nonlinear equations (quadratic programming) employed by the traditional SVM. The other method PPR, which is developed by Friedman (*15*), seeks the "interesting" projections of data from high-dimensional to lower dimensional space and tries to find the intrinsic structural information hidden in the high-dimensional data (*16*).

In this investigation, the descriptors, which were separately calculated by CODESSA and DRAGON (*17, 18*), were combined together. The genetic algorithm-multilinear regression (GA-MLR) method was used to reduce the number of descriptors, select the relevant ones, and build the linear regression model. Afterward, the two nonlinear methods LS-SVM and PPR were used to build the nonlinear models. The aims of this work were to establish a robust QSAR model that could be used for the prediction of fungicidal activity of the drugs against the rice blast and explore the most important structure features to facilitate developing new chemicals in the future. The prediction results of the two nonlinear approaches (LS-SVM and PPR) were in agreement with the experimental data in both the training and test sets compounds. It has been proven that these two approaches are useful and promising tools in predicting the

fungicidal activities against the rice blast. This study provides a new and simple but efficient method, which is helpful to design and screen some new chemicals against the rice blast.

MATERIALS AND METHODS

Data Set. The studied 100 thiazoline derivatives and their corresponding fungicidal activities were taken from the literature (*19*) and listed in **Table 1**. Disease severity was determined by the percentage of infected leaf area, 5 days after the inoculation (*19*). The pots were arranged as a randomized complete block with three replicates per treatment. Three estimates for each treatment were converted into percentage fungal control value (*A*) as the following:

$$A = \%\ \text{control value} = 100[(a - b)/a] \quad (1)$$

where *a* is the area of infection (%) on leaves sprayed with Tween 20 solution alone, *b* is the area of infection (%) on the treated leaves with thiazoline derivatives dissolved in water + dimethyl sulfoxide (99 + 1 by volume) containing Tween 20 (250 mg/L).

To construct a predictive model, the selected thiazoline derivatives were randomly divided into two subsets: a training set and a test set. The training set including 80 compounds was used to select the most important molecular descriptors and construct the regression models; the test set did not take part in the construction of the models but was used to test the stability of them.

Molecular Descriptors Generation. Two-dimensional structures of the compounds were drawn using ISIS Draw 2.3 (*20*). All of the structures were transferred into HyperChem 7.0 (*21*) and pre-optimized using the MM+ molecular mechanics force field. A more precise optimization was performed with the semi-empirical PM3 method in MOPAC, and then the structures with minimum energy were obtained. After these steps, the molecular descriptors can be calculated. In this study, the molecular descriptors consist of two parts: one is calculated by CODESSA (*22*), which contains five kinds of molecular descriptors (*23*), and the other part is calculated by DRAGON 5.4 (*18*), which contains 18 kinds of molecular descriptors (*24*). During the DRAGON calculation process, to delete the redundant and non-useful information, the descriptors with constant or near constant values and the ones that were highly correlated pairwise (the correlation coefficients of these two descriptors were bigger than 0.99) were excluded in the pre-reduction step. Thus, there were a total of 1318 molecular descriptors left for further analysis.

Principle Component Analysis (PCA). To build a regression model, it is important to generate a validated training set, which can represent the whole data set. In this study, the PCA method was used to analyze the diversity of the training and test sets. Using the whole set of the generated descriptors, the PCA method was used to deduce the dimensions of the descriptors by eliminating the redundant information. To perform PCA, the descriptors with constant or missing values should be excluded. After this step, the PCA method was used for analysis, for which PC1, PC2, and PC3 made 21.48, 13.86, and 8.98% contribution to the total PCs, respectively. In all, these three PCs made a total of 44.32% of the variation in the data and played major roles. **Figure 1** illustrates the scores plot of the compounds in the training and test sets based on the three major PCs. From **Figure 1**, it can be concluded that all of the compounds in the training set are well-proportioned, distributing in the 3D spatial space. Thus, the results confirmed that it was feasible for the method of splitting the data set and the compounds in the training set were representative of the whole data set.

Selection of the Structural Descriptors and Model Construction. Once the molecular descriptors are generated, it is important to select the major descriptors for further constructing of the regression model. In this study, the GA-MLR method was used to select the most important descriptors and build the linear regression model based on the reason that the GA method was a promising parameters selection method. At last, five molecular descriptors (see **Table 2**) were selected. On the basis of these selected descriptors, the nonlinear regression methods LS-SVM and PPR were used to build two nonlinear regression models.

GA. After descriptor calculation, the GA method was performed to search the feature space and select the major descriptors relevant to the fungicidal activities against rice blast. This method can deal with large search space efficiently and has less chance to become a local optimal solution than the other algorithms. Basic theories and applications about GA have been found in many references (*25, 26*). Here, we only briefly summarize the main procedure of GA. The first step of GA is to generate a set of solutions (chromosomes) randomly, which is called an initial population. Then, a fitness function is deduced from the gene composition of a chromosome. The Friedman LOF function was used in our study as the fitness function, which was defined as follows:

$$\text{LOF} = \{\text{SSE}/(1 - (c + dp/n))\}^2 \quad (2)$$

where SSE is the sum of squares of errors, *c* is the number of the basis function (other than the constant term), *d* is the smoothness factor (default 0.5), *p* is the number of features in the model, and *n* is the number of data points from which the model is built. Unlike the R^2 error, the LOF measure cannot always be reduced by adding more terms to the regression model. By limiting the tendency to simply add more terms, the LOF measure resists overfitting of a model. Then, crossover and mutation operations are performed to generate new individuals. In the subsequent selection stage, the fittest individuals evolve to the next generation. These steps of evolution continue until the stopping conditions are satisfied. The MLR method is a simple and classical regression method, which can provide explicit equations. In the current work, the models were built using the simple MLR method with the selected variables from GA, called GA-MLR.

