Free and Open Source Chemistry Software in Research of Quantitative Structure-Toxicity Relationship of Pesticides

Vesna Rastija[1,*], *Dejan* Agić[1], *Kristian* Brlas[1], *Vijay* Masand[2]

[1] Faculty of Agriculture, J. J. Strossmayer University of Osijek, Vladimira Preloga, Osijek, Croatia
[2] Department of Chemistry, Vidya Bharati Mahavidyalaya, Camp, CK Naidu Road, Amravati, 444 602, Maharashtra, India

Abstract. Pesticides are toxic chemicals aimed for the destroying pest on crops. Numerous data evidence about toxicity of pesticides on aquatic organisms. Since pesticides with similar properties tend to have similar biological activities, toxicity may be predicted from structure. Their structure feature and properties are encoded my means of molecular descriptors. Molecular descriptors can capture quite simple two-dimensional (2D) chemical structures to highly complex three-dimensional (3D) chemical structures. Quantitative structure-toxicity relationship (QSTR) method uses linear regression analyses for correlation toxicity of chemical with their structural feature using molecular descriptors. Molecular descriptors were calculated using open source software PaDEL and in-house built PyMOL plugin (PyDescriptor). *PyDescriptor* is a new script implemented with the commonly used visualization software PyMOL for calculation of a large and diverse set of easily interpretable molecular descriptors encoding pharmacophoric patterns and atomic fragments. *PyDescriptor* has several advantages like free and open source, can work on all major platforms (Windows, Linux, MacOS). QSTR method allows prediction of toxicity of pesticides without experimental assay. In the present work, QSTR analysis for toxicity of a dataset of mixtures of 5 classes of pesticides comprising has been performed.

1 Introduction

Pesticides are used extensively to control agricultural pest and to improve crop yields. However, small fraction of the pesticides is moving up from surface into stream, rivers and lakes and cause of considerable environmental concern as a result from application drift, rainfall runoff, or residue leaching through the soil into groundwater [1]. The contamination of water by pesticides increasing around the world, so the knowledge of eco-toxicological effects for aquatic organisms for the environmental risk assessment is essential.

Before pesticides are registered they must undergo laboratory testing on animals for short-term (acute) and long-term (chronic) health effects. Laboratory animals are purposely fed doses high enough to cause toxic effects. Small planktonic crustaceans *Daphnia,* fish, and algae are the most common organisms tested for the evaluation of toxic effects of pesticides. In order to reduce expensive and time-consuming experiments and reduce animal testing quantitative structure-toxicity relationship (QSTR) method is valuable [2]. Two-dimensional (2D) and three-dimensional (3D) molecular structure considerable influence on properties of pesticides, such as, absorption, distribution, metabolism, and excretion (ADME). QSTR method allows prediction of environmental toxicity derived from the molecular structure and fills an important gap in risk assessment studies (REACH) [3].

QSTR method involves representations of molecules or molecular patterns in the form of numerical descriptors that capture the structural features and properties of molecules, generally known as molecular descriptors. Molecular descriptors describe: chemical properties (electrophilicity, hydrogen bonding), physical-chemical properties (lipophilicity, polar surface area), 2D structure (topological, information, connectivity, information indices, 2D frequency fingerprints), 3D structure (RDF, WHIM, GETAWEY, geometrical descriptors). Correlation of toxicity of molecule and molecular descriptors is most often expressed by linear equation calculated by multiple linear regression (MLR), or partial least squares (PLS) [4]. Computational neural networks (CNN) is usually performed if there is an assumption about a nonlinear and a highly complex relationship between the structure and the observed toxicity [5].

There are many commercial and free academic packages developed for calculation of molecular descriptors. Most of the molecular descriptors can be calculated by using commercial software packages such as CODESSA [6] and DRAGON [7]. Limitations of most of those packages are high price and hardly interpretable calculated molecular descriptors in terms of structural features. To overcome this, we have developed, *PyDescriptor*, a new script implemented with the commonly used visualization software PyMOL for calculation of a large and diverse set of easily interpretable 1D- to 3D- descriptors. They are also easy interpreting in terms of structural moieties, applicable for representing local environment or structure, simple to understand, independent of experimental properties, sensitive to changes in conformation molecule. PyMOL is free open source molecular graphics tool for 3D visualization of proteins, small molecules, density, surfaces, and trajectories [8]. *PyDescriptor* is a useful addition to the currently existing molecular descriptor calculation software. It has several advantages like free

* Corresponding author: vrastija@pfos.hr

and open source and it is able to works on all major platforms (Windows, Linux, MacOS). The script is freely available for academic use [9].

