Journal of Molecular Structure 1099 (2015) 510–515

Contents lists available at ScienceDirect

Journal of Molecular Structure

journal homepage: http://www.elsevier.com/locate/molstruc

Predicting adsorption of aromatic compounds by carbon nanotubes based on quantitative structure property relationship principles

Mehdi Rahimi-Nasrabadi [a, *], Reza Akhoondi [b], Seied Mahdi Pourmortazavi [c], Farhad Ahmadi [d, e]

[a] Nano-Science Center, Imam Hossein University, P.O. Box.: 16575-347, Tehran, Iran
[b] Instrumental Analysis laboratory, Iranian Catalyst Development Company, Varamin, Iran
[c] Faculty of Material and Manufacturing Technologies, Malek Ashtar University of Technology, Tehran, Iran
[d] Department of Medicinal Chemistry, School of Pharmacy-International Campus, Iran University of Medical Sciences, Tehran, Iran
[e] Department of Medicinal Chemistry, Faculty of Pharmacy, Kermanshah University of Medical Sciences, Kermanshah, Iran

ARTICLE INFO

Article history:
Received 6 March 2015
Received in revised form
27 June 2015
Accepted 30 June 2015
Available online 8 July 2015

Keywords:
Quantitative structure property relationship
Carbon nanotubes
Adsorption
Aromatic compounds

ABSTRACT

Quantitative structure property relationship (QSPR) models were developed to predict the adsorption of aromatic compounds by carbon nanotubes (CNTs). Five descriptors chosen by combining self-organizing map and stepwise multiple linear regression (MLR) techniques were used to connect the structure of the studied chemicals with their adsorption descriptor (K_∞) using linear and nonlinear modeling techniques. Correlation coefficient (R^2) of 0.99 and root-mean square error (RMSE) of 0.29 for multilayered perceptron neural network (MLP-NN) model are signs of the superiority of the developed nonlinear model over MLR model with R^2 of 0.93 and RMSE of 0.36. The results of cross-validation test showed the reliability of MLP-NN to predict the K_∞ values for the aromatic contaminants. Molar volume and hydrogen bond accepting ability were found to be the factors much influencing the adsorption of the compounds. The developed QSPR, as a neural network based model, could be used to predict the adsorption of organic compounds by CNTs.

1. Introduction

Carbon nanotubes (CNTs) are hollow nanosized tubes made of graphite sheets that are rolled into a cylindrical shape and closed by two caps [1–3]. Owing to particular electronic and structural characteristics, the CNTs are potential to be applied as catalysis [4], field-effect transistors [5], hydrogen storage media [6] and electrochemical sensor [7]. They can also enter the body and cause adverse effects due to their own poisonous nature as well as toxic effect of the substances adsorbed by them [8,9]. However, the unique extremely large surface area and highly hydrophobic surfaces cause CNTs to have strong adsorption affinities to a wide range of organic compounds [10–12]. They can be considered as promising means to adsorb organic contaminants from the aquatic environments [13,14].

To find the effect of the structure of an organic compound on its adsorption by CNTs and also to provide some insights into the mechanism of adsorbing organics by CNTs, developing chemoinformatics methods such as quantitative structure property relationship (QSPR) seems to be useful and instructive. The assumption is that the variation in the adsorption of an adsorbate can be correlated with changes in the molecular structures of desired organics. In QSPR studies many statistical techniques such as multiple linear regression (MLR), partial least squares regression (PLS), support vector machine and various types of artificial neural networks (ANNs) can be used to derive correlation models between descriptors of molecular structures and properties [15–18]. With the help of QSPR method as a time and money saving technique, it is possible to predict which organic contaminant can be adsorbed by CNTs in aqueous solution without resorting to experimentation. There are a limited number of reports on the development of predictive models for adsorption of contaminants. Recently, a linear salvation energy relationship (LSER) model has been developed to predict the adsorption of 28 various small molecules on CNTs [19]. The authors found that lipophilicity is the predominant descriptor and hydrogen bond basicity is the second most important factor in the LSER model. In a more recently published study, Apul et al. developed QSPR models using 33 molecular connectivity indices and also LSER models to describe the relationship between the structures of aromatic contaminants and their adsorption on CNTs

* Corresponding author.
E-mail address: rahiminasrabadi@gmail.com (M. Rahimi-Nasrabadi).

http://dx.doi.org/10.1016/j.molstruc.2015.06.085
0022-2860/

[20]. They found that the molecular volume of the LSER model is the most significant factor governing the adsorption process.

Although some theoretical models have been constructed for predicting the adsorption of aromatic compounds by CNTs, neural network based QSPR models for this purpose are still at elementary steps. In the present study, we have decided to develop new QSPR models using MLR and multilayer perceptron neural network (MLP-NN) to improve the models obtained by Apul et al. [20]. With the help of MLP-NN, a nonlinear relationship was discovered between adsorption data and chemical structures of desired compounds. Results obtained by MLP-NN were compared with those given by MLR and the role and contribution of the selected descriptors were finally investigated on adsorption of the desired compounds on CNTs.

