Science of the Total Environment 538 (2015) 934–941

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Suspect screening of large numbers of emerging contaminants in environmental waters using artificial neural networks for chromatographic retention time prediction and high resolution mass spectrometry data analysis

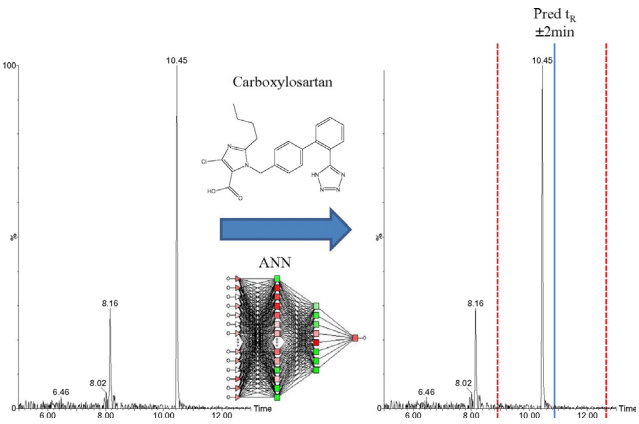
Richard Bade [a], Lubertus Bijlsma [a], Thomas H. Miller [b], Leon P. Barron [b], Juan Vicente Sancho [a], Felix Hernández [a,*]

[a] Research Institute for Pesticides and Water, University Jaume I, Avda. Sos Baynat, E-12071 Castellón, Spain
[b] Analytical & Environmental Sciences Division, Faculty of Life Sciences and Medicine, King's College London, 150 Stamford Street, London SE1 9NH, United Kingdom

HIGHLIGHTS

- A retention time predictor based on artificial neural networks is presented.
- Using 550 compounds, 90% could be predicted within 2 min.
- A notable number of peaks (false positives) could be discarded for further research.
- Useful in the tentative identification of metabolites and transformation products
- This approach highly facilitates wide-scope screening of organic contaminants.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 6 July 2015
Received in revised form 14 August 2015
Accepted 14 August 2015
Available online xxxx

Editor: D. Barcelo

Keywords:
Retention time prediction
Artificial neural networks
Time-of-flight high resolution mass spectrometry
Screening of emerging contaminants

ABSTRACT

The recent development of broad-scope high resolution mass spectrometry (HRMS) screening methods has resulted in a much improved capability for new compound identification in environmental samples. However, positive identifications at the ng/L concentration level rely on analytical reference standards for chromatographic retention time (t_R) and mass spectral comparisons. Chromatographic t_R prediction can play a role in increasing confidence in suspect screening efforts for new compounds in the environment, especially when standards are not available, but reliable methods are lacking. The current work focuses on the development of artificial neural networks (ANNs) for t_R prediction in gradient reversed-phase liquid chromatography and applied along with HRMS data to suspect screening of wastewater and environmental surface water samples. Based on a compound t_R dataset of >500 compounds, an optimized 4-layer back-propagation multi-layer perceptron model enabled predictions for 85% of all compounds to within 2 min of their measured t_R for training (n = 344) and verification (n = 100) datasets. To evaluate the ANN ability for generalization to new data, the model was further tested using 100 randomly selected compounds and revealed 95% prediction accuracy within the 2-minute elution interval. Given the increasing concern on the presence of drug metabolites and other transformation products (TPs) in the aquatic environment, the model was applied along with HRMS data for preliminary identification of

* Corresponding author.
E-mail address: felix.hernandez@uji.es (F. Hernández).

http://dx.doi.org/10.1016/j.scitotenv.2015.08.078

pharmaceutically-related compounds in real samples. Examples of compounds where reference standards were subsequently acquired and later confirmed are also presented. To our knowledge, this work presents for the first time, the successful application of an accurate retention time predictor and HRMS data-mining using the largest number of compounds to preliminarily identify new or emerging contaminants in wastewater and surface waters.

1. Introduction

The number of emerging contaminants in the aquatic environment is increasing, due to urbanization and subsequent societal and industrial needs (Pal et al., 2014). The development of liquid chromatography–high resolution mass spectrometry (LC–HRMS) technologies has revolutionized the analysis of emerging contaminants in environmental waters, and especially for screening of large numbers of compounds (Agüera et al., 2013; Gómez et al., 2010; Hernández et al., 2011; Hogenboom et al., 2009). HRMS instruments allow the recording of full-scan spectra with high mass accuracy and resolution, thus making it possible to search for any given compound based on its exact mass.

There has been much interest in improving the confidence in the identification of small molecules with HRMS; from potential positives through to detection and finally confirmation (Hernández et al., 2015a; Schymanski et al., 2014). The main distinguishing factor between these levels is the (non-) availability of reference standards. Suspect screening refers to compounds tentatively identified based solely on HRMS data and comparable spectral libraries. Confirmation requires reference standards. An additional tool to increase the confidence in the tentative identification of compounds for which standards are unavailable is reliable and accurate t_R prediction. This is of particular relevance in the case of degradation/transformation products (TPs), which can reach the aquatic environment in high concentrations, but commonly for which reference standards are less accessible. Chemical risk assessment is therefore significantly challenging for such compounds.

