Ecotoxicology and Environmental Safety 129 (2016) 10–15

Contents lists available at ScienceDirect

Ecotoxicology and Environmental Safety

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

Developing a support vector machine based QSPR model for prediction of half-life of some herbicides

Kobra Samghani, Mohammad HosseinFatemi *

Chemometrics Laboratory, Faculty of Chemistry, University of Mazandaran, Babolsar, Iran

ARTICLE INFO

Article history:
Received 16 November 2015
Received in revised form
28 February 2016
Accepted 3 March 2016
Available online 9 March 2016

Keywords:
Herbicide
Quantitative structure–activity relationship
Half-life
Support vector machine
Multiple linear regression
Applicability domain

ABSTRACT

The half-life ($t_{1/2}$) of 58 herbicides were modeled by quantitative structure–property relationship (QSPR) based molecular structure descriptors. After calculation and the screening of a large number of molecular descriptors, the most relevant those ones selected by stepwise multiple linear regression were used for developing linear and nonlinear models which developed by using multiple linear regression and support vector machine, respectively. Comparison between statistical parameters of linear and nonlinear models indicates the suitability of SVM over MLR model for predicting the half-life of herbicides. The statistical parameters of R^2 and standard error for training set of SVM model were; 0.96 and 0.087, respectively, and were 0.93 and 0.092 for the test set. The SVM model was evaluated by leave one out cross validation test, which its result indicates the robustness and predictability of the model. The established SVM model was used for predicting the half-life of other herbicides that are located in the applicability domain of model that were determined via leverage approach.

The results of this study indicate that the relationship among selected molecular descriptors and herbicide's half-life is non-linear. These results emphases that the process of degradation of herbicides in the environment is very complex and can be affected by various environmental and structural features, therefore simple linear model cannot be able to successfully predict it.

1. Introduction

Herbicides are chemicals used to inhibit or reduce the normal growth rate of weeds. They are used in agricultural, aquatic, forest, and wild-land ecosystems to reduce the density of weeds to permit the growth of desirable species (Holt, 2013). Decomposition of herbicides occurs in soil, air, water, plants, animals, and microorganisms by photochemical, chemical, or microbiological means (Cobb and Reade, 2010; Holt, 2013). Some herbicides remain active in soil for a long time and can leach into the surface water and ground water and/or can get into crops which, lead to a significant bio accumulation of toxins in the food chains (Piotrowicz-Cieślak and Adomas, 2012). This subject may be more serious as persistence of these compounds increases in environment. Half-life ($t_{1/2}$) of herbicide indicates the time required for degradation of 50% of the herbicide, which can be considered as an important criterion of persistence (Piotrowicz-Cieślak and Adomas, 2012). Benefits from the use of herbicides must be in balance with the concerns about their health and the impact of global warming (Gehin et al., 2005).

For any herbicide to be registered by the environmental protection agency (EPA),very safety and environmental tests must be carried out to characterizing the chemical and physical properties, environmental fate, amounts of the pesticides and their residues in feed and food crops, toxicological properties, and their effects on non target plants and animals (Chan et al., 2007; Gehin et al., 2005; Holt, 2013).

Since carrying these experiments is difficult and time-consuming, therefore the development of some theoretical methods to estimate these environmental parameters is very important. One of these methods is quantitative structure properties relationship (QSPR) methodology, which developed based on theoretical derived molecular descriptors. This methodology was used for prediction of the herbicide's properties in some investigations. For example; in 2001 Martins and coworkers investigated the quantitative relationship between the structures of 20 sulfonylureas and their herbicidal activities by using regression analysis (Martins et al., 2001). They used hammett's electronic parameter for representing the electronic properties of substituents and end up; the obtained parameters of R^2 and standard error for their developed model were; 0.77, 0.37, respectively. Yu et al. (2007) synthesized seventy-four sulfonylurea derivatives and performed 3D-QSAR analyses for modeling their inhibitory activity on *E. coli* acetohydroxyacid synthase (AHAS, EC 2.2.1.6) isoenzyme II. These

