Journal of Chromatography A, 1420 (2015) 98–109

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Genetic programming based quantitative structure–retention relationships for the prediction of Kovats retention indices

Purva Goel[a], Sanket Bapat[b], Renu Vyas[a], Amruta Tambe[a], Sanjeev S. Tambe[a,*]

[a] Chemical Engineering and Process Development Division, CSIR-National Chemical Laboratory, Dr. Homi Bhabha Road, Pune, 411008, India
[b] Digital Information Resource Centre (DIRC), CSIR-National Chemical Laboratory, Dr. Homi Bhabha Road, Pune, 411008, India

ARTICLE INFO

Article history:
Received 4 August 2015
Received in revised form
25 September 2015
Accepted 25 September 2015
Available online 1 October 2015

Keywords:
Gas chromatography
Kovats retention index
Quantitative structure–retention relationships
Artificial intelligence
Genetic programming
Molecular descriptors

ABSTRACT

The development of quantitative structure–retention relationships (QSRR) aims at constructing an appropriate linear/nonlinear model for the prediction of the retention behavior (such as *Kovats retention index*) of a solute on a chromatographic column. Commonly, multi-linear regression and artificial neural networks are used in the QSRR development in the gas chromatography (GC). In this study, an artificial intelligence based data-driven modeling formalism, namely *genetic programming* (GP), has been introduced for the development of quantitative structure based models predicting *Kovats retention indices* (KRI). The novelty of the GP formalism is that given an example dataset, it searches and optimizes both the form (structure) and the parameters of an appropriate linear/nonlinear data-fitting model. Thus, it is not necessary to pre-specify the form of the data-fitting model in the GP-based modeling. These models are also less complex, simple to understand, and easy to deploy. The effectiveness of GP in constructing QSRRs has been demonstrated by developing models predicting KRIs of light hydrocarbons (case study-I) and adamantane derivatives (case study-II). In each case study, two-, three- and four-descriptor models have been developed using the KRI data available in the literature. The results of these studies clearly indicate that the GP-based models possess an excellent KRI prediction accuracy and generalization capability. Specifically, the best performing four-descriptor models in both the case studies have yielded high (>0.9) values of the *coefficient of determination* (R^2) and low values of *root mean squared error* (RMSE) and *mean absolute percent error* (MAPE) for training, test and validation set data. The characteristic feature of this study is that it introduces a practical and an effective GP-based method for developing QSRRs in gas chromatography that can be gainfully utilized for developing other types of data-driven models in chromatography science.

1. Introduction

Gas chromatography (GC) is a powerful analytical tool used widely in the separation and identification of components in a mixture. The retention time of a solute in a GC column depends on the various interactions that it makes with the stationary phase. The structure and properties of the solute and the stationary phase decide the kind of interactions that take place during separation. The identification of the separated compounds is based on the comparison of retention times of the reference standard and the sample. Since retention times vary with the instrumental conditions, Kovats introduced a system-independent and universal scheme termed *Kovats retention index* (KRI) [1] for reporting the retention times in a conventional one-dimensional (1D) GC separation. It uses n-alkanes as the standard references and the retention index (KRI) of an n-alkane is assigned a value equal to 100 times its carbon number. The index normalizes the instrumental variables in the GC, allowing the retention data generated on different systems to be compared. KRI of a solute at an isothermal column temperature can be calculated from the following equation:

$$\mathrm{KRI} = 100\left[n + (N-n)\left(\frac{\log t'_r\,(\mathrm{unknown}) - \log t'_r\,(n)}{\log t'_r\,(N) - \log t'_r\,(n)}\right)\right] \qquad (1)$$

where N is the number of carbon atoms in the larger alkane molecule, n denotes the number of carbon atoms in the smaller alkane molecule, $t'_r\,(n)$ represents the adjusted retention time of the smaller alkane, and $t'_r\,(N)$ refers to the adjusted retention time of the larger alkane.

Quantitative structure–retention relationships (QSRRs) [2] represent the mathematical/statistical models, which using the solutes' structural parameters as inputs, predict their chromatographic retention times or related parameters. These correlations

[*] Corresponding author. Tel.: +91 (20) 2590 2156; fax: +91 (20) 2590 2621.
E-mail addresses: ss.tambe@ncl.res.in, tambe.sanjiv@gmail.com (S.S. Tambe).

http://dx.doi.org/10.1016/j.chroma.2015.09.086
0021-9673/© 2015 Elsevier B.V. All rights reserved.

predict the retention data from the structure and properties of the solute molecules and also help in understanding the possible mechanisms of absorption and elution in the gas chromatography [3,4]. The knowledge of the relationship between the structure and the corresponding retention of isomers can significantly assist in accurately assigning a structure to an unknown compound from numerous isomeric alternatives [5].

The common approach to build a QSRR consists of the following steps: (i) development (or selection) of the descriptors for the molecular structure of the solutes, (ii) usage of an appropriate mathematical method to set up the model, and (iii) evaluation of the prediction and generalization ability of the developed model [6]. Studies on QSRRs form an important continuing activity in the chromatographic thermodynamics. In the contemporary gas chromatography, the QSRR approach is typically employed for the modeling of Kovats and linear temperature-programmed retention indices [7]. A representative list of the QSRR studies pertaining to the gas chromatography and presenting KRI prediction models for a variety of solutes and stationary phases is provided in Table 1 (also see Heberger [3], Bermejo et al. [8] and Kaliszan [48–50]).

