Journal of Chromatography A, 1333 (2014) 25–31

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Application of random forests method to predict the retention indices of some polycyclic aromatic hydrocarbons

N. Goudarzi [a,*], D. Shahsavani [b], F. Emadi-Gandaghi [a], M. Arab Chamjangali [a]

[a] Faculty of Chemistry, Shahrood University of Technology, Shahrood, Iran
[b] Faculty of Mathematics, Shahrood University of Technology, Shahrood, Iran

ARTICLE INFO

Article history:
Received 24 September 2013
Received in revised form 15 January 2014
Accepted 17 January 2014
Available online 30 January 2014

Keywords:
Random forest (RF)
Artificial neural network (ANN)
Quantitative structure–retention relationship (QSRR)
Polycyclic aromatic hydrocarbons (PAHs)

ABSTRACT

In this work, a quantitative structure–retention relationship (QSRR) investigation was carried out based on the new method of random forests (RF) for prediction of the retention indices (RIs) of some polycyclic aromatic hydrocarbon (PAH) compounds. The RIs of these compounds were calculated using the theoretical descriptors generated from their molecular structures. Effects of the important parameters affecting the ability of the RF prediction power such as the number of trees (n_t) and the number of randomly selected variables to split each node (m) were investigated. Optimization of these parameters showed that in the point $m = 70$, $n_t = 460$, the RF method can give the best results. Also, performance of the RF model was compared with that of the artificial neural network (ANN) and multiple linear regression (MLR) techniques. The results obtained show the relative superiority of the RF method over the MLR and ANN ones.

1. Introduction

Polycyclic aromatic hydrocarbons (PAHs) are defined to be composed of two or more fused aromatic rings containing only carbon and hydrogen atoms. They are formed mainly by the incomplete combustion of fossil fuels and coal, petrochemical cracking processing and the degradation of lubricating oils and dyes [1]. Also, they are released from burning oil, gasoline, trash, tobacco, wood or other organic substances such as charcoal-broiled meat. They can occur naturally when they are released from forest fires and volcanoes, and can also be manufactured. Other activities that release PAHs include driving, agricultural burning, roofing or working with coal tar products, coating pipes, steel making, and paving with asphalt. Currently, more than 200 types of PAHs have been found, some of which are highly carcinogenic. PAH contamination of food results not only from cooking processes such as smoking, grilling, baking and frying but also from the air through deposits, soil through transformation and water through deposition and transfer [2,3]. When ingested, PAHs can be absorbed by the gastrointestinal tract and distributed throughout the body, resulting in acute or chronic injury [4,5]. Animal tests have shown that long-term intake of PAHs can cause cancer and immature egg cell death [6]. Exposure to large amounts of coal tar creosote may result in convulsions, unconsciousness, and even death. Breathing vapors of coal tar, coal tar pitch, and creosote can irritate the respiratory tract. Eating large amounts of herbal supplements that contain creosote leaves may cause liver damage [7].

PAHs affect organisms through various toxic actions. The mechanism of their toxicity is considered to be interferenced by the function of cellular membranes as well as the enzyme systems that are associated with the membrane. Identification and determination of PAHs are very important for investigation of the environmental pollution level [8]. High-resolution gas chromatography is a commonly applied method for the analysis of PAHs in various matrices. Correlations of chromatographic retention with the physico-chemical and structural characteristics of substances are the basis for the choice of appropriate chromatographic systems, and are of great significance for solving problems involving identification of the components of complex mixtures. In some cases, the highly hazardous properties or unavailability of the PAHs may limit a chemist's access to a pure standard for comparison to unknowns. Quantitative structure–retention relationship (QSRR) On the other hand, it is hard work to determine experimentally the retention indices (RIs) for all compounds of a specific class. Therefore, alternative methods have overcome during the last years to improve the performance of substitute approaches to obtain theoretical RIs. QSRR (quantitative structure–retention relationship) represents one of the most successful tools, so that the principal aim of this work is to predict the RI data from a set of molecular properties. This technique correlates the variation of a dependent variable (RI) to the variations of a set of descriptors, and the

* Corresponding author. Tel.: +98 273339544.
E-mail addresses: goudarzi@shahroodut.ac.ir, goudarzi10@gmail.com (N. Goudarzi).

0021-9673/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chroma.2014.01.048

