Journal of Chromatography A, 1216 (2009) 5302–5312

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Comparative multiple quantitative structure–retention relationships modeling of gas chromatographic retention time of essential oils using multiple linear regression, principal component regression, and partial least squares techniques

Li-Tang Qin [a], Shu-Shen Liu [a,*], Hai-Ling Liu [b], Juan Tong [a]

[a] Key Laboratory of Yangtze River Water Environment, Ministry of Education, College of Environmental Science and Engineering, Tongji University, Shanghai 200092, China
[b] State Key Laboratory of Pollution Control and Resource Reuse, College of Environmental Science and Engineering, Tongji University, Shanghai 200092, China

ARTICLE INFO

Article history:
Received 20 March 2009
Received in revised form 3 May 2009
Accepted 8 May 2009
Available online 15 May 2009

Keywords:
QSRR
Retention time
Essential oils
E-state index
TSAR

ABSTRACT

Quantitative structure–retention relationships (QSRR) models were built for a data set consisting of 96 essential oils and used to predict their gas chromatographic (GC) retention times (t_R). Multiple linear regression (MLR), principal component regression (PCR), and partial least squares (PLS) have been applied to build different QSRR models by using 13 nonzero E-state indexes and 56 descriptors calculated from TSAR software. The three chemometric methods (MLR, PCR, and PLS) for evaluation of GC t_R values of essential oils have been compared. The best model based on the whole data set derived from MLR model (model M2) appears to be the best predictive power ($r^2 = 0.9689$ and $q^2 = 0.9631$) for this data set. The whole data set was splitted into a training set consisting of 72 compounds and a test set consisting of 24 compounds. The model based on the training set derived from MLR offered the highest r^2 of 0.9756 and q^2 of 0.9693. The best model base on the training set obtained from PLS not only showed a good internal predictive power ($r^2 = 0.9703$ and $q^2 = 0.9633$) but also offered the highest external predictive power ($R^2 = 0.9588$ and $q^2_{ext} = 0.9572$). The results showed that two E-state indexes (sssCH and sOH) and five molecular connective indices ($^1\chi_B$, $^2\chi_p$, $^3\chi_C$, $^4\chi_C$, and $^6\chi_p$) closely relate to the GC t_R values of essential oils. The applicability domain of the QSRR models were defined by control leverage values (h^*) and the models can be used to predict the unknown compounds falling in this domain.

1. Introduction

Essential oils, a new approach to prevent the proliferation of microorganism or protectetion of food from oxidation, are ubiquitously used as antibacterial [1–3], antifungal [3,4], and antioxidant [5] and made them useful as natural additives in the food industry. They are also used to control human diseases of microbial origin and to cure such diseases as atherosclerosis and cancer [6]. These essential oils have been used in the folk medicine for thousands of years as antimicrobial [7,8]. Therefore, the assessment of gas chromatographic (GC) retention times (t_R) of essential oils is a matter of great importance in the health of human being.

The evaluation and characterization of essential oils such as the extraction with organic solvents and simultaneous steam distillation were analyzed by GC analysis. Following GC analysis, identification of the ingredients in the essential oils can be carried out. Many researchers have studied on the GC t_R values of essential oils through experimental determination. Capillary one-dimensional gas chromatography (1D-GC) and multi-dimensional gas chromatography (MD-GC) have routinely been used to analyze the volatile constituents of essential oils [9–21]. The 1D-GC methods, however, extend run times of the complex nature of essential oil samples. The MD-GC methods are only able to analyze a few discrete and critical regions of the chromatogram. Moreover, its analysis time is also further extended. The experimental determination of GC t_R values are time-consuming and expensive. Thus, it is important to use convenient way such as quantitative structure–retention relationships (QSRR) model for the prediction of GC t_R values.

QSRR can be used for prediction of t_R values of chemicals. The process of QSRR model development for GC t_R values can be generally divided into three stages: data preparation, data analysis, and model validation [22]. The first stage includes collection of GC t_R values and calculation of molecular descriptors. The second stage includes an application of statistical approaches for QSRR model development. The last step is validation of QSRR model being built. The study of GC t_R values of different kinds compounds by QSRR have been found in several literature [23–26]. Seldom QSRR researches, however, have been found for the essential oils. Riahi et al. [27] studied on the retention indices of the components of the 44 essential oils by use of QSRR analysis based on genetic algorithm. Liao et al. [28] applied molecular electronegativity-distance

* Corresponding author. Tel.: +86 021 65982767; fax: +86 021 65982767.
E-mail address: ssliuhl@263.net (S.-S. Liu).

0021-9673/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2009.05.016

vector (MEDV) descriptors for the GC t_R values of 69 components from essential oil of *Paulownia tomentosa* flowers. The MEDV-based QSRR model, however, was not strictly validated by external samples and only offered r^2 of 0.9293 and q^2 of 0.9101. Therefore, it is still necessary to further study on the GC retention time of essential oils by using QSRR model.

The main theme of this study is the application of QSRR for GC t_R values of the essential oils and compares the results of multiple linear regression (MLR), principal component regression (PCR), and partial least-squares regression (PLS) for the evaluation of GC t_R values. The selection of optimum descriptors was performed by using variable selection and modeling method based on prediction (VSMP) [29], which was developed in our laboratory. The model was validated by internal validation (leave-one-out (LOO) cross-validation) and external validation (validate by external samples). The criteria recommended by Golbraikh and Tropsha [22] were used to assess the actual predictive power of the models. The domain of QSRR application was defined by leverage [30] and the models can be used to predict GC t_R values for untested compounds that belong to this domain.

