QSPR Modeling is Able to Predict Retention Times of Fatty Acids Using Simple Molecular Descriptors

Mariene H. Duarte, Federal University of Lavras, Lavras, Brazil
Matheus P. Freitas, Department of Chemistry, Federal University of Lavras, Lavras, Brazil
Cleiton Nunes, Department of Food Science, Federal University of Lavras, Lavras, Brazil

ABSTRACT

A QSPR modeling was carried out to predict the chromatographic retention times of a series of fatty acid methyl esters (FAME) widely used as standard in the characterization of lipids from agricultural and food products. Number of carbons, total double bonds, position of double bonds and geometric isomerism were used as descriptors to generate a Multiple Linear Regression (MLR) model. The best model yielded an RMSE = 0.167 and $R^2 = 0.999$ for the calibration set, and RMSE = 0.151 and $R^2 = 1.000$ for the test set. Number of carbons and total double bonds were the most important descriptors, according to the regression coefficients, but position of double bonds and isomerism cannot be neglected as they provide relevant information to improve the accuracy of the predicted property.

KEYWORDS

Agriculture, Chemometrics, Composition, Food, Lipid, Multiple Linear Regression, Multivariate Calibration, Prediction

INTRODUCTION

Identification and quantification of fatty acids is a typical analytical procedure carried out over various agricultural and food matrices. The information obtained from these analyses can be used to access a number of intrinsic characteristics of the lipid portion in these products, such as chemical quality, healthiness, stability, origin, authentication and adulteration (Laroussi-Mezghani et al., 2015; Yun & Surh 2012; Yang et al., 2013).

In order to identify and quantify fatty acids in food, there are some standardized methods, from which the more common are those reported by The American Oil Chemists’ Society (AOCS), International Standard Organization (ISO) and International Union of Pure and Applied Chemistry (IUPAC). In general, these methods are based on preparation of methyl esters from fatty acids by esterification in alkaline or acid conditions followed by separation using gas chromatography (GC) with flame ionization detection. A crucial step of this analytical procedure is the identification of the fatty acids, which is commonly performed by comparison of the retention times of the fatty acids methyl esters (FAME) in the sample with the FAME in a standard mixture. This FAME standard mixture can be either prepared in laboratory or can be commercially available. However, the possibility of identification of a FAME in a sample is restricted by the FAME present in the standard mixture. For instance, if a chromatographic peak of a given sample does not correspond to any signal of the standard FAME mixture, its identification would not be possible, unless FAME candidates are analyzed using trial and error or the retention time of the possible FAME is somehow found/predicted.
Accordingly, some studies report the prediction of chromatographic retention times using Quantitative Structure-Property Relationship (QSPR) approaches. QSPR was employed to model retention times of 368 pesticide residues in animal tissues separated by GC using number of nitrogen atoms, solvation connectivity index-Chi 1, Balaban Y index, Moran autocorrelation-lag 2/weighted by atomic Sanderson electronegativity, total absolute charge and radial distribution function-6.0/unweighted as descriptors (Dashtbozorgi et al., 2013). Similarly, first component WHIM index (E1v), highest eigenvalue n.7 of burden matrix/weighted by atomic van der waals volume (BEHv7), average connectivity index Chi-2 (X2a), 3D-MoRSE signal 23 weighted by atomic Sanderson electronegativity (MoR23m) and principal moments of inertia B (PMIB) were also used as descriptors to predict the retention times of pesticides in a QSPR study (Hadjmohammadi et al., 2007). The retention times of peptides separated by liquid chromatography (LC) were predicted by QSPR based on several classes of theoretical descriptors (Golmohammadi et al., 2015). QSPR models able to predict retention times for volatile organic compounds analyzed by GC (Sarkhosh et al., 2012) and for herbicides eluted by LC (Torrens & Castellano 2012) have been reported. The results of these studies revealed the reliability and good predictability of QSPR models to predict the retention times of molecules eluted by GC or LC. In some cases, chromatographic retention times of e.g. ethers (Golmohammadi et al., 2015), peptides (Jiao et al., 2014), and even general organic compounds (Luan et al., 2008) have been shown to be better described using non-linear techniques such as Support Vector Machines (SVM) or Artificial Neural Network (ANN).

In this work, a QSPR analysis was carried out to model the GC retention times of fatty acids in a mixture of FAME’s widely used as standard for the characterization of lipids in agricultural and food products, consisting of saturated and unsaturated FAME’s with number of carbons ranging from 10 to 24, differing also in the number and position of double bonds, as well as in geometric isomerism. Information on the number of carbons, total double bonds, position of double bonds and geometric isomerism were used as descriptors to build Multiple Linear Regression (MLR) model.

**MATERIAL AND METHODS**

The data set used in this work comprises a series of 34 fatty acids with experimentally determined retention times. The experimental retention times were obtained using a Supelco 37 Component FAME Mix eluted by GC (Shimadzu GC-2010 plus) equipped with an SPTM-2560 capillary column (100 mm x 0.25 mm x 0.2 µM) and an ionization detector (FID). The chromatographic conditions include: slit ratio of 1:100; initial column temperature of 140 °C for 5 min., heat from 140 °C to 240 °C at 4 °C/min, maintenance of 240 °C for 30 min.; helium as the carrier gas at 1 mL/min; and detector injector temperature of 260 °C.

Based on the structural characteristics of this series of compounds, eight molecular descriptors were generated: number of carbons (C), total double bonds (TDB) and double bond position (DB1, DB2, DB3, DB4, DB5, DB6) in the carbon chain. Table 1 shows the retention times for the series of fatty acids and the corresponding descriptors. The descriptors matrix (X) was regressed against the retention time column vector (Y) by multiple linear regression (MLR), using 80% of the compounds in the training set and 20% in the test series (randomly selected). The Chemoface program was used for the calculations (Nunes et al, 2012).

The MLR modeling was performed using 3 approaches: 1) descriptors matrix without preprocessing; 2) autoscaled descriptors matrix; 3) only the most relevant descriptors for the retention time.

The quality of the models was evaluated by the statistical calibration parameters R² (squared correlation coefficients in calibration) and RMSEC (root mean square error of calibration). Subsequently, the model was validated using 10-fold Y-randomization [analyzed using R²y-rand and RMSE (y-rand)] and external validation (analyzed using R²pred and RMSEP) procedures, with the latter performed on the validation set of compounds not previously used to build the model. The former
Multilayer Perceptron Model for Predicting Acute Toxicity of Fungicides on Rats

Checking the Efficacy of Two Basic Descriptors With a Set of Properties of Alkanes

Predictive Quantitative Structure Toxicity Relationship Study on Avian Toxicity of Some Diverse Agrochemical Pesticides by Monte Carlo Method: QSTR on Pesticides