QSTR model for toxicity of OPs based on MEDV

Ruqiong Qin

Guangxi Polytechnic of Construction, Nanning 530003, China

qinruqiong@163.com

Keywords: MEDV, QSTR, OPs, pLD_{50}

Abstract. Organophosphorus pesticides (OPs) are some of the most prevalent pollutants in the total environment and receive more and more concerns as a group of ubiquitous potential persistent organic pollutants. Using the variable selection and modeling based on prediction (VSMP), the molecular electronegativity distance vector (MEDV) derived directly from the molecular topological structures was employed to develop a quantitative structure-toxicity relationship (QSTR) model between the toxicity and MEDV descriptors of 22 OPs. The QSTR model showed a good estimation ability with a correlation coefficient (r^2) of 0.9781 and a high stability with a leave-one-out cross-validation correlation coefficient (q^2) of 0.9519. Main structural factors influencing the toxicity of OPs are the substructures expressed by three atomic groups —CH3、>N—、—O—（or—S—）. The QSTR model was Internal validation and External validation, and was shown that the QSTR model was bootstrapping and high predictive ability.

1. Introduction

Organophosphorous pesticides (OPs) are widely used in agricultural production. Their high acute toxicity and widespread bioactivity have, however, a great impact on the environment and human health [1, 2]. In order to investigate for their potential toxicity, it is essential to determine their toxicity in mammal species.

Several researchers have been investigated on the quantitative structure-toxicity relationship (QSTR) of OPs. Devillers et al.[3] developed a partial least squares (PLS)-based QSTR model for acute toxicity of rat for 51 organicphosphorus compounds. Sazonovas et al.[4] analyzed the quantitative relationship between the structures of organicphosphorus compounds. Knaak et al.[5] developed a quantitative structure-activity relationship (QSAR) model between 31 organicphosphorus compounds and the toxicity of the rat with quantum chemical descriptors. Although several QSTR models have been developed for the organicphosphorus compounds, most of the model was not strictly internal and external validations [6, 7], which can not ensure the robustness and predicative ability of the models and may leading to an inaccurate prediction to the toxicity of compounds.

The present study applied the molecular electronegativity distance vector [8] (MEDV) to effectively characterize molecular structures of 22 organic phosphorus compounds. The QSTR model between the acute toxicity (96h pLD_{50}) of rat of organic phosphorus compounds and its molecular structure was built by using the variable selection and modeling method based on the prediction (VSMP). The QSTR model was validated by internal and external validation methods. The results show that the QSTR model has high statistical parameters and can accurately predict the toxicity of organicphosphorus compounds. Therefore, the model may serve as a theoretical basis for predicting the toxicities of organophosphorus compounds.

2. Materials and methods

2.1 Data set

The experimental toxicity values of 22 kinds of OPs compounds that extracted from PPDB database [9]. The parent structure of these compounds is shown in Fig. 1, where R1~R3 stand for

different substituents. The diversity of their structures can be represented effectively with MEDV descriptor.

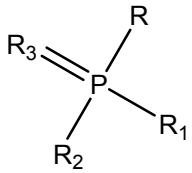

Fig.1 The structure of OPs

2.2 Theoretical methods

Molecular structures of OPs are represented by MEDV [8], and the MLR is applied in establishing the QSTR model between the $p\mathrm{LD}_{50}$ toxicity of rat and MEDV descriptors of OPs compounds.

To ensure the reliability and validity of the regression model, the LOO, LMO [10,11] cross-validation, *y*-randomization test [12], and the bootstrapping method [13] are used for internal validation. The $p\mathrm{LD}_{50}$ values of the whole data were sorted into ascending order [14], and then 16 OPs were equidistantly picked as a training set and the remaining 6 OPs made up a test set. The QSTR model established with training set is validated with the testing set, and its external predictive ability is evaluated with statistics Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, and $\bar{r}_m^2$ [15-16].

