Quantitative Structure-Retention Relationship Models for the Prediction of the Reversed-Phase HPLC Gradient Retention Based on the Heuristic Method and Support Vector Machine

Hongying Du, Jie Wang, Xiaojun Yao, and **Zhide Hu***
Department of Chemistry, Lanzhou University, Lanzhou 730000, China

Abstract

The heuristic method (HM) and support vector machine (SVM) were used to construct quantitative structure–retention relationship models by a series of compounds to predict the gradient retention times of reversed-phase high-performance liquid chromatography (HPLC) in three different columns. The aims of this investigation were to predict the retention times of multifarious compounds, to find the main properties of the three columns, and to indicate the theory of separation procedures. In our method, we correlated the retention times of many diverse structural analytes in three columns (Symmetry C18, Chromolith, and SG-MIX) with their representative molecular descriptors, calculated from the molecular structures alone. HM was used to select the most important molecular descriptors and build linear regression models. Furthermore, non-linear regression models were built using the SVM method; the performance of the SVM models were better than that of the HM models, and the prediction results were in good agreement with the experimental values. This paper could give some insights into the factors that were likely to govern the gradient retention process of the three investigated HPLC columns, which could theoretically supervise the practical experiment.

Introduction

Chromatographic techniques have been widely applied to many drug separations and analysis research. Reversed-phase liquid chromatography (RPLC) is one of the most popular chromatographic techniques in the separation science. Extensive studies have been carried out over several decades to improve the understanding of the solute retention mechanism, but much remains to be illustrated. There have been several approaches used to interpret the solute retention mechanism, including classical thermodynamics, kinetics of molecular interaction (1), and quantitative structure–retention relationships (QSRR) (2–3).

QSRR studies quantify the relationships between the chromatographic behavior determined by a representative series of analytes in given separation systems and the molecular structural parameters (i.e., physico-chemical properties or molecular descriptors) accounting for the structural differences among the various analytes (4). In the last two decades of the twentieth century, QSRR has often been applied to the following aims (5): (*i*) To predict retention for a new solute; (*ii*) To identify the most important structural descriptors relevant to the retention behavior of a solute; (*iii*) To gain insight into the molecular mechanism of separation operating in a given chromatographic system; (*iv*) To evaluate properties of stationary phases.

The QSRR approach provides a promising method for the estimation of retention behavior of solutes based on the descriptors derived from the molecular structures alone to fit experimental data. The advantages of this approach lie in the fact that it mainly requires the information of chemical structures and few experimental data; therefore, it saves much time and money. The main steps of this method include data collection, molecular descriptor generation and selection, model development, and finally, model evaluation.

In this investigation, the software CODESSA, developed by Katritzky group, was used to calculate a large number of descriptors. These include constitutional descriptors, topological descriptors, electrostatic descriptors, and quantum chemistry descriptors. The CODESSA combines diverse methods for quantifying the structural information about the molecule with advanced statistical analyses to establish molecular structure-property/activity/retention relationships. The heuristic method (HM) implemented in the framework of the CODESSA program was used to perform a complete search for the best multi-linear correlations with a multitude of descriptors. So far, this software has been successfully applied to many similar analyses (6–8).

In the early studies, most of the models were built based on multi-linear regression (9–13). In recent years, artificial neural network (14–15) has become a very popular and powerful chemometrics tool to solve some chemical problems, including optimization of chromatographic analysis (16–19). Even though the artificial neural network method can offer high accuracy in most cases, it can also suffer from over-fitting the data and poor

* Author to whom correspondence should be addressed: email huzd@lzu.edu.cn.

Table I. The Compounds and Predicted Results of the Retention Time (min)