* Corresponding author.
E-mail address: mhfatemi@umz.ac.ir (M. HosseinFatemi).

http://dx.doi.org/10.1016/j.ecoenv.2016.03.002
0147-6513/

QSAR models were developed by using comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) as 3D-QSAR models. The obtained parameter of R^2 was 0.954 and 0.913 for CoMFA and CoMSIA model, respectively. Moreover, in 2003 some of organic chemicals were classified into herbicide and non herbicide compounds by González et al. (2003). They developed a linear classification function to discriminate herbicides from non herbicide chemicals which showed a good global classification rate of 91%. In another work, quantitative structure–toxicity relationship models were developed to estimate the toxicity of some sulphonylurea and phenylurea herbicides to rats by Can et al. (2013). The toxicity of these chemicals in different matrices was predicted by using multiple-linear regression. According to this model, to design the less toxic sulphonylurea and phenylurea pesticides, molecules should be highly polar, water-soluble, and having low molecular mass and refractivity.

In the present work, quantitative structure–property relationship models were developed for predicting the half-life of some herbicides. Multiple linear regression (MLR) and support vector machine (SVM) were applied as linear and non-linear feature mapping techniques.

2. Methodology

2.1. Data set

Data set consists of 58 herbicides that their half-lifes in the soil, were reported in literature, which are shown in Table 1 (Mackay et al., 2010). The experimental values of the logarithm of half-life for these herbicides lie in the range of 0–2.7 for propanil and prometon, respectively. The chemicals in the data set were sorted according to their half-life values and divided into two groups, namely, the training and test sets by desirable distance from each other (*y*-ranking method).

2.2. Descriptor generation and screening

The structures of herbicides were drawn and optimized with HyperChem software (version 7) using the AM1 semi-empirical method. The theoretical descriptors for these chemicals were calculated by Dragon (version 3), CODESSA (version 2.7.2) and Accelrys Materials Studio (version 4.3) software. These descriptors belonged to the topological, geometrical and quantum chemical molecular descriptors. The prescreening of descriptors was performed by elimination of constant or near-constant descriptors. Since correlated descriptors encode similar information about the chemicals, therefore, descriptors that show a higher correlation ($R > 0.90$) with each other were identified and only one of them that have higher correlation with the logarithm of half-life was remained. Then the method of stepwise multiple linear regression was performed on the remaining descriptors to select the most relevant ones. Nine selected descriptors were used as independent variable to developing MLR and SVM model. The names and chemical class of these descriptors are shown in Table 2. The Pearson correlation among these descriptors is shown in Table S1. As can be seen in this table, there is not any high correlation between selected molecular descriptors.

2.3. Support vector machine

In support vector regression, the basic idea is to map the descriptors into a higher dimensional space F via a nonlinear mapping function, then linear regression do in this new space (Cortes and Vapnik, 1995) according to Eq. (1):

Table 1
Data set and their experimental and SVM calculated half-life, $t_{1/2}$ (day) in logarithmic scale.

