Middle-East Journal of Scientific Research 13 (8): 993-997, 2013
ISSN 1990-9233
© IDOSI Publications, 2013
DOI: 10.5829/idosi.mejsr.2013.13.8.3521

Nonlinear Chemometrics Model for Prediction Retention Behavior Petroleum Hydrocarbons

Hadi Noorizadeh and Abbas Farmany

Faculty of Science, Islamic Azad University, Ilam Branch, Ilam, Iran

Abstract: Total petroleum hydrocarbons" (TPHs) or "petroleum hydrocarbons" (PHCs) are one of the most widespread soil pollutants in Iran, Canada, North America and worldwide. We performed studies upon an extended series of petroleum hydrocarbons, with retention time (RT), using quantitative structure-retention relationship (QSRR) methods that imply analysis of correlations and representation of models. A suitable set of molecular descriptors was calculated and the genetic algorithm (GA) was employed to select those descriptors that resulted in the best-fit models. The kernel partial least squares PLS (KPLS) was utilized to construct the linear QSRR model. The proposed method will be of importance in this research and could be expected to apply to other similar research fields. This is the first research on the QSRR of the petroleum hydrocarbons compounds against the RT using chemometrics model.

Key words: Petroleum hydrocarbons · Environmental forensics · Biogenic organic compounds · QSRR · Chemometrics

INTRODUCTION

Crude oils consist of complex mixtures of compounds, most of which are hydrocarbons. Petroleum hydrocarbon (PHC) is often used as a general termto describe amixture of various organic com-pounds, mostly pure hydrocarbons, but also including low-polarity hetero-substituted saturates and aromatics, found in crude oil, bitumen and coal. Petroleum hydrocarbons are typically grouped by structure: saturates, olefins, aromatics, polar compounds (a wide variety of compounds containing sulfur, oxygen and nitrogen) and asphaltenes [1]. When PHCs are released to soil they create a wide variety of problems related to their environmental toxicity, risks to human health, mobility and persistence.

Petroleum hydrocarbons have been widely recognized as one of the most widespread soil contaminants in Canada, the USA, Iran and worldwide. About 60% of Canada's thousands of contaminated soil sites, for example, involve PHC contamination [2] which impairs the quality and use of both land and water. Annually, costs for clean-up of PHC-contaminated soils and sediments are very high.

Typical methods for setting these regulations include the use of risk-based exposure and toxicity models [3]. In Canada; many jurisdictions use the PHC guidelines of the Canadian Council of Ministers of the Environment. Under these regulations, PHC values are required by the CCME Reference Method for the Canada-Wide Standard (CWS) for Petroleum Hydrocarbons in Soil – Tier 1 Method [4]. This method, similar to the International Organization for Standardization (ISO) 16703 method for determination of PHC in soil [5], uses solvent extraction, sample clean-up to remove interference components such as humic substances and polar compounds and then gas chromatographic techniques to determine theweights and concentrations of all hydrocarbons in soil based on the carbon ranges. Soils exceeding the CCME criteria must be managed as petroleum-contaminated materials.

In comparison to PHCs, however, the composition and abundances of biogenic compounds in soil are less well understood. Biogenic hydrocarbon is a general term

Corresponding Author: Hadi Noorizadeh, Faculty of Science, Islamic Azad University, Ilam Branch, Ilam, Iran. Tel: + 989183452507.

used to describe the mixture of organic compounds (BOC) such as plant alkanes, sterols and sterones, fatty acids and fatty alcohols and waxes and wax esters, biosynthesized by living organisms. BOCs are also produced during the early stages of diagenesis in recent aquatic sediments [6]. BOC sources include vascular plants, algae, bacteria and animals. Plants and algae produce biogenic hydrocarbons as protectivewax coating that are released back into the sediment at the end of their life cycle. BOCs are natural components of thriving plant communities and are a significant component of many organic soils and sediments.

