Available online at www.sciencedirect.com
ScienceDirect
Analytica Chimica Acta 598 (2007) 12–18
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

Quantitative structure–retention relationships for organic pollutants in biopartitioning micellar chromatography

Binbin Xia [a], Weiping Ma [a], Xiaoyun Zhang [a,*], Botao Fan [b]

[a] *Department of Chemistry, Lanzhou University, Lanzhou 730000, Gansu, PR China*
[b] *Université Paris 7-Denis Diderot, ITODYS 1, Rue Guy de la Brosse, 75005 Paris, France*

Received 16 March 2007; received in revised form 9 June 2007; accepted 5 July 2007
Available online 30 July 2007

Abstract

Quantitative structure–retention relationship (QSRR) models have been successfully developed for the prediction of the retention factor ($\log k$) in the biopartitioning micellar chromatography (BMC) of 66 organic pollutants. Heuristic method (HM) and radial basis function neural networks (RBFNN) were utilized to construct the linear and non-linear QSRR models, respectively. The optimal QSRR model was developed based on a 6-17-1 radial basis function neural network architecture using molecular descriptors calculated from molecular structure alone. The RBFNN model gave a correlation coefficient (R^2) of 0.8464 and root-mean-square error (RMSE) of 0.1925 for the test set. This paper provided a useful model for the predicting the $\log k$ of other organic compounds when experiment data are unknown.

Keywords: Quantitative structure–retention relationship; Heuristic method; Radial basis function neural networks; Biopartitioning micellar chromatography

1. Introduction

In the last two decades, environmental pollution has far reaching negative consequences in the lives of humans. Various kinds of organic pollutants have been detected in the environment. Among these, benzene derivatives, phenolic derivatives and organic acids are important environmental contaminants because of their high toxicity and widespread occurrence, and these organic pollutants have the capability of long-distance transfer, precipitation and accumulation in environment [1]. They affect the growth and decay of plants, and the health of human beings and all animals. The adverse effect on human health, some of which are highly detrimental, has already been documented elsewhere [2–4]. They can elicit various toxicities, such as dermal lesion, body weight loss, hepatotoxicity, immunossuppression, reproductive and developmental toxicity, endocrine disruption, neurotoxicity, and carcinogenicity. Moreover, they have begun to threaten biological diversity and ecosystem integrity [5]. As they are important items of many industries and agriculture, it is not practically feasible to eliminate these from the environment. Therefore, it is very important to evaluate the toxicity of organic pollutants for risk assessment.

However, it is very difficult to evaluate the toxicity of organic pollutants in vivo. For a substance to provoke a biological response when administered to an organism, a number of processes must occur such as absorption, transport and distribution [6]. Foremost, the substances must have access to the organism, and then internally dissolve in a body fluid. Finally, it must be transported to the site of action or receptor where it must bind at a certain concentration to produce biological response. However, biopartitioning micellar chromatography (BMC) is helpful in constructing good models to describe the biological behavior of different kinds of compounds because the characteristics of the BMC systems are similar to biological barriers and extracellular fluids.

Mostly, polioxyethylene (23) lauryl ether (Brij35) micellar mobile phases and C_{18} reversed stationary phase comprise the BMC. As a mode of micellar liquid chromatography, BMC is a simple and reproducible approach in emulating the partitioning of chemicals in biomembranes [7]. The reason could be generalized as follows: first, the stationary phase modified by hydrophobic adsorption of Brij35 surfactant monomers structurally resembles the ordered array of the membranous hydrocarbon chains. Second, the hydrophilic/hydrophobic char-

* Corresponding author. Tel.: +86 931 891 2578; fax: +86 931 891 2582.
E-mail address: xyzhang@lzu.edu.cn (X. Zhang).

0003-2670/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2007.07.016

acter of the adsorbed surfactant monomers resembles the polar membrane regions. In addition, Brij35 micellar mobile phases prepared at the specific physiological conditions could also mimic the environment of partitioning [8]. So the retention data in BMC can not only reflect the penetration of the chemicals in organic tissues but also indicate the degree of ecotoxicity indirectly.

