APPLICATION OF RETENTION TIME PREDICTION MODELS FOR SUSPECT AND NON-TARGET HRMS SCREENING OF EMERGING CONTAMINANTS IN THE AQUATIC ENVIRONMENT

AALIZADEH R., THOMAIDIS N.S., BLETSOU A.A. and GAGO FERRERO P.

Laboratory of Analytical Chemistry, Department of Chemistry, National and Kapodistrian University of Athens, Panepistimiopolis 157 71, Athens, Greece
E-mail: ntho@chem.uoa.gr

Quantitative structural-retention time relationship (QSRR) could be effectively used to support confirmation of the presence of suspect and unknown compounds in environmental samples [1]. However, accurate and validated models with large applicability domain should be developed for this purpose. In this regard, we provided a data set of 304 compounds in negative and 528 compounds in positive electrospray ionization mode for the high resolution mass spectrometric screening of suspect and non-target compounds in environmental samples. To building the models, first we optimized their geometrical angle based on molecular mechanic (MM+) and semi-empirical techniques (AM1). Then, the molecular descriptors based on Dragon package were obtained. Subsequently, the data set for each ionization mode was split into training and test set (to externally evaluate the validity of models) using Principle Component Analysis (PCA) and k-Nearest Neighborhood (kNN) clustering. To select the relevant group of molecular descriptors with no inter-correlation and high correlation with retention time, which can explain the chemical behavior of a compound over retention time values, stepwise (SW) and genetic algorithm (GA) were used. Based on the selected descriptors, Multiple Linear Regression (MLR), Artificial Neural Networks (ANN) and Support Vector Machines (SVM) were used to correlate linearly and non-linearly the molecular structures with the experimental retention times. The results of each model were evaluated rigorously by cross-validation analyses and a test set of compounds. The results indicated that the kNN-GA-SVM model was the most accurate one among others for both ionization modes for prediction purposes. The possible presence of outliers was assessed carefully using William plot in order to detect the origin of outliers (chemical structures or reported retention time) in developed models. This model (kNN-GA-SVM) was used further as a diagnostic tool for suspect emerging contaminants, and the results showed that the model is reliable and can be widely used for future HRMS screening applications.

Keywords: QSRR, MLR, SVM, ANN, Suspect Screening

ACKNOWLEDGMENTS

REFERENCES

1. Schymanski EL, Singer HP, Longree P, Loos M, Ruff M, Stravs MA, Ripolles Vidal C, Hollender J (2014), Strategies to Characterize Polar Organic Contamination in Wastewater: Exploring the Capability of High Resolution Mass Spectrometry. Environmental Science & Technology 48 (3):1811-1819. doi:10.1021/es4044374