Chemometrics and Intelligent Laboratory Systems 154 (2016) 7–15

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Artificial neural network-based equation to predict the toxicity of herbicides on rats

Mabrouk Hamadache [a,*], Salah Hanini [a], Othmane Benkortbi [a], Abdeltif Amrane [b], Latifa Khaouane [a], Cherif Si Moussa [a]

[a] Laboratoire des Biomatériaux et Phénomènes de Transport (LBMPT), Université de Médéa, Quartier Ain D'heb, 26000 Medea, Algeria
[b] Ecole Nationale Supérieure de Chimie de Rennes, Université de Rennes 1, CNRS, UMR 6226, 11 allée de Beaulieu, CS 50837, 35708 Rennes Cedex 7, France

ARTICLE INFO

Article history:
Received 16 October 2015
Received in revised form 2 March 2016
Accepted 4 March 2016
Available online 21 March 2016

Keywords:
Acute oral toxicity
ANN-based equation
Domain applicability
Herbicides
Prediction

ABSTRACT

The use of herbicides is increasing around the world. The benefits achieved by the use of these herbicides are indisputable. Despite their importance in agriculture, herbicides can be dangerous to the environment and the human health, depending on their toxicity, and the degree of contamination. Also, it is essential and evident that the risk assessment of herbicides is an important task in the environmental protection. The objective of this work was to investigate and implement an Artificial Neural Network (ANN) model for the prediction of acute oral toxicity of 77 herbicides to rats. Internal and external validations of the model showed high Q^2 and $\overline{r_m^2}$ values, in the range 0.782–0.997 for the training and the test. In addition, the major contribution of the current work was to develop artificial neural network-based equation to predict the toxicity of 13 other herbicides; the mathematical equation using the weights of the network gave very significant results, leading to an R^2 value of 0.959. The agreement between calculated and experimental values of acute toxicity confirmed the ability of ANN-based equation to predict the toxicity for herbicides that have not been tested as well as new herbicides.

1. Introduction

Herbicides are widely used in agriculture. They are indispensable to the farmer in his fight against plant pests and diseases. They are also used to slow the spread of insects. The benefits achieved by the use of herbicides are indisputable. Despite these advantages, several environmental dangers and some potential risk have emerged from the excessive use of these compounds. For nearly fifty years, they have been detected in the water of rivers and groundwater [1–10]. They are also found in agricultural and animal products (wheat, corn, fruits, vegetables, cereals, tea, fish, milk, eggs, meat, honey and medicinal herbs, etc.) [11–14]. As a result, this contamination could give rise to serious health and safety problems for consumers.

Herbicides have a major drawback such as toxicity. Long-term exposure to herbicides can cause harm to human life and can disrupt the functioning of various organs in the body. This significant relationship between exposure to herbicides and some chronic diseases has been the subject of several scientific publications. Exposure to these persistent pesticides has been associated with health effects including cancer, headache, skin and eye irritation, immune system problems, stomach, kidney, Parkinson and Alzheimer's disease, reproductive difficulties, birth defects, diabetes, cataracts and anemia [15–22]. As seen, humans and the environment are exposed to hundreds of herbicides. The pollution caused by these compounds has become an important issue affecting the survival and the development of humans. It is evident that the risk assessment for herbicides can provide a precaution against the corresponding pollution. In environmental risk assessment, knowledge of the acute toxicity and chronic toxicity is a basic need [23–25].

Development of in silico predictive methods that are designed to reduce and replace the use of animals to predict biological activity of chemical compounds is a widely explored area of predictive toxicology [24]. This pathway is imposed for several reasons: economic considerations, reduction of time constraints, and pressure of public opinion [26]. These methods, which include Quantitative Structure–Activity Relationship (QSAR) has been used in medicinal chemistry and computational toxicology for a long time, find growing applications in chemical risk assessment and are indispensable tools for ecotoxicological risk assessment [27,28]. Of the fact that is a promising technique, an increasing interest in the use of QSAR for environmental risk assessment and for predicting toxicity [29,30] is observed.

Quantitative Structure–Activity Relationship (QSAR) models are increasingly used in toxicology, ecotoxicology, and pharmacology for predicting the activity of the molecules from their physicochemical

* Corresponding author. Tel.: +213 7 78 12 37 50; fax: +213 25 58 12 53.
E-mail addresses: mhamdeche@yahoo.fr (M. Hamadache), s_hanini2002@yahoo.fr (S. Hanini), benkortbi_oth@yahoo.fr (O. Benkortbi), abdeltif.amrane@univ-rennes1.fr (A. Amrane), latifa_khaouane@yahoo.fr (L. Khaouane), simoussa_cherif@yahoo.fr (C.S. Moussa).

http://dx.doi.org/10.1016/j.chemolab.2016.03.007
0169-7439/

properties and/or their structural characteristics. A QSAR model is a mathematical relationship between the chemical's quantitative molecular descriptors and its toxicological, biological, and physicochemical activities. These descriptors are then correlated with a toxicological response of interest through a suitable statistical approach such as linear multiple regression, discriminant analysis and artificial neural networks [31]. The establishment of QSAR models involves a number of steps and conditions: accuracy of the input data, obtain and select the relevant descriptors capable to reflect the structure of the compounds, selection of appropriate statistical tools and checking the validity and stability of the suggested model. Reference books dealing with fundamental concepts of QSAR modeling and their basic concepts for applications in risk assessment are currently available in the literature [32,33].

QSAR studies conducted by the use of artificial neural network (ANN) modeling approaches have been developed for a large number of toxic endpoints with varying methodologies and varying degrees of success. Their applications encompass both the human health effects and the environmental impact of chemicals [34]. In recent years, researchers have used different modeling techniques such as artificial neural networks (ANN) to reduce the numbers of expensive, complicated and time-consuming tests. Predictive models based on ANN have been studied extensively in many areas of medicine [35]. Advantageously, a neural network (NN) model has a distinctive ability of learning nonlinear functional relationships. It does not require any prior structural knowledge of relationships between important variables and processes to be modeled.

There are many reports about QSAR prediction of pesticides toxicity [36–38]; however, among this abundant literature, studies specifically dedicated to QSAR prediction of herbicides acute oral toxicity appear rather limited. So far, no artificial neural network-based equation has been developed to predict acute oral toxicity of herbicides on rats. Currently, testing for acute oral toxicity is still required in the toxicological assessment of chemicals and agrochemicals worldwide [39]. Consequently, the aim of this study is to develop an ANN-based equation to predict acute oral toxicity of herbicides on rats.

The step one of this work is to develop a QSAR model that could be used to predict oral acute (LD_{50}) toxicity of a diverse set of 77 herbicides on rats. The QSAR model established by using artificial neural networks and molecular descriptors satisfies the guidelines required by the Organisation for Economic Cooperation and Development (OECD). The basic requirements to develop a QSAR model were respected. The first work is to use herbicides with toxicity data with high quality obtained under the same experimental conditions (i.e., the same protocol). Selection of non-redundant and non-correlated descriptors is the second requirement. Third, the statistical tool used to derive the QSAR can be in some cases a source of mistakes and hence the commercial software Statistica was used. Finally, the model is evaluated both in terms of her robustness as well as in terms of her prediction performances and its applicability domain (AD).

The second step of this study is to calculate the oral acute (Lethal Dose: LD_{50}) toxicity of other 13 herbicides based on the developed mathematical equation using the weights of the network. The accuracy of this formula based on ANNs was investigated and the results were very encouraging.

2. Materials and methods

2.1. Rat LD_{50} data

It is well known that high-quality experimental data are essential for the development of high quality QSAR models [40]. If they are unreliable, the model will be unreliable. The rat lethal dose 50 (LD_{50} — rat, male via oral exposure) values were retrieved from Pesticide Properties DataBase (PPDB) [41]. The LD_{50} correspond to the concentration (mg/kg) of pesticide that leads to the death of 50% of rat. The LD_{50} is one way to measure the short-term poisoning potential (acute toxicity) of a material. All values of oral acute toxicity were first converted into mmol/kg body weight and the $1/LD_{50}$ [$(mmol/kg)^{-1}$] as the endpoint was examined. The initial database that included 146 herbicides was rigorously reviewed and "cleaned" by removing pesticides whose LD_{50} was not experimentally determined or whose LD_{50} was not determined in the same experimental conditions. A total of 90 herbicides with experimental data were selected to form the final database and was divided into two sets. The first set with 77 herbicides (Table 1) was dedicated to develop the QSAR model (64 herbicides for training, and 13 herbicides for test set. The second set which included 13 herbicides that had not been used for the development of the QSAR model, was left for the prediction of oral acute LD_{50} based on the developed mathematical formula using the weights and the bias of the network.

2.2. Descriptor calculation

All descriptors were obtained from the online program E-Dragon 1.0 (www.vcclab.org). The structure files of compounds under study, which are the input files for Dragon calculation, cannot be generated in Dragon. The structures have been drawn in SMILES (Simplified Molecular-Input Line-Entry System) notation. SMILES notations were obtained from the Pesticide Properties DataBase (University of Hertfordshire, 2007–2013). Herbicides compounds represented by SMILES format was used as input for calculation of 1666 molecular descriptors with the online software, E-DRAGON. The software converted the molecules from SMILES notation to 3-dimensional structures using the algorithm derived from CORINA [42]. Twenty types of descriptors were calculated by the Dragon software, like: (1) constitutional descriptors; (2) topological descriptors; (3) walk and path counts; (4) connectivity indices; (5) information indices; (6) two dimensional (2D) autocorrelations; (7) edge adjacency indices; (8) Burden eigenvalue descriptors; (9) topological charge indices; (10) eigenvalue-based indices; (11) Randic molecular profiles; (12) geometrical descriptors; (13) RDF descriptors; (14) 3D-MORSE descriptors; (15) WHIM descriptors; (16) GETAWAY descriptors; (17) functional group counts; (18) atom-centered fragments; (19) charge descriptors; and (20) molecular properties.

2.3. Selection of relevant descriptors

An important step in QSAR model is to select robust and informative descriptors from a variety of descriptors. Several methods to simplify a database are used; for example the Principal Component Analysis (PCA), curvilinear component analysis, or the method of Gram-Schmidt orthogonalization can be used. The method used to select the most significant descriptors was described previously [43,44]. In the first step, invariant descriptors, namely those with absent values (represented by the code "999"), were manually removed. Next, any descriptor that had identical values for >75% of the samples and any descriptors with a relative standard deviation < 0.05 were removed. Finally, half of the descriptors showing an absolute value of the Pearson correlation coefficient > 0.75 were also removed. The number of descriptors obtained after the selection was 76. For relevant descriptors selection, stepwise regression was then used [45]; in this procedure, a variable that entered the model in the earlier stages of selection may be deleted at the later stages. Stepwise addition of further descriptors was continued to find the best multi-parameter regression models with the optimal values of statistical criteria (highest values of correlation coefficient R^2). Stepwise regression were performed by the STATISTICA software (STATISTICA 8.0, Tulsa; StatSoft, Inc., OK, USA.) and XLSTAT software. Eighteen descriptors were selected with stepwise regression. However, it was important to reduce the number of descriptors [46,47]. Finally, the number of descriptors used to develop the model was 8: HATS1e, HATS1v, ISH, MATS1m, Gats3p, R8u, Gats6m and H-046.