Of particular interest inmany PHC soil contamination studies is the carbon range C16–C34 fraction (CCME F3), which is regulated atamaximumlevel of 300µg/g for coarse soils and 1300µg/g for fine soils in most of Canada. This fraction contains many of the highly toxic 3- to 5-ring aromatic contaminants, but is also the range in which many of the biogenic organic compounds are also found. In 2005, sediment and plant samples were collected from twenty-nine storm-water management (SWM) ponds and wetlands located in urban municipalities of the Canadian provinces of Alberta, British Columbia, Ontario, Saskatchewan and Manitoba. The study was a joint project of Environment Canada, the Ministry of the Environment of Ontario and twenty-two Canadian municipalities. The SWM ponds and wetlands were designed to improve storm water-runoff quality by gravitational setting and storage of storm-water pollutants in the facility basins. Routine sediment removal is required to maintain treatment efficiencies and flood control capabilities. Removed sediments must be evaluated for PHC contamination in accordance with the previously described CCME PHC analytical protocols. This analytical method, however, is designed to measure total hydrocarbons in soils, ignoring the possible presence or absence of "background" BOCs in soils. Consequently, BOC-enriched soils can be mistakenly identified as petroleum-contaminated materials, causing false PHC soil toxicity criteria exceedances, particularly for the C16–C34 range (Fraction3). This has significant implications regarding SWM facility sediment evaluation and disposal requirements and could trigger unnecessary and costly soil biodegradation and landfill disposal requirements, while also wasting valuable landfill space. In this type of scenario, therefore, it is critically important to distinguish and quantify PHC and BOC in soils or wet sediments. Develop reliable GC method for forensic identification and differentiation of BOC from PHC inmultiply-contaminated soils and sediments and to estimate the content of both PHCs and BOCs in the same sample. To its end, a new reliable GC–MS method, in combination with a derivatization technique, for characterization of various biogenic sterols and other major biogenic compounds such as fatty acids and fatty alcohols has been developed to "fingerprint" BOC distributions in soils.

A multi-criteria approach has been applied to positively identify and distinguish biogenic compounds from petroleum hydrocarbons. Thirteen biogenic sterols, nineteen fatty carboxylic acids and fourteen fatty alcohols in a wide carbon range have been positively identified and quantified in over 30 complex real-world sediment samples collected from municipal SWM ponds and wetlands across Canada [4, 5].

Mathematical modeling of interactions in chromatography helps chemists to find a model that can be used to obtain a deep understanding about the mechanism of interaction and to predict the RT of new or even unsynthesized compounds. Building retention prediction models may initiate such theoretical approach and several possibilities for retention prediction in GC. Among all methods, quantitative structure-retention relationships (QSRR) are the most popular. In QSRR, the retention of given chromatographic system is modeled as a function of solute (molecular) descriptors. A number of reports, dealing with QSRR retention index calculation of several compounds, have been published in the literature [6-8].

There is a trend to develop QSRR from a variety of methods. In particular, genetic algorithm (GA) is frequently used as search algorithms for variable selection in chemometrics and QSRR. GA is a stochastic method to solve the optimization problems defined by fitness criteria, applying the evolution hypothesis of Darwin and different genetic functions, i.e. crossover and mutation [9, 10]. Kernel partial least square (KPLS) is the most commonly used multivariate calibration method [11, 12]. In the present study, GA-KPLS was employed to generate QSRR models that correlate the structure of petroleum hydrocarbons; with observed RT.

