International Journal of Mass Spectrometry 434 (2018) 172–178

Contents lists available at ScienceDirect

International Journal of Mass Spectrometry

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

Quantitative structure-retention relationships model for retention time prediction of veterinary drugs in food matrixes

Hamada A.A. Noreldeen[a,b,c], Xingyu Liu[a], Xiaolin Wang[a], Yanqing Fu[a,b], Zaifang Li[a,b], Xin Lu[a], Chunxia Zhao[a,*], Guowang Xu[a,*]

[a] CAS Key Laboratory of Separation Science for Analytical Chemistry, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian, 116023, China
[b] University of Chinese Academy of Sciences, Beijing, 100049, China
[c] Marine Chemistry Lab., Marine Environment Division, National Institute of Oceanography and Fisheries, Hurghada, 84511, Egypt

ARTICLE INFO

Article history:
Received 30 May 2018
Received in revised form 6 September 2018
Accepted 20 September 2018
Available online 25 September 2018

Keywords:
Quantitative structure-retention relationships
Liquid chromatography–mass spectrometry
Illegal additives
Food matrixes

ABSTRACT

Quantitative structure-retention relationships (QSRR) is a technique used in the prediction of the retention time of compounds based on their structure and chromatographic behavior. In this study, an easy and usable QSRR model was established based on multiple linear regression (MLR) to predict three kinds of illegal additives in food matrixes. For this purpose, 95 drugs were chosen, including a training set of 62 drugs, a test set of 30 drugs, and a real sample set of 3 drugs. The molecular descriptors for each compound were obtained by free softwares of advanced chemistry development (ACD) and toxicity estimation software tool (TEST). After that, the MLR-based QSRR model was established, both internal and external validation was used for validation of this model. The result indicated that the following descriptors have great influence on the predicted retention time: ACDlogP, ALOGP, ALOGP2, Hy, Ui, ib, BEHp1, BEHp2, GATS1m, GATS2m. The correlation coefficient for fitting model revealed a strong correlation between the drug retention time and selected molecular descriptors (R^2 = 0.966). Moreover, the four validation methods (leave-one-out, k-fold cross-validation, test set, and real sample set) indicated the high reliability of this model. In conclusion, this method provided a more suitable and usable model for research work in several branches of analytical chemistry, especially in the field of food safety to improve the ability of retention time prediction for illegal additives.

1. Introduction

Nowadays the development of quantitative structure-retention relationships (QSRR) technique can greatly improve the prediction accuracy and running efficiency. The primary aim of QSRR model is to obtain the mathematical relation between the molecular descriptors and chromatographic retention time of the different compounds. Many QSRR studies have been done on the gas chromatography and liquid chromatography [1–6]. To assess the quality of QSRR models and other related methods, such as quantitative structure-property relationships (QSPR) and quantitative structure-activity relationships (QSAR), the correlation coefficient (R^2) and the standard error measures were commonly used. A high R^2 and low error in an external validation step usually indicate the applicability and reliability of a developed model [7–10].

In QSRR model many molecular descriptors can be obtained for one analyte. As a result, the number of molecular descriptors should be reduced and optimized by suitable variable selection. One of these variable selection techniques, "leaps" is R's package which has often been used for variable selection [11]. As well, the modelling has been achieved by several techniques such as artificial neural network (ANN) [12,13], partial least square (PLS) regression [12,14], support vector machine (SVM) [15], and multiple linear regression (MLR) [12,14,16,17]. MLR is the most common technique in QSRR modelling due to its simplicity and ease of interpretation [12].

MLR is widely used in QSRR studies, the general purpose of multiple regressions is to quantitate the relationship between molecular descriptors and retention time. A multilinear model can be represented as:

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \beta 3x3 + \ldots + \beta kxk + \varepsilon \quad (1)$$

where k is the number of molecular descriptors (independent variables), β is the regression coefficients and y is the retention time (dependent variable) [17]. To avoid the over-fitting of MLR model, model prediction power can be tested by leave-one-out, cross-validation, and external test set [18].

* Corresponding authors.
E-mail addresses: zhaocx@dicp.ac.cn (C. Zhao), xugw@dicp.ac.cn (G. Xu).

https://doi.org/10.1016/j.ijms.2018.09.022
1387-3806/

Recently food safety issues attract more and more attention from the public. Phenomena of illegal additions have been frequently occurred, such as the abuse of inedible substances, overuse of veterinary drugs and pesticides in food products, and so on [19–23]. All these illegal additives have a harmful effect on the human health [19–23]. However, in many cases these illegal additions and potential risks are unknown, new screening and identification methods need to be developed to reduce the food security risk and provide necessary technical support for risk warning of food. Based on the advantages of QSRR, developing a more accurate MLR-based QSRR model to predict the retention time of unknown compounds will help to identify illegal additions and avoid health damage caused by these illegal additives [19,20].

This study aimed to establish MLR-based QSRR model and predict the retention time of illegal additions in food matrices. To optimize the performance of MLR and overcome its limitation in dealing with big data set, we reduced the huge number of descriptors data set to a set of fewer than 50 descriptors based on the best correlation with the experimental retention time. For descriptors calculation, two free softwares of advanced chemistry development (ACD) and toxicity estimation software tool (TEST) were used in this study. Furthermore, the model discrimination was exhibited excellent in the internal validation set and the external validation set.