No.	Model analytes	Chromolith			Symmetry C18			SG-MIX		
		t_{Rexp}	SVM t_{Rcalc}	HM t_{Rcalc}	t_{Rcalc}	SVM t_{Rcalc}	HM t_{Rcalc}	t_{Rexp}	SVM t_{Rexp}	HM t_{Rexp}
1	Benzamide	4.75	5.43	5.70	8.12	8.18	9.91			
2*	4-Cyanophenol	6.00	5.42	6.06	9.78	8.71	9.44	11.60	9.17	9.82
3	Indazole	7.55	5.96	7.07	11.30	9.69	10.00	11.55	12.49	12.34
4	Benzonitrile	7.28	7.25	7.42	11.23	11.17	11.93	10.92	11.06	11.61
5	Indole	8.24	8.34	8.37	12.23	12.17	12.43	12.75	13.38	13.24
6	2-Naphthol	9.33	9.55	9.44	13.12	13.18	12.80	13.72	13.00	12.81
7	Anisole	9.12	9.20	9.22	13.37	13.43	12.70	12.57	12.17	12.83
8	Benzene	9.12	9.20	9.52	13.57	13.63	14.31	12.12	12.21	13.10
9*	1-Naphthylacetonitrile	9.65	9.69	8.84	13.47	13.97	13.42	14.08	13.80	13.43
10	Benzyl chloride	10.08	10.75	10.49	14.23	15.94	15.37	13.88	14.00	14.10
11	Naphthalene	11.31	11.34	11.18	15.57	15.51	15.40	15.13	15.13	14.94
12	Biphenyl	12.08	12.00	11.92	16.35	16.41	16.19	15.88	15.46	15.24
13*	Phenanthrene	12.61	12.77	13.00	17.25	17.51	18.15	16.58	16.81	16.94
14	Pyrene	13.39	13.31	13.88	18.87	18.81	19.33	17.23	17.25	18.35
15	2,29-Dinaphthyl ether	14.00	13.92	14.69	19.62	19.68	20.51	17.80	17.78	17.88
16	Toluene	10.51	10.43	10.45	14.82	14.76	14.45	13.78	13.80	14.07
17	Ethylbenzene	11.33	11.40	11.18	15.63	15.59	15.64	14.78	14.91	14.69
18	*n*-Propylbenzene	12.08	12.16	11.90	16.45	16.39	16.02	15.62	15.42	15.01
19	*n*-Butylbenzene	12.69	12.77	12.59	17.30	17.27	16.86	16.27	16.25	15.60
20	*n*-Amylbenzene	13.23	13.20	13.20	18.28	18.28	17.62	16.78	16.73	16.06
21	*n*-Hexylbenzene	13.81	13.62	13.89	19.48	19.42	18.37	17.23	17.20	16.59
22	Cumene	11.89	11.97	11.69	16.18	16.28	16.26	15.42	15.60	15.14
23	2-Ethyltoluene	11.97	11.69	11.34	16.33	16.47	16.42	15.58	15.49	15.12
24	1,2,3-Trimethylbenzene	12.03	11.78	11.57	16.35	16.19	17.11	15.52	16.81	16.05
25*	1,3,5-Trimethylbenzene	12.27	11.86	11.62	16.77	16.17	15.59	15.73	15.39	15.00
26	Anthracene	12.72	13.11	13.45	17.55	17.61	17.02	16.63	17.26	17.31
27	1-Methylnaphthalene	12.03	11.95	11.71	16.40	16.34	16.01	15.78	15.96	15.66
28	1-Bromonaphthalene	12.35	12.14	11.77	16.87	16.91	17.50	16.25	16.23	16.03
29	*o*-Xylene	11.25	10.76	10.51	15.63	15.43	15.76	14.75	14.77	14.60
30	*m*-Xylene	11.41	10.75	10.53	15.78	15.45	15.04	14.88	13.97	14.08
31*	*p*-Xylene	11.41	11.23	11.05	15.22	15.27	14.97	15.18	14.88	14.68
32	3-Cyanobenzoic acid	6.56	6.13	6.03	10.28	9.54	10.13	10.93	10.39	10.14
33*	3-Fluorobenzoic acid	8.21	7.51	7.47	12.05	11.55	11.38	11.85	11.06	11.05
34	*o*-Toluic acid	8.69	8.01	7.73	12.45	12.39	12.48	12.07	12.05	11.41
35	*p*-Toluic acid	8.88	8.80	8.49	12.58	12.37	12.28	12.35	12.30	12.13
36	4-Ethylbenzoic acid	9.87	9.95	9.27	13.62	13.68	13.28	13.42	13.39	12.91
37	3-Hydroxybenzoic acid	5.47	5.55	5.60	9.08	9.14	9.14	8.77	8.75	8.63
38	4-Hydroxybenzoic acid	4.59	5.92	6.52	7.97	9.02	9.42	8.02	8.37	9.64
39	Benzoic acid	7.55	7.24	7.37	11.32	11.38	11.70	10.80	10.82	10.81
40	1-Naphthylacetic acid	9.52	12.29	11.96	13.20	13.26	14.75	13.38	15.40	15.15
41	Acetylsalicylic acid	6.99	7.07	7.20	10.57	10.63	10.07	10.45	10.47	10.63
42	Naproxen	10.35	10.27	9.79	14.07	14.01	15.09	14.17	14.15	14.10
43	Fenbufen	10.51	10.59	9.87	14.17	14.23	13.75	14.37	14.39	13.92
44	Diclofenac	11.47	11.39	11.80	15.32	15.38	15.55	15.55	16.15	16.56
45	2-Chloroaniline	7.36	7.28	7.52	11.92	11.49	12.11	11.50	12.09	11.84
46	2-Methoxyaniline	7.49	7.49	7.62	11.65	11.59	10.49	12.08	11.77	11.74
47*	3,4-Dichloroaniline	9.04	8.43	7.92	13.18	13.18	13.54	13.70	13.12	12.87
48	3,5-Dichloroaniline	9.79	8.57	8.15	13.92	13.13	12.60	14.18	12.28	12.23
49	3,5-Dimethylaniline	8.11	8.19	8.34	12.40	12.50	12.26	12.02	12.36	12.25
50	3-Chloroaniline	7.33	7.41	7.43	11.73	11.79	11.99	11.67	11.49	11.61
51	3-Methylaniline	6.29	7.04	7.54	10.88	11.39	11.69	10.07	11.51	11.45
52*	4-Chloroaniline	7.17	7.46	7.38	11.65	11.82	12.06	11.58	11.55	11.81
53*	*n*-Ethylaniline	8.48	8.43	8.26	12.98	12.93	11.76	12.23	12.09	12.08

* Compounds in the test set.

Table I. (continued) The Compounds and Predicted Results of the Retention Time (min)

No.	Model analytes	Chromolith			Symmetry C18			SG-MIX		
			SVM	HM		SVM	HM		SVM	HM
		t_{Rexp}	t_{Rcalc}	t_{Rcalc}	t_{Rcalc}	t_{Rcalc}	t_{Rcalc}	t_{Rexp}	t_{Rexp}	t_{Rexp}
54*	Coumarin	6.56	7.77	8.37	10.88	10.08	12.95	11.40	11.24	12.28
55	Phthalimide	4.91	4.90	5.26	9.53	9.47	9.03	9.65	9.67	10.14
56	Phthalonitrile	5.55	5.47	5.33	9.65	9.71	10.41	10.30	10.28	10.24
57	1,4-Naphthoquinone	7.57	7.65	7.53	12.03	12.09	11.38	12.08	12.06	11.36
58	Phenylacetylene	9.49	10.26	10.19	13.80	14.19	13.72	13.45	13.37	13.75
59*	Carbazole	10.40	10.83	9.90	14.53	14.54	14.64	15.10	15.78	14.79
60	9,10-Anthraquinone	10.93	10.85	9.90	15.17	15.11	14.41	15.17	15.15	14.46
61	Xanthene	12.43	12.51	12.13	17.77	17.71	17.07	16.25	16.27	15.97
62*	Hexachlorobutadiene	12.93	12.79	12.74	17.53	16.82	17.38	16.48	17.47	17.30

* Compounds in the test set.

reproducibility of results. This is due to random initialization of the network, variation of stopping the criteria, and lack of information regarding the classification produced (20). As mentioned previously, there is still a need to apply more accurate and informative techniques to QSRR analytes.

