Journal of Molecular Structure 1098 (2015) 191–198

Contents lists available at ScienceDirect

Journal of Molecular Structure

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

Combining radial basis function neural network with genetic algorithm to QSPR modeling of adsorption on multi-walled carbon nanotubes surface

Zeinabe Hassanzadeh [a], Mohsen Kompany-Zareh [b], Raouf Ghavami [a, *], Somayeh Gholami [b], Atefe Malek-Khatabi [b]

[a] Department of Chemistry, Faculty of Science, University of Kurdistan, P.O. Box 416, Sanandaj, Iran
[b] Department of Chemistry, Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan 45137-66731, Iran

ARTICLE INFO

Article history:
Received 24 February 2015
Received in revised form
27 May 2015
Accepted 29 May 2015
Available online 5 June 2015

Keywords:
RBFN
GA
QSPR
wsGA-RBFN
MWCNTs

ABSTRACT

The configuring of a radial basis function neural network (RBFN) consists of optimizing the architecture and the network parameters (centers, widths, and weights). Methods such as genetic algorithm (GA), K-means and cluster analysis (CA) are among center selection methods. In the most of reports on RBFN modeling optimum centers are selected among rows of descriptors matrix. A combination of RBFN and GA is introduced for better description of quantitative structure-property relationships (QSPR) models. In this method, centers are not exactly rows of the independent matrix and can be located in any point of the samples space. In the proposed approach, initial centers are randomly selected from the calibration set. Then GA changes the locations of the initially selected centers to find the optimum positions of centers from the whole space of scores matrix, in order to obtain highest prediction ability. This approach is called whole space GA-RBFN (wsGA-RBFN) and applied to predict the adsorption coefficients (log*k*), of 40 small molecules on the surface of multi-walled carbon nanotubes (MWCNTs). The data consists of five solute descriptors [R, π, α, β, V] of the molecules and known as data set1. Prediction ability of wsGA-RBFN is compared to GA-RBFN and MLR models. The obtained Q^2 values for wsGA-RBFN, GA-RBFN and MLR are 0.95, 0.85, and 0.78, respectively, which shows the merit of wsGA-RBFN. The method is also applied on the logarithm of surface area normalized adsorption coefficients (logK_{SA}), of organic compounds (OCs) on MWCNTs surface. The data set2 includes 69 aromatic molecules with 13 physicochemical properties of the OCs. Thirty-nine of these molecules were similar to those of data set1 and the others were aromatic compounds included of small and big molecules. Prediction ability of wsGA-RBFN for second data set was compared to GA-RBF. The Q^2 values for wsGA-RBFN and GA-RBF are obtained as 0.89 and 0.80, respectively.

1. Introduction

Quantitative structure-property relationships (QSPRs) have been used as an efficient tool to predict properties of chemical compounds in various fields of science such as; chemistry, biology, pharmacology and environmental studies [1]. QSPR approach is based on the fact that macroscopic properties of a chemical compound depends on one or more distinct molecular features (descriptors) which are solely extracted from the molecular topology or geometry and are independent of the experimental properties. The main goal of the QSPR studies is to find a mathematical relationship between the interested properties and one or more descriptors derived from the structure of the molecule [2,3]. Linear and nonlinear methods have been used to construct QSPR models. The most commonly used linear methods are multiple linear regression (MLR), principal component regression (PCR) and partial least-squares regression (PLSR).

When a dependent variable is a nonlinear function of independent variables, it cannot be evaluated by a simple linear relationship. In this case, linear methods have some limitations and give poor statistical results. Flexible nonlinear artificial intelligent

* Corresponding author. Department of Chemistry, Faculty of Science, University of Kurdistan, P.O. Box 416, Iran.
E-mail addresses: rghavami2000@yahoo.com, r.ghavami@uok.ac.ir (R. Ghavami).

http://dx.doi.org/10.1016/j.molstruc.2015.05.039
0022-2860/

based algorithms, such as artificial neural network (ANN) were used to perform nonlinear mapping of the physicochemical descriptors to the corresponding property and using ANN the predictions precision was improved [4,5]. RBFN is one of the most popular neural network models and is a powerful nonlinear regression technique, which has an input, a hidden and one output layer [6–8]. Because of simple network structures, faster learning algorithm and better approximation capability, the number of publications available on RBFNs has been increased [9–12]. The configuration of an RBFN is commonly performed by initialization of RBFN (determination of the architecture and optimization of the centers and the widths) and then training the weights between the hidden and output layers. Which optimizes the various systems parameters and the system is able to give high performance.

Various methods such as K-means clustering [13], hybrid particle swarm optimization (HPSO) algorithm [14], genetic algorithm (GA) [15,16], particle swarm optimization (PSO) [17], have been used for initialization of RBFN. Sarimveis et al. proposed training methodology based on a fuzzy partition of the input space and a combination of self-organized and supervised learning [18]. The applications of GA in optimizing the architecture, the centers and the width of RBF network have been reviewed by Harpham et al. [19]. In all of the above mentioned methods, the selected centers of RBFN are the rows of the independent variable matrix. One question that arises here is: "Can optimal centers of RBFN belong to the whole space of the independent variables, if they are not exactly the rows of the independent variables?" To the best of our knowledge, GA has not been used to continuously change the numerical values of RBFN centers element by element in the whole space of data.

