Chemosphere 156 (2016) 334–340

Contents lists available at ScienceDirect

Chemosphere

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

Linear and nonlinear models for predicting fish bioconcentration factors for pesticides

Jintao Yuan [a, b], Chun Xie [c], Ting Zhang [b], Jinfang Sun [b], Xuejie Yuan [c], Shuling Yu [d], Yingbiao Zhang [e], Yunyuan Cao [a], Xingchen Yu [a], Xuan Yang [a], Wu Yao [a, *]

[a] School of Public Health, Zhengzhou University, Zhengzhou, 450001, China
[b] Key Laboratory of Environmental Medicine Engineering, Ministry of Education, School of Public Health, Southeast University, Nanjing, 210009, China
[c] Shangqiu Medical College, Shangqiu, Henan Province 476100, China
[d] Key Laboratory of Natural Medicine and Immune-Engineering of Henan Province, Henan University, Kaifeng, Henan 475004, China
[e] Shenzhen Prevention and Treatment Center for Occupational Diseases, Shenzhen, Guangdong 518001, China

HIGHLIGHTS

- Three linear and nonlinear methods were used to study bioconcentration factors for pesticides.
- Applicability domain was analyzed, and outliers were determined.
- Comparison with reported model, new models is simpler.
- Relative significance of the descriptors to each model was studied.
- Selected descriptors were served as molecular information to study bioconcentration factors.

ARTICLE INFO

Article history:
Received 6 September 2015
Received in revised form
27 April 2016
Accepted 2 May 2016
Available online 13 May 2016

Handling Editor: I. Cousins

Keywords:
Bioconcentration factor
Multiple linear regression
Multilayer perceptron neural network
Projection pursuit regression

ABSTRACT

This work is devoted to the applications of the multiple linear regression (MLR), multilayer perceptron neural network (MLP NN) and projection pursuit regression (PPR) to quantitative structure–property relationship analysis of bioconcentration factors (BCFs) of pesticides tested on Bluegill (*Lepomis macrochirus*). Molecular descriptors of a total of 107 pesticides were calculated with the DRAGON Software and selected by inverse enhanced replacement method. Based on the selected DRAGON descriptors, a linear model was built by MLR, nonlinear models were developed using MLP NN and PPR. The robustness of the obtained models was assessed by cross-validation and external validation using test set. Outliers were also examined and deleted to improve predictive power. Comparative results revealed that PPR achieved the most accurate predictions. This study offers useful models and information for BCF prediction, risk assessment, and pesticide formulation.

1. Introduction

Bioconcentration factor (BCF) is the equilibrium ratio of the concentration of a substance in an exposed organism to the concentration of a dissolved substance bioavailable in the surrounding aquatic environment (Mackay and Fraser, 2000). The European regulation on classification, labeling, and packaging requires the establishment of BCFs for all compounds to regulate chemical substances (European Commission Environment Directorate General, 2007). For fish, BCF value facilitates the estimation of pesticide daily intake by the daily consumption of fish and by establishing safe limits for water pesticide concentration. Pesticides are used in agricultural treatments to improve crop yields (Cooper and Dobson, 2007). Nevertheless, the use of pesticides should be controlled because an important fraction of pesticides are released into the environment, which presents a potential hazard (Mnif et al., 2011; Landrigan et al., 1999). Therefore, investigating BCFs of pesticides is important.

However, experimental determination of BCF is expensive and time-consuming. Thus, a few estimation methods have been

* Corresponding author. School of Public Health, Zhengzhou University, 100 Science Avenue, Zhengzhou, 450001, China.
E-mail address: yaowu@zzu.edu.cn (W. Yao).

http://dx.doi.org/10.1016/j.chemosphere.2016.05.002
0045-6535/

reported. Several studies have reported on the linear correlation between BCF values and n-octanol/water partition coefficient (log K_{ow}) (Devillers et al., 1996; Mackay, 1982; Garg and Smith, 2014). Dimitrov et al. (2002) proposed a nonlinear model for the log BCF/log K_{ow} relationship based on a large set of narcotic compounds to analysis the specific role of the internal water phase on the formation of the body bioconcentration. Gissi et al. (2013) integrated the most used CAESAR and Meylan models for predicting BCF of 851 compounds from the ANTARES BCF dataset to demonstrate integrated model with more reliability of the predictions. Moreover, a chromatographic retention factor (Bermúdez-Saldaña et al., 2005) and an artificial membrane accumulation index (Fujikawa et al., 2009) have been reported. Piir et al. (2010) summarized some modeling techniques that use different compounds and algorithms. However, only one study (Jackson et al., 2009) predicted the BCF of pesticides using log K_{ow} and E-state descriptors. We propose here three alternative models for predicting BCFs of pesticides and providing useful information for risk assessment and pesticide research.