As a new and powerful modeling tool, the support vector machine (SVM) has recently gained much interest in pattern recognition and function approximation applications. This nonlinear method has proven to be very effective for addressing the general purpose of classification and regression problems (21–26). In most of these cases, the performance of SVM modeling is significantly better than the traditional machine learning approaches, including artificial neural networks. Compared with traditional regression and neural network methods, SVM has some advantages, including global optimization, good generalization ability, and dimensional independence (27–29). Its flexibility in classification and the ability to approximate continuous function make SVM very suitable for quantitative structure–activity relationship and quantitative structure–property relationship studies.

In this investigation, SVM was used for the prediction of the retention time of a diverse set of solutes in three different RP-high-performance liquid chromatography (HPLC) columns using the molecular descriptors calculated by the software CODESSA and selected by the HM method. The structural factors affecting the compounds' retention behaviors on these columns and the characteristics of the three columns were also investigated. This study provides a new method to investigate the relationship between the characteristics of columns and the retention behavior of the analytes.

Method

Data set

The data set was collected from ref. 30, including benzene derivatives, organic acid derivatives, aniline derivatives, and other compounds. The following columns were employed: Symmetry C18, 15.0 × 0.46 cm i.d., particle size 5 µm (Waters, Milford, MA), packed with octadecylbonded silica; Chromolith, 10.0 × 0.46 cm i.d. (Merck KGaA, Darmstadt, Germany) made of a highly porous monolithic rod of silica; and SG-MIX column, 25.0 × 0.40 cm i.d., particle size 5 µm (Nicolaus Copernicus University, Toruń, Poland). All the mobile phases contained methanol and TRIS buffer of pH 2.5 or 7.2 for the suitable separation.

A complete list of these compounds' names and corresponding experimental retention times are given in Table I. The data set was randomly divided into two subsets: the training set and the test set. The training set of 50 compounds was used to build regression models, and the test set of 12 compounds was used to evaluate the prediction capability of the model.

Descriptor calculation

To obtain a QSRR model, compounds were often represented by the molecular descriptors. All of the structures of the molecules were drawn with the software ISIS Draw, and then the structures were imported into the HYPERCHEM program and pre-optimized using the MM+ molecular mechanics force field. A more precise optimization was done with the semi-empirical PM3 method. Then the molecular structures were exported into the software MOPAC and the quantum properties were calculated using the PM3 method. The MOPAC output files were used by the CODESSA program to calculate five classes of descriptors: constitutional (number of various types of atoms and bonds, number of rings, molecular weight); topological (Wiener index, Randic indices, Kier-Hall shape indices, etc.); geometrical (moments of inertia, molecular volume, molecular surface area, etc.); electrostatic (minimum and maximum partial charges, polarity parameter, charged partial surface area descriptors, etc.); and quantum chemical (reactivity indices, dipole moment, HOMO and LUMO energies, etc.) (31). In this study, a total of 420 descriptors were generated by the CODESSA program to represent the compound structures.

The selection of the descriptors based on HM (31)

After molecular descriptors were generated, the HM was also used to select the most important descriptors based on the training set. It then built a linear regression model at the same time. The advantages of this method are its rapid selection and no matter for the data size. Furthermore, the method can

demonstrate which descriptors have bad or missing values, which ones were insignificant, and which ones were highly intercorrelated. This information is helpful in reducing the number of descriptors involved in a search of the best QSRR model.

First of all, the descriptors were checked to ensure that the values of each descriptor were available for each structure, and that there was a variation in these values. Descriptors for values that were not available for every structure were discarded. Descriptors having a constant value for all structures in the data set were also discarded. Thereafter, all possible one-parameter regression models were tested and insignificant descriptors were removed. As a next step, the program calculated the pair correlation matrix of descriptors and further reduced the descriptors pool by eliminating highly correlated descriptors. All two-parameter regression models with remaining descriptors were subsequently developed and ranked by the regression correlation coefficient. A stepwise addition of further descriptor scales was performed to find the best multi-parameter regression models with the optimum values of statistical criteria (highest values of R^2, the cross-validated R^2_{CV}, and the F-value).

SVM

The foundation of SVM was developed by Vapnik, and they are gaining popularity due to many attractive features and promising empirical performance (27–29). Compared with traditional neural networks, SVM possess prominent advantages: (*i*) A strong theoretical background provides SVM with high generalization capability and avoids local minima; (*ii*) SVM always has a solution, which can be quickly obtained by a standard algorithm (quadratic programming); (*iii*) SVM need not determine network topology in advance, which can be automatically obtained when the training process ends; (*iv*) SVM builds a result based on a sparse subset of training samples, which reduces the workload (32).

Originally, SVM was developed for the pattern recognition problems. With the introduction of ε-insensitive loss function, SVM has been extended to solve nonlinear regression estimation and time series prediction (33). Theories of support vector classification and regression can be found in the tutorials for SVM (28). For this reason, we will only briefly describe the main idea of support vector regression here.

SVM can be applied to regression problems by the introduction of an alternative loss function. In support vector regression, the input x (descriptor) is fist mapped into a higher dimensional feature space by the use of kernel function. Thus a non-linear feature mapping will allow the treatment of non-linear problems in a linear space. The prediction or approximation function used by a basic SVM is:

$$f(x) = \sum_{i=1}^{l} \alpha_i K(x,x_i) + b$$

where x_i is a feature vector corresponding to a training object, $K(x,x_i)$ is a kernel function, α_i is a coefficient. The component of vector α and the constant b represent the hypothesis and are optimized during the training. $K(x,x_i)$ is a kernel function which value is equal to the inner product of two vectors x and x_i in the feature space $\phi(x)$ and $\phi(x_i)$. That is, $K(x,x_i) = \phi(x)\cdot\phi(x_i)$. The elegance of using kernel function lies in the fact that one can deal with feature spaces arbitrary dimensionality without having to compute the map $\phi(x)$ explicitly, and it may be useful to think of the kernel, $K(x,x_i)$, as comparing patterns, or as evaluating the proximity of objects in their feature space. Thus, a test point is evaluated by comparing it to all training points.