Carbon nanotubes (CNTs) are nanomaterials with large surface area and have attracted special attention. Because of their large surface area and high surface hydrophobicity, CNTs have strong adsorption affinity to a wide range of synthetic organic compounds, including polar/non-polar aliphatic and aromatic OCs [20,21]. Compared to conventional activated carbons (ACs), CNTs have been considered as potential adsorbents for microorganisms, volatile organic compounds (VOCs), natural organic matter (NOM) and toxins from water sources [22–26]. Because of these, studying the adsorption mechanism of organic compounds on CNTs has attracted the researchers' attention. It is expensive and time consuming to obtain the experimental adsorption data. Therefore, developing the predictive models for the adsorption of OCs by CNTs is very important. Some QSPR studies have been performed on adsorption of OCs on MWCNTs. Linear solvation energy relationship (LSER) model was developed to characterize the interactions between small organic molecules adsorbed onto the nanoparticles [27]. Wang et al. developed 3D-QSPR model for adsorption of aromatic compounds by carbon nanotubes based on physicochemical properties of adsorbed compounds and compared MLR, ANN and support vector machine (SVM) methods [28]. Apul et al. developed LSER and QSAR model of adsorption of aromatic contaminants by MWCNTs [29]. In the present investigation, a combination of RBFN and GA is employed to QSPR studies of adsorption of OCs on MWCNTs surface. RBFN is used to construct QSPR model and GA is used to optimize the numerical values of RBFN centers.

2. Materials and methods

2.1. Computer hardware and software

All calculations were run on a 2.3 GHz Intel Core i7 3610QM with 6 GB of RAM using all 7 available cores under Windows 7 operating system. Genetic algorithm of MATLAB 2010 was utilized and other calculations were also performed in the MATLAB environment.

2.2. Database sets

The adsorption properties of nanomaterials are assumed to be controlled by some interactions between nanomaterials and adsorbed compounds [30,31]. The interactions were described by solvatochromic descriptors (R, π, α, β, V) which developed by Abraham. R is the excess molar refraction representing the molecular force of lone-pair electrons, π is the effective solute dipolarity and polarizability, α is the effective solute hydrogen-bond acidity, β is the effective solute hydrogen bond basicity, and V is the McGowan characteristic volume that represents London dispersion [32]. The adsorption of molecules on nanomaterials surface is not only dependent on the solvatochromic descriptors but also is significantly influenced by the physicochemical properties of adsorbed compounds, such as hydrophobicity, $\pi-\pi$ interaction, H-bonding donation/acceptance and electrostatic interactions [33–35]. The physicochemical properties of compounds are closely related to their structures. Two data sets are used in this study. One of them is based on solvatochromic descriptors and the other is based on 3D molecular descriptors of the adsorbed compounds.

2.2.1. Data set1

Logarithm of adsorption coefficients (log$k$) of 40 small OCs are selected as dependent variable and solvatochromic descriptors are selected as independent variables from literature [27]. Dependent variable is a 40 dimensional vector ($\boldsymbol{y}$) and independent variable matrix, $\mathbf{X}$, is $m \times n$ dimensional matrix, where m (= 40) is the number of organic molecules and n (= 5) is the number of descriptors. The log$k$ values of the probe compounds on MWCNTs and their solvation descriptors [R, π, α, β, V] are given in the Supplementary information Table S1.

2.2.2. Data set2

The logarithm of surface area normalized adsorption coefficients (logK_{SA}), of 69 OCs compounds on MWCNTs is considered as dependent variables and thirteen 3D molecular descriptors of these compounds are used as independent variables. 39 of these compounds are data set1 compounds and the others are aromatic compounds including small and big molecules [28]. Dependent variable forms an m dimensional vector ($\boldsymbol{y}$) and independent variables form $m \times n$ dimensional matrix, $\mathbf{X}$. Here m and n are 69 and 13, respectively. The logK_{SA} values of the compounds on MWCNTs and their physicochemical descriptors are given in Supplementary information Table S2.

2.3. Genetic algorithm

Genetic algorithm (GA) solves complicated optimization problems that mimic the process of natural selections [36–38]. GA starts with a random population of possible solutions (chromosomes) consisting of a predefined number of individuals (rows) and variables (columns). The fitness of each chromosome is measured by computing the corresponding value of a fitness function. Then new generation are produced by techniques inspired by natural evolution, such as mutation, selection and crossover. As the algorithm proceeds, the members of the population are gradually improved. The stopping criterion in this study is minimization of Q^2_{F2} of RBFN.

2.4. RBFN

RBFN is a three layer artificial neural network which uses a set of d-dimensional radial activation functions [39,40]. Gaussian function is a typical choice for the radial basis functions. The equations for RBF activation is: