Journal of Cereal Science 79 (2018) 303–310

Contents lists available at ScienceDirect

Journal of Cereal Science

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

A retention index-based QSPR model for the quality control of rice

Cristian Rojas [a, b, *], Piercosimo Tripaldi [b], Andrés Pérez-González [b], Pablo R. Duchowicz [a, **], Reinaldo Pis Diez [c]

[a] Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), CONICET, UNLP, Diag. 113 y 64, C.C. 16, Sucursal 4, 1900 La Plata, Argentina
[b] Vicerrectorado de Investigaciones, Universidad del Azuay, Av. 24 de Mayo 7-77 y Hernán Malo, Cuenca, Ecuador
[c] CEQUINOR, Centro de Química Inorgánica (CONICET, UNLP), Departamento de Química, Facultad de Ciencias Exactas, UNLP, C.C. 962, 1900 La Plata, Argentina

ARTICLE INFO

Article history:
Received 29 May 2017
Received in revised form
3 November 2017
Accepted 5 November 2017
Available online 11 November 2017

Keywords:
Rice
Volatile organic compounds
Contaminants
Dragon descriptors

ABSTRACT

The purpose of work presented here was to calibrate and validate a mathematical model based on a quantitative structure-property relationship for modeling the retention indices (*I*) of 137 volatile organic compounds (VOCs) measured in the headspace of rice using a Divinylbenzene-Carboxen-Polydimethylsiloxane (DVB-CAR-PDMS) fiber in the solid-phase microextraction-gas chromatography-mass spectrometry (SPME-GC-MS) analysis. The dataset was split into training, validation and test sets according to the Balanced Subsets Method (BSM). The study was divided into three different steps. In the first step, 1753 conformation-independent descriptors were considered for modeling. In the second step, 1145 conformation-dependent descriptors were taken into account to obtain a model. Finally, in the last step both conformation-independent and conformation-dependent descriptors were used to build the model. A three-descriptor model was retained as the optimal one in all cases. Conformation-dependent descriptors led to models with no appreciable improvement over those obtained with conformation-independent descriptors. The final conformation-independent QSPR model was used as a tool for the quality control of volatile contaminants of rice by predicting the retention indices in a set of 46 rice contaminants.

1. Introduction

Rice (*Oriza sativa* L.) is one of the cereal grain most produced around the world, and it is a staple food in several countries (Fukuda et al., 2014). For this reason, it is necessary to improve the quality control of crops in order to ensure the optimal organoleptic characteristics of rice to make it acceptable by consumers (Grimm et al., 2002). In fact, in cooked rice, it has been observed that small variations in the sensory properties, especially aroma, produces changes in the consumers' acceptance (Fukuda et al., 2014).

The aromatic profile of rice is produced by the presence of different fragrance compounds even in low concentrations. Sensory studies demonstrated that consumers were able to discriminate between different kinds of rice as a function of their aroma (Bryant and McClung, 2011). Desirable fragrances are produced by strong-smelling organic chemicals, which elicits a common characteristic of a pleasant odor. A fragrance substance is usually used as a food additive to enhance the aromatic profile of processed foods. Among the rice fragrances, 2-acetyl-1-pyrroline (2-AP) is the main compound generated during the growth of the plant. However, during post-harvest and storage, the concentration of 2-AP decreases and the aromatic profile of rice changes (Bryant and McClung, 2011).

Since rice fragrances are volatile organic compounds (VOCs), that determine its aromatic profile and therefore, its sensory quality, researchers have focused on the analytical identification of such compounds. To this end, solid-phase microextraction (SPME) in conjunction with gas chromatography-mass spectrometry (GC-MS) has been proven to be an efficient method for analyzing the aromatic profiles of different varieties of rice (Bryant and McClung, 2011; Grimm et al., 2002). In fact, the SPME-GC-MS technique is generally used for quantitative determination of the aromatic profile and impurities of fragrances, as well as for quality control in order to provide details of aromatic profiles in few minutes (Bryant

* Corresponding author. Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), CONICET, UNLP, Diag. 113 y 64, C.C. 16, Sucursal 4, 1900 La Plata, Argentina.
** Corresponding author.
E-mail addresses: crojasvilla@gmail.com (C. Rojas), pabloducho@gmail.com (P.R. Duchowicz).

https://doi.org/10.1016/j.jcs.2017.11.004
0733-5210/

and McClung, 2011). Fragrances exhibiting high molecular weight are better retained on fibers containing polydimethylsiloxane (PDMS) as a stationary phase, while a stationary phase containing Carboxen (CAR) or divinylbenzene (DVB) is better for the retention of smaller VOCs. Thus, the fiber containing the DVB-CAR-PDMS combination has been proved to be adequate for analyzing complex components or fragrances with different polarities, as well as to deal with high temperatures in the SPME-GC-MS analysis (Bryant and McClung, 2011; Grimm et al., 2011).

