Chemosphere 152 (2016) 190–195

Contents lists available at ScienceDirect

Chemosphere

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

Quantitative modeling of bioconcentration factors of carbonyl herbicides using multivariate image analysis

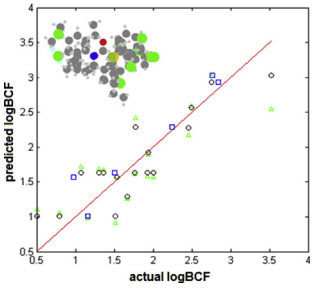
Mirlaine R. Freitas, Stephen J. Barigye, Joyce K. Daré, Matheus P. Freitas*

Department of Chemistry, Federal University of Lavras, P.O. Box 3037, 37200-000, Lavras, MG, Brazil

HIGHLIGHTS

- Image-based descriptors describe the bioconcentration factors of carbonyl herbicides.
- A three-parameter linear equation yielded a predictive and validated QSPR model.
- Interpretation of built model provided information on structural features affecting BCF.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 21 October 2015
Received in revised form
3 March 2016
Accepted 4 March 2016
Available online xxx

Handling Editor: I. Cousins

Keywords:
Bioconcentration
Carbonyl herbicides
QSPR
Image analysis

ABSTRACT

The bioconcentration factor (BCF) is an important parameter used to estimate the propensity of chemicals to accumulate in aquatic organisms from the ambient environment. While simple regressions for estimating the BCF of chemical compounds from water solubility or the *n*-octanol/water partition coefficient have been proposed in the literature, these models do not always yield good correlations and more descriptive variables are required for better modeling of BCF data for a given series of organic pollutants, such as some herbicides. Thus, the logBCF values for a set of carbonyl herbicides comprising amide, urea, carbamate and thiocarbamate groups were quantitatively modeled using multivariate image analysis (MIA) descriptors, derived from colored image representations for chemical structures. The logBCF model was calibrated and vigorously validated ($r^2 = 0.79$, $q^2 = 0.70$ and $r^2_{test} = 0.81$), providing a comprehensive three-parameter linear equation after variable selection (logBCF = 5.682 − 0.00233 × **X9774** − 0.00070 × **X813** − 0.00273 × **X5144**); the variables represent pixel coordinates in the multivariate image. Finally, chemical interpretation of the obtained models in terms of the structural characteristics responsible for the enhanced or reduced logBCF values was performed, providing key leads in the prospective development of more eco-friendly synthetic herbicides.

1. Introduction

The excessive use of herbicides in agriculture can be an important cause of health risks and hazards in animal and human life (Corsonlini et al., 2005). Some herbicides are persistent organic pollutants and, therefore, they can be strongly sorbed in soil and accumulate in the trophic food chain. On the other hand, herbicides

Abbreviations: BCF, Bioconcentration factor; MLR, Multiple Linear Regression; LOOCV, Leave-one-out Cross-validation; MIA, Multivariate Image Analysis; RGB, Red-Green-Blue; RMSE, Root Mean Square Error; QSPR, Quantitative Structure-Property Relationships.

* Corresponding author.
E-mail address: matheus@dqi.ufla.br (M.P. Freitas).

http://dx.doi.org/10.1016/j.chemosphere.2016.03.011
0045-6535/

can also reach water resources, and thus affecting the aquatic life, due to their propensity to be absorbed through the aquatic organisms' respiratory and dermal surfaces (Gobas and Morrison, 2000). This process is denominated as bioconcentration and is quantified in terms of the equilibrium ratio of the concentration of the chemical substance in the organism to that in the surrounding aquatic environment, known as the bioconcentration factor (BCF) (Mackay, 1982; Chiou et al., 1977; Arnot and Gobas, 2006). Due to the enormous costs incurred in performing standard bioconcentration tests as established by OECD 305 (OECD, 1996), many efforts have been invested in developing *in silico* QSPR models for estimating the BCFs of chemical compounds. Several *in silico* BCF models have been proposed in the literature ranging from simple regression equations using the physicochemical parameters solubility or *n*-octanol/water partition coefficient to more complex models based on topological molecular descriptors and advanced statistical and machine learning methods (Devillers et al., 1998). However, each approach comes along with inherent advantages and limitations. While simple and interpretable models are desirable, they usually possess lower predictive power and robustness and vice versa. The present manuscript is an attempt to strike a tradeoff between the two extremes, using the Multivariate Image Analysis applied to quantitative structure-property relationships (MIA-QSPR) method approach, with the ultimate goal of providing key leads on the structural features necessary for developing environmentally friendly herbicides with a reduced tendency to bioconcentrate in aquatic organisms.

The multivariate image analysis applied to quantitative structure-property relationships (MIA-QSPR) method has proved to be a successful technique in the prediction of the soil sorption of herbicides when this parameter does not adequately correlate with the octanol/water partition coefficient (Freitas et al., 2014, 2015). This is an alignment-based method centered on the notion that 2D images of chemical structures contain important topostructural and topochemical information, which may be employed in explaining the variance in physicochemical, chemical and biological properties of molecules (Freitas et al., 2005, 2008). From a digital image processing perspective, images are simply an array (or grid) of pixels and, therefore, in order to extract relevant chemical information from the molecular structure images, these pixels (*i.e.* their numeric values and coordinate positions) are considered as descriptors for the corresponding chemical structure.