LS-SVM. The LS-SVM, which was a modified algorithm of SVM, was described clearly by Suykens et al. (*27, 28*) and used to build the nonlinear model. Here, we only briefly describe the main idea of LS-SVM for function estimation. In principle, LS-SVM always fits a linear relation ($y = wx + b$) between the regressors (x) and the dependent variable (y). The best relation can be obtained by minimizing the cost function (Q) containing a penalized regression error term

$$Q_{\text{LS-SVM}} = \frac{1}{2}w^T w + \gamma \sum_{k=1}^{N} e_k^2 \quad (3)$$

subject to

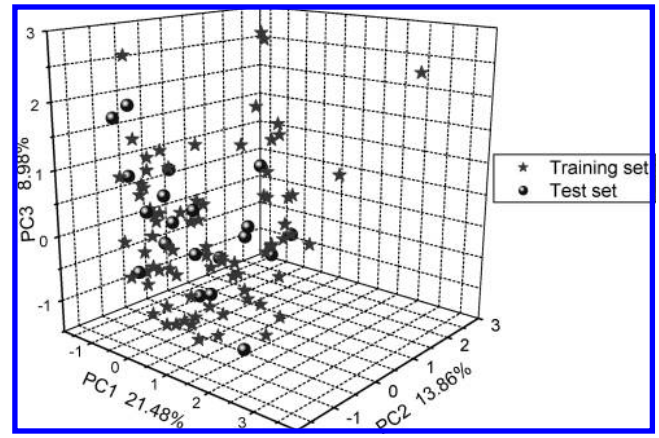


Figure 1. PCA of the training and test sets.

Table 2. Involved and Statistical Parameters of the MLR Model[a]

abbreviation	descriptors	coefficient	confidence intervals	*t* test	t_p
Con.	constant	−497.049	74.749		
MaxRHN	maximum n−n repulsion for a H−N bond	9.309	1.391	8.515	1.37×10^{-12}
RDF065u	RDF065u	−0.018	0.003	−5.504	5.10×10^{-7}
Mor02m	Mor02m	−0.028	0.004	−6.683	3.81×10^{-9}
L2s	L2s	−0.177	0.024	−10.151	1.11×10^{-15}
R1v+	R1v+	−5.690	1.074	−3.595	5.82×10^{-4}

[a] *n*, 80; R^2, 0.8666; adjusted R^2, 0.8575; R_{cv}^2, 0.8345; *F*, 96.11 (95% confidence level).

$$y_k = w^T \varphi(x_k) + b + e_k, \quad k = 1, ..., N \tag{4}$$

where $\varphi: R^n \rightarrow R^m$ is the feature map mapping the input space to a usually high-dimensional feature space, γ is the relative weight of the error term, and e_k is error variables taking noisy data into accurate and avoiding poor generalization.

LS-SVM considers this optimization problem to be a constrained optimization problem and uses a language function to solve it. By solving the Lagrange style of eq 3, the weight coefficients (w) can be written

$$w = \sum_{k=1}^{N} \alpha_k x_k^T x + b \quad \text{with } \alpha_k = 2\gamma e_k \tag{5}$$

By substituting eq 4 into the original regression line ($y = wx + b$), the following result can be obtained:

$$y = \sum_{k=1}^{N} \alpha_k x_k^T x + b \tag{6}$$

It can be seen that the Lagrange multipliers can be defined as

$$\alpha_k = (x_k^T x + (2\gamma)^{-1}(y_k - b) \tag{7}$$

Finding these Lagrange multipliers is very simple as opposed to the SVM approach, in which a more difficult relation has to be solved to obtain these values. In addition, it easily allows for a nonlinear regression as an extension of the linear approach by introducing the kernel function. This leads to the following nonlinear regression function:

$$f(x) = \sum_{k=1}^{N} \alpha_k K(x, x_k) + b \tag{8}$$

where $K(x, x_k)$ is the kernel function. The value is equal to the inner product of two vectors x and x_k in the feature space $\Phi(x)$ and $\Phi(x_k)$; that is, $K(x, x_k) = \Phi(x)^T \Phi(x_k)$. The choices of a kernel and its specific parameters together with γ have to be tuned by the user. The radial basis function (RBF) kernel $K(x, x_k) = \exp(-|x_k - x|^2/\sigma^2)$ is commonly used, and then leave-one-out (LOO) cross-validation was used to tune the optimized values of the two parameters γ and σ.

All computations implementing LS-SVM were performed using the Matlab/C toolbox (*29*).

PPR. PPR developed by Friedman and Stuetzle (*30*) is a powerful tool for seeking the interesting projections from high-dimensional data into lower dimensional space by means of linear projections. Therefore, it can overcome the curse of dimensionality because it relies on estimation in at most trivariate settings. At present, it has been successfully applied to tackle some chemical problems (*31, 32*). Friedman and Stuetzle's concept of PPR avoided many difficulties experienced with other existing nonparametric regression procedures. It does not split the predictor space into two regions, thereby allowing, when necessary, more complex models. In addition, interactions of predictor variables are directly considered because linear combinations of the predictors are modeled with general smooth functions. Another significant property of PPR is that the results of each interaction can be depicted graphically. The graphical output can be used to modify the major parameters of the procedure: the average smoother bandwidth and the terminal threshold. The basic theory of PPR can be found in refs *16, 33, and 34*. Here, we only give a brief description. Given the ($k \times n$) data matrix X, where k is the number of observed variables and n is the number of units, and an m-dimensional orthonormal matrix A ($m \times k$), the ($m \times n$) matrix $Y = AX$ represents the coordinates of the projected data onto the m-dimensional ($m < k$) space spanned by the rows of A. Because such projections are infinite, it is important to have a technique to pursue a finite sequence of projections that can reveal the most interesting structures of the data. Projection pursuit (PP) is such s powerful tool that combines both ideas of projection and pursuit (*13, 32*). In a typical regression problem, PPR aims to approximate the regression pursuit function $f(x)$ by a finite sum of ridge functions with suitable choices of α_i and g_i

$$g^{(p)}(x) = \sum_{i=1}^{p} g_i(\alpha_i^T x) \tag{9}$$

where α_i values are $m \times n$ orthonormal matrices and p is the number of ridge functions.

All calculation programs implementing PPR were written in R-file under R2.3.1 environment (*35*) running operating system on a Pentium IV with 512 M RAM.

Evaluation of Regression Models. Once the models are generated, it is important to evaluate the availability of them. In this study, the root-mean-square error (RMSE) is used to assess the predictive ability and accuracy of the models and the relative standard error (RSE) is used to estimate the relative error of the predictors. The representations of the two methods are defined below

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n_s} (y_{ke} - y_{kp})^2}{n_s}} \tag{10}$$

$$\text{RSE} = \sqrt{\frac{\sum (y_{ke} - y_{kp})^2}{\sum y_{kp}^2}} \times 100\% \tag{11}$$

where k represents the kth molecule, y_{ke} is the desired output (experimental property), y_{kp} is the actual output of the models, and n_s is the number of compounds in the analyzed set.

RESULTS AND DISCUSSION

Results of the GA-MLR Method. A variety of subset sizes of descriptors were investigated to determine the optimum number of descriptors in the regression model. If adding another descriptor did not significantly improve the statistics of the model, it was determined that the optimum subset had been achieved. The influences of the number of the descriptors on the coefficients of determination (R^2) and RMSE to the training and test sets are shown in parts **a** and **b** of **Figure 2**, respectively. The higher the values of R^2 for the training and test sets and the lower the RMSE, the better the results. From **Figure 2**, it was clear to conclude that the five descriptors were the best selection. The involved descriptors and the statistical parameters of this model are summarized in **Table 2**, and the correlation matrix of these selected descriptors is shown in **Table 3**. The statistical results of the MLR model for the training and test sets are listed in **Table 4**, and the predicted fungicidal activities are listed in **Table 1**. **Figure 3** shows the predicted log A versus the experimental values for all of the compounds in the training and test sets.

The developed QSAR models should not only offer a reliable prediction capability but also gain insight into the factors that are likely to influence the fungicidal activities of thiazoline derivatives by interpreting the meaning of the selected descriptors. The five selected molecular descriptors were divided into two kinds: one is a quantum chemical descriptor (MaxRHN), and the other class is conformational (3D) descriptors, including one geometry, topology, and atom-weights assembly (GETAWAY) descriptor (R1v+), one 3D molecule representation of structures based on electron diffraction (MoRSE) descriptor (Mor02m), one radial distribution functions (RDF) descriptor (RDF065u), and one weighted holistic invariant molecular (WHIM) descriptor (L2s).

According to the t-test values of the selected descriptors, the most important descriptor is L2s. It is the second-component size directional WHIM index/weighted by atomic electrotopological states. It belongs to WHIM descriptors, which are 3D molecular indices that represent different sources of chemical information. The WHIM descriptors contain information about