Since the first pioneering studies of the applications of the quantitative structure-property relationships (QSPR) to chromatographic retention indices (I), there has been an increasing interest of researchers to use the approach known as Quantitative Structure-(Chromatographic) Retention Relationships (QSRR) (Kaliszan, 2007). The QSRR models have been proven to be useful to: (1) predict the retention index of un-evaluated and un-synthesized compounds and to select drug candidates; (2) prepare chromatography experiments and to optimize the separation of complex mixtures; (3) understand the molecular mechanism of retention phenomena; and, (4) design in a rational way new phases with predefined properties (Kaliszan, 2007).

Recently, Fatemi and Malekzadeh (2014) had developed a QSRR model to predict the I values of 96 volatile compounds identified in three glutinous rice varieties during four different cooking stages. Retention indices had been measured by combined GC-MS with a modified headspace solid-phase microextraction method using the DB Wax capillary column. The SMILES string notation and a graphical molecular representation had been used to calculate molecular descriptors and to perform a linear model with the CORAL software. The data set was split into training set ($n = 70$, $R^2 = 0.972$ and $RMSD = 79.5$), calibration set ($n = 13$, $R^2 = 0.971$ and $RMSD = 125.6$) and test set ($n = 13$, $R^2 = 0.952$ and $RMSD = 191.6$). Moreover, the model was also validated by means of the leave-one-out cross-validation procedure ($R^2_{loo} = 0.932$), the Y-randomization procedure ($0.0003 \leq R^2 \leq 0.271$), and some other criteria.

Therefore, the aim of this work was to build a quantitative structure-property relationship by using the retention indices of 137 VOCs observed in the headspace of rice, keeping in mind the five principles defined by the Organization for Economic Co-operation and Development (OECD) to make it applicable (Organisation for Economic Co-operation and Development, 2007). To the best of our knowledge, there is no available a retention index-based QSPR model for the quality control of rice as well as its use as a tool for predicting the I of contaminants of raw rice. Molecular geometries were optimized by means of the PM7 semiempirical method. Subsequently, the replacement method (RM) variable subset selection was used to search for an optimal QSPR model. Moreover, both internal and external validation procedures were carried out in order to guarantee the predictive capability of the model, and its applicability domains (AD) were properly defined. Finally, an explanation of the chemical information of each molecular descriptors in modeling the I is presented.

2. Materials and methods

2.1. Dataset description and data filtering

Experimental retention indices for 138 main volatile organic compounds were retrieved from the literature (Grimm et al., 2002). Among these compounds, some of them do not really belong to the aromatic profile of rice and are contaminants adsorbed in containers during the rice storage. Experimental retention indices were measured by solid-phase microextraction-gas chromatography-mass spectrometry (SPME-GC-MS) using the Divinylbenzene-Carboxen-Polydimethylsiloxane (DVB-CAR-PDMS) fiber for dealing with high temperatures.

The chemical name, CAS number and the retention index of VOCs were merged using KNIME (Berthold et al., 2008). Subsequently, the SMILES (simplified molecular input line entry system) strings were obtained from both the CAS number and chemical name using the Chemical Identifier Resolver node. SMILES structures were verified for the correct match between CAS and structure. Compounds exhibiting different SMILES were manually checked on public databases: PubChem, ChemSpider and NIST Chemistry WebBook.

During the filtering of the dataset, the compound 2,2,4-trimethylheptane (Cas Number 14720-74-2) was identified as a duplicate with trimethylheptane. Therefore, the trimethylheptane compound was excluded and the average I of 878.5 was used as the retention index for 2,2,4-trimethylheptane. Consequently, 137 compounds were used to build the QSPR model. Details for the filtering dataset are given in Table A.1.

2.2. Molecular representation and geometry optimization

Compounds were initially optimized by means of the molecular mechanics force field (MM+) which was subsequently refined using the PM7 semiempirical method as implemented in the MOPAC package (Stewart, 2016). Geometries were considered optimized when the maximum element of the gradient vector of the total energy with respect to the atomic coordinates became less than 1 kcal (Å mol)$^{-1}$.

2.3. Molecular descriptors

Molecular descriptors are used as the structural representation of optimized molecules in order to develop the QSPR model. Descriptors are the final result of a logical and mathematical procedure that transforms chemical information encoded within a symbolic representation of a molecule into a numerical quantity or into the result of some standardized experiment (Todeschini and Consonni, 2009). Thus, 5239 molecular descriptors were calculated by means of Dragon software (2016). Such descriptors were grouped into twenty nine blocks: constitutional indices, functional group counts, atom-centered fragments, molecular properties, ring descriptors, topological indices, walk and path counts, connectivity indices, information indices, 2D matrix-based descriptors, 2D autocorrelations, Burden eigenvalues, P VSA-like descriptors, edge adjacency indices, CATS2D, 2D atom pairs, atom-type E-state indices, ETA indices, Randić molecular profiles, geometrical descriptors, RDF descriptors, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors, charge descriptors, 3D matrix-based descriptors, 3D autocorrelations, 3D atom pairs, and CATS3D.