3. Results and discussion

3.1 QSTR model based on the whole data set

These three descriptors were selected by the VSMP method. Using x_1, x_7 and x_{10} descriptors as independent variables and the median lethal concentration of $p\mathrm{LD}_{50}$ as the dependent variable, the QSTR model (model M) for the training set was established with the MLR method. The model M is shown as follows:

$$y = (1.7599 \pm 0.05475) - (0.2366 \pm 0.0100) \cdot x_1 + (0.1974 \pm 0.0172) \cdot x_7 + (0.0515 \pm 0.0055) \cdot x_{10}$$

$$n=16,\ m=3,\ R^2=0.9815,\ RMSE=0.09,\ F=212.679 \text{ (modeling)}$$

$$n=16,\ m=3,\ Q^2=0.9675,\ RMSEV=0.12 \text{ (LOO validating)}$$

where, *n* is the number of samples, *m* is the number of optimal variables, *F* is the Fisher's statistics, *RMSE* is the root-mean-square error estimated, *RMSEV* is the root-mean-square error of LOO cross-validation, values before and after "±" represent the regression coefficient of the model and its corresponding standard deviation. It can be seen from the above equation that the QSTR model has a good estimation ability with high R^2 and Q^2 values.

3.2 Validation of QSTR model

The whole data set was divided into the training set of 16 compounds and the test set of 6 compounds. The correlation of the experimental and predictive $p\mathrm{LD}_{50}$ values of the training set and the test set is shown in Fig.2.

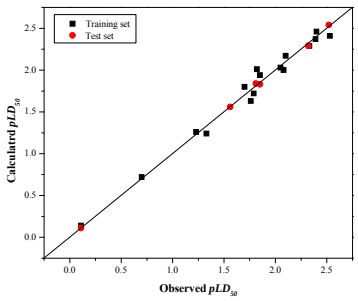

Fig.2 Plot of observed versus calculated $p\mathrm{LD}_{50}$ resulted from the model M

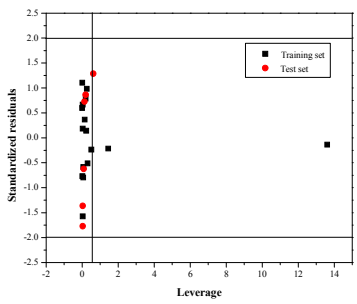

Fig.3 LOO standardized residuals analysis for the model M

By internal validation, the LOO validation shown that R^2=0.9815, Q^2_{Loo}=0.9641, R^2-Q^2=0.0174<0.3, shows that model M has no over-fitting. LOO cross-validation is not sufficient to explain the robustness of model M. The correlation coefficient of LMO cross-validation (Q^2_{LMO}) is 0.7447. The result of bootstrapping method is: R^2_{bstr}=0.9785, Q^2_{bstr}=0.6120. R^2 and Q^2 in LOO and LMO cross-validations and the bootstrapping method are all meet the condition of R^2>0.6, Q^2>0.5, which shows that model M has good robustness[26]. The result of *y*-randomization test is: R^2_{yrand}=0.1133 and Q^2_{yrand}= - 0.7653, showing that there is no chance correlation between the independent variable and dependent variable of model M2.

The relevant statistical parameters of model M for test set are: Q^2_{F1},= 0.9677, Q^2_{F2}, =0.9674, Q^2_{F3}=0.9493, *CCC*=0.9986 ; $\bar{r}_m^2$ =0.9354, r^2_m=0.9482, $r_m'^2$ =0.9225, △ r^2_m= - 0.0256 ; (R^2-R^2_0)/R^2=0.00064, (R^2-R'^2_0)/ R^2=0.00270; k=0.9796, k'=1.0155. All the relevant statistical parameters of the external validation meet the conditions suggested by Chirico and Gramatica, Roy, and Golbraikh and Tropsha. It can be concluded that model M has a good external predictive ability.

Moreover, it can be seen from the residual values of model M (see Fig. 3) that there is no abnormal value in the data set. There is no sample present in the area of absolute value of LOO standard residual greater than 2.0 and leverage value greater than 0.5625, which shows the whole data set is within the application domain. It can be seen from the relationship between MEDV descriptors and the sub structure of compounds that three descriptors x_1, x_7, and x_{10} are corresponding to the structural fragments —CH3、>N—、—O—（或—S—）. These substructures are the main factors affecting the toxicity of OPs compounds to rat.