For a given dataset, the kernel function and the regularity parameter C must be selected to specify one SVM. Any function that satisfies Mercer's condition can be used as the kernel function. In support vector regression, the Gaussian kernel $K(u,v) = \exp(-\gamma^* |u - v|^2)$ is most commonly used.

All calculation programs implementing SVM were written in R-file based on R script for SVM. All scripts were compiled using R 1.7.1 compiler running operating system on a Pentium IV with 512 M RAM.

Results and Discussion

HM

As 420 descriptors were generated by using the CODESSA program, the next step was to use an efficient method to pick out the most popular descriptors. As one of the most powerful tools, HM investigated a variety of subset sizes to determine the optimum numbers of all the calculated descriptors. When the addition of another descriptor did not significantly improve the statistics of a model, it was determined that the optimum subset size had been achieved. To avoid the "overparametrization" of the model, an increase of the R^2 value of less than 0.02 was chosen as the breakpoint criterion. The root-mean-square error (RMSE) was used as an error function which was defined as in the following:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (d_i - o_i)^2}{n}}$$

where d_i are the desired outputs in the training set, o_i are the actual outputs obtained from the method, and n is the number of the samples in the training set.

The influences in the number of the descriptors (N) for the three different columns on the correlation coefficient (R^2), the cross-validated coefficient (R^2_{CV}), and the squared standard error (s^2) are shown in Figure 1. In the different columns, three multilinear regression models were constructed respectively. The statistical parameters (R^2 and t-test) of the models and the corresponding physical-chemical meanings to the selected descriptors are summarized in Table II. Table III gives the correlation matrix of each selected descriptor for every column. The linear correlation coefficient value of each two descriptors was < 0.80 (Table III), which meant that the descriptors were independent of each other in this multi-linear analysis. Tables I and IV show the results of the HM models using the selected descriptors. For example, on the Chromolith column, the best regression model had a correlation coefficient of R^2 = 0.9264, F = 141.53, and the cross-validated coefficient of R^2_{CV} = 0.9095. The cross-validated coefficient result confirmed the predictive capability of this model. This model gave an RMSE of 0.7124 retention units for the whole set.

In all of the three models, the topological descriptor Kier and Hall index (KHI1 or KHI3) is involved. This descriptor means that the structure of the molecules is a very important factor for RP-

HPLC. According to the values of the *t*-test, this descriptor is also the most important factor in RP-HPLC. It represents the size of the hydrophobic segment and contained group contributions from all non-hydrogen atoms in the fragment. It is defined as:

$$KHI3 = \sum_{i=1}^{N} (\delta_{i1}\delta_{i2}\delta_{i3}\delta_{i4})^{-1/2} \quad \text{where} \quad \delta_i = \frac{Z_i^v - H_i}{Z_i - Z_i^v - 1}$$

where Z_i is the total number of electrons in the ith atom, Z^v_i is the number of valance electrons, and H_i is the number of hydrogen directly attached to the ith atom. The positive values indicate that the shapes of the molecules are in favor of retention times. The electrostatic descriptors CHAS and HACA-2/TMSA are related to the hydrogen bonding capacity of the compounds. Furthermore, CHAS encodes the hydrophilicity of the compounds. An increase in this descriptor strengthens the hydrophilicity of the molecule, decreases the interaction between the solute and stationary phase, and then favors the elution process. The other descriptor HACA-2/TMSA is hydrogen

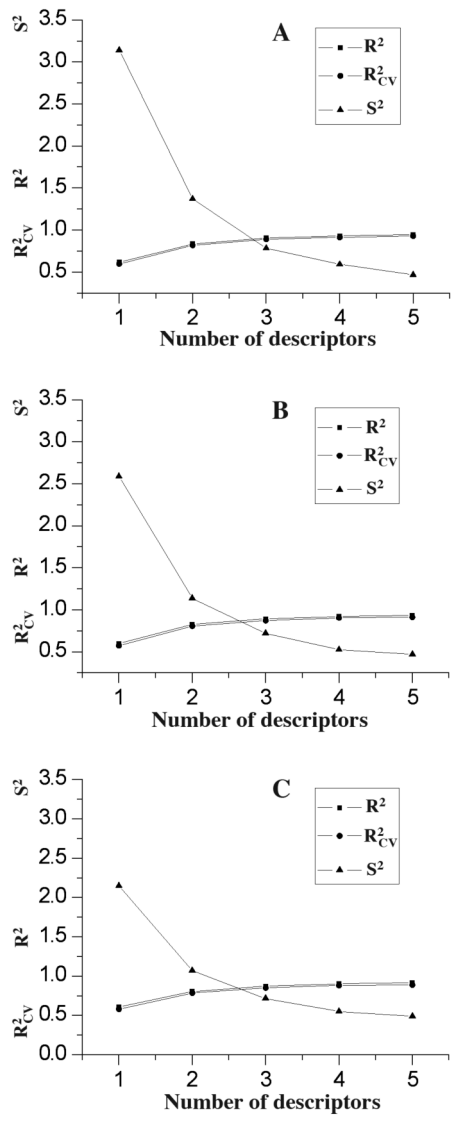

Figure 1. *R^2, R^2_{CV}, and s^2 vs.* the number of descriptors for the three regression models [Symmetry C18 (A), Chromolith (B), SG-MIX (C)].

