Wide-scope QSRR models to support suspect and non-target screening of polar compounds in HILIC - ESI(+) - LC-HRMS

R. Aalizadeh and N.S. Thomaidis

Laboratory of Analytical Chemistry, Department of Chemistry, National and Kapodistrian University of Athens, Panepistimiopolis Zografou, 15771 Athens, Greece

E-mail: ntho@chem.uoa.gr

The aquatic environment has been increasingly challenged by the continuous use and discharge of polar compounds, like new generation of pesticides, pharmaceuticals, illicit and new psychoactive drugs and their transformation products or metabolites. Therefore, robust and fast methods with large applicability domain are required to identify these compounds in environmental samples. Nowadays, liquid chromatography coupled with high resolution mass spectroscopy plays a significant role for the identification of polar and mid-polar compounds. Hydrophilic interaction chromatography (HILIC) is an alternative separation technique that is used increasingly for their identification by liquid chromatography – high resolution mass spectrometry (LC-HRMS). However, well-established retention time prediction models with wide applicability domain have not been presented so far to support suspect and non-target HRMS screening of samples. In this work, a large dataset consisted of 685 polar compounds were analyzed by HILIC-LC-QToFMS, and their retention time in positive electrospray ionization mode were derived. Quantitative structure-retention time method was used to correlate their chemical structures with the observed retention times. The molecular descriptors were generated using Dragon and Marvin software, and then, the prepared dataset was split into training and test set based on principle components analysis and k-medoids clustering technique. Genetic algorithm was used for selecting the relative molecular descriptors among the generated variables. After selection of relevant group of descriptors, Kohonen Self-Organizing Maps (SOMs) was used to evaluate the accuracy of classification and also selection. Multiple linear regression (MLR) and support vector machine (SVM) were used as regression tools and then validated by several validation techniques. The applicability domain of the proposed models was studied carefully by a new display of Williams plot and Monte Carlo sampling method. The results indicated that LogD, AlogP, and maximum negative charge that a molecule can carry have major relative importance between modeling variables. This work provides the first, validated and robust model for estimating retention time of compounds in HILIC-LC-HRMS platforms for large number of polar compounds to improve their identification by suspect and non-target screening workflows.