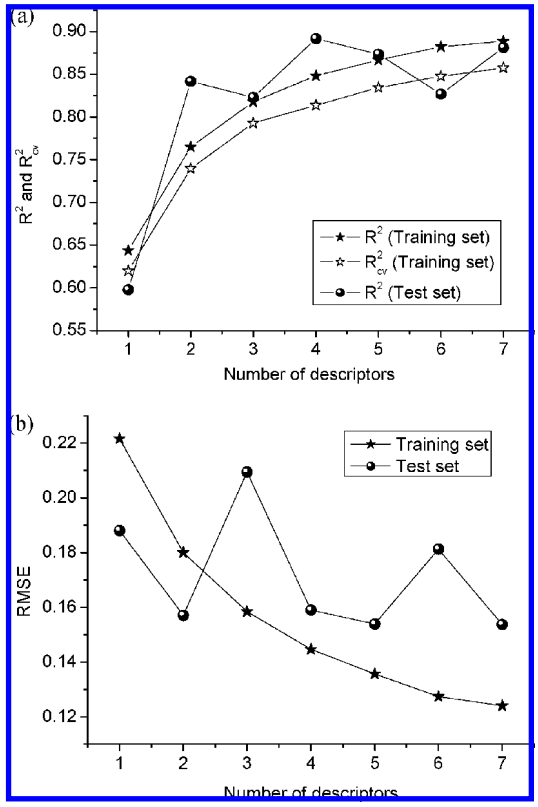


Figure 2. Procedure of the selection of descriptors. (a) R^2 for the training set and R_{cv}^2 for the training set versus the number of the descriptors. (b) RMSE for the training and test sets versus the number of descriptors.

Table 3. Correlation Matrix of the Selected Molecular Descriptors

	MaxRHN	RDF065u	Mor02m	L2s	R1v+
MaxRHN	1.0000				
RDF065u	−0.1778	1.0000			
Mor02m	0.2575	−0.2815	1.0000		
L2s	0.0333	−0.0864	0.6338	1.0000	
R1v+	−0.0372	−0.4773	−0.0157	0.1040	1.0000

Table 4. Comparison of R^2, RMSE, and RSE for Different QSAR Models

methods	data set	R^2	RMSE	RSE
GA-MLR	trainging set	0.8666	0.1356	7.57
	test set	0.8736	0.1538	8.56
	whole set	0.8505	0.1395	7.78
LS-SVM	trainging set	0.9412	0.0903	3.33
	test set	0.9392	0.1496	8.38
	whole set	0.9173	0.1049	5.85
PPR	trainging set	0.9576	0.0768	2.75
	test set	0.9431	0.1268	7.03
	whole set	0.9395	0.0890	4.96

the whole 3D molecular structure in terms of size, shape, symmetry, and atom distribution. They are calculated from (x, y, and z) coordinates of a molecule within different weighting schemes in a straightforward manner and represent a very general approach to describe molecules in a unitary conceptual framework. This descriptor L2S is calculated from the electro-topological weights on the hydrogen-depleted structures and could be used to analyze the shape and symmetry of the chemical structures. It can be used to distinguish different conformations of the same molecule and especially for different geometric isomers (*36*).

The other one important descriptor is MaxRHN (*37*). It is a quantum chemical descriptor. This descriptor describes the

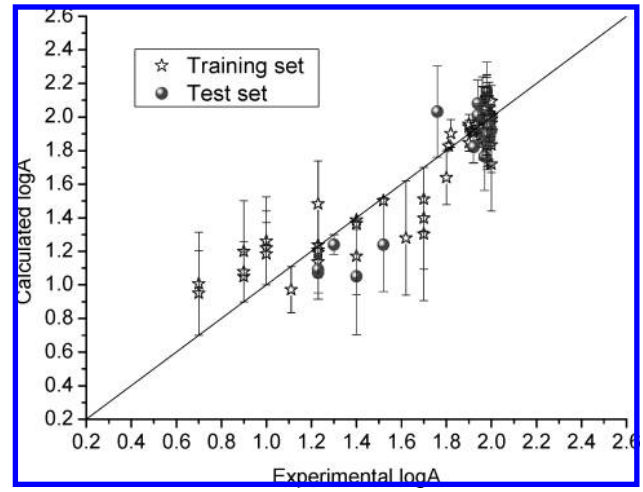


Figure 3. Plot of experimental log A values versus predicted values $\pm$ S (note: $S = |\text{Exp.} - \text{Pred.}|$; Exp., the experimental values from ref *19*; Pred., the predicted values by GA-MLR).

nuclear repulsion energy between two given atomic species (atoms A and B) in the molecule and was calculated as follows:

$$E_{nn}(AB) = Z_A Z_B / R_{AB} \quad (12)$$

where Z_A and Z_B are the nuclear (core) charges of atoms A and B, respectively, and R_{AB} is the distance between them. This energy describes the nuclear repulsion driven processes in the molecule and may be related to the conformational (rotational and inversion) changes or atomic reactivity in the molecule.

The 3D MoRSE (*38*) descriptor, Mor02m (*39*), is the meaning of signal 2/weighted by atomic masses atomic mass weighted, and it contributes negatively to the fungicidal activity. The MoRSE descriptor calculated by summing atomic weight is viewed by different angular scattering functions. The values of these functions are calculated at 32 evenly distributed values of scattering angles in the range of 0−32 $Å^{-1}$ from the 3D atomic coordinates of a molecule. It is calculated using the following function:

$$\text{Mor}(s, w) = I(s, w) = \sum_{i=2}^{n} \sum_{j=1}^{i-1} w_i w_j \frac{\sin(sr_{ij})}{sr_{ij}} \quad (13)$$

where s is the scattering angle, r_{ij} is the Euclidean distance between the atoms i and j, and w_i and w_j are the weights of the atoms i and j, respectively. The notation m in the descriptor Mor02m represents the after digital value used for atomic weights, which was contributed especially through atomic masses.