In the present paper we have generated QSTR models using molecular descriptors calculated by *PyDescriptors* for estimation of toxicity of 43 pesticides obtained on aquatic vertebrates bluegill sunfish (*Lepomis macrochirus*) [1].

	Chemical	Exp. endpoint	Prediction fitting
		$\log LC_{50}$ mol/L	$\log LC_{50}$ mol/L
1	bensulfuron	2.43	-0.26
2	chlorimuron	0.59	1.39
3	chlorsulfuron*	0.92	0.86
4	flumetsulam	2.97	-0.36
5	halosulfuron	2.22	0.01
6	imazapyr*	3.19	-0.03
7	imazaquin	3.13	0.11
8	imazethapyr	3.17	-0.14
9	metsulfuron*	2.61	-0.35
10	nicosulfuron	3.39	0.36
11	primisulfuron*	2.34	-0.99
12	prosulfuron	2.57	-0.96
13	sulfometuron	2.22	-0.48
14	triasulfuron	2.40	-0.49
15	atrazine*	2.22	-0.19
16	cyanazine	1.97	-0.28
17	metribuzin	2.65	-0.33
18	prometon	2.12	0.29
19	prometryn	1.62	0.36
20	simazine	2.53	-0.68
21	acetochlor	0.74	0.25
22	alachlor*	1.31	0.03
23	metolachlor	1.64	-0.51
24	propachlor	0.86	0.55
25	azinphos-methyl	-1.19	0.40
26	chlorpyrifos	-1.25	0.26
27	diazinon	-0.13	-0.53
28	disulfoton	-0.34	-0.24
29	ethoprophos	1.19	-0.16
30	fonofos	-1.35	-0.22
31	malathion	-0.45	0.49
32	parathion	-0.24	-0.47
33	parathion methyl	0.86	0.51
34	phorate*	-1.82	1.25
35	terbufos	-1.89	0.92
36	butylate*	1.39	-0.62
37	carbaryl	1.69	0.25
38	carbofuran	0.45	0.85
39	EPTC	2.13	-0.40
40	molinate	1.87	0.25
41	pebulate*	1.54	0.35
42	thiobencarb	0.84	-0.52
* member of the test set			

Table 1. Experimentally obtained toxicity endpoint and estimated values by eq. (1) of pesticides for *Lepomis acrochirus*.

2 Methods

2.1 Toxicity data

Toxicity data for aquatic vertebrates bluegill sunfish (*Lepomis macrochirus*) were retrieved from literature. Toxicity of 43 pesticides is expressed as LC_{50} (lethal concentration that kills 50 % of the animals in a test population / $molL^{-1}$). LC_{50} were converted in the form of a logarithm ($\log LC_{50}$) (Table 1).

2.2 Calculation of molecular descriptors

Molecular descriptors were calculated using open source software PaDEL [9] and a new in-house built PyMOL plugin (PyDescriptor) [8] followed by extensive objective and subjective feature selection to avoid redundant descriptors.

2.3 Regression analysis and validation of models

For model building, the dataset was divided into training (80%) and test (20%) sets. The best QSAR models were obtained using a Genetic Algorithm using QSARINS v 2.2 [11].