Experimental

Data Set: Retention time of the petroleum hydrocarbons which contains 45 compounds was taken from literature [13]. n-Alkanes, PAHs and petroleum biomarkers: Characterizations of n-alkanes, PAHs and petroleum

biomarkers were performed on an Agilent 6890 GC system interfaced to an Agilent 5973 mass spectrometer. An aliquot of the final extract (225µL) was taken and spiked with 25µL of the internal standard solution IS-F1 for GC–MS analyses of n-alkanes, terpanes and steranes, diamondoids and sesquiterpanes, respectively. Another aliquot of the final extract (225µL) was taken and spiked with 25µL of the internal standard terphenyl-d14 solution (10µg/mL) for GC–MS analysis of petroleum-characteristic alkylated PAH homologues and other EPA (US Environmental Protection Agency) priority PAHs. The GC separation of target compounds was achieved using a HP-5MS capillary column (30m×0.25mm I.D., 0.25µm film thickness) with a GC oven temperature program: 50 °C for 2min heated to 300 °C at 6 °C/min and hold for 15min at 300 °C. Samples were injected in splitless mode (injector temperature at 280 °C) with He as carrier gas at the flow rate of 1mL/min. The mass-selective detector (MSD) was operated at an electron impact (70 eV) in the SIM mode. The system control and data acquisition were achieved with the Agilent Enhanced MSD ChemStation. In order to evaluate the generated models, we used leave-group-out cross validation (LGO-CV). This methodology systematically removed one group data at a time from the data set. A QSRR model was then constructed on the basis of this reduced data set and subsequently used to predict the removed data set. This procedure was repeated until a complete set of predicted was obtained.

Descriptor Calculation: All structures were drawn with the HyperChem software (version 6). Optimization of molecular structures was carried out by semi-empirical AM1 method using the Fletcher- Reeves algorithm until the root mean square gradient of 0.01 was obtained. Since the calculated values of the electronic features of molecules will be influenced by related conformation. In the current research an attempt was made to use the most stable conformations. Some electronic descriptors such as polarizability, dipole moment and orbital energies of LUMO and HOMO were calculated by the HyperChem software. Also optimized structures were used to calculate 1497 descriptors by DRAGON software Version 3.

One of the challenging parts in developing models is choosing suitable parameters encoding different aspects of the molecular structure. A large number of structural descriptors can be calculated using existing software's such as Dragon. However, nowadays the main problem is choosing the most adequate and interpretable parameters needed for developing the models among a large number of them. To reduce the original pool of descriptors to an appropriate size, objective descriptor reduction was performed using various criteria. Reducing the pool of descriptors eliminates those descriptors which contribute either no information or whose information content is redundant with that of other descriptors present in the pool. Any descriptor that had identical or zero values for greater than 90% of the compounds was eliminated. These methods reduced the pool of descriptors to 831.

Genetic Algorithm: Genetic algorithm has been proposed by John Holland in the early 1970s but it was possible to apply them with reasonable computing times only since 1990s, when computers became much faster. GA is a stochastic method to solve the optimization problems, defined by fitness criteria applying to the evolution hypothesis of Darwin and different genetic functions, i.e., crossover and mutation. Compared to the traditional search and optimization procedures, GA is robust, global and generally more straightforward to apply to situations where there is little or no a priori knowledge about the process to be controlled. Since GA does not require derivative information or a formal initial estimate of the solution region and because of the stochastic nature of the search mechanism, it is capable to search the entire solution space with a greater probability of finding the global optimum. In GA, each individual of the population, defined by a chromosome of binary values as the coding technique, represented a subset of descriptors. The number of the genes at each chromosome was equal to the number of the descriptors. The population of the first generation was selected randomly. A gene was given the value of one, if its corresponding descriptor was included in the subset; otherwise, it was given the value of zero.

Software and Programs: A Pentium IV personal computer (CPU at 3.06 GHz) with windows XP operational system was used. Geometry Optimization was performed by HyperChem (Version 7.0 Hypercube, Inc.), Dragon software was used to calculate of RI. MLR analysis was performed by the SPSS Software (version 13, SPSS, Inc.) by using enter method for model building. MINITAB software (version 14, MINITAB) was used for the simple PLS analysis. Cross validation, GA-PLS, GA-MLR, L-M ANN and other calculation were performed in the MATLAB (Version 7, Mathworks, Inc.) environment.