2. Materials and methods

2.1. Materials

Acetonitrile (ACN, HPLC grade) and methanol (MeOH, HPLC grade) were purchased from Merck (Darmstadt, Germany). Formic acid (FA) was acquired from Sigma-Aldrich (St. Louis, MO, USA), and Ultrapure water (H_2O) was prepared by a Milli-Q system (Millipore, Billerica, MA, USA).

2.2. Sample preparation

All ninety-five compounds of three kinds of veterinary drugs were prepared at 100 ng/mL by different dissolved solvents as described in previous work [19]. For real sample preparation, 200 mg (±1 mg) of homogenized fish sample was weighed and placed into a 2 mL centrifuge tube and 1 mL of ACN with 1% FA (v/v) were added into the sample with the addition of zirconia bead into the tube simultaneous homogenization and extraction were carried out in a mixed grinding apparatus (MM400, Retsch, Germany) under the condition of 20 Hz for 1 min. After that, centrifugation was done at 14,000 rpm for 10 min at 4 °C in a Sorvall Biofuge Stratos centrifuge system (Thermo Fisher Scientific). Then the supernatant was transferred to the 2 mL Eppendorf tube and freeze-dried in a lyophilizer (Kansas, U.S.A.). The residue was redissolved with 200 μL 20% ACN (v/v), vortexed 1 min, and centrifuged at 14,000 rpm for 10 min at 4 °C. Three compounds were added, adjusting the concentration to be 100 ng/mL. Finally, the mixture is ready for LC/MS analysis.

2.3. Liquid chromatography–mass spectrometry (LC–MS) analysis

The UHPLC-MS analysis was performed on Acquity UHPLC system coupled with a LTQ Orbitrap XL hybrid mass spectrometer (Thermo Fisher). ZORBAX SB-Aq column (50 mm × 2.1 mm, 1.8 μm, Agilent, USA) was used for chromatographic analysis of both the three kinds of drug compounds with and without fish tissue matrix. The mobile phase consisted of two solvents: 0.1% formic acid in water (v/v, A) and 0.1% formic acid acetonitrile (v/v, B). The gradient started with 2% B and held for 1 min, increased to 40% B within 7 min, increased further to 95% B linearly within 1 min and held for another 2 min, then dramatically decreased to 2% B within 0.1 min, and finally held at 2% B within 0.9 min to equilibrate the column. The total run time for each sample was 12 min, the injection volume was 5 μL, and the flow rate was 0.35 μL/min. The column temperature was set at 45 °C. The ion source was an electrospray in positive mode [19].

For MS signal acquisition, the scan range was set from 50 to 1000 Da. The MS was operated with a resolution of 30 K. The ion source parameters were set as follows: the capillary temperature was set at 325 °C, the flow rates of sheath gas were 45 arbitrary units, with the auxiliary gas set at 10 arbitrary units.

2.4. QSRR model design

Fig. 1 illustrated the QSRR model developed in this study. Firstly, the data set was created and molecular descriptors were calculated. Secondly, data was split into a training set and test set with a ratio of 2:1, then the training set is used to select the most significant descriptors related to the retention time. Thirdly, MLR model was built and its power was checked by external validation (test set) and internal validation methods [7]. Finally, real sample set was (external validation) used as well for a model validation.

2.4.1. Calculation of the molecular descriptors

In this study, two free softwares were used to calculate the molecular descriptors of drug compounds. The benefit of these free softwares is the availability at costless and can be download from the website at any time and easy to install on a personal computer. Firstly, the free version of ACD software (Advanced Chemistry Development, Toronto, Canada) was employed to calculate the log P. Then toxicity estimation software tool (TEST, v4.2, Cincinnati, OH, USA) was employed to calculate the rest molecular descriptors, such as highest eigenvalue 1 of Burden matrix/weighted by atomic polarizabilities (BEHp1), highest eigenvalue 2 of Burden matrix/weighted by atomic polarizabilities (BEHp2), Geary autocorrelation-lag 1/weighted by atomic masses (GATS1m), Geary autocorrelation-lag 2/weighted by atomic masses (GATS2m) and so on. After that, suitable molecular descriptors obtained from the two free softwares were optimized and selected. The descriptors including highly correlated descriptors, descriptors not available for all compounds, and descriptors with near constant value were removed, then the data set became ready to build QSRR model [24,25].

2.4.2. QSRR model building and optimization

The dataset was divided into two subsets: one for the training set and the other for the test set. Firstly, the training set was used to build QSRR models for three kinds of veterinary drugs. The variable selection for MLR was done by free package “leaps”, (V. 3.0) [11]. QSRR modelling was carried out by multiple linear regression based on R language. QSRR optimization was also done by adjusting “leaps” package at n = 10, n = the maximum number of the variables and run the MLR for 10 times, finally, the best equation based on R^2, the root mean squared error (RMSE), and mean absolute error (MAE) values were selected.

2.4.3. Model validation

Validation of QSRR model is the most important and essential part, because of that, the validation in this study was carried out by two basic principles, external and internal validation. For internal validation, the free package “boot”, (V. 1.3–20) [26] was used to carry out K-fold cross-validation (CV) at K = 10 and leave-one-out (LOO) cross-validation, RMSE and MAE were considered to calibrate and evaluate the accuracy, quality, and determine the error between the experimental and predicted retention time in the QSRR model [24,25,27].