Chemometrics and Intelligent Laboratory Systems 151 (2016) 219–227

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Extending proteochemometric modeling for unraveling the sorption behavior of compound–soil interaction

Watshara Shoombuatong[a], Sunanta Nabu[a], Saw Simeon[a], Virapong Prachayasittikul[b], Maris Lapins[c], J.E.S. Wikberg[c], Chanin Nantasenamat[a,*]

[a] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok10700, Thailand
[b] Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok10700, Thailand
[c] Department of Pharmaceutical Biosciences, Uppsala University, Uppsala751 24, Sweden

ARTICLE INFO

Article history:
Received 24 August 2015
Received in revised form 29 December 2015
Accepted 4 January 2016
Available online 14 January 2016

Keywords:
Phthalic acid esters
Soil sorption
Quantitative structure–property relationship
Proteochemometrics
Data mining

ABSTRACT

Contamination of ground water by industrial chemicals presents a major environmental and health problem. Soil sorption plays an important role in the transport and movement of such pollutant chemicals. In this study, proteochemometric (PCM) modeling was used to unravel the origins of interactions of 17 phthalic acid esters (PAEs) against 3 soil types by predicting the organic carbon content normalized sorption coefficient ($\log K_{oc}$) values as a function of fingerprint descriptors of 17 PAEs and physical and textural properties of 3 soils. The results showed that PCM models provided excellent predictivity ($R^2 = 0.94, Q^2 = 0.89, Q^2_{Ext} = 0.85$). In further validation of the model, our proposed PCM model was assessed by leave-one-compound-out ($Q^2_{LOCO} = 0.86$) and leave-one-soil-out ($Q^2_{LOSO} = 0.86$) cross-validations. The transparency of the PCM model allowed interpretation of the underlying importance of descriptors, which potentially contributes to a better understanding on the outcome of PAEs in the environment. A thorough analysis of descriptor importance revealed the contribution of secondary carbon atoms on the hydrophobicity and flexibility of PAEs as significant properties in influencing the soil sorption capacity.

1. Introduction

Phthalic acid esters (PAEs) are widely used as plasticizers [1] and defoaming agents [2] in industrial production. Owing to their aromatic nature, PAEs are commonly used in consumer products [3]. In spite of their wide usage, they are nonetheless, hazardous to human health. Because PAEs are not covalently bound to plastic polymers, they can be easily leached to the environment during and after the manufacturing process. Several studies have shown that PAEs disturb the hormonal system as well as vital organs [4,5]. Owing to its negative impacts on mammal health, understanding the soil sorption behavior of PAEs has become an important environmental issue.

Soil sorption is the process of removing solute from an undersaturated solution to a solid phase, and as such plays a dominant role in determining the destiny of PAEs in the ecosystem [6]. The soil sorption coefficient, otherwise known as the normalized soil organic carbon-water partition coefficient ($\log K_{oc}$), is a valuable parameter for assessing the soil sorption behavior. Two factors that govern the soil sorption behavior of PAEs include the physiochemical properties of PAEs and the properties of soils [7]. Particularly, several studies have shown that physicochemical properties of PAEs affect soil sorption such as the alkyl side chain length [8] and polarity [9] on soil sorption of PAEs. As for soils, low carbon content and pH [10], low salinity of soils [11] and the presence of dissolved organic matter (DOM) [12] have all been shown to decrease soil sorption.

Quantitative-structure activity/property relationship (QSAR/QSPR) is a powerful and robust approach for predicting the biological activity and chemical property as a function of molecular descriptors for compounds of interest [13,14,15,16]. QSPR modeling has been extensively used for predicting the soil sorption behavior of compounds [17,18,19,20,21,22,23,24,25,26,27,28]. However, the inherent limitation of these constructed QSPR models is that they only take the physicochemical property of the compounds into consideration while not considering soil properties.

Recently, Yang et al. [28] proposed a QSPR model using multiple linear regression and descriptors derived from soil sorption behavior whereby the response value was obtained by averaging the $\log K_{oc}$ from the three types of soil under investigation. This work provided an acceptable predictive result. However, it is well known that conventional QSAR [28] could not simultaneously analyze the information for a series of compounds against a series of soils. In spite of this, there are ample opportunities for further improvements

* Corresponding author.
E-mail address: chanin.nan@mahidol.ac.th (C. Nantasenamat).

http://dx.doi.org/10.1016/j.chemolab.2016.01.002
0169-7439/

by finding ways to unravel the origins of interactions of 17 phthalic acid esters (PAEs) against 3 soil types by predicting the logK_{oc} values.

Proteochemometrics (PCM) is a computational approach that can alleviate these limitations because it is capable of quantifying the relationship of several ligands and several proteins in one unified model [29,30]. PCM had previously been employed to study the interaction of ligands against a wide range of proteins including among others the melanocortin receptors [31], G-protein coupled receptors [32], HIV-1 proteases [33], major histocompatibility complex proteins [34], cytochrome P450 enzymes [35], protein kinases [36] and green fluorescent proteins [37].

This study extends the use of PCM from the typical protein–ligand space onto the compound-soil space and in doing so allows the simultaneous consideration of all 3 soil types in a single unified model instead of three separate QSAR models as previously proposed. The response variable logK_{oc} was predicted as a function of the properties of 17 PAEs and 3 soils. Important features were identified and used to provide insights into the underlying basis of soil sorption behavior.