From Table 1, it is noticed that the *multilinear regression* (MLR) and *artificial neural networks* (ANNs) are the two most commonly utilized methods for developing QSRRs. The MLR is a linear modeling technique, while ANNs perform a nonlinear function approximation. In situations where the relationship between the molecular descriptors and the KRI is nonlinear, the MLR method performs poorly and a nonlinear modeling method such as the multilayer perceptron (MLP) neural network needs to be explored. Although an efficient nonlinear modeling method, for a typical laboratory worker the drawback of the MLP is its complexity [38]. Also, the "black-box" nature of the ANN-based models poses significant difficulties in interpreting the model parameters (network weights) in terms of the data used in the model construction. Owing to their complex architecture, the MLP-based mathematical models are often found difficult to understand and deploy in a practical setting.

The field of artificial intelligence (AI) comprises a data-driven modeling paradigm, namely "*genetic programming* (GP)." It was proposed [51,52] as a systematic method for getting computers to automatically solve a pre-defined problem starting from a high-level statement of what needs to be done. The GP formalism has another important application, namely *symbolic regression* (SR), which is of interest to this study. Upon provided with (a) an example data set consisting of the values of the dependent and independent (predictor) variables, and (b) the form of the data-fitting function, the conventional linear and nonlinear regression analyses estimate the parameters associated with the function. The SR formalism is different than these types of regression analyses. It possesses following characteristics [53–55]: (i) SR searches the space of the mathematical expressions, while minimizing various error metrics, (ii) it simultaneously searches both, an appropriate linear or a nonlinear form (structure) and the associated parameters of a function that fits the given example data optimally, (iii) SR makes no assumptions regarding the form of the probable data-fitting functions, (iv) the optimal expressions searched and optimized by SR are of low complexity and therefore easy to deploy, (v) it is capable of identifying the key predictors and their combinations in the data, and (vi) the obtained models are amenable to the human interpretation and help explicate the observed phenomena underlying the example data. Despite its several attractive properties and significant potential, the GP-based SR has not been utilized as frequently as ANN and support vector regression (SVR) formalisms in the various science, engineering and technology branches. Some of the applications of the GP in chemical sciences and engineering, are soft-sensor development for biochemical systems [55], fermentation modeling [56], electronic nose [57], synthesis of heat-integrated complex distillation systems [58], classification of Raman spectra [59], optimization of a controlled release pharmaceutical formulation [60], modeling of a nanofiltration process [61], prediction of higher heating values of biomasses [62] and multiple alignment of liquid chromatography–mass spectrometry data [63].

An exhaustive literature search (also see Table 1) has revealed that the GP formalism has not been used in the development of QSRRs; it has also been rarely employed in the chromatography science. Accordingly, the objective of this paper is to introduce the GP technique as an attractive alternative to MLR, ANN, and other data-driven modeling formalisms for developing QSRRs. The effectiveness of the GP has been demonstrated by developing QSRRs for the prediction of Kovats retention indices for two sets of compounds namely, light hydrocarbons (case study-I) and adamantane derivatives (case study-II). In both the case studies, GP-based KRI models have been developed using two, three and four molecular descriptors as inputs. The results of these case studies clearly reveal that the GP-based QSRRs possess an excellent KRI prediction accuracy and generalization capability. For affording a rigorous comparison of the performance of the GP-based QSRRs, KRI predicting nonlinear models have been developed using the MLP neural networks. This comparison indicates that although the KRI prediction accuracies of the GP- and MLP-based QSRRs are comparable, the models belonging to the former category possess a superior generalization capability.

The remainder of this paper is structured as follows. In the second section titled "Methods," an overview of the genetic programming based symbolic regression is presented first, followed by its stepwise procedure. Section 3 titled "Results and Discussion" begins with the explanation of the molecular descriptors used in the development of the GP-based KRI prediction models; next, this section provides the details of the two case studies wherein the GP-based QSRRs have been developed for the prediction of KRIs in respect of the light hydrocarbons and adamantane derivatives. Both these case studies also present results of the sensitivity analysis of the seven molecular descriptors used in the GP-based modeling. Finally, "Conclusion" section summarizes the main findings of this study.

2. Methods

2.1. GP-based symbolic regression

Given a multiple input–single output (MISO) example dataset, D, consisting of N_{pat} number of input–output patterns, the task of the GP-based symbolic regression is to obtain an appropriate linear/nonlinear form and the associated parameters of a function (f) that best-fits the input–output data. Each pattern in the dataset contains L inputs ($x_1, x_2, \ldots, x_L$) and a single output (y), and the generalized form of the equation to be fitted is given by:

$$y = f(x_1, x_2, \ldots, x_l, \ldots, x_L;\ \beta_1, \beta_2, \ldots \beta_j, \ldots, \beta_J) \tag{2}$$

where β_j (j = 1,2,. . ., J) denotes a function parameter [62].

The GP-based symbolic regression is an iterative procedure. It begins with generating a population consisting of a pre-specified number of randomly formed equations (candidate/probable solutions). These compete to model—in the most parsimonious way—the given example data set consisting of the input (descriptor/predictor/independent) and dependent (output) variables. The candidate solutions are commonly coded in the form of "tree structures." The SR method forms a new generation of solutions using four steps, namely, *fitness evaluation*, *selection*, *crossover* and *mutation*. Here, new candidate equations are generated by recombining the previous equations (crossover) and probabilistically varying their sub-expressions (mutation).