The pioneering MIA-QSPR framework considered binary (black and white) wire-frame graphs of chemical structures. However, despite the reasonably good correlations obtained in modeling experiments with this framework, it exhibited redundancy due to the inability to discriminate heteroatoms/atom types or to use atomic property weighting schemes. Consequently, the **aug-MIA-QSPR** (acronym of **aug**mented **MIA-QSPR**) approach was introduced, distinguished by the incorporation of information on atom types and groups, atomic size, as well as different atomic properties (such as van der Waals radii and electronegativity), by means of different color schemes carefully defined according to the RGB system of colors. This method, which was recently implemented in QSAR and QSPR experiments (aug-MIA-QSAR/QSPR), yields more robust and, most importantly, interpretable models (Nunes and Freitas, 2013; Duarte et al., 2015a, 2015b). For a detailed review on this method, see Barigye and Freitas (2016).

Carbonyl herbicides, such as amides, ureas, carbamates and thiocarbamates are extensively used in the agricultural industry (Grover and Cessna., 1990). For instance, thiocarbamates have been shown to form reactive sulfoxide and sulfone intermediates, which may be involved in toxic reactions through covalent modification of cysteine and serine sites of enzymes (Zimmerman et al., 2004). Thus, it is important to comprehend the structural characteristics responsible for increased or decreased logBCF values. In this sense, the aim of the present report is to build aug-MIA-QSAR based models for the logBCF of a series of carbonyl herbicides to provide insight on structural moieties responsible for the observed BCF profiles within the series of compounds.

2. Materials and methods

The experimental logBCF values of 25 carbonyl herbicides pertaining to the amide, urea, carbamate and thiocarbamate classes were obtained from the literature (see Table 1 for identity and Supplementary Material for the corresponding structures) (Higgins et al., 2010; Grisoni et al., 2016; Mackay et al., 1997; Bermúdez-Saldaña et al., 2005; Jackson et al., 2009; Asano et al., 1989; Vogs et al., 2015).

The MIA-QSPR methodology is a simple and computationally cheap procedure briefly defined as follows: for a given dataset of chemical compound images, these are aligned with respect to a common basic scaffold, and then each image (grid of pixels) is unfolded into a single row pixel values, and thus the entire dataset yields a bi-dimensional matrix of descriptors which is then employed to explain the variance of the modeled property (Freitas et al., 2014, 2015, 2005, 2008; Nunes and Freitas, 2013; Duarte et al., 2015a, 2015b; Barigye and Freitas, 2016). In this sense, to generate the aug-MIA descriptors for carbonyl herbicides employed in the present study, firstly, the chemical structures of the herbicides were drawn using the GaussView program (Dennington et al., 2008) and saved separately as bitmap files, keeping the common substructure (the carbonyl group) aligned. Fig. 1 shows the 25 chemical images superposed to give insight on the structural variance that explains the variance in the logBCF block. The chemical structures were drawn by considering atoms as spheres with sizes proportional to the corresponding van der Waals radii. In addition, the atom types were colored differently to distinguish them, since different numbers are assigned to each color pixel, consistent with the RGB system of colors. According to the RGB model, the entire color spectrum is obtained from the contribution of red (255), green (255) and blue (255) components, thus varying from 0 (black, absence of color) to 765 (white, the sum of all three original components). In this sense, the images in this series of compounds comprised the following pixel values: 765 (blank space, white), 688 (fluorine, light blue), 615 (chemical bond, light grey), 612 (hydrogen, light grey), 493 (sulfur, yellow), 426 (carbon, grey), 289 (chlorine, green), 279 (nitrogen, blue) and 229 (oxygen, red). Since images are simply 2D grids of pixels, each image was expressed as a matrix of pixel values and unfolded to form a row vector, therefore, yielding a 25 row dataset matrix. Fig. 2 summarizes the procedure followed in generating the descriptors according to the aug-MIA-QSAR strategy. In general, the MIA-QSPR approach yields thousands of descriptors. However, many of these descriptors are in fact redundant given that they do not usually vary for a given series of compounds (particularly due to blank spaces or identical substructures), are strongly correlated, or simply do not explain the variance in the Y-block, and are thus not exactly useful in modeling tasks. In this sense, the obtained data matrix was pruned by removing columns with zero variance (representing blank spaces and identical substructures) and using a step-wise variable selection procedure, based on supervised and unsupervised filters (Urias et al., 2015).

The chemical dataset was split into training (*ca.* 75%, 19 compounds) and test sets (*ca.* 25%, 6 compounds) using random sampling, and subsequently a search for the best 3 variable model for the logBCF was performed using the Multiple Linear Regression method coupled with the Genetic Algorithm (MLR-GA). The built QSPR model was validated using leave-one-out cross-validation