Table II. The Linear Models of Training Set for Three Columns

Descriptor	Chemical meaning	Coefficient	*t*-Test
Column: Symmetry C18*			
(constant)	Intercept	0.238	0.118
CHAS	Count of H-acceptor sites (Zefirov's PC)	–1.923	–14.359
KHI3	Kier and Hall index (order 3)	3.038	16.002
MSPBO	Max SIGMA-PI bond order	–74.268	–6.580
HLEG	HOMO–LUMO energy gap	1.208	6.184
Column: Chromolith†			
(constant)	Intercept	5.018	8.342
CHAS	count of H-acceptor sites (Zefirov's PC)	–1.216	–7.903
KHI1	Kier and Hall index (order 1)	1.372	14.763
TDM	Tot dipole of the molecule	–0.548	–7.282
M1ERIC	Min 1-electron react. index for a C atom	–74.003	–4.707
Column: SG-MIX‡			
(constant)	Intercept	9.990	17.810
HACA-2/TMSA	HACA-2/TMSA (Zefirov's PC)	-569.800	-7.112
KHI3	Kier and Hall index (order 3)	2.067	13.540
TDM	Tot dipole of the molecule	-0.428	-5.638
M1ERIC	Min 1-electron react. index for a C atom	-73.586	-4.616

* R^2 = 0.9389, F = 172.83, s^2 = 0.5701, R^2_{CV} = 0.9228, n = 62.
† R^2 = 0.9264, F = 141.53, s^2 = 0.5296, R^2_{CV} = 0.9095, n = 62.
‡ R^2 = 0.9118, F = 113.66, s^2 = 0.5548, R^2_{CV} = 0.8890, n = 61.

Table III. Correlation Matrix of the Four Selected Descriptors of Each Column's Model Used in This Work*

	CHAS	KHI3	MSPBO	HLEG
Column: Symmetry C18				
CHAS	1			
KHI3	–0.109	1		
MSPBO	0.381	–0.001	1	
HLEG	–0.049	–0.633	–0.072	1
	CHAS	**KHI1**	**TDM**	**M1ERIC**
Column: Chromolith				
CHAS	1			
KHI1	–0.100	1		
TDM	0.610	–0.107	1	
M1ERIC	0.374	0.246	0.232	1
	HACA-2/TMSA	**KHI3**	**TDM**	**M1ERIC**
Column: SG-MIX				
HACA-2/TMSA	1			
KHI3	–0.178	1		
TDM	0.577	–0.137	1	
M1ERIC	0.294	0.276	0.230	1

* The definitions of all the descriptors are shown in Table II.

Table IV. The Results of the HM

	HM (training set)				Test set
Column	R^2	R^2_{CV}	F	s^2	R^2
Symmetry C18	0.9389	0.9228	172.83	0.5701	0.8770
Chromolith	0.9264	0.9095	141.53	0.5296	0.8968
SG-MIX	0.9118	0.8890	113.66	0.5548	0.8940

bond acceptor charged surface area/total molecular surface area. These descriptors account sufficiently for the electrostatic and hydrogen bonding influence on the retention of the compounds and it relates to the hydrogen-bonding interactions. The descriptors refer to the area-weighted surface charge of hydrogen bonding acceptor atoms. This descriptor describes the hydrophilicity of the solutes. From the values of the *t*-test, it can be concluded that the hydrophobicity was the main characteristic for RPLC, especially in the C18 Symmetry column. The negative values of t indicated that the larger the descriptor, the smaller the retention.

In the Chromolith and SG-MIX models, the descriptor Tot dipole of the molecule (TDM) (35) was involved. TDM calculates the total dipole moment of a molecule, and describes the dipole moment and divisive ability of the compounds. The dipole moment of the molecule is defined as

$$\mu = -\sum_{i=1(V)}^{\infty} \int_{(F)} \phi_i \hat{r} \phi_i dv + \sum_{\alpha=1}^{M} Z_\alpha \vec{R}_\alpha$$

where ϕ is molecular orbital, $\hat{r}$ is electron position operator, Z_α is the αth atomic nuclear charge and R_α is position vector of αth atomic nucleus. It can describe the polar interactions from permanent or induced dipoles between solute, stationary-phase, and mobile phase molecules. The values of the *t*-test indicated that TDM plays an important role in the Chromolith and SG-MIX columns, which was in good agreement with reference 30. By comparing the values of *t*, we found that the lower polarity of the SG-MIX was in comparison to the Chromolith column. The negative values indicated that these structural features made a negative contribution to the retention times.

In the Symmetry C18 model, two other quantum chemical descriptors were involved. The descriptor max SIGMA-PI bond order (36) relates to the strength of intramolecular bonding interactions and characterize the stability of the molecules, their conformational flexibility, and other valency-related properties. The SIGMA bond describes the hydrophilicity of the molecule, and the PI bond describes the hydrophobicity of the compound. The higher the descriptor, the stronger the hydrophilicity of the compound. The last descriptor is quantum-chemical descriptor HOMO-LUMO energy gap ($\varepsilon_{HOMO} - \varepsilon_{LUMO}$) (37). It is an approximate estimate of the first electron excitation energy in the UV/visible spectra of the molecule or of the bandgap between valence and empty zone in solids.

Other descriptors have small *t*-test values, so the roles of these descriptors were not significant. They were not discussed in this work.

From the previously mentioned discussions, we can conclude that the selected descriptors can account for the structural features which are responsible for the retention behaviors of analytes. Based on the discussion of the meaning of the descriptors, it can be concluded that there are several factors influencing the retention of solutes: (*i*) Steric interactions between the solute and the stationary phase; (*ii*) Hydrogen bond interactions, especially in Symmetry C18; (*iii*) Polar interactions from permanent or induced dipoles between solute, stationary-phase, and mobile phase molecules, especially in the Chromolith column. As mentioned previously, it could be seen that the QSRR method provides a new way to research the properties of the columns. All of the conclusions are in agreement with references 9 and 34.

Support vector regression model

Selection of the kernel function and parameter of the SVM

The performances of non-linear approach SVM for regression depend on the combination of several parameters, such as capacity parameter C, ε of the ε-insensitive loss function, the kernel type *K*, and its corresponding parameters. C is the regularization parameter that controls the trade-off 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. But reference 38 indicated that prediction error was scarcely influenced by C. To make the learning process stable, a large value should be set for C (e.g., C = 100).

The kernel type is an important one. For regression tasks, the Gaussian kernel is commonly used. The form of the Gaussian function in R is as follows:

$$K(u,v) = \exp(-\gamma^* |u - v|^2)$$

Where γ is a constant, the parameter of the kernel; u and v are two independent variables; and γ controls the amplitude of the Gaussian function and, therefore, controls the generalization ability of SVM. We had to optimize γ. Each RMSE based on the LOO cross-validation of training set was plotted versus γ (Figure 2) on the Chromolith column. As shown in Figure 2, the optimal value is 0.05.