The models have been assessed by: fitting criteria; internal cross-validation using leave-one out (LOO) method and Y-scrambling; and external validation. Fitting criteria included: the coefficient of determination (R^2), adjusted (R^2_{adj}), cross-validate R^2 using leave-one-out method (Q^2_{LOO}), global correlation among descriptors (*Kxx*), difference between global correlation between molecular descriptors and *y* the response variable, and global correlation among descriptors (ΔK), standard deviation of regression (*s*), and Fisher ratio (*F*). Internal and external validations also included the following parameters: root-mean-square error of the training set ($RMSE_{tr}$); root-mean-square error of the training set determined through cross validated LOO method ($RMSE_{cv}$), root-mean-square error of the external validation set ($RMSE_{ex}$), concordance correlation coefficient of the training set (CCC_{tr}), test set using LOO cross validation (CCC_{cv}), and of the external validation set (CCC_{ex}), mean absolute error of the training set (MAE_{tr}), mean absolute error of the internal validation set (MAE_{cv}) and mean absolute error of the external validation set (MAE_{ex}) [12], predictive residual sum of squares determined through cross-validated LOO method ($PRESS_{cv}$) in the training set and in the external prediction set ($PRESS_{ex}$). The analysed external validation parameters also include the coefficient of determination (R^2_{ex}). Robustness of QSAR models was tested by Y-randomisation test. New parallel models were developed based on fit to randomly reordered Y-data (Y scrambling), and the process was repeated several times (2000 iterations) [12]. Investigation of the applicability domain of a prediction model was performed by leverage plot or Williams plot (plotting residuals vs. leverage of training compounds). Detection of outliers was carried out for compounds that have values of standardized residuals greater than two standard deviation units using QSARINS. The leverage *h*

of a compound is the measure of its influence on the model.

3 Result and discussion

The best three-descriptor based QSTR model for prediction of toxicity for the *Lepomis acrochirus* is:

$$\log LC_{50} = 1.948 - 0.588\ ALogP + 1.223\ FP747 - 0.375\ fPH3A \quad (1)$$

$N_{\text{training set}} = 34 \quad N_{\text{prediction set}} = 9$

The statistical results of the obtained QSTR model are presented in Table 2. Satisfaction of fitting criteria implies the following: the closer R^2 values are to unity, the more similar calculated values are to the experimental ones, that is, $R^2 \geq 0.60$. Also, larger F statistic and lower standard deviation means that the model is more significant. In order to avoid overfitting, inter-correlation between the descriptors included in the equation is detected based on Kxx and ΔK. Low Kxx and $\Delta K \geq 0.05$ implies no chance correlation between descriptors. The minimum acceptable statistical parameters for internal and external predictivity include the following conditions: $R^2_{\text{ext}} \geq 0.60$; $CCC_{\text{ext}} \geq 0.85$; $RMSE_{\text{cv}}$ and MAE_{cv} close to zero; and $RMSE_{\text{tr}} < RMSE_{\text{cv}}$. Robust QSAR models should have low $R^2_y\text{scr}$ and low $Q^2_y\text{scr}$ values and $R^2_y\text{scr} > Q^2_y\text{scr}$. In order to investigate the applicability of a prediction model and detect possible outliers, the applicability domain of the selected model was evaluated by a leverage analysis expressed as Williams plot, in which residuals and the leverage values were plotted. Williams plot is given in Figure 1. A scatter plot of experimentally obtained toxicity calculated by QSTR model versus values calculated by Eq. (1) is presented in Figure 2.

Table 2. Statistical parameters of the obtained QSAR models.

	Statistical parameters	Value
Fitting criteria	R^2	0.87
	R^2_{adj}	0.86
	F	68.24
	Kxx	0.35
	ΔK	0.19
	$RMSE_{\text{tr}}$	0.51
	MAE_{tr}	0.43
	CCC_{tr}	0.93
Internal cross-validation	Q^2_{loo}	0.84
	$RMSE_{\text{cv}}$	0.56
	MAE_{cv}	0.48
	$PRESS_{\text{cv}}$	0.92
	CCC_{cv}	0.92
Y-scrambling	$R^2_y\text{scr}$	8.67
	$Q^2_y\text{scr}$	-19.05
External validation	R^2_{ext}	0.79
	$RMSE_{\text{ext}}$	0.66
	MAE_{ext}	0.52
	$PRESS_{\text{ext}}$	3.95
	CCC_{ext}	0.85