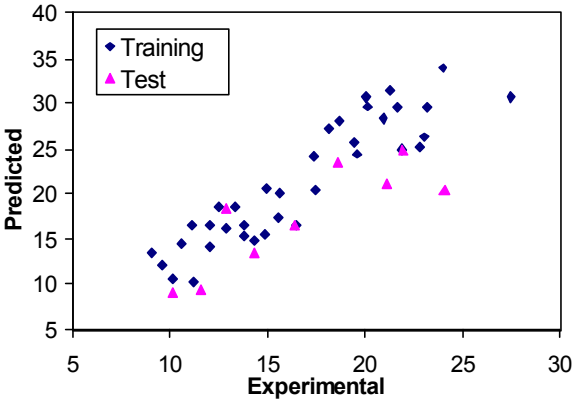

Fig. 1: Predicted vs. experimental RT by GA-KPLS

RESULTS AND DISCUSSION

GA-KPLS Analysis: With the aim of improving the predictive performance of nonlinear QSRR model, GA-KPLS modeling was performed. The leave-group-out cross validation has been performed. In this paper a radial basis kernel function, $k(x,y)= \exp(\|x-y\|^2/c)$, was selected as the kernel function with $c = rm\sigma^2$ where r is a constant that can be determined by considering the process to be predicted (here r set to be 1), m is the dimension of the input space and σ^2 is the variance of the data [14]. It means that the value of c depends on the system under the study. The best model is selected on the basis of the highest square correlation coefficient (R^2) and root mean square error (RMSE) of prediction and simplicity of the model. These parameters are probably the most popular measure of how well a model fits the data. The best GA-KPLS model contains 13 selected descriptors in 5 latent variables space. The R^2 and RMSE for training and test sets were (0.828, 0.763) and (0.780, 0.849), respectively. The Q^2 for training set was 0.801. The predicted values of RT are plotted against the experimental values for training and test set in Fig. 1. Obviously, there is a close agreement between the experimental and predicted RT and the data represent a very low scattering around a straight line with respective slope and intercept close to one and zero. Each of the statistical parameters mentioned above were used for assessing the statistical significance of the QSRR model. The result indicates that the GA-KPLS model have good statistical quality with low prediction error.

The Q^2, which is a measure of the model fit to the cross validation set, can be calculated as:

$$R^2_{cv} \equiv Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y_i})^2} \quad (1)$$

where y_i, $\hat{y}_i$ and $\bar{y}$ were respectively the experimental, predicted and mean RT values of the samples. The accuracy of cross validation results is extensively accepted in the literature considering the Q^2 value. In this sense, a high value of the statistical characteristic ($Q^2 > 0.5$) is considered as proof of the high predictive ability of the model [15]. However, several authors suggest that a high value of Q^2 appears to be a necessary but not sufficient condition for a model to have a high predictive power and consider that the predictive ability of a model can only be estimated using a sufficiently large collection of compounds that was not used for building the model [16].

We believe that applying only LGO-CV is not sufficient to evaluate the predictive ability of a model. Thus we employed a two-step validation protocol which contains internal (LGO-CV) and external (test set) validation methods. The data set was randomly divided into training (calibration and prediction sets) and test sets after sorting based on the RT values. The training set consisted of 36 molecules and the test set, consisted of 9 molecules. The training set was used for model development, while the test set in which its molecules have no role in model building was used for evaluating the predictive ability of the models for external set. Inspection of the results reveals a higher R^2 and lower RMSE for GA-KPLS model for the training and test sets. This clearly shows the strength of GA-KPLS as a nonlinear feature selection method. Result indicates that the RT of petroleum hydrocarbons possesses some nonlinear characteristics.

CONCLUSION

In this study, an accurate QSRR model for estimating the retention time (RT) of petroleum hydrocarbons was developed by employing the one nonlinear model (GA-KPLS). A model has good predictive capacity and excellent statistical parameters. It is easy to notice that there was a good prospect for the GA-KPLS application in the QSRR modeling. This indicates that RT of petroleum hydrocarbons possesses some nonlinear

characteristics. It can also be used successfully to estimate the RT for new compounds or for other compounds whose experimental values are unknown.