2. Materials and methods

2.1. Data set

The data for adsorption of 59 aromatic compounds by multi walled CNTs (MWCNTs) were taken from the recent study by Apul et al. [20]. To obtain experimental values of single point adsorption descriptor (K_∞) the isotherm data had been used. The descriptor is defined as $K_\infty = qe/Ce$, where qe and Ce are solid and liquid phase equilibrium concentrations at infinite dilution conditions. We employed K_∞ as dependent variable to develop linear and nonlinear QSPR models. The values of K_∞ range from 13.0×10^{-2} to 10.2×10^{3} for benzylalcohol and pyrene, respectively. The values were transformed into logarithmic scale and the relationship between log K_∞ and molecular descriptors was examined. The components in the data set were severing randomly into the training, test and validation sets consisting of 43, 5 and 11 members, respectively. The training set participated in the generation of the model and adjusting its parameters and independent set of samples in validation set was used to assess the performance of the model. In case of MLR modeling both test and validation sets were considered as prediction set. Diversity analysis was performed on the data set in order to confirm that the whole data set is represented by the structures of the subsets [21,22]. In this way, the mean distances of one sample to the remaining ones were computed from descriptor space matrix. The details of diversity analysis are described in Supplementary content. The experimental values of K_∞ for all molecules studied in this work are listed in Table 1.

2.2. Molecular descriptors calculation, screening and selection

The simple mathematical representation of a molecule could be carried out via molecular descriptors which are used to encode the significant structural features of different molecules. The descriptors are calculated by the Hyperchem program (ver. 7) [23]. The program constructs all molecular structures. The Austin Model 1 (AM1) as the semiempirical method was used to optimize the molecular geometry. Thereafter, Dragon[1] was utilized to calculate a wide range of molecular descriptors. Four Abraham descriptors were added to the descriptor pool in order to study the possible nonlinear relationship between log K_∞ and these descriptors. These Abraham descriptors had been used previously as independent parameters to develop linear solvent energy relationships [20]. The following Abraham descriptors were included the ability of hydrogen bond donating (A), the ability of hydrogen bond accepting (B), polarizability/polarity term (P)and also molar volume (V). The molecular descriptors were obtained by the Absolv module of ADME Suite 5.0 software. The computed descriptors were analyzed to investigate the presence of a constant or near constant values and the descriptors with low variation were kept obsolete from further investigation. The remained descriptors were arranged in a 59×189 matrix, while 59 and 189 are respectively the number of the compounds and molecular descriptors. Then, the SOM network was accomplished with the transpose of data matrix in order to reduce the descriptor space and the results lead to the reduction of the descriptors number to 42. Numerous stepwise MLR models were developed to choose the most significant descriptors between the remained molecular descriptors. These established models were investigated based on the multiple correlation coefficients (R), standard error (SE), F-statistic, the number of included descriptors and the prediction ability. Based on the results, five descriptors finally were selected and utilized for the generation of nonlinear and linear models.

2.3. Multiple linear regressions

MLR equation, *i.e.*, $\boldsymbol{y} = b_0 + b_1\boldsymbol{x}_1 + b_2\boldsymbol{x}_2 + \ldots + b_n\boldsymbol{x}_n$ is used to create linear model relating a dependent variable **y** (here log K_∞) to independent variables $\mathbf{x_i}$, *i.e.*, molecular descriptors. The coefficient vector $\boldsymbol{b}$ is computed by the descriptor matrix $\boldsymbol{X}$, including a further column with ones in order to compute coefficientb_0, according to$\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. It should be noted that the MLR model is appropriately fitted, while the sum of squares correspond to the differences of predicted and experimental values are minimized.

2.4. Multilayered perceptron neural network

ANNs could be used in non-linear mapping complex patterns. Usually, a neural network contains several layers including an input layer, one or several hidden layers, and finally an output layer. The inputs in MLP-NN are completely joined to the hidden layer while the hidden layer neurons are completely joined to the outputs [24,25]. During learning step, the sequenced input patterns offered to the network are propagated to the onward direction as layer by layer until the final output layer is computed. The error term is defined as the difference between the resulted output and the target value. The computed errors are used as the input pattern for feedback joints in order to adjust the synaptic weights in backward direction layer by layer. The performance of MLP-NN model is influenced by the kind of transfer function and the neuronnumber in the hidden layer. In order to obtain an optimal MLP-NN model, different networks with 4–8 neurons in the hidden layer and with three logistic, exponential and identity functions in each hidden and output layer were trained using Quasi-Newton algorithm. The trained optimal network was utilized as an analytical tool for the prediction of log K_∞ value correspond to the compounds.

2.5. Applicability domain

The ability domain of MLP-NN model was appraised by Leverage approach to investigate the applicability of the developed QSPR model for estimation of the adsorption of new molecules without any experimental data [26,27]. The warning leverage was computed by$h^* = 3(p)/n$; where, p is representative for the number of model variable plus one. The n represents the number of training materials. The leverage value is obtained as:

$$h_i = x_i^T\left(X^TX\right)^{-1}x_i \quad i = 1, \ldots, n \tag{1}$$

In this equation,h_i is the material leverage (i) in the descriptor space, x_i is the descriptor raw-vector for the query material, while$X$

[1] http://www.disat.unimib.it/chem.