Quantitative Structure Activity Relationship Study of Volatile Organic Compounds in Waste Water by Chemometrics Methods

A. Yazdanipour
Chemistry Department, Arak branch, Islamic Azad University, Arak, Iran

Abstract

A quantitative structure-activity relationship (QSAR) study is suggested for the prediction of retention time of volatile organic compounds in waste water. Modeling of the retention time of volatile compounds as a function of molecular structures was established by means of the chemometrics methods such as partial least squares (PLS) and least squares support vector machines (LS-SVM). These models were applied for the prediction of the retention times of these compounds, which were not in the modeling procedure. The predictive quality of the QSAR models were tested for an external prediction set of 8 compounds randomly chosen from 59 compounds. The resulted model showed high prediction ability with root mean square error of prediction of 0.0335 for LS-SVM.

Keywords: QSAR; Retention times; Volatile organic compounds; LS-SVM; PLS.

Introduction

An important property that has been extensively studied in quantitative structure activity relationship (QSAR) is the chromatographic retention time. A quantitative structure retention relationship (QSRR) study involves the prediction of chromatographic retention parameters using molecular structure. QSRR studies we widely investigated in gas chromatography (GC) and high-performance liquid chromatography (HPLC). The chromatographic parameters are expected to be proportional to a free energy change that is related to the solute distribution on the column. Chromatographic retention is a physical phenomenon that is primarily dependent on the interactions between the solute and the stationary phase. Molecular group contribution methods are widely employed to estimate gas chromatographic retention parameter.

The anthropogenic pollution of environmental water goes into the global scales, thus representing significant public health risk. Volatile organic compounds (VOCs) have been shown to affect a wide number of biological and environmental systems, they influence various atmospheric processes, some are carcinogens and/or mutagens, while others are persistent and show bioaccumulation effects. [1] In addition, many VOCs exhibit toxic effects on aquatic organisms. As regards the water, VOCs are among the most commonly found contaminates in groundwater. Their volatility is the reason they are not often found in concentrations above a few $\mu g\ L^{-1}$ in surface waters, but in groundwater their concentrations can be hundreds or thousands of times higher.[2].

Among the investigation of QSAR, one of the most important factors affecting the quality of the model is the method to build the model. Many multivariate data analysis methods such as multiple linear regression (MLR) , [3] partial least squares (PLS) [4, 5] and artificial neural network (ANN) [6] have been used in QSAR studies. MLR, as most

commonly used chemometrics method, has been extensively applied to QSAR investigations. However, the practical usefulness of MLR in QSAR studies is rather limited, as it provides relatively poor accuracy. ANN offers satisfactory accuracy in most cases but tends to overfit the training data. The support vector machine (SVM) is a popular algorithm developed from the machine learning community. Due to its advantages and remarkable generalization performance over other methods, SVM has attracted attention and gained extensive applications [7, 8]. As a simplification of traditional of SVM, Suykens and Vandewalle [9, 10] have proposed the use of least-squares SVM (LS-SVM). LS-SVM encompasses similar advantages as SVM, but its additional advantage is that it requires solving a set of only linear equations (linear programming), which is much easier and computationally more simple.[11-13] In the present study, the PLS and LS-SVM methods were applied in QSAR for modeling the relationship between the retention time of 59 volatile organic compounds by using structural molecular descriptors.

Materials and computational methods

Data set and methods

The QSRR model fro the estimation of the retention times of various volatile organic compounds is established in the following steps: the molecular structure input and generation of the files containing the chemical structures is stored in a computer-readable format; quantum mechanics geometry is optimized with a semi-empirical (AM1) method; structural descriptors are computed; and the structural-retention time model is generated by the chemometrics methods and statistical analysis. The retention time of 59 volatile organic compounds was collected from.[14]

Computer hardware and software

All calculations were run on a Pentium IV personal computer with windows XP operating system. ChemDraw Ultra version 9.0 (ChemOffice 2005, CambridgeSoft Corporation) software was used to draw the molecular structures and optimization by the AM1. Descriptors were calculated utilizing Dragon software (Milano Chemometrics and QSAR research group, http://www.disat.unimib.it/chm/). These descriptors are calculated using two-dimensional representation of the molecules and therefore geometry optimization is not essential for calculating these types of descriptors.