2. Materials and methods

2.1. Dataset

A dataset describing the sorption behavior of 17 PAEs against 3 soil types was obtained from the work of Yang et al. [28] as shown in Fig. 1. The PAEs are based on the 1,2-benzenedicarboxylic acid chemotype containing different substituents at the ortho position. Soils were from different regions of China, namely NanChang Honggutan (NH), NanJing Xianlin (NX), and JaiXing Xiuzhou (JX). The soil organic carbon content normalized sorption coefficients (logK_{oc}) were determined by batch equilibration experiments to approximate the behavior of soil sorption. The dataset used in this study is publicly available on figshare at https://dx.doi.org/10.6084/m9.figshare.2058933.v1.

2.2. Description of compounds

As fingerprints are easy to calculate, informative and interpretable [38], a set of 307 substructure fingerprint count was calculated using the PaDEL software [39] to represent the 17 PAEs. In situations when two or more descriptors were highly correlated by more than 0.9, only one of them was retained. Finally, the remaining descriptors (Fig. 2) consisting of SubFPC2 (secondary carbon), SubFPC307 (chiral center specified), SubFPC295 (C ONS bond), SubFPC3 (tertiary carbon) and SubFPC5 (alkene) were further used for multivariate analysis.

2.3. Description of soils

Yang et al. [28] previously determined the following 8 properties for the 3 soil types in their investigation on the sorption behavior of PAEs: organic carbon content (SOC), pH, clay content (clay), cation exchange capacity (CEC), sand content (sand), silt content (silt), soil nitrogen content (TN), and moisture (MC). These properties are used in this study as soil descriptors.

2.4. Computation of interaction cross-terms

Descriptor blocks for both compounds (C) and soils (S), as derived from the aforementioned section, are comprised of 5 and 8 descriptors, respectively. Cross-terms for compound–soil interaction ($C \times S$) were obtained by computing the products of compound and soil descriptors; thereby giving rise to $5 \times 8 = 40$ cross-terms.

In addition, cross-terms for self-interaction of compounds ($C \times C$) and soils ($S \times S$) were computed according to the lower triangular matrix as described below:

$$\frac{N \times (N-1)}{2} \tag{1}$$

where N is the number of compound or soil descriptors. Applying the above equation resulted in $8 \times (8-1) \times 0.5 = 28S \times S$ cross-terms. Meanwhile, after removing non-informative descriptors (having standard deviation equal to zero) from the total number of cross-terms for self-interaction of compounds resulted in 4 $C \times C$ cross-terms.

2.5. PCM modeling

Five descriptor blocks (e.g., C, S, $C \times S$, $C \times C$ and $S \times S$) containing a total of 85 descriptors were utilized for multivariate analysis. These descriptors were subjected to mean centering followed by scaling to unit variance.

The PCM modeling performed herein is based on partial least squares (PLS) regression. PLS establishes the correlation between the matrix of predictors or independent variables **X** (i.e., all descriptors and cross-terms) that have high variance and great correlation with the response variable **Y** (logK_{oc}). The approximation of the correlation is achieved by simultaneously projecting the **X** and **Y** matrices onto lower dimensional spaces that are represented by PLS components. More details of PLS modeling is provided elsewhere [40,41].

A total of 10 PCM models for predicting the logK_{oc} value was formulated using various combinations of descriptor blocks as summarized in Tables 1 and 2. The constructed model using all descriptor blocks can be expressed as follows:

$$\log K_{oc} = \sum_{c=1}^{5}(coeff_c \times x_c) + \sum_{c=1}^{8}(coeff_s \times x_s) + \sum_{c=1,s=1}^{40}(coeff_{c,s} \times x_c \times x_s) + \sum_{c1=1,c2=1}^{4}(coeff_{c1,c2} \times x_{c1} \times x_{c2}) + \sum_{s1=1,s2=1}^{28}(coeff_{s1,s2} \times x_{s1} \times x_{s2}) + \epsilon \tag{2}$$

where ϵ is the intercept term. Owing to size heterogeneity for each descriptor block and cross-term, block-scaling $\left(\frac{1}{\sqrt{N}}\right)$ was applied on the five descriptor blocks in order to avoid a situation where the block having the largest number of descriptors outweighs the small ones. The PCM models were implemented using the R package plsdepot [42].

2.6. Validation of PCM models

Validation of predictive models are crucial for any empirical modeling. The goodness-of-fit and goodness-of-prediction are commonly used to evaluate the robustness of a PCM model. The former is characterized by the coefficient of determination (R^2_{Tr}) and root mean square error ($RMSE_{Tr}$) while the latter is characterized by the coefficient of determination (Q^2_{CV}) and root mean square error ($RMSE_{CV}$) where both can be obtained from cross-validation [14]. Although, a high Q^2_{CV} value is frequently used as one of the criterion for robust models, it is not a definitive condition for obtaining a robust model [43]. External validation is important for assessing the ability of any models to afford predictions for unknown data (i.e., uncharacterized compound) [43]. Thus, the dataset in this study was divided into internal (75%) and external (25%) sets where the former was used to construct a model and subjected to a conventional five-fold cross-validation (5-fold CV) while the external set was used to externally assess the model using Q^2_{Ext} and $RMSE_{Ext}$. The external and internal sets were subjected to 20 rounds of random splits. To further