NO	Name	CAS #	log $t_{1/2,exp}$	log $t_{1/2,\ pre}$	Residual
1	Alachlor	15972-60-8	1.18	1.15	−0.03
2	Ametryn	834-12-8	1.78	1.72	−0.06
3*	Atrazine	1912-24-9	1.78	1.71	−0.07
4	Barban	101-27-9	0.70	0.73	0.03
5	Benefin	1861-40-1	1.60	1.60	0.00
6	Bifenox	42576-02-3	0.85	0.88	0.03
7	Bromacil	314-40-9	1.78	1.81	0.03
8	Butachlor	23184-66-9	1.08	1.28	0.20
9	Butylate	2008-41-5	1.11	1.14	0.03
10*	Chlorbromuron	13360-45-7	1.60	1.59	−0.01
11	Chlorpropham	101-21-3	1.48	1.45	−0.03
12	Chlorsulfuron	64902-72-3	1.60	1.57	−0.03
13	Cyanazine	21725-46-2	1.15	1.26	0.11
14*	2,4-DB	94-82-6	0.70	0.79	0.09
15	Diallate	2303-16-4	1.48	1.45	−0.03
16	Dichlobenil	1194-65-6	1.78	1.81	0.03
17	Dichlorprop	120-36-5	1.00	0.97	−0.03
18	Diclofop-methyl	51338-27-3	1.48	1.51	0.03
19	Dinitramine	29091-05-2	1.48	1.51	0.03
20*	Dinoseb	88-85-7	1.48	1.45	−0.03
21	Diphenamid	957-51-7	1.48	1.50	0.02
22	Diuron	330-54-1	1.95	2.06	0.11
23	EPTC	759-94-4	0.78	0.81	0.03
24	Ethalfluralin	55283-68-6	1.78	1.81	0.03
25	Fenoprop	93-72-1	1.32	1.35	0.03
26	Fenuron	101-42-8	1.78	1.81	0.03
27*	Fluometuron	2164-17-2	1.93	1.87	−0.06
28	Fluridone	59756-60-4	1.32	1.35	0.03
29	Isopropalin	33820-53-0	2.00	1.92	−0.08
30*	Linuron	330-55-2	1.78	1.66	−0.12
31	Metolachlor	51218-45-2	1.95	1.59	−0.36
32	Metribuzin	21087-64-9	1.60	1.57	−0.03
33	Molinate	2212-67-1	1.32	1.07	−0.25
34	Monolinuron	1746-81-2	1.78	1.81	0.03
35	Monuron	150-68-5	2.23	2.07	−0.16
36	Napropamide	15299-99-7	1.85	1.82	−0.03
37	Neburon	555-37-3	2.08	2.05	−0.03
38	Nitrofen	1836-75-5	1.48	1.49	0.01
39	Norflurazon	27314-13-2	1.48	1.45	−0.03
40*	Oryzalin	19044-88-3	1.30	1.56	0.26
41	Pebulate	1114-71-2	1.15	1.12	−0.03
42	Pendimethalin	40487-42-1	1.95	1.92	−0.03
43	Profluralin	26399-36-0	2.04	1.78	−0.26
44	Prometon	1610-18-0	2.70	2.67	−0.03
45	Prometryn	7287-19-6	1.78	1.81	0.03
46	Pronamide	23950-58-5	1.78	1.78	0.00
47	Propachlor	1918-16-7	0.80	0.84	0.04
48	Propanil	709-98-8	0.00	0.03	0.03
49*	Propazine	139-40-2	2.13	1.94	−0.19
50	Propham	122-42-9	1.00	1.03	0.03
51	Pyrazon	1698-60-8	1.32	1.29	−0.03
52	Simazine	122-34-9	1.78	1.75	−0.03
53	Terbacil	5902-51-2	2.08	2.05	−0.03
54	Terbutryn	886-50-0	1.62	1.67	0.05
55*	Thiobencarb	28249-77-6	1.32	1.38	0.06
56	Triallate	2303-17-5	1.91	1.94	0.03
57*	Trifluralin	1582-09-8	1.78	1.62	−0.16
58*	Vernolate	1929-77-7	1.08	1.1	0.02
59**	Fluroxypyr	69377-81-7	–	1.15	–
60**	Imazapic	104098-48-8	–	1.79	–
61**	Imazapyr	81334-34-1	–	1.67	–
62**	TCDD	1746-01-6	–	1.63	–
63**	2,4,5-T	93-76-5	–	0.74	–
64**	Glyphosate	1746-01-6	–	0.98	–
65**	MCPA	94-74-6	–	0.92	–
66**	Picloram	1918-02-1	–	2.09	–
67**	Dithiopyr	97886-45-8	–	1.34	–
68**	Acetochlor	3456-82-1	–	1.12	–
69**	Asulam	3337-71-1	–	1.33	–
70**	Benfluralin	1861-40-1	–	1.29	–
71**	Dimethenamid	87674-68-8	–	1.69	–
72**	Aminopyralid	150114-71-9	–	2.14	–
73**	MCPB	94-81-5	–	1.32	–

* Refer to test set compounds and;
** Refer to compounds with no reported half-life.