Results and discussion

Retention times of 59 volatile organic compounds including halogenated was taken from the literature,[14] and are presented in Table 1. A major step in constructing QSAR models is finding one or more molecular descriptors that represent variation in the structural property of the molecules by a number. A wide variety of descriptors have been reported to be used in QSAR analysis.[15]

No.	Substance	Retention time (s)	No.	Substance	Retention time (s)
1 [t]	d_1-Chloroform	5.05	31 [t]	Chlorobenzene	12.98
2 [t]	1,1-Dichloroethene	1.75	32 [t]	Ethylbenzene	13.07
3 [t]	Dichloromethane	2.49	33 [t]	*m*-/*p*-Xylene	13.35
4 [t]	*trans*-1,2-Dichloroethene	2.57	34 [p]	*o*-Xylene	13.07
5 [p]	1,1-Dichloroethane	3.28	35 [t]	Styrene	14.29
6 [t]	2,2-Dichloropropane	3.94	36 [t]	1,1,1,2-Tetrachloroethane	13.38
7 [t]	*cis*-1,2-Dichloroethene	4.23	37 [t]	Isopropylbenzene	15.06

Table 1continued…

8 [t]	1,1,1-Trichloroethane	4.68	38 [t]	Bromoform	15.37
9 [t]	Carbon tetrachloride	4.72	39 [t]	Bromobenzene	15.93
10 [t]	Bromochloromethane	4.81	40 [t]	Propylbenzene	15.99
11 [t]	1,1-Dichloropropene	4.96	41 [p]	2-Chlorotoluene	16.26
12 [t]	Chloroform	4.99	42 [t]	1,3,5-Trimethylbenzene	16.44
13 [t]	1,2-Dichloroethane	6.29	43 [t]	4-Chlorotoluene	16.59
14 [t]	d-Benzene	5.56	44 [t]	1,2,3-Trichloropropane	16.77
15 [t]	Benzene	5.61	45 [t]	1,1,2,2-Tetrachloroethane	16.88
16 [t]	Pentafluorobenzne	5.24	46 [t]	*tert*-Butylbenzene	17.13
17 [t]	Trichloroethene	6.89	47 [t]	1,2,4-Trimethylbenzene	17.30
18 [p]	1,2-Dichloropropane	7.85	48 [t]	*sec*-Butylbenzene	17.59
19 [t]	Dibromomethane	8.29	49 [p]	4-Isopropyltoluene	17.99
20 [p]	Dibromodichloromethane	8.72	50 [t]	d4-1,4-Dichlorobenzene	18.28
21 [t]	Tetrachloroethene	10.51	51 [t]	1,3-Dichlorobenzene	18.01
22 [t]	*cis*-1,3-Dichloropropene	9.56	52 [t]	1,4-Dichlorobenzene	18.30
23 [t]	*trans*-1,3-Dichloropropene	11.05	53 [t]	1,2-Dichlorobenzene	19.19
24 [t]	1,1,2-Trichloroethane	11.49	54 [t]	Butylbenzene	18.97
25 [t]	1,3-Dichloropropane	11.71	55 [t]	1,2,4-Trichlorobenzene	22.45
26 [t]	Dibromochloromethane	12.20	56 [t]	1,2,3-Trichlorobenzene	23.30
27 [t]	1,2-Dibrommethane	12.22	57 [t]	Naphthalene	22.98
28 [t]	d8-Toluene	9.64	58 [p]	Hexachlorobutadiene	22.45
29 [p]	Toluene	9.74	59 [t]	1,2-Dibrom-3-	21.48
30 [t]	d5-Chlorobenzene	12.95		chloropropane	

[t] training set, [p] prediction set

Table 1 Retention time of volatile organic compounds in the present study.

A pool containing molecular descriptors is derived to property characterize the chemical structure of the VOCs, involving variables of the type Constitutional, Topological Geometrical, Charge, GETAWAY (GEometry, Topological, Atoms-Weighted AssemblY), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), Molecular Walk Counts, BCUT descriptors, 2D-Autocorrealtions, Aromaticity Indices, Randic molecular profiles, Radial Distribution Functions, Functional Groups and Atom-Centered Fragments. These variables are calculated by means of the software Dragon version 5.4. For the evaluation of the predictive ability of a different model, the root mean square error of prediction (RMSEP) and relative standard error of prediction (RSEP) can be used.

PLS analysis

The factor-analytical multivariate calibration method is a powerful tool for modeling, because it extracts more information from the data and allows building more robust models. According to retention time data (Table 1), data randomly classified to training and prediction sets. The optimum number of factors to be included in the calibration model was determined by computing the prediction error sum of squares (PRESS) fro cross-validated models using a high number of factors (half of the number of total training set + 1). The cross-validation method employed was to eliminate only one compound at a time and then PLS calibrated the remaining of training set. The retention time of the left-out sample was predicted by using this calibration. This process was repeated until each compound in the training set had been left out once. According to Haaland suggestion [16], the optimum number of factor was selected.

LS-SVM analysis

The all descriptors were used as the input to develop nonlinear model by LS-SVM. The quality of LS-SVM for regression depend on γ and σ^2 parameters. In this work, LS-SVM was performed with radial basis function (RBF) as a kernel function. To determine the optimal parameters, a grid search was performed based on leave-one-out cross-validation on the original training set for all parameter combinations of γ and σ^2 from 1 to 100 and 1 to 100, respectively, with increment steps of 1. Table 2 shows the optimum γ and σ^2 parameters for the LS-SVM and RBF kernel, using the calibration sets.