The RDF descriptor (3D), RDF065u, is the meaning of radial distribution function 6.5/unweighted. The RDF descriptors are based on the distance distribution in the molecule. The radical distribution function of an ensemble of n atoms can be interpreted as the probability distribution of finding an atom in a spherical volume of radius R. A typical RDF descriptor is denoted by RDF_{sw}, where s takes the values $10 \le s \le 155$ in units of 5 and$w \in \{u, m, v, e, p\}$, and is defined as the following expression:

$$\text{RDF}(R, w) = f \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_i w_j e^{-\beta(R - R_{ij})^2} \quad (14)$$

where f is a scaling factor, r_{ij} is the Euclidean distance between the atoms i and j, w_i and w_j are the weights of the atoms i and j, respectively, and β is the smoothing parameter, which defines

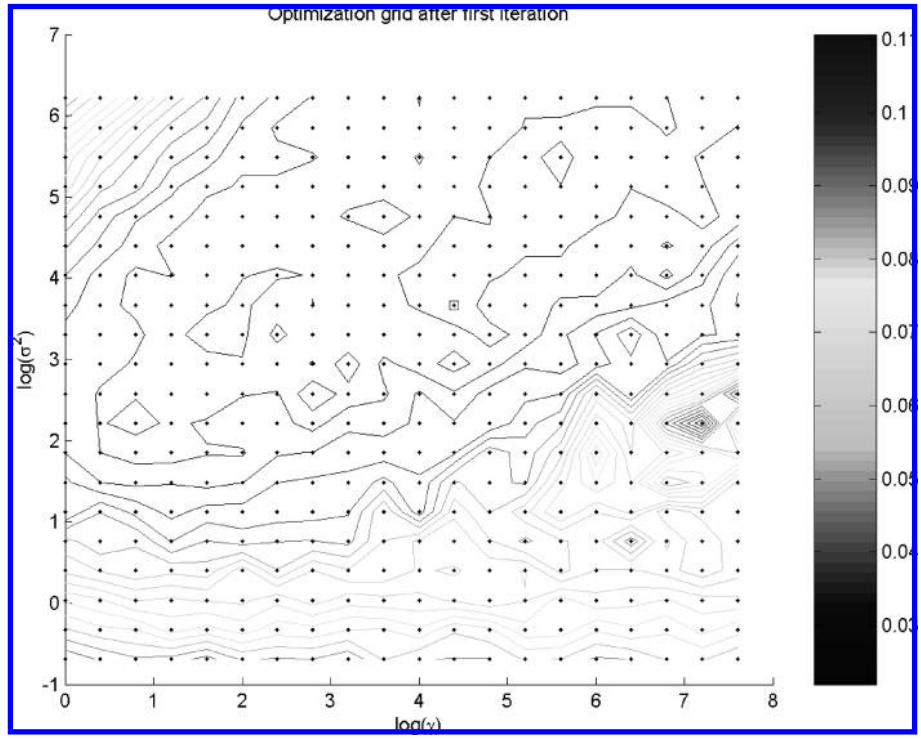


Figure 4. Contour plot of the optimization error for LS-SVMs when optimizing the parameters σ and γ in the regression problem. Note that the small square indicates the selected optimal settings.

the probability distribution of the individual interatomic distance. β can be interpreted as the temperature factor that defines the movement of the atoms. The negative coefficient of the descriptor indicates that it was also negative to the fungicidal activity.

The last descriptor is a GETAWAY descriptor, R1v+. It is the abbreviation of R maximal autocorrelation of lag 1/weighted by atomic van der Waals volumes. The GETAWAY descriptors are chemical structure descriptors encoding the 3D information of the molecule derived from a new representation of molecular structure, the molecular influence matrix (MIM), which was denoted by **H** and defined as the following:

$$\mathbf{H} = \mathbf{M}(\mathbf{M}^{\mathbf{T}}\mathbf{M})^{-1}\mathbf{M}^{\mathbf{T}} \qquad (15)$$

where **M** is the molecular matrix constituted by the centered Cartesian coordinates x, y, and z of the molecule atoms (including hydrogens) in a chosen conformation and the superscript **T** refers to the transposed matrix. Thus, it has a strong relationship with the shape and size of the molecule. The descriptor R1v+ also relates to the geometry of the molecule.

From the above discussion, it can be seen that all of the descriptors involved in the model have physical and chemical meanings. They can also account for the structural features responsible for the fungicidal activity of thiazoline derivatives. On the basis of the coefficients and the values of the t test, it could be easier to see the kind of thiazoline with smaller values of RDF065U, Mor02m, L2s, and R1v+ and a bigger value of MaxRHN will be more effective against *M. grisea*. At last, it can be concluded that the fungicidal activity of the thiazoline derivatives depends upon the 3D and conformational structures of the molecule.

Results of the LS-SVM Method. To construct a more accurate model, the LS-SVM method was also performed to build a nonlinear prediction model with the same features after the GA-MLR model was generated. In this study, RBF kernel was used as the kernel function. Thus, γ (the relative weight of the regression error) and σ (the kernel parameter of the RBF kernel) need to be optimized. Here, the optimal parameters are found by an intensive grid search method. The result of this grid search is an error-surface spanned by the model parameters. A robust model is obtained by selecting those parameters that give the lowest error in a smooth area. To find the optimized combination of the parameters γ and σ, a process of 10-fold cross-validation of the whole training set was performed.