2.4. Model development

2.4.1. Molecular descriptor selection in MLR

A crucial step in QSPR modeling is the selection of an optimal set of molecular descriptors to construct the mathematical multivariate model. The supervised methods of molecular descriptor selection aim to search the most informative descriptors among thousands of them in order to facilitate the interpretation and prediction of QSPR models. In this work, the replacement method (RM) (Duchowicz et al., 2006) variable subset selection was used. In brief, RM is a sequential method that generates d subsets of descriptors from a D pool containing thousands of descriptors. This selection of descriptors was performed in order to minimize

(optimize) the root-mean-square deviation (*RMSD*) in multiple linear regression (MLR) models.

2.4.2. Model validation

The QSPR model was validated in order to determine its predictive ability by using a validation set as well as by predicting the retention index of compounds in a test set. The split of a dataset should be done in order to achieve similar structure-property relationships in the three sets, that is, molecules in the training set should be representative of both validation and test set compounds. In this work, the split of the dataset was carried out by means of the Balanced Subsets Method (BSM) (Rojas et al., 2015a) based on the *k*-means cluster analysis (*k*-MCA). This procedure has been applied elsewhere under similar situations (Rojas et al., 2015a, b). In brief, *k*-MCA creates *k*-clusters or groups of compounds in terms of distance metrics (e.g. Euclidean distance), in such a way that compounds in the same cluster are very similar, and compounds in different clusters are very different. The BSM partition considerers the experimental property and conformation-independent molecular descriptors only (after the exclusion of the linearly correlated descriptors). This was done in order to consider conformation-independent structure-property relationships during the clustering procedure and to avoid geometry optimization biases. The steps involved in the BSM partition were:

a) prepare a matrix (C_1) that included the experimental retention indices for the 137 VOCs and the 1753 conformation−independent molecular descriptors.
b) remove the linearly dependent descriptors from C_1. The new size of the reduced matrix C_2 was 137 × 136.
c) standardize matrix C_2 for centering and scaling its matrix elements.
d) create N^0_{train} clusters with the 137 compounds through the *k*−MCA method, for which the C_2 standardized matrix was used together with the Euclidean metrics, and 5000 runs for optimizing (i.e., minimizing the Euclidean distance) the *k*−MCA algorithm. This step calculated N^0_{train} cluster centroid locations with dimensions of 1 × 136. $N^0_{train} = N_{train} - N_{\min\,\max}$, where N_{train} was the number of molecules in the training set and $N_{\min\,\max}$ was the number of compounds having the maximum or minimum retention index.
e) the training set (N_{train}) was designed by including one compound per cluster (i.e., the nearer molecule to the centroid in each cluster). The $N_{\min\,\max}$ molecules were also included in the training set in order to avoid model extrapolations.
f) Create N_{val} clusters with the remaining $N - N_{train}$ molecules through the *k*−MCA method with same numerical conditions as described above. This step calculated N_{val} cluster centroid locations.
g) the validation set (N_{val}) was designed by including one molecule per cluster (i.e., the nearer compound to the centroid in each cluster).
h) Finally, the test set (N_{test}) included the remaining $N - N_{train} - N_{val}$ VOCs.

During the RM variable selection procedure, molecules in the training set were used to calibrate the model, whereas the validation set was used for the cross-validation of the model in order to avoid the presence of overfitting. Finally, the predictive ability of the selected QSPR model was checked by predicting the retention index of compounds in the test set.

The QSPR model was also validated through the cross-validation technique of leave-one-out (loo) and leave-many-out (lmo). In the leave-one-out cross-validation technique, each molecule was excluded from the model at a time, and then the model was constructed and used to predict its property. On the other hand, in the leave-many-out approach a user-defined percentage of the molecules (20%) are randomly excluded, and the remaining molecules (80%) are used to calibrate the model and then used to predict the property of the removed molecules. The leave-many-out procedure is based on 50000 iterations.

The Y-randomization approach (Rücker et al., 2007) was applied in order to evaluate the risk of chance correlation in the model. This approach depends on randomly scrambling the experimental property values in such a way that they do not correspond to the respective compounds. After analyzing a certain number of cases (e.g. 10000) of Y-randomization, the quality of the model (R^2_{rand} or $RMSD_{rand}$) must be of lower quality than model parameters (R^2_{train} or $RMSD_{train}$).

2.4.3. Applicability domain assessment

The merit of a QSPR model is related to the reliability of its predictions. The applicability domain (AD) is defined as a theoretical space that depends on the nature of the molecular descriptors and the experimental properties of the molecules (Gramatica, 2007). In other words, the model is confined to a chemical space, which is defined by the chemical information provided by the molecules of the training set. The applicability of such a model to molecules in the test set is then restricted to those compounds that are structurally similar to compounds present in the training set. The best way to characterize the AD of a MLR model is the leverage approach (Eriksson et al., 2003), which is based on the calculation of a leverage value (h_i) for each *i*th compound, and then to compare its value to a theoretical warning leverage (h^*) value. Consequently, only molecules falling within this theoretically space are considered reliable predictions or model interpolations ($h_i \leq h^*$); otherwise, retention indices of the molecules are considered model extrapolations or unreliable predictions ($h_i > h^*$).