Selecting a combination of variables that produce the best result for QSAR study is also one of the most important problems. Feature selection methods such as simulated annealing, genetic algorithm, replacement method (RM), enhanced replacement method (ERM), and many more are used (Gonzáleza et al., 2008). Among these techniques, a modified ERM, inverse enhanced replacement method (IERM), is simple and of low computational cost (Morales et al., 2006). Thus, IERM was employed to select the best DRAGON descriptors for QSAR models. Multiple linear regression (MLR), multilayer perceptron neural network (MLP NN) and projection pursuit regression (PPR) have attracted attention and have been extensively applied (Torrecilla et al., 2009; Goodarzi et al., 2010; Parinet et al., 2015; Yan et al., 2015; Du et al., 2008). Therefore, MLR, MLP NN and PPR were also employed in this paper.

This study was performed to develop linear and nonlinear QSAR models for predicting the log BCF of pesticides. The presented QSAR models were validated using a test set and the proposed parameters by Tropsha (2010). The performances of the different models were compared. In addition, the descriptors used in the QSAR models were used to identify molecular characters related to the log BCF of pesticides.

2. Materials and methods

2.1. Data set

Molecular structures and activities of pesticides were derived from Jackson et al. (2009). One duplicate compound and a salt (the neutral form of this salt could not be optimized by the SYBYL program) were removed. A total of 107 pesticides were evaluated in this study. BCF values were measured on the same Bluegill (*Lepomis macrochirus*) following a standardized protocol (US EPA, 1996). BCF values are presented as log values in Table S1 (Table S1 in the Supplementary material). The modeled log BCF ranged from −0.92 to 4.00. The compounds were divided into training ($n = 80$) and test ($n = 27$) sets using DUPLEX algorithm (Snee, 1977; Puzyn et al., 2011), which allows maintenance of a comparable diversity in both sets; the latter are therefore similar in terms of representativeness (for more details, see the Supplementary material and elsewhere (Snee, 1977; Ritota et al., 2010)). The training set was used to build models, and the independent test set was used to evaluate the predictive ability of the models.

2.2. DRAGON descriptors

The three-dimensional structures of 107 pesticides were constructed using Sybyl-x 1.3. The energy of each molecule was minimized using gradient descent method and by employing Tripos force field and Gasteiger–Huckel charges. The molecular descriptors were obtained by encoding the optimized molecular structures into the DRAGON program, and 4885 different types of molecular descriptors were calculated to describe the structural diversity of the chemicals. Then, the initial pool of descriptors was reduced by applying the DRAGON built-in variable exclusion procedure. Three types of descriptors were excluded, namely, constant (relative standard deviation < 0.001), near-constant (all values are equal except one), and highly correlated descriptors. For each pair of highly correlated descriptors, R > 0.9, the descriptor with the largest mean correlation coefficient with the rest of the descriptors is removed. Thus, 1230 molecular descriptors were maintained.

2.3. Variable selection

We used IERM to select the best subset of DRAGON descriptors. The RM and ERM determined the optimal subset of d ($d \ll$ D) descriptors from a large group of D descriptors with a minimum standard deviation (S) (Mercader et al., 2010, 2011). By contrast, IERM uses an initial set of very high S determined by an "inverse RM" (RM was used to maximize S), as follows:

$$S = \frac{1}{(N-d-1)} \sum_{i=1}^{N} res_i^2 \quad (1)$$

In Eq. (1), N is the number of molecules in the training set, and res_i represents the residual for molecule i. The residual is the difference between experimental and predicted properties. $S(\mathbf{d}_n)$ is a distribution in a space of $D!/[d!(D-d)!]$ points. Full search (FS) can arrive at the global minimum by calculating $S(\mathbf{d}_n)$ for all space points, but FS is too difficult to perform if D is very large. However, IERM is more efficient than FS in reaching the global minimum. The two-step IERM was performed as follows. First, an initial set of descriptors, $\mathbf{d}_k$, was chosen by an inverse RM. One of the descriptors, X_{ki}, was replaced iteratively by each remaining $D - d$ descriptor, and the set with the smallest value of S was retained. Second, from the resulting set, the descriptor with the greatest value of S in its coefficient was chosen and replaced using all descriptors, except the one replaced in the previous iteration. This step was repeated until the set remained unchanged. IERM variable selection was implemented in MATLAB.

2.4. MLP NN

MLP NN, which is a supervised neural network, was used to investigate the nonlinear relationship among the selected molecular descriptors and log BCF of pesticides. This network generally consists of several artificial neurons arranged in three layers (topology of NN), namely, input, hidden, and output layers. The three-layer topology with a single hidden layer is sufficient for solving similar or more complex problems (Torrecilla et al., 2009). Thus, three-layer MLP was used in this work. The descriptors of the MLR model were used as inputs for the network. The Broyden–-Fletcher–Goldfarb–Shanno learning algorithm was used to develop MLP NN models. Different networks with 4–12 neurons in the hidden layer were trained to determine the optimal number of neurons in the hidden layer. Overfitting was avoided by repeating the learning process and verifying the sum-of-squares error. The sum-of-squares error is simply given by the sum of differences between the target and prediction outputs defined over the entire training set. The calculation process in each neuron in the hidden and output layers is performed by successive activation and