The parameter (σ) of the RBF kernel in the style of σ^2 and the parameter γ were tuned simultaneously in a grid 20 × 20 ranging from 1 to 2000 and from 0.5 to 500, respectively. In this way, parameter optimization was performed in different orders of magnitude. Because the grid search has been performed over just two parameters, a contour plot of the optimization error can be visualized easily (**Figure 4**). This is an advantage of LS-SVM compared to the traditional SVM, in which three parameters have to be optimized. From **Figure 4**, the optimal parameter settings can be selected from a smooth subarea with a low prediction error. The selected optimal values of γ and σ^2 are 81.4934 and 39.238 respectively, marked by the small square in the figure. The cost value of the 10-fold cross-validation is 0.017 271.

The prediction results of the optimal LS-SVM model are shown in **Table 1** and **Figures 5** and **6**. The statistical parameters of the LS-SVM are listed in **Table 4**.

Results of the PPR Method. To compare the results of different nonlinear chemometrics methods, the PPR method was applied to build the other nonlinear model with the same five selected descriptors. In the PPR approach, there are several parameters needed to be adjusted. The parameters "nterms" and "max.terms" represent the number of terms to include in the final model and the number of maximum terms to choose from when building the model, respectively. The "df" defines the smoothness of each ridge term by the requested equivalent degrees of freedom, if "sm.method" is "spline". The levels of optimization (argument "optlevel") differ in how thoroughly the

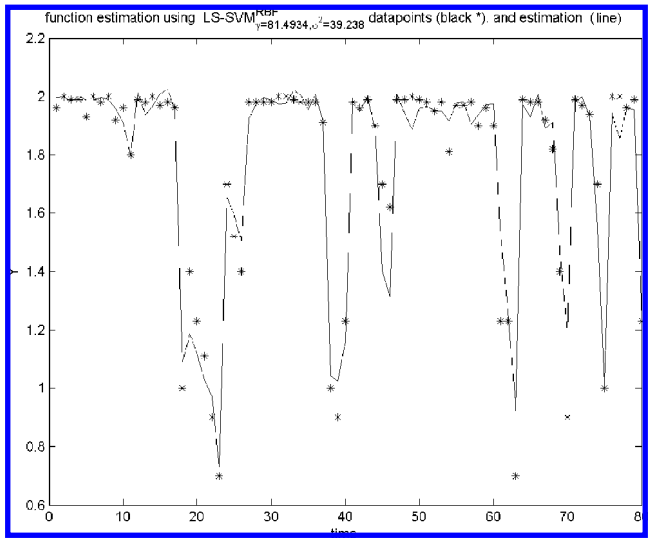


Figure 5. Illustration of fitted results of the LS-SVM method based on the RBF kernel function with training data. Note that the solid line indicates the estimated outputs, the dotted line represents the true underlying function, and the stars indicate the training data points.

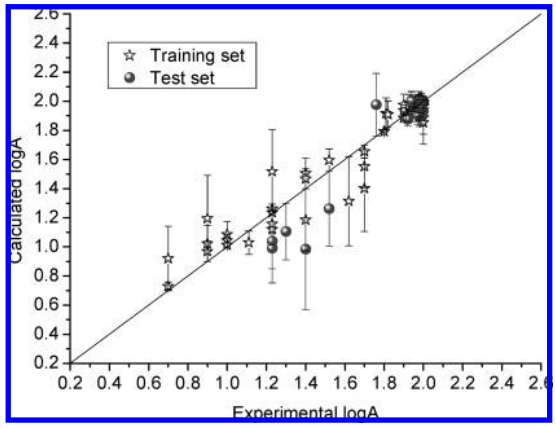


Figure 6. Plot of experimental log *A* values versus predicted values ± *S* (Note: *S* = |Exp. − Pred.|; Exp., the experimental values from ref *19*; Pred., the predicted values by LS-SVM).

models are refitted during this process. At level 0, the existing ridge terms are not refitted. At level 1, the projection directions are not refitted but the ridge functions and the regression coefficients are refitted. Levels 2 and 3 refit all of the terms and are equivalent for one response; level 3 is more careful to rebalance the contributions from each regressor at each step and therefore is a little less likely to converge to a saddle point of the sum of squares criterion. In this investigation, the four parameters "nterms", "max.terms", "optlevel", and "df" were determined as 3, 15, 3, and 5, respectively. The predicted results and the statistical parameters of the optimal PPR model were shown in **Tables 1** and **4**, respectively. The scatter plot was given in **Figure 7**. From **Figure 7** and **Table 4**, it can be seen that the predicted values were in agreement with the experimental log *A* values for almost all of the compounds.

Comparison of the Results Obtained by Different Approaches. To check the superiority of three different methods (GA-MLR, LS-SVM, and PPR), the predicted accuracy for different data sets (training set, test set, and whole set) were collected together and shown in **Table 4**. As seen from this table, the nonlinear regression methods LS-SVM and PPR show better predictive capability and the corresponding prediction results were in better agreement with the experimental values.