2.4.4. Descriptor interpretation

Another important issue to be addressed in QSPR studies is how descriptors included in a MLR model are related in their properties. Since MLR models provide numerical coefficients for each *j*th descriptor, the degree of contribution of the selected descriptors was found by standardizing their regression parameters (b^s_j). Consequently, the larger the absolute value of b^s_j for a given descriptor, the greater the importance of such a descriptor in modeling the experimental property (Draper and Smith, 1981).

2.5. Software

Open Babel (O'Boyle et al., 2011) was used to handle molecular file formats, while a KNIME workflow (Berthold et al., 2008) was used for data filtering. The MOPAC package (Stewart, 2016) was used for the optimization of structures, and molecular descriptors were computed using Dragon version 7 (2016). Partition of the data set by means of the BSM, variable selection was accomplished through the RM approach, and model fitting along with validation were carried out in MatLab (The MathWorks Inc.), by using toolboxes and functions written by the authors.

3. Results and discussion

In order to evaluate the contribution of conformational descriptors in modeling the *I* property, three datasets were constructed. The first dataset contained conformation-independent descriptors, the second dataset included only conformation-

dependent descriptors, and the third dataset included both conformation-independent and conformation-dependent descriptors. Initially, non-informative molecular descriptors were excluded; that is, descriptors with constant values (descriptors with all values equal), descriptors with near-constant values (descriptors with only one value different from the remaining ones), and descriptors affected by missing values.

The BSM was used to split the original dataset of 137 compounds into a training set with 46 molecules, a validation set with 46 compounds, and a test set formed by 45 molecules. This partition guarantees a balanced structure-property design in each of the three groups. Subsequently, the supervised RM variable subset selection explored a pool containing (a) 1753 conformation-independent molecular descriptors, (b) 1145 conformation-dependent molecular descriptors, and (c) 2898 descriptors that combined both conformation-dependent and conformation-independent descriptors.

In the three datasets, the RM procedure explored descriptor pools for models containing from 1 to 6 molecular descriptors, and the selection of the optimal model was done by optimizing *RMSD* in the validation set, as well as by keeping the model's size as small as possible according to the principle of parsimony (Ockham's razor) (Hoffmann et al., 1996). Thus, a three-parametric quantitative structure-property relationship was retained as the optimal one in each dataset.

It was found that the parameters for training and validation sets exhibited a negligible variation among the models of the same size (d). In fact, when the conformation-dependent and the conformation-independent descriptors were analyzed together (see Table A.2), the optimal model became identical to the best conformation-independent model (see Table 1). In addition, the best conformation-dependent QSPR model (see Table A.3) did not reflect any further improvement with respect to the conformation-independent model. In fact, it was demonstrated elsewhere that the retention indices measured in non-polar and polar stationary phases was well predicted by models that include conformation-independent descriptors only (Rojas et al., 2015a, b). This fact can be considered an important finding since conformation-independent QSPR models avoid ambiguities generated by incorrect geometry selection of the molecules that exist in various conformations. Thus, the following three-parametric conformation-independent QSPR model was selected as the best one for modeling the I of volatile organic compounds presented in the headspace of rice:

$$I = 729.9 - 826.1\,X0A_v + 29.6\,XMOD + 492.2\,MATS1p \qquad (1)$$

$$N_{train} = 46,\ \ d = 3,\ \ R^2_{train} = 0.98,\ \ RMSD_{train} = 67.3,\ \ R^2_{ij\ max} = 0.32$$

$$N_{val} = 46,\ \ R^2_{val} = 0.97,\ \ RMSD_{val} = 79.9$$

$$N_{test} = 45,\ \ R^2_{test} = 0.97,\ \ RMSD_{test} = 80$$

$$o(3\,S) = 1,\ \ R^2_{loo} = 0.97,\ \ RMSD_{loo} = 74,\ \ R^2_{lmo} = 0.98,\ \ RMSD_{lmo} = 86,\ \ RMSD_{rand} = 353.8$$

The goodness-of-fit of the present model for the training, validation and test sets was 98%, 96% and 97%, respectively. Moreover, the model exhibited good stability in cross-validation leave-one-out (97%) and leave-many-out, after 50000 iterations for random data removal (97%). The similar performance in calibration, cross-validation and prediction indicated the absence of overfitting in the QSPR model. In addition, the Y-randomization procedure demonstrated the absence of change correlation in the model ($RMSD_{rand} < RMSD_{train}$). Other recommended validation criteria (Golbraikh and Tropsha, 2002) were also evaluated in order to thoroughly validate the model and to avoid the proposal of an overoptimistic and perhaps erroneous, "predictive" QSPR model:

$$R^2_{loo} > 0.5\ \ (0.97)\ \text{ and }\ R^2_{test} > 0.6\ \ (0.97)$$

$$1 - R^2_0 \big/ R^2_{test} < 0.11\ \ (0.000)\ \text{ or }\ 1 - R'^2_0 \big/ R^2_{test} < 0.1\ \ (0.000)$$

$$0.85 \le k(1.03) \le 1.15\ \text{ and }\ 0.85 \le k'(0.97) \le 1.15$$

$$R^2_m > 0.5\ \ (0.95)$$

All these parameters are defined in Table A.4, and indicate that a stable and predictive MLR was achieved for the I of VOCs presented in the headspace of rice. Numerical data of the predicted retention indices provided by Eq. (1) are shown in Table A.1, while descriptor values for the training, validation and test sets are provided in Table A.5. Fig. 1 shows the predicted retention index calculated with Eq. (1) as a function of the experimental I. Moreover, Fig. 2 presents the dispersion plot between the residuals of the calculated and experimental retention indices. Both figures indicate that a quantitative structure-property relationship with good predictive power was achieved.