4. Conclusion

MEDV can effectively represent molecular structures of 22 organophosphorus compounds. The three optimum descriptors are highly correlated with the toxicity of organicphosphorus compounds. Model M based on 16 OPs in the training set can successfully estimate its toxicity, and shows a good internal robustness and external predictive ability. All the statistics parameters meet the conditions proposed, which indicates model M has a good predictive ability. Therefore, model M may be useful in predicting unknown toxicity values of the OPs.

References

[1] Schulz R, Liess M. A field study of the effects of agriculturally derived insecticide input on stream macroinvertebrate dynamics[J]. Aquatic Toxicology, 3-4, 46(1999), 155-176

[2] Lai K, Stolowich N J, Wild J R. Characterization of PS bond hydrolysis in organophosphorothioate pesticides by organophosphorus hydrolase[J]. Archives of Biochemistry and Biophysics, 7, 318(1995), 59-64

[3] Devillers, J. Prediction of mammalian toxicity of organophosphorus pesticides from QSTR modeling. SAR QSAR Environ. Res.. 15, (2004), 501-510.

[4] Sazonova, N. A.; DasBanerjee, T.; Middleton, F. A.; Gowtham, S.; Schuckers, S. and Faraone, S. V. Transcriptome - wide gene expression in a rat model of attention deficit hyperactivity disorder symptoms: Rats developmentally exposed to polychlorinated biphenyls. Medical Genetics. 156(2011), 898-912.

[5] Knaak, J., Kozbelt, S. J., and Sullivan, L. Metabolism of 2-ethylhexyl sulfate by the rat and rabbit. Toxicol. Appl. Pharmacol.. 8(1966), 369-379.

[6] Zhao J, Yu S. Quantitative structure–activity relationship of organophosphate compounds based on molecular interaction fields descriptors[J]. Environmental Toxicology and Pharmacology, 2, 35(2013), 228-234

[7] Zvinavashe E, Du T, Griff T, Berg H H J Soffers A E M F, Vervoort J. Quantitative structure-activity relationship modeling of the toxicity of organothiophosphate pesticides to Daphnia magna and Cyprinus carpio[J]. Chemosphere, 11, 75(2009), 1531-8

[8] Liu S S, Liu H L, Yin C S, Wang L S. VSMP: a novel variable selection and modeling method based on the prediction[J]. Journal of chemical information computer sciences, 3, 43(2003), 964-969

[9] Information on：http://sitem.herts.ac.uk/aeru/ppdb/en/index.htm

[10] Kiralj R, Ferreira M M. Basic validation procedures for regression models in QSAR and QSPR studies: theory and application[J]. Journal of the Brazilian Chemical Society, 4, 20(2009), 770-787

[11] BesalúE. Fast computation of cross-validated properties in full linear leave-many-out procedures[J]. Journal of Mathematical Chemistry, 3, 29(2001), 191-204

[12] Rücker C, Rücker G, Meringer M. y-Randomization and its variants in QSPR/QSAR[J]. Journal of Chemical Information and Modeling, 6, 47(2007), 2345-2357

[13] Wehrens R, Putter H, Buydens L. The bootstrap: a tutorial[J]. Chemometrics Intelligent Laboratory Systems, 1, 54(2000), 35-52

[14] Tropsha A, Gramatica P, Gombar V K. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models[J]. QSAR & Combinatorial Science, 1, 22(2003), 69-77

[15] Schüürmann G, Ebert R U, Chen J, Wang B, Kuhne R. External validation and prediction employing the predictive squared correlation coefficient test set activity mean vs training set activity mean[J]. Journal of Chemical Information and Modeling, 11, 48(2008), 2140-2145

[16] Consonni V, Ballabio D, Todeschini R. Comments on the definition of the Q^2 parameter for QSAR validation[J]. Journal of Chemical Information and Modeling, 7, 49(2009),1669-1678