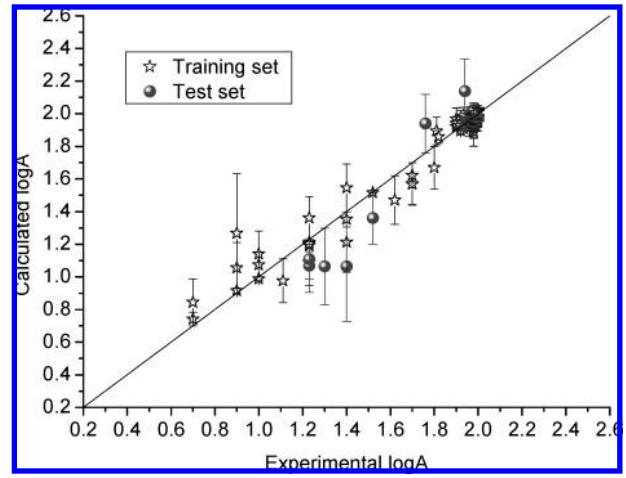


Figure 7. Plot of experimental log *A* values versus predicted values ± *S* (Note: *S* = |Exp. − Pred.|; Exp., the experimental values from ref *19*; Pred., the predicted values by PPR).

In summary, the three machine learning methods GA-MLR, LS-SVM, and PPR were used to develop the linear and nonlinear QSAR models for predicting the fungicidal activities of thiazoline derivatives against rice blast inhibitors. The obtained models clearly demonstrate that there are strong correlations between the structural information and fungicidal activities of these compounds. The prediction results indicate that the LS-SVM and PPR methods are powerful and promising tools for QSAR analysis. The models developed in this study identify and provide insight into the structural features related to the biological activity of these compounds. Furthermore, this study provides instruction for further design of thiazoline derivatives with higher inhibitory activity for the protection of rice blast disease.

ABBREVIATIONS USED

GA-MLR, genetic algorithm-multilinear regression; LS-SVM, least-squares support vector machine; PPR, project pursuit regression; QSAR, quantitative structure−activity relationship; MLR, multiple linear regression; PLS, partial least-squares; ANN, artificial neural network; PCA, principle component analysis; RMSE, root-mean-square error; RSE, relative standard error; MaxRHN, maximum n−n repulsion for a H−N bond; GETAWAY, geometry, topology, and atom-weights assembly; MoRSE, molecule representation of structures based on electron diffraction; RDF, radial distribution function; WHIM, weighted holistic invariant molecular.

ACKNOWLEDGMENT

The authors thank the R Development Core Team for affording the free R2.3.1 software and also express their gratitude to Jeanette Bradley (Yale Medical School) for proofreading, with special thanks to the anonymous reviewers and the editor for their professional and intensive comments.

LITERATURE CITED

(1) Jordan, D. B.; Basarab, G. S.; Liao, D.-I.; Johnson, W. M. P.; Winzenberg, K. N.; Winkler, D. A. Structure-based design of inhibitors of the rice blast fungal enzyme trihydroxynaphthalene reductase. *J. Mol. Graphics Modell.* **2001**, *19*, 434–447.

(2) Yamaguchi, I.; Kubo, Y. Target sites of melanin biosynthesis inhibitors. In *Target Sites of Fungicide Action*; CRC Press: London, U.K., 1992; pp 101−118.