Although there are many experimental techniques developed to determine the retention data, quantitative structure–retention relationship (QSRR) provides a promising method for the estimation of the retention based on the descriptors calculated from the molecular structure [9]. QSRR, as an extension of linear free energy relationships (LFERs), has been successfully employed for the prediction of chromatographic behavior [10–13]. The main steps involved in QSRR include the following: data collection, molecular descriptors obtaining and selection, correlation model development and finally model evaluation.

In this work, the advantage of QSRR over other methods lies in the fact that the descriptors used to build the models can be calculated from the structure alone and are not dependent on any experimental properties. At present, many types of molecular descriptors such as constitutional, topological, geometrical, electrostatic and quantum chemical descriptors have been proposed to describe the structural features of molecules [14–16]. One of the important problems is selection of the descriptors that can represent the retention mechanism most exactly. After the calculation of the molecular descriptors, linear methods, such as multiple linear regression (MLR), principal component regression (PCR), partial least squares (PLS) or non-linear methods, such as different types of artificial neural networks (ANN) and genetic algorithms (GA), can be employed to derive correlation models between the molecular structures and retention.

Recently, neural networks have gained great popularity in QSAR/QSPR research due to their flexibility in modeling non-linear problems [17]. They have been applied to predict many physicochemical properties. The radial basis function neural network (RBFNN) has advantages of small training times and free parameters to be adjusted by fast linear methods. The optimization of its topology and learning parameters is easy to implement [18]. Many problems in chemistry and chemical engineering have been successfully solved by the use of RBFNN [19–23]. Nevertheless, to the best of our knowledge, there is no prediction of retention factor ($\log k$) of BMC by the QSRR approach based on RBFNN.

In the present work, the heuristic method (HM) and RBFNN were used to predict $\log k$ of 66 organic pollutants in BMC using descriptors calculated and selected by the software CODESSA. The aim was to establish a robust QSRR model that could be used for the prediction of $\log k$ of wide variety of pollutants, to show the flexible modeling ability of RBFNN, and, at the same time, to seek the important structure features related to the retention behavior of pollutants.

2. Methodology

2.1. Data set

In this study, the data set was taken from the reference [6]. The 66 organic pollutants belong to different families and many of them are phenolic derivatives and organic acids containing chlorine and sulfur atoms. A complete list of the pollutants' names and their corresponding experimental retention data ($\log k$) was given in Table 1.

The entire set of organic pollutants was divided into two subsets randomly—a training set of 44 compounds and a test set of 22 compounds. The training set was used to build the models and the test set was used to evaluate the models once they were built.

2.2. Descriptor generation

All structures of the organic pollutants were drawn with the HyperChem 4.0 program [24]. The pre-optimized of all molecules were performed using MM+ molecular mechanics force field. A more precise optimization was done with the semi-empirical PM3 method in HyperChem. The molecular structures were optimized using the Polak-Ribiere algorithm until the root mean square gradient was 0.01. The output files were transferred into the Microsoft Windows version of the CODESSA program developed by Katritzky et al. [25], to calculate molecular descriptors. Five types of molecular descriptors were calculated: constitutional, topological, geometric, electrostatic and quantum-chemical descriptor.

2.3. Theory of heuristic method

After calculating a large number of descriptors, a feature selection step was carried out to reduce the large set of descriptors to a suitable number without losing any important information [26]. The HM in CODESSA was employed

Table 1
The selected descriptors, regression coefficients and *t*-values for the linear model

Descriptor	Chemical meaning	Coefficient	DX	*t*-test
Constant	Intercept	−0.0032	0.1883	−0.0170
KHI	Kier Hall index (order 0)	0.1441	0.0131	11.0200
MNACc	Max net atomic charge for a C atom	−1.9554	0.2897	−6.7500
HDCA-1	HA dependent HDCA-1 [Zefirov's PC]	−0.1348	0.0460	−2.9310
HDSA-2/TMSA	HA dependent HDSA-2/TMSA [quantum-chemical PC]	5.3329	3.5634	1.4970
RNS	Relative number of S atoms	−6.9348	2.4311	−2.8530
RPCG	Relative positive charge (QMPOS/QTPLUS) [Zefirov's PC]	2.8931	0.6260	4.6210