Compound *2-Hexyl-1-octanol* exhibited a residual greater than the limiting value of three standard deviations. After a careful control of both the chemical structure and the I value from the source, we are confident that this information is correct. Consequently, the irregular behavior of this outlier may be attributed to the wide chemical diversity of the VOCs considered in the present dataset, as well as to specific analytical aspects during the retention index measurement. For example (Rojas et al., 2015b), the nature of the sample (e.g. chemical properties, preparation and mechanism to introduce it in the equipment); the interaction between the analyzed volatile organic compound and the wall interfaces of the fiber; the equipment characteristics (e.g. sensitivity, stationary phase properties and conditions used to measure the I value); and data processing (e.g. variations associated with peak integration

Table 1
The best conformation-independent QSPR models selected by the RM variable subset selection. The optimal model appears in bold.

d	R^2_{train}	$RMSD_{train}$	R^2_{val}	$RMSD_{val}$	$R^2_{ij\ max}$	molecular descriptors
1	0.95	97.9	0.94	99.8	0.00	SpPos_B(m)
2	0.97	83.3	0.95	90.2	0.04	XMOD, AVS_B(e)
3	**0.98**	**67.3**	**0.96**	**79.9**	**0.32**	**X0Av, XMOD, MATS1p**
4	0.98	60.7	0.96	78.2	0.18	X1sol, SM3_B(s), VE2sign_B(s), MATS1v
5	0.99	55.4	0.96	81.5	0.80	Psi_i_t, X1sol, Xindex, VE2_Dz(Z), MATS1p
6	0.99	49.9	0.96	78.7	0.54	nR06, X1sol, Xindex, BIC0, ATSC1s, nFuranes