Prediction of t_R plays an important role in the qualitative identification of emerging contaminants. Many different approaches to t_R prediction exist and range from the simple (Kern et al., 2009; Nurmi et al., 2012) to the complex (Goryński et al., 2013; Ji et al., 2009; Kaliszan et al., 2003; Ukić et al., 2014a). For example, logK_{ow} models can be derived using freely accessible data from chemical databases such as ChemSpider and PubChem, as well as freeware prediction sources such as VCCLABS. Its use in t_R prediction is extremely simple to implement. It is frequently used in environmental studies for the description of the fate of various pollutants and as a simple t_R predictor for TPs (Kern et al., 2009) and emerging contaminants (Bade et al., 2015; Nurmi et al., 2012). Alongside simple algorithms, other and more complex in silico approaches now exist which are based on quantitative structure–retention relationship (QSRR) modeling, including artificial neural networks, support vector machines and random forests (Giaginis and Tsantili-Kakoulidou, 2012; Héberger, 2007). The principal aim of QSRR is to predict retention data from the molecular structure and its physicochemical properties, using a range of input descriptors and measured t_R data. One QSRR method gaining recent attention for broad screening using high resolution techniques is the use of artificial neural networks (ANNs), a predictive computing technique that has shown itself as a promising t_R predictor with potentially higher accuracy than classical models (Miller et al., 2013; Ukić et al., 2014b). The design of ANNs were inspired by the human brain and differ from classical computer programs in that they generally employ non-linear learning techniques using a set of case examples (i.e. a training dataset) (Kaliszan et al., 2003). In the training phase, the ANN requires a range of suitable molecular descriptors as well as the true output value (in this case, measured t_R) to use for comparison with predicted values. At the same time, a second dataset of case examples is often used for verification and to assess overall ANN predictive error. The true output values in the verification set are generally not employed for learning, but the number of training cycles can be stopped by the user or the software when the overall measured error across all cases is at its minimum. Therefore, ANN learning is generally an iterative process and once an acceptable number of training cycles is reached, the optimized ANN can be applied to predict the output where experimentally derived data are unavailable (Miller et al., 2013). In some cases, a third dataset can be used after the model has been finalized to 'blind test' the predictive power of the network. Its use is even more pertinent for analyses where large number of new analytes are expected to occur and with potentially high variance from sample to sample, such as in environmental and municipal water samples. Therefore, since information from the sample includes chromatographic t_R as well as HRMS data, it makes this interpretation of suspect occurrence more accessible in the first instance.

The aim of this work was to develop and evaluate ANN for predictions of unknown chromatographic t_R in suspect screening of environmental waters. To the best of our knowledge, this method includes the largest range of physicochemically diverse compounds for this purpose ($n = 544$ in total) and includes both neutral and charged compounds eluted under gradient reversed-phase LC conditions. Lastly, this work aimed to improve upon a recent logK_{ow}-based t_R prediction approach (Bade et al., 2015) using the ANNs as an alternative. This work, for the first time, presents the use of ANN for identification of additional suspect compounds (including metabolites and TPs) in wastewater and surface water samples both with and without reference standards.

2. Experimental

2.1. Reagents and chemicals

A total of 544 analytical grade reference materials were used for preparation of model solutions at 25 μg/L or 50 μg/L (diluted from mixed standard solutions in methanol or acetonitrile with water) for ANN modeling of t_R. These included pesticides, drugs of abuse, human/veterinary pharmaceuticals and mycotoxins (See Supplementary Information (SI) Table S1 for all compounds used in this study). These covered a large range of molecular hydrophobicity (logK_{ow} −3 to 9). Information relating to 595 standards was available (Bade et al., 2015), however after transforming the compounds using SMILES codes, some errors were observed, leading to incomplete data, and a further 42 were removed from the initial ANN method development (Section 3.1) to use in a subsequent blind test (Section 3.2). Further details relating to these compounds can be found elsewhere (Bade et al., 2015; Hernández et al., 2015b).

2.2. Water samples for suspect identification

A total of 44 composite (24-h) influent and effluent wastewater (IWW and EWW) samples and grab surface water (SW) samples were used to demonstrate the application of the developed ANN model. All these samples were previously used in different studies performed at our lab using the same analytical instrumentation for analysis (Hernández et al., 2015a). All measured t_R data herein were generated using ultra-high pressure liquid chromatography coupled to quadrupole-time of flight mass spectrometry (UHPLC–QTOF-MS).

2.3. UHPLC-QTOF MS

A Waters Acquity UPLC system (Waters, Milford, MA, USA) was interfaced to a hybrid quadrupole-orthogonal acceleration-TOF mass