Geoderma 268 (2016) 1–6

Contents lists available at ScienceDirect

Geoderma

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

Aug-MIA-SPR/PLS-DA classification of carbonyl herbicides according to levels of soil sorption

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

ARTICLE INFO

Article history:
Received 12 August 2015
Received in revised form 6 January 2016
Accepted 13 January 2016
Available online 21 January 2016

Keywords:
Carbonyl herbicides
Soil sorption
Aug-MIA-SPR
PLS-DA

ABSTRACT

A major challenge in the design of new herbicides lies in the development of highly active, environmentally friendly compounds. Soil sorption is an ecotoxicological parameter used to probe the prospective environmental fate of persistent organic pollutants, such as some herbicides. This parameter, described in terms of logK_{OC} (the logarithm of the soil/water partition coefficient normalized to organic carbon), is usually estimated using the octanol/water partition coefficient (logP, easily calculated or determined experimentally). However, estimations obtained with the logP are not always accurate. Thus, this work reports the use of molecular descriptors derived from multivariate image analysis of carbonyl herbicides to achieve a predictive classification model based on the partial least squares-discriminant analysis (PLS-DA) method. This model yields 80% accuracy in calibration, 75% in leave-one-out cross-validation and 100% in external validation. In addition, the Y-randomization test reveals that the obtained model is stable from fortuitous correlation, since the accuracy in calibration after shuffling the classes block is only 0.5%. Chemical interpretation in terms of the structural features that affect soil sorption is performed, based on the weights of the selected variables in the classification model. Finally, novel herbicides are rationally designed, based on the inferences arrived at in the structural interpretation experiment and predictions of their qualitative and quantitative soil sorption profiles performed, using the built aug-MIA-SPR and Wang's models, respectively.

1. Introduction

Herbicides play an important role in weed control, but may persist in the environment and, consequently, accumulate in the food chain, causing serious health problems (Corsonlini et al., 2005). While some herbicides accumulate in water, thus affecting principally aquatic organisms through bioconcentration (Kehrig et al., 2011), others interact strongly with soil, presenting high soil sorption, which is described in terms of the logarithm of the soil/water partition coefficient normalized to organic carbon — logK_{OC}. This physicochemical parameter can be estimated using the easily calculable or measurable octanol/water partition coefficient, logP. However, univariate models based on logP for the prediction of logK_{OC} have not shown to be useful in the prediction of the soil sorption of many classes of herbicides (Sabljić et al., 1995; Freitas et al., 2014a, 2014b). Thus, more descriptive variables should be used to estimate/predict the logK_{OC} of chemical structures with herbicidal activity.

In this sense, the multivariate image analysis (MIA) method has emerged in the last decade as a successful source of descriptors, which have been used to model bioactivities and physicochemical parameters of a variety of compounds (Antunes et al., 2008; Nunes and Freitas, 2013a; Goodarzi et al., 2009; Goodarzi and Freitas, 2008, 2009). The descriptors in MIA are pixels (numerical data) of superimposable bidimensional chemical images (of a congeneric series of molecules), whose variance from an image to another (in terms of pixel coordinates representing different groups or substituent positions) explains the fluctuation in the response for a series of compounds. The MIA descriptors have been upgraded from molecules drawn as wireframes (Freitas et al., 2005) to colored and more realistic chemical representations (Nunes and Freitas, 2013b). Such a scheme has been recently used to model the soil sorption of aromatic herbicides, allowing for the comprehension of the structural characteristics responsible for high/low soil sorption (Freitas et al., 2015).

Recently, the so called aug-MIA descriptors have been extended to classification tasks using principal component analysis (PCA), hierarchical cluster analysis (HCA) and partial least squares-discriminant analysis (PLS-DA), giving rise to predictive aug-MIA-SAR models (Duarte et al., 2015). This qualitative analysis can be useful when quantitative response variables are not accessible or significantly inaccurate (in which case averaged values yield relatively high residuals), rendering regression for quantitative analysis impractical. In the present aug-MIA-SPR modeling, carbonyl herbicides (amides, carbamates, thiocarbamates and ureas) were evaluated because of their wide range of applications and the poor

Abbreviations: Aug-MIA-SPR, augmented multivariate image analysis applied to structure–property relationships; PLS-DA, partial least squares-discriminant analysis; logK_{OC}, logarithm of the soil/water partition coefficient normalized to organic carbon; LOOCV, leave-one-out cross-validation.

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

http://dx.doi.org/10.1016/j.geoderma.2016.01.013
0016-7061/

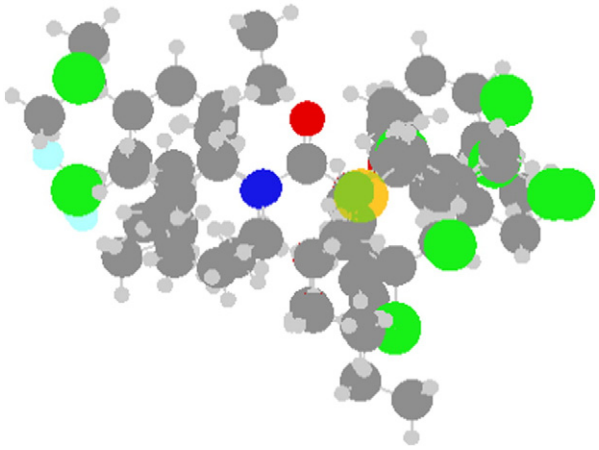

Fig. 1. Superposed chemical images of herbicides used in the aug-MIA-SPR modeling.

correlation of logK_{OC} with logP for some of these compounds (Sabljić et al., 1995). Despite the high predictive ability of some QSPR models (Schüürmann et al., 2006; Wang et al., 2009), the chemical interpretation indispensable to rationally design new herbicide candidates has not been straightforward, which can be easily assessed using our proposed SPR model.