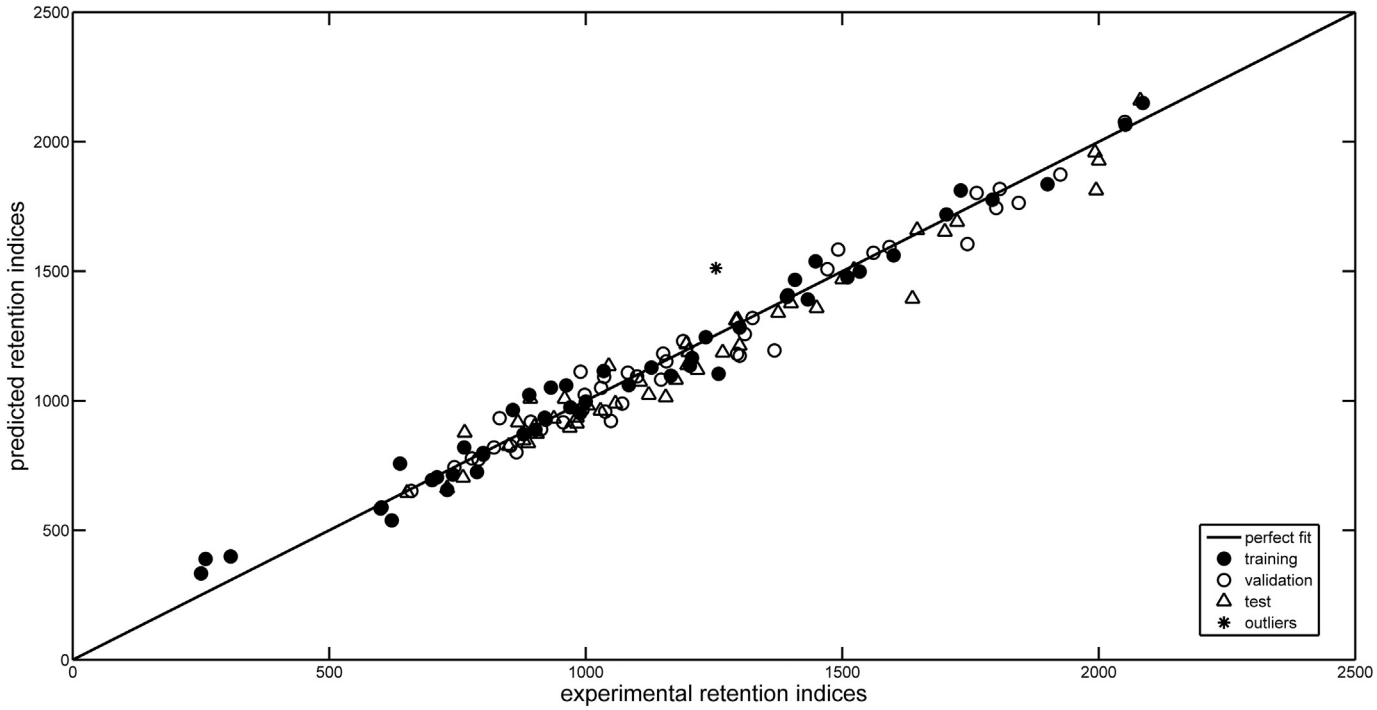

Fig. 1. Experimental *versus* predicted retention indices for VOCs in the headspace of rice. Training molecules are marked with black circles, molecules of the validation set are marked with white circles, and triangles indicate molecules of the test set.

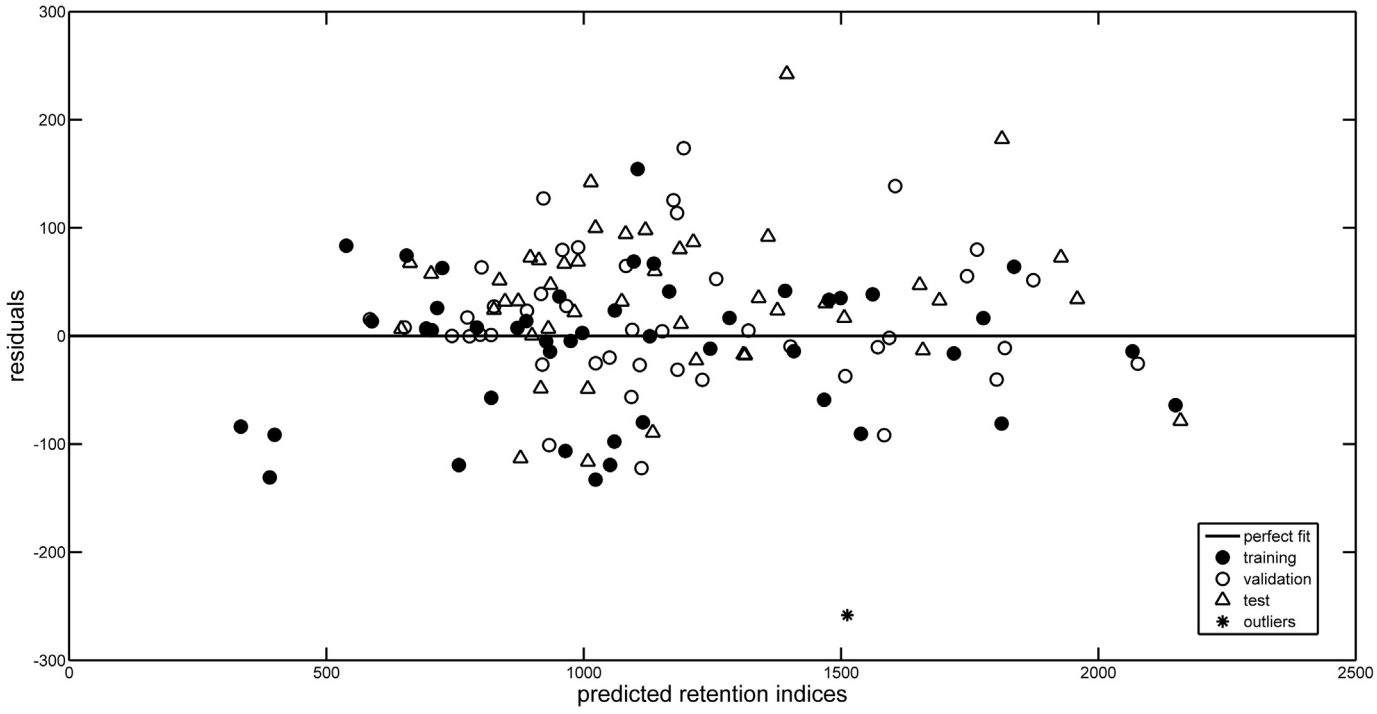

Fig. 2. Dispersion plot of residuals for the QSPR model. Training molecules are marked with black circles, molecules of the validation set are marked with white circles, and triangles indicate molecules of the test set.

and reproducibility of the data).

Eq. (1) shows that the model is described by two connectivity index descriptors (*XMOD* and *X0Av*) and a 2D autocorrelation descriptor (*MATS1p*). The maximum coefficient of determination ($R^2_{ij\ max} = 0.32$) between *X0Av* and *MATS1p* shows a low correlation, indicating that those descriptors are not collinear and each one describes different aspects of the retention index mechanism. Moreover, the contribution of each descriptor in predicting the *I* of VOCs in the DVB/CAR/PDMS fiber was evaluated by standardizing the regression coefficients of the three descriptors: 0.95 (*XMOD*) > 0.19 (*X0Av*) > 0.13 (*MATS1p*).

The modified Randić index (*XMOD*) is a descriptor calculated by

means of a Randić-like formula on an H-depleted graph, which considers valence electrons and connectivity (Lohninger, 1993). The Randić connectivity index measures the degree of branching and compactness of molecules and has been shown to be well-correlated with chromatographic retention times (Rojas et al., 2015b). Thus, compounds containing a high degree of branching (i.e., compacted molecules) are related to larger values of *XMOD* (synergistic effect). This relationship was previously described by Yan et al. (2013). On the other hand, the average valence connectivity index of order 0 (*X0Av*) describes the presence of heteroatoms in compounds as well as double and triple bonds. This descriptor has an antagonistic influence on the prediction of the retention index, and consequently, the *I* decreases when increasing the presence of heteroatoms or double and triple bounds in the molecule. This inverse relationship between *X0Av* and *I* was also described by Riahi et al. (2008). Finally, the Moran autocorrelation of lag 1 weighted by polarizability (*MATS1p*) is a descriptor calculated by applying the Moran coefficient (Moran, 1950) to the H-filled molecular graph weighted by atomic polarizabilities (*p*). This descriptor provides information regarding the distribution of polarizability along the topological structure of the volatile organic compounds. In other words, high retention indices are related to positive values of the Moran coefficient (positive spatial autocorrelations), that is, VOCs containing atoms with similarly polarizability at lag 1. Fig. 3 shows the distribution of the volatile organic compounds within the chemical space defined by three molecular descriptors and the relationship with the retention index property.

The applicability domain assessment provides information regarding the limitation of the proposed QSPR model; that is, the limitations of the three molecular descriptors and the retention index space. Therefore, predictions of the retention index are restricted only for chemicals exhibiting a leverage value below the warning leverage of the model ($h^* = 0.130$). There are no test molecules with leverage values above the warning leverage of the model, and therefore their predicted retention index can be considered reliable.

During the analysis of VOCs in the headspace of rice, some contaminants are detected, such as plasticizer (phthalic acid esters) and antioxidants (e.g. BHA), which may migrate from the packing materials used during rice transportation (Grimm et al., 2002). Other common contaminants detected during rice analysis are the polycyclic aromatic hydrocarbons (PAHs) (Escarrone et al., 2014; Liu and Korenaga, 2001; Tao et al., 2006), which are chemicals originated from an incomplete combustion of fossil fuels. Since PAHs are widely distributed in the air, soil and water (environmental pollution), it is inevitable the human exposure to PAHs; particularly in diet (e.g. cereals and vegetables). In addition, we consider pesticides used in rice cultivars, particularly pyrethroid and carbamate derivatives (Berg, 2001). Thus, we use the QSPR developed here in order to predict the retention index of 46 common volatile contaminants of rice (see Table 2).

There are 14 compounds exhibiting leverage values higher than the warning leverage, and consequently lying outside the applicability domain (i.e., they are considered as extrapolations of the QSPR model). Such compounds are: Tetrachloroethylene, DEHP plasticizer, three PAHs (*Indeno[1,2,3-cd]pyrene*, *Dibenz[a,h]anthracene*, *Benzo[g,h,i]perylene*), two fungicides (*Validamycin A*, *Propiconazole*), two herbicides (*Fenoxaprop-P-Ethyl*, *Pyrazosulfuron Ethyl*), and five insecticides (*lambda-cyhalothrin*, *Deltamethrin*, *Alphamethrin*, *Fipronil*, *Etofenprox*). On the other hand, 32 contaminants belong to the AD of the model, i.e., their predicted retention indices are reliable, and could be used in order to identify such contaminants in rice samples by means of the GC technique using the DVB-CAR-PDMS fiber. For instance, if we have $I = 2137.4$, we could suspect the presence of either *Perylene* or *Benzo[b]fluoranthene* PAHs. Finally, the developed QSPR model could support chromatographers as a fast identification tool for other contaminants for which their experimental retention indices are not available and their molecular structures are known.