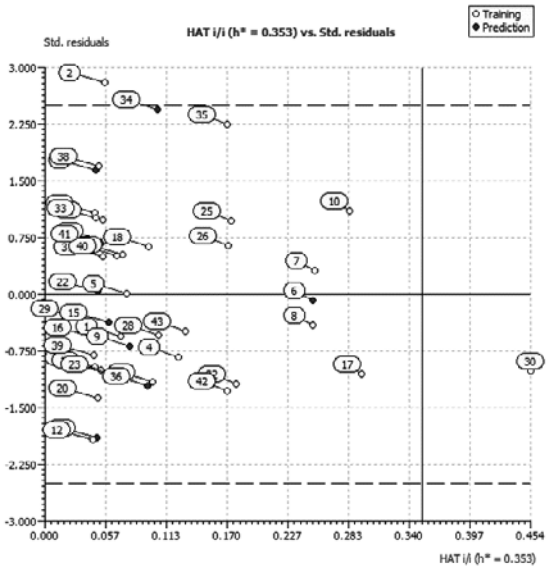


Fig. 1. Applicability domain of the QSAR model for logIC_50 expresses by eq. (1).

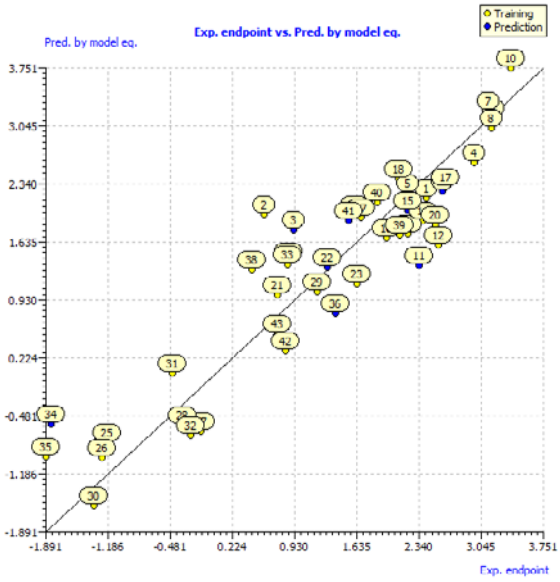


Fig. 2. A scatter plot of experimentally obtained toxicity calculated by QSTR model versus values calculated by eq. (1).

Obtained model has satisfactory results of fitting parameters and internal validation and low collinearity between the three descriptors. The results of Y-scrambling demonstrated that model wasnot obtained by chance correlation. Model 1 may be considered as predictive due to the high values of R^2_{ext} and CCC_{ext}, as well as small difference between $RMSE_{tr}$ and $RMSE_{ex}$, and between MAE_{tr} and MAE_{ex}. As can be seen from the Williams plot (Figure 1), toxicity of pesticides **30** (fonofos) predicted my model (1) must be used with reserve, because its leverage value is greater than the warning leverage ($h^* = 0.353$). Also, the same model has generated one outliers, pesticides **2** (chlorimuron) because its standardized residual is greater than ± 2.5.

The best QSTR model obtained include the following descriptors: lipophilicity (ALog*P*), PaDEL fingerprint

descriptor *FP747* and PyMOL descriptor *fPH3A*. Considering the negative coefficient of ALog*P* in Eq. (1) highly toxic compounds have a high lipophilicity. High lipophilic compounds my easily pass lipidous membranes and accumulate in fat tissue, therefore cause enhanced toxic effect [13]. Negative coefficient of PyDescriptor *fPH3A* implies that frequency of occurrence of hydrogen within 3 Å from phosphorus positively influence on increased toxicity of pesticides. QSAR study of toxicity of phoshorhydrazide (PHA) derivates revealed that the NH–P(X) moiety has a much higher inhibitory activity than the NH–C(X) moiety. The presence of the electron acceptor substituent around the P=X group increases the inhibitory potential of the PHA derivatives [14]. Obtained results are in accordance with previous findings of QSTR modeling of toxicity of organic molecules to *Daphnia magna* [4]. Obtained PLS models suggests that higher lipophilicity and electrophilicity, and hydrogen bond donor groups are responsible for greater toxicity.