The optimal value for ε depends on the type of noise present in the data, which is usually unknown. Even if enough knowledge

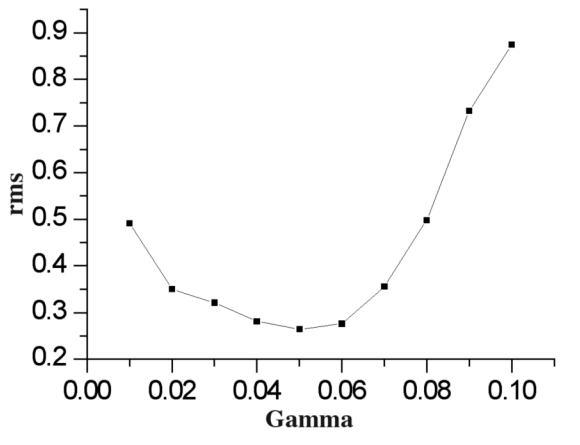

Figure 2. The gamma vs. RMSE on LOO cross-validation in the Chromolith model (C = 100, epsilon = 0.03).

of the noise is available to select an optimal value for ε, there is the practical consideration given to 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. Choosing the appropriate value of ε is critical from theory. To find an optimal value of ε, the RMSE of LOO cross-validation for the training set on different ε was calculated. The curve of RMSE versus the epsilon is shown in Figure 3 (on the Chromolith column), respectively. The optimal value of ε was found as 0.03.

On the Symmetry and SG-MIX columns, we used the same method to select the optimum values of γ and ε. On the Symmetry C18 column, the γ and ε were 0.1 and 0.02, respectively. On the SG-MIX column, the γ and ε were 0.01 and 0.02, respectively.

The predicted results of SVM

From the above discussion, in the Gaussian-kernel SVM, the optimal γ, ε, and C were 0.1, 0.02, and 100 on the Symmetry C18 column; 0.05, 0.03, and 100 on the Chromolith column; and 0.01, 0.02, 100 on the SG-MIX column, respectively. The predicted results of the optimal SVM are shown in Tables I and V, and in Figure 4. The proposed models were statistically stable and fitted the data well. For example, on the Chromolith column, the experimental data, the predicting data values of the training data, and the testing data by the SVM model are listed in Table I under the optimal model. The model gave the RMSE 0.5922 for the training set, 0.5149 for the testing set, and 0.5781 for the whole set; the corresponding square correlation coefficients (R^2) were 0.9464, 0.9503, and 0.9468, respectively.

From analysis of the results obtained from linear HM regression models and non-linear SVM models, we conclude that non-linear models can simulate the relationship between the structural descriptors and the chromatography retention of compounds more accurately. SVM can correctly represent the structure-retention relationships of these compounds. The molecular descriptors calculated solely from the structures can

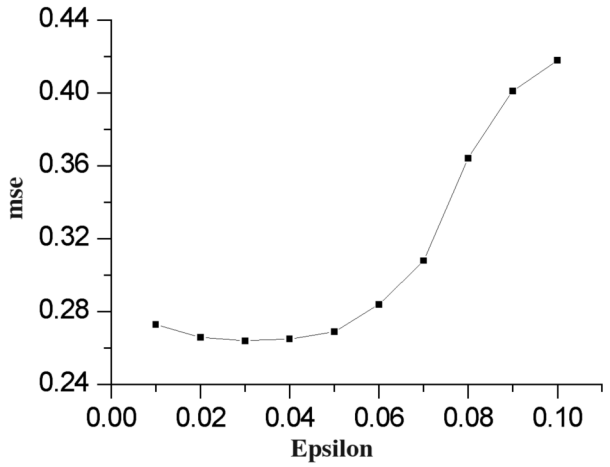

Figure 3. The epsilon vs. RMSE on LOO cross-validation in the Chromolith model (C = 100, gamma = 0.05).

Table V. The Results of the SVM

Column		SVM		
		Training set	Test set	Total
Symmetry C18	R^2	0.9794	0.9695	0.9766
	RMSE	0.4186	0.5205	0.4402
Chromolith	R^2	0.9464	0.9503	0.9468
	RMSE	0.5922	0.5149	0.5781
SG-MIX	R^2	0.9429	0.9274	0.9289
	RMSE	0.5718	0.8510	0.6355