REFERENCES

1. Speight, J.S., 1991. The Chemistry and Technology of Petroleum, 2nd ed., MarcelDekker, New York.
2. Treasury Board of Canada Secretariat, the Federal Contaminated Sites Accelerated Action Plan (FCSAAP), 2007, http://www.tbs-sct.gc.ca/rma/eppiibdrp/hrdb-rhbd/fcsaap-paalcf/description e.asp#g1.
3. Vorhees, D., J. Gustafson and W. Weisman, 1999. Total Petroleum Hydrocarbon Criteria Working Group, Vol. 5: Human Health Risk-Based Evaluation of Petroleum Contaminated Sites: Implementation of the Working Group Approach, Amherst Scientific Publishers, MA.
4. CCME, Reference Method for the Canada-Wide Standard (CWS) for Petroleum Hydrocarbons in Soil – Tier 1Method, the Canadian Council ofMinisters of the Environment (CCME), 2007, http://www.ccme.ca.
5. ISO16703, Soil quality – Determination of content of hydrocarbon in the range C10 to C40 by gas chromatography, International Organization for Standardization, Geneva, 2004.
6. D¹browska, M., M. Starek and J. Skuciński, 2011. Lipophilicity study of some non-steroidal anti-inflammatory agents and cephalosporin antibiotics: A review, Talanta, 86: 35-51.
7. D'Archivio, A.A., A. Incani and F. Ruggieri, 2011. Cross-column prediction of gas-chromatographic retention of polychlorinated biphenyls by artificial neural networks, J. Chromatogr A, 1218: 8679-8690.
8. Noorizadeh, H. and A. Farmany, 2010. QSRR Models to Predict Retention Indices of Cyclic Compounds of Essential Oils, Chromatographia, 72: 563-569.
9. Noorizadeh, H. and M. Noorizadeh, 2011. QSRR-based estimation of the retention time of opiate and sedative drugs by comprehensive two-dimensional gas chromatography, Med Chem Res, in press.
10. Van Dijck, G. and M.M. Van Hulle, 2011. Genetic algorithm for informative basis function selection from the wavelet packet decomposition with application to corrosion identification using acoustic emission, Chemom. Intell. Lab. Syst., 107: 318-332.
11. Noorizadeh, H. and A. Farmany, 2011. Quantitative structure-retention relationship for retention behavior of organic pollutants in textile wastewaters and landfill leachate in LC-APCI-MS, Environ. Sci. Pollut Res, in press.
12. Ribeiro, R.J., F. Augusto, TJS. Salva, R.A. Thomaziello and M.C. Ferreira, 2009. Prediction of sensory properties of Brazilian Arabica roasted coffees by headspace solid phase microextraction-gas chromatography and partial least squares, Anal. Chim. Acta, 634: 172-179.
13. Wang, Z., C. Yang, F. Kelly-Hooper, B.P. Hollebone, X. Peng, C.E. Brown, M. Landriault, J. Sun and Z. Yang, 2009. Forensic differentiation of biogenic organic compounds from petroleum hydrocarbons in biogenic and petrogenic compounds cross-contaminated soils and sediments, j. Chromatogr. A, 1216: 1174-1191.
14. Noorizadeh, H., A. Farmanya, H. Narimani and M. Noorizadeh, 2011. QSRR using evolved artificial neural network for 52 common pharmaceuticals and drugs of abuse in hair from UPLC–TOF-MS, Drug. Test. Anal, in press.
15. Tan, A., I.A. I.M. Lévesque, Lévesque, F. Viel and N. Boudreau, 2011. Analyte and internal standard cross signal contributions and their impact on quantitation in LC–MS based bioanalysis, J. Chromatogr. B, 879: 1954-1960.
16. Kim, K., J.M. Lee and I.B. Lee, 2005. A novel multivariate regression approach based on kernel partial least squares with orthogonal signal correction, Chemom. Intell. Lab. Syst., 79: 22-30.