A QSPR study of GC/MS Retention Data of 85 Volatile Organic Compounds as Air Pollutant Materials by Multivariate Methods

M. Ayati, J.B. Ghasemi

Chemistry Department, Faculty of Sciences, K.N. Toosi University of Technology, Tehran, Iran

Abstract

A quantitative structure-property relationship (QSPR) study is suggested for the prediction of retention times of volatile organic compounds. Various kinds of molecular descriptors were calculated to represent the molecular structure of compounds. Modeling of retention times of these compounds as a function of the theoretically derived descriptors was established by multiple linear regression (MLR) and artificial neural network (ANN). The stepwise regression was used for the selection of the variables which gives the best-fitted models. After variable selection ANN, MLR methods were used with leave-one-out cross validation for building the regression models. This provided a new and effective method for predicting the chromatography retention index for the volatile organic compounds.

Introduction

Volatile organic compounds are compounds that have a high vapor pressure and low water solubility, like organic chemicals in general, there are millions of different compounds which may be classified as volatile organic compounds (VOCs). VOCs are common ground-water contaminants. Many volatile organic compounds are also hazardous air pollutants. VOCs also play a major role in the formation of various secondary pollutants through photochemical reactions in the presence of sunlight and nitrogen oxides. Furthermore, some VOCs could contribute to the atmospheric ozone depletion and the build-up persistent pollutions in remote areas. Therefore, these compounds have been an important environmental issue over the last two decades and have attracted significant attention from different research groups attempt to provide a sensitive and specific analytical method for identifying and measuring the VOCs [1]. The experimental determination of retention time is time-consuming and expensive, and there are some limitations for these methods. Alternatively, quantitative structure–retention relationship (QSRR) provides a promising method for the estimation of retention time based on descriptors derived solely from the molecular structure to fit experimental data. QSRR studies are widely investigated in gas chromatography (GC) and high-performance liquid chromatography (HPLC).

Materials & Methods

All calculations were run on a Toshiba personal computer with a Pentium IV as CPU and windows XP as operating system. The molecular structures of data set were sketched using ChemDraw (Ver. 11, supplied by Cambridge Software Company). The sketched structures were exported to Chem3D module in order to create their 3D structures. Energy minimization was performed using MM+ molecular mechanics and AM1 semi-empirical to obtain the root mean square (RMS) gradient below 0.01 kcal /(mol Å). Here, 477 descriptors were generated for each compound, using Dragon (ver. 3) software, SPSS (ver.16, www.spss.com) were used to stepwise regression. Other calculations were performed in PLS_Toolbox (ver. 4.1, Eigenvector Company) and MATLAB (version 7.8, Math Works, Inc.) environment. The retention times of volatile organic compounds was obtained from reference [2]. Principal components analysis (PCA) was performed

on the calculated structural descriptors to the whole data set for detection of the probable homogeneities/inhomogeneties in the data set, and to separation of the data into training and test sets. According to the results of PCA, 2 molecules were removed as outliers the remaining molecules were divided into a 67 compounds as training set and 16 compounds(according to special location of objects in the score plot) as prediction set to evaluate the models. We have also tested the stability of models by random selection of train and test sets to assure the resulting models are not just chance correlations. With stepwise regression 23 descriptors were selected and a simple "break point" technique was used to control the model expansion in the improvement of the statistical quality of the model. In this way 6 descriptors that have high contribution in the variance of dependent variable (t_R) were selected and used to build the models. X1sol, X5A (topological descriptors), Mor10m, Mor21m (2D autocorrelation descriptors), and ATS2m, MATS2m as 3D-MoRSE descriptors are the six descriptors that were included in the models.

Results

MLR is performed either to study the relationship between the response variable and predictor variables or to predict the response variable based on the predictor variables. The validation of the model was performed by cross-validation method (R^2_{cv}=0.964). Also to show the absence of the chance correlation between independent and dependents variables the MLR modeling was performed on the randomized data. The mean of low chance correlation value (R^2_{CR}=0.091) confirm this result. ANN model used to handle the probable nonlinear relationship between descriptors and retention times. A set of six descriptors that were appeared in the MLR model, were used as input parameter of the network. In this investigation, the logsig function was used as a transfer function of hidden layer, and a linear function for the output layer. We optimized the parameters such as number of nodes, momentum (*a)* and learning rate (*η*). A neural network with 6-6-1 topology was developed with optimum momentum (0.5) and learning rate (0.5). High correlation coefficients (0.972 and 0.977 for MLR and ANN, respectively) and low prediction errors (0.247 and 0.315 for MLR and ANN, respectively) obtained confirm good predictive ability of both models.

Conclusion

Comparison of the linear (MLR) and nonlinear (ANN) methods showed the little superiority of the ANN model for the prediction of the retention time of VOCs. The result suggests that a relation between the molecular descriptor and the retention times is linear. The QSPR models proposed with the simply calculated molecular descriptors can be used to estimate the chromatographic retention times of new compound even in the absence of the standard candidates. The most important selected descriptors are topological which can capture the variance in the retention times what related to the size and shape of the molecules.

References

1) D.W. Sin, Y.C. Wong, W.C. Sham, D. Wang, Development of an Analytical Technique and Stability Evaluation of 143 C3-C12 Volatile Organic Compounds in Summa Canisters by Gas Chromatography Mass Spectrometry, *Analyst*, 126(3), (2001), 310-321
2) EPA Method 8260C: Volatile organic compounds by gas chromatography/mass spectroscopy (GC/MS), 2006