-Prediction of retention time of volatile organic compounds

The predictive ability of these methods (PLS and LS-SVM) were determined using 8 retention time (their structure are given in Table 1). The results obtained by PLS and LS-SVM methods are listed in Table 2.

Substance	Actual retention time (s)	Predicted retention time (s)			
		PLS	Error (%)	LS-SVM	Error (%)
1,1-Dichloroethane	3.28	3.12	-4.88	3.27	-0.30
1,2-Dichloropropane	7.85	8.14	3.69	7.88	0.38
Dibromodichloromethane	8.72	8.32	-4.59	8.69	-0.34
Toluene	9.74	9.24	-5.13	9.75	0.10
o-Xylene	13.07	12.46	-4.67	13.05	-0.15
2-Chlorotoluene	16.26	16.59	2.03	16.21	-0.31
4-Isopropyltoluene	17.99	18.61	3.45	18.04	0.28
Hexachlorobutadiene	22.45	22.06	-1.74	22.41	-0.18
NF [a]		9			
PRESS		1.2364			
γ				5	
σ^2				10	

Table 2 continued…

$Q^{2\,b}$	0.9137	0.9842
RMSEP	0.4383	0.0335
RSEP (%)	3.1947	0.2445

[a] Number of factor (PLS), [b] Q^2 coefficient for the model validation by leave-one-out

Table 2 Actual and predicted values of retention times of VOCs using PLS and LS-SVM models

Table 2 also shows Q^2, RMSEP, RSEP and the percentage error for prediction of retention time of volatile organic compounds. As can be seen, the percentage error was also quite acceptable only for LS-SVM. Good results were achieved in LS-SVM model with percentage error ranges from -0.34 to 0.38 for retention time of VOCs. Also, it is possible to see that LS-SVM presents excellent prediction abilities when compared with other regression. According to the results, structural descriptors are suitable descriptors for describing the retention time of VOCs. When LS-SVM method with all descriptors is used, prediction of retention time in test step, with a small error is possible; this is improved in comparison with other method (PLS). This shows that by using all structural descriptors and also LS-SVM method, the retention time of VOCs is predicted with satisfactory results.

Conclusion

A least squares-support vector machine (LS-SVM) model was established to predict the retention time of some volatile organic compounds in waste water. A proper model with high statistical quality and low prediction errors was obtained. The model could predict the retention time of organic compounds not existed in the modeling procedure accurately. The structural and topological descriptors concerning to the whole molecular properties and those of individual atoms in the molecule were found to be important factors controlling the retention time behavior.

Acknowledgment

The author gratefully acknowledges the support to this work from Islamic Azad University, Arak Branch, research council.

References

1. Glegg,G.A., and Rowland, S.J., *Mar. Pollut. Bull*., **32**, 486 (1996).
2. Golfinopoulos, S.K., Lekkas, T.D., and Nikolaou, A.D., *Chemosphere*, **45**, 275 (2001) .
3. Niazi,A., Jameh-Bozorghi, S., and Nori-Shargh, D., *J. Hazard. Mat.* ,**151**, 603 (2008) .
4. Niazi,A., Jameh-Bozorghi, S., and Nori-Shargh, D., *Turk. J. Chem*., **30**, 619 (2006) .
5. Niazi, A., Azizi, A., *Turk. J. Chem*., **32**, 217 (2008).
6. Hemmateenejad, B., Safarpour, M.A., and Taghavi, F., *J. Mol. Struc. (TheoChem)*, **635**, 183 (2003) .
7. Belousov, A.I., Verzakov, S.A., and Von Frese, J., *J. Chemometr. Intell. Syst.*, **64**, 15 (2002).
8. Burbidge, R., Trotter, M., Buxton, B., and Holden, S., *Comput. Chem*., **26**, 5 (2001).
9. Suykens, J.A.K., and Vandewalle, J., *Neural Process. Lett.,* **9**, 293 (1999) .

10. Suykens, J.A.K., van Gestel, T., de Brabanter, J., de Moor, B., and Vandewalle, J., *Least-Squares Support Vector Machines*, World Scientifics, Singapore, (2002).
11. Niazi, A., Ghasemi, J., and Zendehdel, M., *Talanta* ., **74**, 247 (2007) .
12. Niazi, A., Ghasemi, and Yazdanipour, A., *Spectrochim. Acta.* A. **68**, 523 (2007) .
13. Ke, Y., Yiyu, C., *Chin. J. Anal. Chem.*,**34**, 561 (2006).
14. Safarova, V.I., Sapelnikova, S.V., Djazhenko, E.V., Teplova, G.I., Shajdulina, G.F., and Kudasheva, F.Kh., *J. Chromatogr.,* **800**, 325 (2004).
15. Todeschini, R., and Consonni, V., *Handbook of molecular descriptors*, Wiley-VCH, Weinheim (2000).