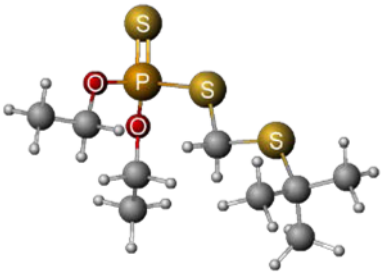
Figure 3a presents a chemical structure of the most toxic compound (**35**), an aliphatic organothiophosphate insecticide, terbufos. Thiophosphates are a very toxic class of organophosphorus compounds, especially if possess reactive functional groups such as: methyl, phosphate ester (P=O type) and unsubstituted phenyl group [15]. QSTR study of some organophosphorus compounds performed by using the quantum chemical and topological descriptors relieved that the sulphur atoms instead of oxygen atoms improved toxicity [16].

Figure 3b shows a structure of minimum toxic compound (**6**) imazapyr, an imidazolinone herbicide. Imazapyr does not contain phosphorus atom. According a positive coefficient of fingerprint descriptor *FP747* in eq. (1) imply that higher values of this descriptors mean lower toxicity.

4 Conclusion

In the present work, we have used an open source molecular descriptor calculation PyMOL plugin *PyDescriptor* for calculation easily interpretable and informative molecular descriptors. Robust QSTR models with good external predictive ability have been developed for the toxicity of pesticides for the fish, bluegill sunfish. The developed models, since, satisfy the threshold values for many statistical parameters could be useful for the prediction of experimentally undermined toxicity of known pesticides, as well as new pesticides. The model can also be employed to better understand the mechanism of toxicity of the various families of pesticides on the aquatic organisms, as well as the identification of potential aquatic pollutant.

Our results indicates that future QSTR analysis of pesticides should apply a specific group of descriptors relates with lipohilicity and structure fragment involved in electron transfer.



a)

b)

Fig. 3. Structure of: a) the most toxic pesticide, terbufos (**35**); b) the least toxic pesticide imazapyr (**6**).

Acknowledgements

Authors are thankful to Josip Juraj Strossmayer University of Osijek, Osijek, Croatia for financial support to the project INGI-2015-20 to complete the present work.

References

1. W. Battaglin, J. Fairchild, Water Sci. Technol. **45**, 95 (2002)
2. S. Cassani, S. Kovarich, E. Papa, P. P. Roy, L. van der Wal, P. Gramatica, J. Hazard. Mater. **258**, 50 (2013)
3. EC Proposal for a Regulation of the European Parliament and of the Council concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency and amending Directive 1999/45/EC and Regulation (EC) on Persistent Organic Pollutants (Brussels, Belgium) (2003)
4. S. Kar, K. Roy, J. Hazard. Mater. **177**, 344 (2010)
5. D. V. Eldered, P. C. Jurs, SAR QSAR Environ. Res. **10**, 75 (1999)
6. A. R. Katritzky, V. Lobanov, M. Karelson, CODESSA Reference Manual. University of Florida, Gainesville, (1996)
7. Dragon Professional Version 5.5, Talete srl, (2007) (http://www.talete.mi.it/dragon.htm).
8. PyMOL, Schrödinger, LLC. (2007) (http://www.pymol.org/).

9. PyDescriptor, http://www.unios.hr/wp-content/uploads/2016/06/PyDescriptors-Tutorial.pdf
10. C. W. Yap, J. Comput. Chem. **35**, 1466 (2010)
11. P. Gramatica, N. Chirico, E. Papa, S. Cassani, S. Kovarich, J. Comput. Chem. **34,** 2121 (2013)
12. P. Gramatica, QSAR Comb. Sci. **26**, 694 (2007)
13. O. G. Kolumbin, L. N. Ognichenko, A. G. Artemenko, P. G. Polischuk, M.A. Kulinskyb, E. N. Muratov, V. E. Kuz'min, V. A. Bobeica, Chem. J. Mold. **8**, 95 (2013)
14. K. Gholivand, L. Asadi, A. A. E. Valmoozi, M. Hodaii, M. Sharifi, H. M. Kashani, H. R. Mahzouni, M. Ghadamyari, A. A. Kalate, E. Davari, S. Salehi, M. Bonsaii, RSC Adv. **6**, 24175 (2016)
15. A.-M. Petrescu, G. Ilia, Ecoterra **14**, 90 (2015)
16. S. A. Senior, M. D. Madbouly, A.M. El Massry, Chemosphere **85**, 7 (2011)