2. Material and methods

2.1. The data set and descriptors

The data set of GC retention times (t_R) of 96 essential oil compounds was taken from the values reported in the literature [31] and listed in Table 1 . Retention times on a 30 m × 0.25 mm internal diameter (I.D.) HP5-MS fused silica with a film thickness of 0.25 μm (Agilent Technologies), were determined by gas chromatography/mass spectrometry (GC/MS). The GC system was operated under temperature programmed conditions: 60 °C for 1 min to 260 °C at 5 °C/min, then held isothermal for 10 min. Helium was used as the carrier gas at an initial pressure of 89.0 kPa for 1 min and then programmed at 1.6 kPa/min to a final pressure of 152.0 kPa, with a hold of 10 min. The t_R values of 96 essential oils cover a wide range from 5.02 (compound **1**, tricyclene) to 45.76 (compound **96**, ethyl palmitate). Nineteen compounds display the t_R values between 5.00 and 10.00, 27 between 10.00 and 20.00, 30 between 20.00 and 30.00, and 20 between 30.00 and 50.00. The distribution of GC t_R values of 96 essential oils are shown in Fig. 1.

The QSRR analyses were performed using E-state index, molecular attributes, and molecular indices. There are 41 original E-state index descriptors for a given compound. The values of 28 E-state index for the essential oils equal to zero and only 13 nonzero E-state indexes were used in this study. Various molecular attributes such as molecular mass, molecular surface area, and log *P* were calculated using TSAR 3.3 software (Accelrys, Oxford Molecular Limited, Oxford, England, 2000). The molecular indices such as molecular connectivity indices [32], shape indices [32], and topology indices (Randic [33], Balaban [34], and Wiener [35]) were also calculated using TSAR 3.3 software. The E-state index, molecular descriptors, and molecular attributes are given in Table 2.

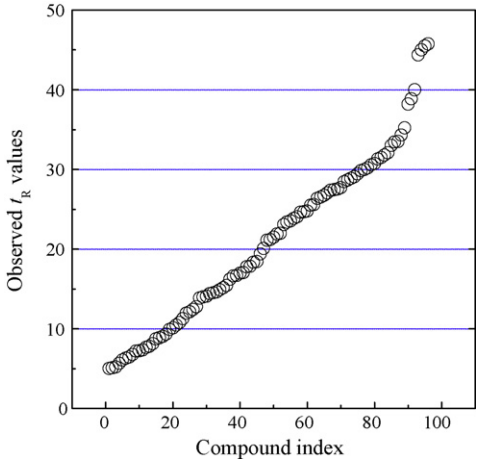

Fig. 1. Distribution of retention time (t_R) values for the whole data set.

2.2. The calculation of E-state index

E-state index, which was developed by Kier and co-workers [36,37], is a 2D topological molecular structures and based on atom levels. It closely related to molecular connectivity, but it only needs to classify the electronic valence of molecular environment and partial topological information of non-hydrogen atom and need not to break up the fragment subgraph in every order.

The calculation of E-state index was summarized according to the original literature of E-state index [37] and our previous study [38]. First, the intrinsic state (I) of every atomic type is calculated:

$$I = [(2/N)^2\delta^v + 1]/\delta \quad (1)$$

where N is the principal quantum number for the valence shell of that atom; δ^v and δ are the molecular connectivity δ values, which are given in Eq. (2).

$$(\delta = \sigma - h,\ \delta^v = \sigma + n + \pi - h) \quad (2)$$

where σ and π are the number of electrons in σ and π orbital, respectively; n is the number of electrons in lone pairs; h is the number of hydrogen atoms bonded to the atom.

Then, the other non-hydrogen of perturbation term (ΔI_i) in molecular topological and the E-state index for atom i, named as S_i, are defined as follows:

$$\Delta I_i = \sum_{j \neq i}^{\text{all}} (I_i - I_j)/(d_{ij} + 1)^2 \quad (3)$$

$$S_i = I_i + \Delta I_i \quad (4)$$

where d_{ij} is the shortest graph distance between atoms i and j.

2.3. Molecular connectivity indices

Hall and co-workers [32] have developed molecular connectivity indices (Chi) that reflect the atom identities, bonding environments and number of bonding hydrogens. Molecules that are drawn without hydrogen atoms can be decomposed into fragments of length m, which may be divided into different categories. Hall and Kier defined four series of fragment categories: path (P), cluster (C), path/cluster (PC), and ring (CH). The spread and numbers of fragment membership for each category is defined by molecule connectivity.

Fragment of length 1 can only be of type path. Kier Chi1 ($^1\chi_p$) is based on a contribution from every molecule edge, i.e., TSAR 3.3 for windows (Reference Guide, Oxford, England, 2000),

$$^1\chi_p = \sum_{j=1}^{T}\sum_{i=1}^{T}[(\delta_j\delta_i)^{-0.5}] \quad (5)$$

where i and j are bonded to give a fragment and there are T fragments of path length 1, in the skeleton.

For Chi of order $m \geq 2$ fragments may be of type path, cluster, or ring (ring/clusters are formed for $m \geq 4$). Each index type is defined as a sum of connectivity terms c_s. For a special category type t, with a membership of T fragments made up of N atoms, the connectivity