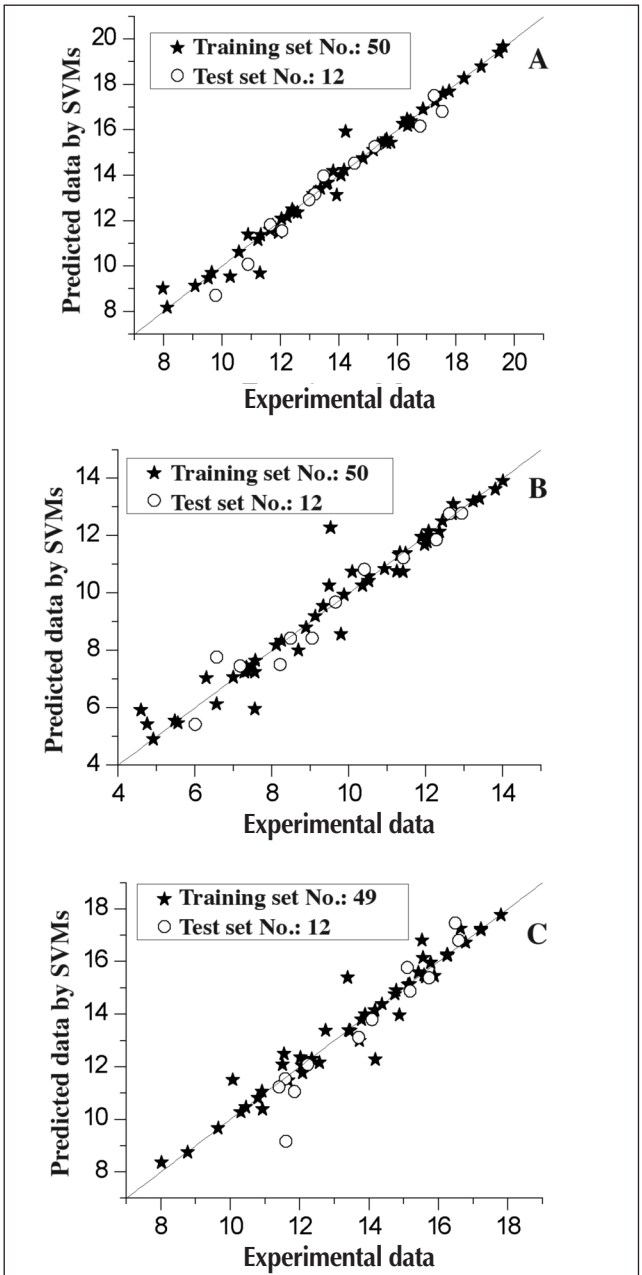

Figure 4. Plot of predicted retention vs. experimental values of the training set and test set. Symmetry C18 column model by SVM (gamma = 0.1, epsilon = 0.02, C = 100) (A); Chromolith column model by SVM (gamma = 0.05, epsilon = 0.03, C = 100) (B); SG-MIX column model by SVM (gamma = 0.05, epsilon = 0.03, C = 100) (C).

describe the structural features of the compounds, which were responsible for their retention behaviors. The characteristics of hydrophobia play an important role in the retention of the compounds on the Symmetry C18, Chromolith, and SG-MIX columns. The retention of the compounds on the three columns are determined by several intermolecular interactions, such as polar interactions between solute, stationary-phase, and mobile phase molecules; hydrogen bond interactions; and steric interactions between the solute and the stationary phase.

Comparison with the literature

In order to make a comparison, the results of HM, SVM, and the original reference (30) are listed in detail in Table VI. From Table VI, it can be concluded that the SVM method is a powerful tool to model the relationship between the retention times of compounds and different columns in RP- HPLC. Moreover, from Table I and Figure 4, it can be seen clearly that the three QSRR models possessed good prediction ability. The predicted values were in good agreement with the experimental results.

Conclusions

Accurate linear and nonlinear QSRR models of the retention times of compounds in the three columns (Symmetry C18, Chromolith, and SG-MIX) were built based on HM and SVM, respectively. The linear and non-linear models gave satisfactory results. From the comparison of the obtained results, the SVM method gave the better results, and the prediction of the retention times was in agreement with the experiment value. We conclude that: (*i*) The linear model constructed by HM could correctly represent the relationship between the retention times and the molecular structures calculated from the chemical structures alone. The selected descriptors can illuminate the features of the compounds which were responsible for their retention behaviors, such as steric interactions, the hydrogen bond interactions, polar interactions, and characteristics of hydrophobic. This information will be helpful to direct and understand the actual experiments. (*ii*) Nonlinear regression models can simulate the HPLC retention phenomena more accurately than the linear model. In summary, this investigation developed a new method to research the characteristics of columns. It can also provide another idea for dealing with other QSRR problems.

Table VI. The Comparison of the Square Correlation Coefficients (R^2) for the Training Set and Test Set Based on HM, SVM, and Reference 30

		Reference	HM	SVM
Symmetry C18	Training set	0.9757	0.9389	0.9794
	Test set	0.8083	0.877	0.9695
Chromolith	Training set	0.9781	0.9264	0.9464
	Test set	0.8029	0.8968	0.9503
SG-MIX	Training set	0.9312	0.9118	0.9429
	Test set	0.6738	0.894	0.9274

Acknowledgment

The authors would like to express their gratitude to Jeanette Bradley and June Watzl (Yale University, School of Medicine) for proofreading. Special thanks to the anonymous reviewers, and to the Editor for their professional, extensive, and useful comments.

References

1. R.P.W. Scott. *Silica Gel and Bonded Phases*. Wiley, New York, NY, 1993.
2. T. Baczek, R. Kaliszan, K. Novotná, and P. Jandera. Comparative characteristics of HPLC columns based on quantitative structure–retention relationships (QSRR) and hydrophobic-subtraction model. *J. Chromatogr. A* **1075:** 109–115 (2005).
3. R. Kaliszan. Quantitative structure-retention relationships applied to reversed-phase high-performance liquid chromatography. *J. Chromatogr. A* **656:** 417–435 (1993).
4. R. Kaliszan. *Structure and Retention in Chromatography. A Chemometric Approach*. Harwood Academic Publishers, Amsterdam, The Netherlands, 1997.
5. R. Kaliszan. Quantitative structure–retention relationships. *Anal. Chem.* **64:** 619–631 (1992).
6. R.J. Hu, H.X. Liu, R.S. Zhang, C.X. Xue, X.J. Yao, M.C. Liu, Z.D. Hu, and B.T. Fan. QSPR prediction of GC retention indices for nitrogen-containing polycyclic aromatic compounds from heuristically computed molecular descriptors. *Talanta* **68:** 31–39 (2005).
7. A.R. Katritzky, V.S. Lobanov, and M. Karelson. QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.* **24:** 279–288 (1995).
8. A.R. Katritzky, M. Karelson, and V. Lobanov. QSPR as a means of predicting and understanding chemical and physical properties in terms of structure. *Pure Appl. Chem.* **69:** 245–249 (1997).
9. R. Kaliszan, M.A. van Straten, M. Markuszewski, C.A. Cramers, and H.A. Claessens. Molecular mechanism of retention in reversed-phase high-performance liquid chromatography and classification of modern stationary phases by using quantitative structure–retention relationships. *J. Chromatogr. A* **855:** 455–486 (1999).
10. M.A. Al-Haj, R. Kaliszan, and A. Nasal. Test analytes for studies of the molecular mechanism of chromatographic separations by quantitative structure-retention relationships. *Anal. Chem.* **71:** 2976–2985 (1999).
11. D. Bolliet, C.F. Poole, and M. Rosès. Conjoint prediction of the retention of neutral and ionic compounds (phenols) in reversed-phase liquid chromatography using the solvation parameter model. *Anal. Chim. Acta* **368:** 129–140 (1998).
12. A. Jakab, G. Schubert, M. Prodan, and E. Forgàcs. Determination of the retention behavior of barbituric acid derivatives in reversed-phase high-performance liquid chromatography by using quantitative structure–retention relationships. *J. Chromatogr. B* **770:** 227–236 (2002).
13. C.G. Georgakopoulos and J.C. Kiburis. Quantitative structure-retention relationships in doping control. *J. Chromatogr. B Biomed. Appl.* **687:** 151–156 (1996).
14. J. Zupan and J. Gasteiger. *Neural Networks in Chemistry and Drug Design*. Wiley-VCH Verlag, Weinheim, Germany, 1999.
15. L. Fausett. *Fundamentals of Neural Networks*. Prentice Hall, New York, NY, 1994.
16. K. Petritis, L.J. Kangas, P.L. Ferguson, G.A. Anderson, L. Paša-Toli, M.S. Lipton, K.J. Auberry, E.F. Srittmatter, Y. Shen, R. Zhao, and R.D. Smith. Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses. *Anal. Chem.* **75:** 1039–1048 (2003).