purpose is to achieve predictive and explanatory elements for a group of structurally analog compounds. It can be noted that RI is one of the parameters that is used as a criterion for the identification and possible separation of compounds. Thus we can predict the possibility of separation and identification by calculating the RI values for an unknown, unmeasured or even unsynthesized compound without performing any experiment. It is applicable for each compound in the presence of other compounds with the specific experimental conditions such as types of the GC column and detector, temperature, etc. Also, we can access this relation between the activity of a compound and the molecular structure properties. Quantitative structure–property/activity relationship (QSPR/QSAR) analysis is a powerful method for relating the physically measurable properties/activities to structurally related quantities of any compound. Recently, Drosos and co-workers have developed a QSRR study for some PAHs using quantum mechanics and other sources of descriptors estimated by different approaches. The B3LYP/6-31G* level of theory was used for geometrical optimization and quantum mechanics related variables. A good linear relationship was found between gas-chromatographic RI and electronic or topologic descriptors by stepwise linear regression analysis [9]. A molecular electronegativity–distance vector (MEDV) has been proposed to describe the structure of PAHs and relate their RIs for the programmed temperature SE-52 capillary-column gas chromatography by Liu and co-workers [10]. They applied multiple linear regression (MLR) in combination with the cross-validation (CV) technique, and a four-parameter QSRR of 209 PAHs was developed with the correlation coefficient (R) of 0.9812 and the root mean square error (RMS) of 15.533 between the estimated and experimental RIs. In 2001, Ferreira used partial least-squares (PLS) regression and electronic, geometric, and topological descriptors to predict some physiochemical properties of 48 PAHs such as boiling temperature, RI, n-octanol/water partition coefficient, and solubility in water [11]. Also, principal component regression (PCR) and PLS with leave-one-out cross-validation were used for building the regression models to predict the octanol–water partition coefficient and retention time indices of some PAH compounds [12]. In 2007, Héberger established QSPR models for prediction of GC RI on polar and non-polar stationary phases [13]. On the other hand, QSPR techniques were used for calculation of retention time of peptides [14], polychlorinated biphenyls (PCBs) [15], and polybrominated diphenylethers (PBDEs) [16]. The other properties such as electrophoretic mobility [17], soil sorption coefficient [18], flash point [19], octanol-water partition coefficient [20–22], impact sensitivities [23], normal boiling points [24], elution conductivity [25], retention factor [26], water solubility [27], and bioconcentration factor [28] were modeled using different QSPR techniques. QSPR analysis consists of mathematical equations relating the chemical structure to a wide variety of physical, chemical, biological, and technological properties. QSPR models, once established, can be used to predict the properties of compounds as yet unmeasured or even unknown. One of the novel and powerful tools used to calculate the rank of variables and predict the property–activity simultaneously is called random forests (RF).

The current study is devoted to explore the application of the RF algorithm in order to predict the RIs of some PAHs. Also, to evaluate the modeling power of the RF model, the results obtained were compared with those obtained using the ANN and MLR techniques.

2. Materials and method

2.1. Data

The experimental RIs of some PAH compounds were taken from Ref. 29. Analyses of PAHs were performed on the Agilent 6890GC-5973MSD. The GC separation was achieved using an HP-5MS capillary column (30 m × 0.25 mm I.D., 0.25 μm film thickness) with a GC oven temperature program 60 °C for 2 min, heated to 258 °C at 6 °C/min, then to 300 °C at 2 °C/min and hold for 4 min at 300 °C. Samples were injected in splitless mode with the injector temperature at 280 °C with helium as carrier gas at constant flow rate (1 mL/min). The mass spectrometer was operated in both the scan and secondary ion mass spectrometry (SIMS) mode for compound identification and quantitation, respectively. The QSPR model for estimation of the RIs of these compounds was established according to the following steps: the molecular structure input and the generated files containing the chemical structures were stored in a computer–readable format; the quantum mechanics geometry was optimized using the semi-empirical AM1 method; the structural descriptors were computed; these structural descriptors were selected; and the structure-RI model was generated by means of the MLR, ANN, and RF techniques. The names of the PAH compounds and their experimental and predicted RI values are shown in Table 1 . The dataset was split randomly into a training set (70%) and a test or prediction set (30%). A prediction set of 25 compounds was selected randomly from the original 83 PAH compounds, with the remaining ones constituting the training set. The molecules included in the training set with the RI values in the range of 200.00–559.90 were used to adjust the parameters of the model, and 25 compounds with the RI values in the range of 253.56–547.71 were used to evaluate the prediction ability.

2.2. Descriptor generation and screening

A major step in constructing the QSPR model is to find a set of molecular descriptors that represents variation in the structural properties of the molecules. The RI values of solutes are related to some of their structural, electronic, and geometric properties. The values for these molecular features can be encoded quantitatively by numerical values named molecular descriptors. These molecular parameters can be used to search for the best QSPR model for the RIs. The process of calculation of the molecular descriptors carried out was as follows: the two-dimensional structures of the molecules were drawn using the Hyperchem 7.5 software [30]. The final geometries were obtained using the semi-empirical AM1 method. The molecular structures were optimized using the Polak-Ribiere algorithm until the root mean square gradient was 0.001 kcal mol^{-1}. The resulting geometry was transferred into the Dragon software package for calculation of 18 classes of descriptors, developed by the Milano Chemometrics and QSPR Group [31,32]. 361 descriptors were removed because they included zero or other constant/near constant values, and did not have enough structural information. Also, only one of any pair of variables with a correlation coefficient greater than 0.90 was considered in developing the model. Finally, 193 descriptors were retained for the variable selection and construction of the model.

2.3. Random forests (RF) model

The fundamental of RF is based upon the method of classification and regression trees (CART), in which the feature space is split into disjoint cuboids (nodes), and then the response is approximated by a simple model like averaging the data in each region. The splitting algorithm is hierarchical and designed in a binary fashion, so that in each step it determines the splitting variable and the split-point along that variable. In each step of the algorithm (each node of a tree), the best splitting variable and the best input data