Chemosphere 168 (2017) 877–884

Contents lists available at ScienceDirect

Chemosphere

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

Application of radial basis function neural network to predict soil sorption partition coefficient using topological descriptors

Mohammad Reza Sabour*, Saman Moftakhari Anasori Movahed

Faculty of Civil Engineering, K.N.Toosi University of Technology, No. 1346, Vali-e-asr Street, 19967-15433, Tehran, Iran

HIGHLIGHTS

- Development of a Neural Network model that is capable of predicting $logK_{oc}$ of organic compounds.
- Employment of Topological descriptors as the inputs of the model.
- The model is adaptive and capable of predicting $logK_{oc}$ for new products.
- The performance of this model is distinctively well, especially in test set compared with previous models.

ARTICLE INFO

Article history:
Received 2 July 2016
Received in revised form
21 October 2016
Accepted 29 October 2016
Available online 9 November 2016

Handling Editor: I. Cousins

Keywords:
Artificial Neural Network
Soil sorption
Generalized Regression Neural Network (GRNN)
Topological descriptors

ABSTRACT

The soil sorption partition coefficient $logK_{oc}$ is an indispensable parameter that can be used in assessing the environmental risk of organic chemicals. In order to predict soil sorption partition coefficient for different and even unknown compounds in a fast and accurate manner, a radial basis function neural network (RBFNN) model was developed. Eight topological descriptors of 800 organic compounds were used as inputs of the model. These 800 organic compounds were chosen from a large and very diverse data set. Generalized Regression Neural Network (GRNN) was utilized as the function in this neural network model due to its capability to adapt very quickly. Hence, it can be used to predict $logK_{oc}$ for new chemicals, as well. Out of total data set, 560 organic compounds were used for training and 240 to test efficiency of the model. The obtained results indicate that the model performance is very well. The correlation coefficients (R2) for training and test sets were 0.995 and 0.933, respectively. The root-mean square errors (RMSE) were 0.2321 for training set and 0.413 for test set. As the results for both training and test set are extremely satisfactory, the proposed neural network model can be employed not only to predict $logK_{oc}$ of known compounds, but also to be adaptive for prediction of this value precisely for new products that enter the market each year.

1. Introduction

Processes concerning with the soil sorption of organic chemicals play an undeniable role in determining the environmental fate, persistence and other behavioral aspects of chemicals such as biodegradation (Scow and Johnson, 1996; Cornelissen et al., 2005). The soil sorption partition coefficient $logK_{oc}$ is an indispensable parameter that can be used in assessing the environmental risk of organic chemicals (Gramatica et al., 2007; Wang et al., 2009; Phillips et al., 2011; Wen et al., 2012; Shao et al., 2014; Wang et al., 2015).

In order to obtain the value of K_{oc} experimental measurements such as HPLC and batch equilibrium method can be used (OECD, 2007, 2001). There is no denying that experimental data are important, however they can be improved. With these methods being expensive and time-consuming, let alone that a simple error in these tests can lead to inaccuracy, it is vital to develop a model that can be accurate, adaptive and efficient. The practicality of this model does not diminish the importance of experimental data but it can be useful for preliminary analysis.

Quantitative structure-activity relationship (QSAR) based models have been proven to be capable in the prediction of K_{oc} values. Several of these models have been published for K_{oc} (Huuskonen, 2003; Schüürmann et al., 2006; Gramatica et al., 2007; Wang et al., 2009; Wen et al., 2012; dos Reis et al., 2014; Shao et al., 2014; Wang et al., 2015). However, there are

* Corresponding author.
E-mail address: sabour@kntu.ac.ir (M.R. Sabour).

http://dx.doi.org/10.1016/j.chemosphere.2016.10.122
0045-6535/

parameters that can be improved. Some of them use octanol/water partition coefficient ($logK_{ow}$) or water solubilities ($logS_w$) of the compounds (Gawlik et al., 1997; Doucette, 2003), but the application of these models is only possible for compounds with specified $logK_{ow}$ range, Another constraint of QSAR models is that these models' validity is based on the number and diversity of their data. However, the number and diversity of the compounds and chemicals entering the market each year is sharply increasing and these models have to be updated (their data) to be valid and useful again. But the present model and the application of neural network make it possible for our model to be adaptive and be able to stay capable and powerful even with the emerging compounds, of course eventually there needs to be an update in the present model's data as well. In this study the model is trained with variety of data and will be able to respond to new emerging compounds. The neural network model learns the complex relationship between topological descriptors and $logK_{oc}$,thus, the model can be more accurate for a longer time.

Many Quantitative Structure-Property Relationship (QSPR) models were also developed, based mainly on small datasets (Briggs, 1981; Gao et al., 1996; Liao et al., 1996; Poole and Poole, 1999; Tao et al., 2001; Thanikaivelan et al., 2000; Baker et al., 2001; Tao et al., 2001; Kahn et al., 2005; Liu and Yu, 2005; Duchowicz et al., 2007; Gramatica et al., 2007; Goudarzi et al., 2009; Bronner and Goss, 2010; Wen et al., 2012). Hence, these models were not capable enough to predict unknown chemical compounds, due to narrow application domain and low generalization capability (Shao et al., 2014).

What distinguishes this study from previous ones are the followings: (a) the model developed for this study is based on molecular topological descriptors that never been exclusively reviewed before. Moreover, topological descriptors can be obtained easily either by the use of data bases or calculation by software. So, the process can be done within seconds and is not at all time-consuming. More importantly, as these descriptors are based on molecular structures they can be calculated even for unknown and new chemicals and compounds that enter the market every year, in order for our model to remain as powerful and capable as it was in the first place. (b) by reducing the number of topological descriptors from 8000 to 6400 molecular descriptors and taking rigorous measurements of OECD guidelines to show the negligible chance correlation along with the application of Radial Basis Function Neural Network (RBFNN), it is ensured that the model learning algorithm uses a diverse data set for training so as to adapt itself quickly for new chemicals and compounds that it was not even trained for (further explained in results and discussion section).

Artificial neural networks have the ability to implicitly detect complex nonlinear relationships between dependent and independent variables as it “learns” the relationship inherent in the data presented to it (Desai and Bharati, 1998). So, unlike QSAR models a non-linear approach will be reached. Furthermore, not only its error in prediction of $logK_{oc}$ is less than previous models but also this model can adapt itself to predict $logK_{oc}$ values for new chemicals and still be accurate.

2. Materials and methods

2.1. Data set

The experimental logK_{oc} values of 800 organic compounds were collected from EPI Suit Software (EPA, 2012) (Version 4.1) (http://www.epa.gov/oppt/exposure) and the literature (Yaws, 1999; Huuskonen, 2003; Razzaque and Grathwohl, 2008; Sun et al., 2010; Pose-Juan et al., 2011; Schenzel et al., 2012; Gellrich and Knepper, 2012; Stenzel et al., 2013). Even though EPI Suite software is very reliable, to further validate the data, random samples were taken and compared with molecular properties provided by Hazardous Substances Data Bank (HSDB, 2015) and also the ChemSpider website (Chemspider, 2015); the results were consistent and the differences were negligible. note that we did not exclude any of the compounds, so that our ANN model can be trained by a variety of compounds and characteristics. Therefore, the model possesses the ability to predict the logK_{oc} of new compounds, too. Out of total data set, 70% was randomly used for training and 30% for prediction. The normal distribution of logK_{oc} values for the total data set and training set with mean of 2.95 and 3.15 and standard deviation (SD) of 1.475 and 1.441 respectively, are shown in Figs. S1 and S2. It implicates that they both have the similar statistical distribution. The name of the compounds and experimental $logK_{oc}$ and topological descriptors values are provided in supplementary material.

2.2. Molecular topological descriptors calculation

The molecular topological descriptors for 800 organic compounds were extracted and calculated through taking advantage of MOLE db (Molecular Descriptors Data Base) and E-Dragon (electronic remote version of the well-known software, DRAGON) respectively. In order to use E-DRAGON the following processes were employed. First we extracted the CAS numbers (Chemical Abstracts Service) for all 800 compounds. Second these CAS numbers were used in Enhanced NCI Database Browser (version 2.2) to obtain the SMILES (simplified molecular-input line-entry system) of each organic compound. Third, to verify the molecular structures of each organic compound, the structures were imported to Chem3D software (version 15.0) (included in ChemOffice Professional package) (http://www.cambridgesoft.com/Ensemble_for_Chemistry/ChemOffice/ChemOfficeProfessional/); both MMFF94 and MM2 methods were employed to optimize the structures with minimizing the energy through molecular dynamics. Finally, the resulted optimized structures were exported and submitted to E-DRAGON software for descriptors calculation. This structural validation was performed for all 800 compounds. However, as topological descriptors mainly focus on molecule size, the changes in topological descriptors calculation were not drastic, hence, the structurally validated results had slight impact on $logK_{oc}$ prediction. (all the changes are indicated by red color in Supplementary data). It is necessary to note that, in neural network modeling every input (neurons) of the model has its individual value and by excluding descriptors the model's efficiency will decrease.

Totally, 6400 topological molecular descriptors were calculated and 8 different topological descriptors were used, all of which are explained in Table 1, as follow: (1) Pol (polarity number), (2) J (Balaban distance connectivity index), (3) ZM1 (first Zagreb index M1), (4) ZM1V (first Zagreb index by valence vertex degrees), (5) ZM2 (second Zagreb index M2), (6) ZM2V, (7) W (Wiener W index), (8) Har (Harary H index). The meaning and detailed calculation of these topological descriptors are provided by the related literature references of Handbook of Molecular Descriptors (Todeschini and Consonni, 2008).

2.3. RBFNN; background and theoretical aspects

Artificial Neural Networks (ANNs) can be used for different purposes, including: monitoring, control, classification and prediction. ANNs have the ability to simulate important parameters based on past observations. There are many types of ANNs for modeling function approximation of the engineering problems (Park et al., 2005). Radial basis networks can require more neurons than standard feedforward backpropagation networks, but often