17. W.Q. Guo, Y. Lu, and X.M. Zheng. The predicting study for chromatographic retention index of saturated alcohols by MLR and ANN. *Talanta* **51:** 479–488 (2000).
18. Y.L. Loukas. Artificial neural networks in liquid chromatography: efficient and improved quantitative structure–retention relationship models. *J. Chromatogr. A* **904:** 119–129 (2000).
19. F. Ruggieri, A.A. D'Archivio, G. Carlucci, and P. Mazzeo. Application of artificial neural networks for prediction of retention factors of triazine herbicides in reversed-phase liquid chromatography. *J. Chromatogr. A* **1076:** 163–169 (2005).
20. D.T. Manallack and D.J. Livingstone. Neural networks in drug discovery: have they lived up to their promise? *Eur. J. Med. Chem.* **34:** 195–208 (1999).
21. J. Wang, H.X. Liu, S. Qin, X.J. Yao, M.C. Liu, Z.D. Hu, and B.T. Fan. Study on the structure–activity relationship of new anti-HIV nucleoside derivatives based on the support vector machine method. *QSAR Comb. Sci.* **26:** 161–172 (2007).
22. H.X. Liu, X.J. Yao, R.S. Zhang, M.C. Liu, Z.D. Hu, and B.T. Fan. Accurate quantitative structure–property relationship model to predict the solubility of C60 in various solvents based on a novel approach using a least-squares support vector machine. *J. Phys. Chem. B* **109:** 20565–20571 (2005).
23. C.X. Xue, R.S. Zhang, H.X. Liu, X.J. Yao, M.C. Liu, Z.D. Hu, and B.T. Fan. An accurate QSPR study of O-H bond dissociation energy in substituted phenols based on support vector machines. *J. Chem. Inf. Comput. Sci.* **44:** 669–677 (2004).
24. H.X. Liu, R.S. Zhang, X.J. Yao, M.C. Liu, Z.D. Hu, and B.T. Fan. Prediction of the isoelectric point of an amino acid based on GA-PLS and SVMs. *J. Chem. Inf. Comput. Sci.* **44:** 161–167 (2004).
25. Y.D. Cai, X.J. Liu, X.B. Xu, and K.C. Chou. Support vector machines for predicting HIV protease cleavage sites in protein. *J. Comput. Chem.* **23:** 267–274 (2002).
26. C.W. Morris, A. Autret, and L. Boddy. Support vector machines for identifying organisms—A comparison with strongly partitioned radial basis function networks. *Ecol. Model.* **146:** 57–67 (2001).
27. C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Disc.* **2:** 121–167 (1998).
28. V.N. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, 1998.
29. N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, U.K., 2000.
30. R. Kaliszan, T. Baczek, A. Buci ski, B. Buszewski, and M. Sztupecka. Prediction of gradient retention from the linear solvent strength (LSS) model, quantitative structure-retention relationships (QSRR), and artificial neural networks (ANN). *J. Sep. Sci.* **26:** 271–282 (2003).
31. A.R. Katritzky, V.S. Lobanovn, and M. Karelson. *Comprehensive Descriptors for Structural and Statistical Analysis, Reference Manual*. Version 2.0. 1994.
32. S.R. Gunn, M. Brown, and K.M. Bossley. Network performance assessment for neurofuzzy data modelling. *Lecture Notes Comput. Sci.* **1280:** 313–323 (1997).
33. H.X. Liu, R.J. Hu, R.S. Zhang, X.J. Yao, M.C. Liu, Z.D. Hu, and B.T. Fan. The prediction of human oral absorption for diffusion rate-limited drugs based on heuristic method and support vector machine. *J. Comput. Aid. Mol. Des.* **19:** 33–46 (2005).
34. L.R. Snyder, J.W. Dolan, and P.W. Carr. The hydrophobic-subtraction model of reversed-phase column selectivity. *J. Chromatogr. A* **1060:** 77–116 (2004).
35. A.R. Katritzky, R. Petrukhin, H.F. Yang, and M. Karelson. *CODESSA PRO*. User's Manual. University of Florida, FL, 2002.
36. A.B. Sannigrahi. Ab initio molecular orbital calculations of bond index and valency. *Adv. Quant. Chem.* **23:** 301–351 (1992).
37. A.R. Katritzky, V.S. Lobanov, and M. Karelson. *CODESSA: Reference Manual*, Version 2. University of Florida, FL, 1994.
38. W.J. Wang, Z.B. Xu, W.Z. Lu, and X.Y. Zhang. Determination of the spread parameter in the Gaussian kernel for classification and regression. *Neurocomputing* **55:** 643–663 (2003).

Manuscript received May 25, 2007;
Revision recieved November 11, 2007.