2. Material and methods

A series of 26 carbonyl herbicides pertaining to amides, carbamates, thiocarbamates and ureas was obtained from the literature (Mackay et al., 1997) and the corresponding chemical structures drawn using the GaussView program (Dennington et al., 2008), with the carbonyl group considered as the common basic moiety for 2D alignment. It is worth mentioning that more than 20 compounds are recommended for SPR purposes (Young, 2009) and, therefore, the present dataset fulfills this requirement. Fig. 1 presents the superposed 26 chemical images to illustrate the variance in the chemical space, as well as the carbonyl group as the congruent substructure for alignment purposes. The spheres representing atoms in molecules were designed to be proportional to the respective van der Waals radii and each chemical structure (image) was saved as bitmaps (bmp) files in a workspace of 612 × 441 pixel dimension using the Microsoft Windows Paint application. The images were numerically transformed according to the RGB color system, in which each atom color is the result of a contribution from red (255), green (255) and blue (255) components; thus, the colors vary from black (zero) to white (765, the sum of all three components). This procedure was performed using the Chemoface program (Nunes et al., 2012), giving a 612 × 441 data matrix for each molecule. The 26 matrices were merged to give a 26 × 612 × 441 tridimensional array. Later, the 3D-array was unfolded to yield a [26 × (612 × 441)] matrix.

A step-wise variable selection procedure, considering both relevance (in Shannon's entropy terms) and orthogonality, was applied to the unfolded matrix yielding a lower dimensionality data matrix, capable of capturing the essence of the system using a minimum number of variables and allowing for greater model interpretability. Given the disparity in the logK_{OC} values reported for the same herbicides in the literature, a pattern recognition model using PLS-DA was exploited rather than a quantitative one. This method is based on the transformation of the original variables in latent variables, and these are linearly independent (Chevallier et al., 2006; Neto and Moita, 1998). These latent variables are linear combination of original variables. The preference of PLS-DA to linear discriminant analysis (LDA) is to profit from the score plots obtained with the former as these allow for a straightforward assessment of the contribution of the different variables in the stratification of compounds according to their properties. Later, these variables may be examined for the chemical information codified in terms of the functional groups or substructures relevant for the studied property. Note that when the variables constituting an LDA model are orthogonal, the PLS-DA and LDA techniques yield the same results (Wold et al., 2001).

Based on the average logK_{OC} values obtained from the literature (Mackay et al., 1997; Kenaga and Goring, 1980; Liu and Qian, 1995),

Table 1
Herbicides and the respective classification according to the soil sorption. Compounds marked with asterisk (*) pertain to the test set (external validation) and were selected through Kennard–Stone sampling.

Cpd no.	Herbicide	Class	Cpd no.	Herbicide	Class
1	Chlortoluron	2	14	Butachlor	2
2	Diuron	2	15	Metolachlor	2
3	Fenuron	1	16	Propachlor*	2
4	Fluometuron	2	17	Propanil*	2
5	Isoproturon	2	18	Diphenamid*	2
6	Monuron	1	19	Pronamide	2
7	Neburon	3	20	Butylate	2
8	Monolinuron	2	21	Diallate	3
9	Linuron	2	22	Triallate*	3
10	Barban	3	23	EPTC*	2
11	Chlorpropham	2	24	Pebulate	2
12	Propham	1	25	Molinate	2
13	Alachlor	2	26	Vernolate*	2

Table 2
Statistical results for the aug-MIA-SPR/PLS-DA modeling.

Parameter	Value
Variables	3
Calibration success (%)	80
Y-randomization success (%)	0.5
Leave-one-out success (%)	75
External validation success (%)	100

Table 3
Selected variables and the corresponding pixel values according to the RGB color system encoding atom types (289 = chlorine; 426 = carbon; 612 = hydrogen; 765 = blank space).

Cpd no.	X3100	X4543	X4148	Class[a]
1	765	765	289	2
2	765	765	289	2
3	765	765	765	1
4	765	765	765	2
5	765	765	765	2
6	765	765	765	1
7	765	426	289	3
8	765	765	765	2
9	765	765	289	2
10	765	765	289	3
11	765	765	289	2
12	765	765	765	1
13	765	765	765	2
14	765	765	765	2
15	765	765	765	2
16	765	765	765	2
17	765	765	289	2
18	765	765	765	2
19	426	289	765	2
20	612	426	612	2
21	765	426	765	3
22	765	426	765	3
23	612	426	612	2
24	612	426	765	2
25	612	765	765	2
26	612	426	612	2

[a] Classes: 1 = low soil sorption; 2 = medium soil sorption; 3 = high soil sorption.