(3) Talbot, N. J. Having a blast: Exploring the pathogenicity of *Magnaporthe grisea*. *Trends Microbiol.* **1995**, *3*, 9–16.
(4) Bell, A. A.; Wheeler, M. H. Biosynthesis and functions of fungal melanins. *Annu. Rev. Phytopathol.* **1986**, *24*, 411–415.
(5) Nakasako, M.; Motoyama, T.; Kurahashi, Y.; Yamaguchi, I. Cryogenic X-ray crystal structure analysis for the complex of scytalone dehydratase of a rice blast fungus and its tight-binding inhibitor, carpropamid: The structural basis of tight-binding inhibition. *Biochemistry* **1998**, *37*, 9931–9939.
(6) Camargo, A. B.; Marchevsky, E.; Luco, J. M. QSAR study for the soybean 15-lipoxygenase inhibitory activity of organosulfur compounds derived from the essential oil of garlic. *J. Agric. Food Chem.* **2007**, *55* (8), 3096–3103.
(7) Jalali-Heravi, M.; Asadollahi-Baboli, M.; Shahbazikhah, P. QSAR study of heparanase inhibitors activity using artificial neural networks and Levenberg−Marquardt algorithm. *Eur. J. Med. Chem.* **2008**, *43*, 548–556.
(8) Jalali-Heravi, M.; Kyani, A. Comparison of shuffling-adaptive neuro fuzzy inference system (shuffling-ANFIS) with conventional ANFIS as feature selection methods for nonlinear systems. *QSAR Comb. Sci.* **2007**, *26*, 1046–1059.
(9) Pripp, A. H. Quantitative structure−activity relationship of prolyl oligopeptidase inhibitory peptides derived from β-casein using simple amino acid descriptors. *J. Agric. Food Chem.* **2006**, *54* (1), 224–228.
(10) Hansch, C.; Kurup, A.; Garg, R.; Gao, H. Chem-bioinformatics and QSAR: A review of QSAR lacking positive hydrophobic terms. *Chem. Rev.* **2001**, *101*, 619–672.
(11) Draper, N. R.; Smith, H. *Applied Regression Analysis*, 3rd ed.; John Willy and Sons: New York, 1998.
(12) Lindberg, W.; Persson, J. A.; Wold, S. Partial least-squares method for spectrofluorimetric analysis of mixtures of humic acid and ligninsulfonate. *Anal. Chem.* **1983**, *55*, 643–648.
(13) Hemmateenejad, B.; Safarpour, M. A.; Taghavi, F.; Jamali, M. Application of ab initio theory for the prediction of acidity constant of 1-hydroxy-9,10-anthraquinone derivatives using neural network modeling method. *J. Mol. Struct.* **2003**, *635*, 183–190.
(14) Suykens, J. A. K.; Vandewalle, J. Least squares support vector machine classifiers. *Neural Process. Lett.* **1999**, *9*, 293–300.
(15) Friedman, J. H. Exploratory projection pursuit. *J. Am. Stat. Assoc.* **1987**, *82*, 249–266.
(16) Huber, P. J. Projection pursuit (with discussion). *Ann. Stat.* **1985**, *13*, 435–525.
(17) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *CODESSA: Reference Manual*; University of Florida: Gainesville, FL, 1994.
(18) Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. DRAGON−Software for the calculation of molecular descriptors, version 5.4 for Windows, Talete srl, Milan, Italy, 2006.
(19) Song, J. S.; Moon, T.; Nam, K. D.; Lee, J. K.; Hahn, H.-G.; Choi, E.-J.; Yoon, C. N. Quantitative structural−activity relationship (QSAR) study for fungicidal activities of thiazoline derivatives against rice blast. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 2133–2142.
(20) ISIS Draw2.3 (1990−2000) MDLInformation Systems, Inc.
(21) HyperChem, Release 6.0 for Windows, Hypercube, Inc., 2000.
(22) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. CODESSA, version 2.0. Reference Manual, 1995−1997.
(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) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.
(25) Lucasius, C. B.; Kateman, G. Genetic algorithms for large-scale optimization in chemometrics: An application. *Trends Anal. Chem.* **1991**, *10*, 254–261.
(26) Hou, T. J.; Wang, J. M.; Liao, N.; Xu, X. J. Applications of genetic algorithms on the structure−activity relationship analysis of some cinnamamides. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 775–781.
(27) Suykens, J. A. K.; Vandewalle, J. Least squares support vector machine classifiers. *Neural Process Lett.* **1999**, *9*, 293–300.
(28) Suykens, J. A. K.; van Gestel, T.; de Brabanter, J.; de Moor, B.; Vandewalle, J. *Least-Squares Support Vector Machines*; World Scientific Publishing Company: Singapore, 2002.
(29) Pelckmans, K.; Suykens, J. A. K.; Van Gestel, T.; De Brabanter, D.; Lukas, L.; Hamers, B.; De Moor, B.; Vandewalle, J. LS-SVMlab: A Matlab/C toolbox for least squares support vector machines. Internal Report 02-44, ESATSISTA, K. U. Leuven, Leuven, 2002.
(30) Friedman, J. H.; Stuetzle, W. Projection pursuit regression. *J. Am. Stat. Assoc.* **1981**, *76*, 817–823.
(31) Du, Y. P.; Liang, Y. Z.; Yun, D. Data mining for seeking an accurate quantitative relationship between molecular structure and GC retention indices of alkenes by projection pursuit. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1283–1292.
(32) Du, H. Y.; Wang, J.; Zhang, X. Y.; Yao, X. J.; Hu, Z. D. Prediction of retention times of peptides in RPLC by using radial basis function neural networks and projection pursuit regression. *Chemom. Intell. Lab. Syst.* **2008**, *92*, 92–99.
(33) Donoho, D.; Johnstone, I. M. Discussion on projection pursuit. *Ann. Stat.* **1985**, *13*, 496–500.
(34) Diaconis, P.; Shahshahani, M. On nonlinear functions of linear combinations. *SIAM J. Sci. Stat. Comput.* **1984**, *5*, 175–191.
(35) Birattari, M.; Bontempi, G. R manuals, The R Development Core Team, 2003.
(36) Todeschini, R.; Gramatica, P. SD-modeling and prediction by WHIM descriptors. Part 5. Theory development and chemical meaning of WHIM descriptors. *QSAR Comb. Sci.* **1997**, *16*, 113–119.
(37) Strouf, O. *Chemical Pattern Recognition*; Research Studies Press: Hertfordshire, U.K., 1986.
(38) Schuur, J. H.; Selzer, P.; Gasteiger, J. The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure−spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334–344.
(39) Gupta, R. A.; Gupta, A. K.; Soni, L. K.; Kaskhedikar, S. G. Exploration of physicochemical properties and molecular modeling studies of furanylamide analogs as antituberculosis agents. *QSAR Comb. Sci.* **2007**, *26*, 897–907.

Received for review July 19, 2008. Revised manuscript received September 22, 2008. Accepted September 23, 2008.

JF8022194