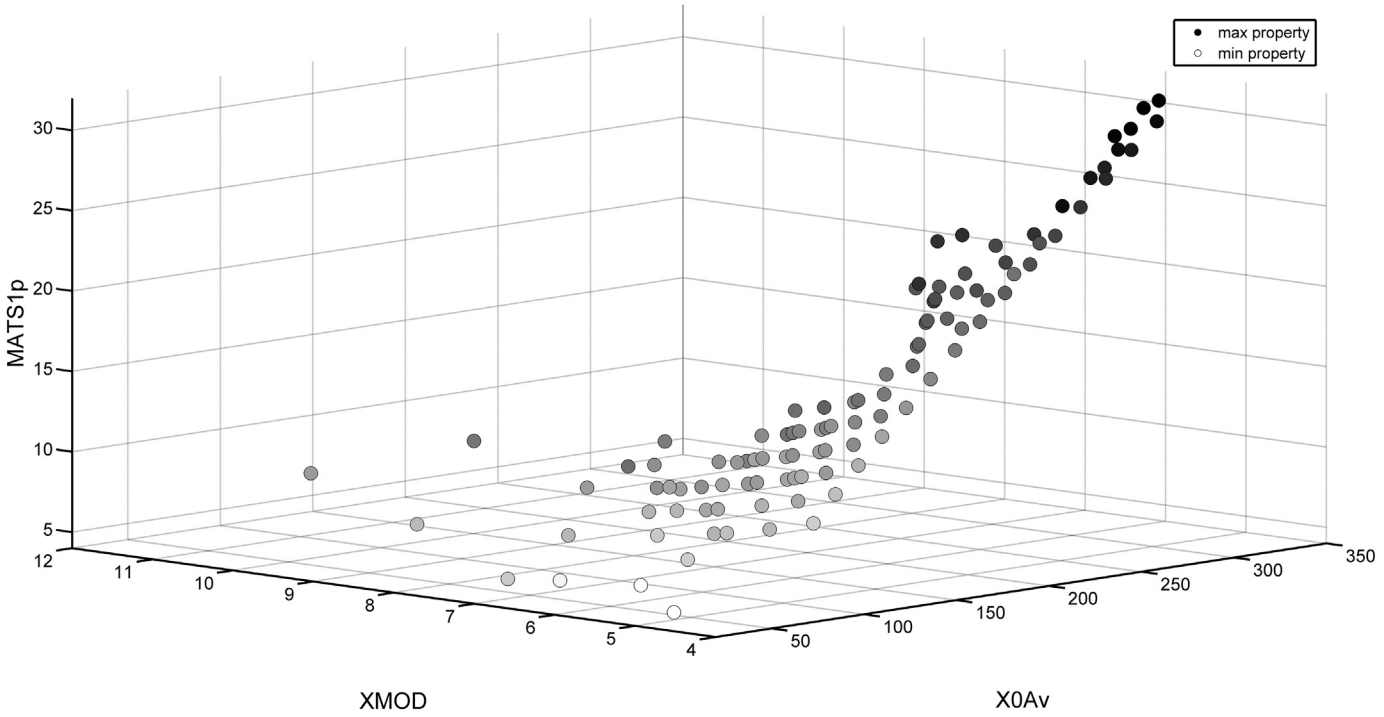

Fig. 3. Distributions of VOCs detected in the headspace of rice in the chemical space resulting from the molecular descriptors involved in the QSPR model (Eq. (1)) and the relationship with the retention index property.

Table 2
Common contaminants detected in rice: name, CAS number, source, predicted retention index and leverage values.

Name	CAS number	Source	Predicted I	h_i	Reference
Tetrachloroethylene	127-18-4	Unknown	678.1	0.297[a]	(Grimm et al., 2002)
oxime, methoxy-phenyl-	not available		1360.4	0.055	
1,4-dichlorobenzene	106-46-7		1077.1	0.016	
Diethyl Phthalate	84-66-02	Plasticizer	1690.4	0.052	
Dibutyl Phthalate	84-74-2		2044.1	0.080	
DEHP	117-81-7		2715.4	0.195[a]	
BHA	25013-16-5	Antioxidant	2443.0	0.129	
Acenaphthylene	208-96-8	PAHs	1422.7	0.077	(Escarrone et al., 2014; Liu and Korenaga, 2001; Tao et al., 2006)
Fluorene	86-73-7		1502.0	0.071	
Phenanthrene	85-01-8		1594.7	0.075	
Anthracene	120-12-7		1591.7	0.075	
Fluoranthene	206-44-0		1782.8	0.089	
Pyrene	129-00-0		1779.8	0.089	
Benz[a]anthracene	56-55-3		1949.0	0.096	
Chrysene	218-01-9		1952.0	0.096	
Perylene	198-55-0		2137.4	0.117	(Liu and Korenaga, 2001; Tao et al., 2006)
Benzo[b]fluoranthene	205-99-2		2137.4	0.117	(Escarrone et al., 2014; Tao et al., 2006)
Benzo[k]fluoranthene	207-08-9		2134.4	0.117	
Benzo[a]pyrene	50-32-8		2134.4	0.117	
Indeno[1,2,3-cd]pyrene	193-39-5		2316.6	0.140[a]	
Dibenz[a,h]anthracene	53-70-3		2304.9	0.134[a]	
Benzo[g,h,i]perylene	191-24-2		2316.6	0.140[a]	
Naphthalene	91-20-3		1230.6	0.067	(Escarrone et al., 2014)
Validamycin A	37248-47-8	Fungicides	3549.9	0.439[a]	(Berg, 2001)
Propiconazole	60207-90-1		2386.0	0.137[a]	
Hexaconazole	79983-71-4		2199.0	0.088	
Isoprothiolane	50512-35-1		2063.1	0.070	
Iprodione	36734-19-7		2252.1	0.109	
Cyproconazole	94361-06-5		2150.9	0.086	
2,4-D	94-75-7	Herbicides	1630.2	0.044	
Pretilachlor	51218-49-6		2171.3	0.100	
Fenclorim	3740-92-9		1667.4	0.036	
Fenoxaprop-P-Ethyl	71283-80-2		2602.0	0.190[a]	
MCPA	94-74-6		1528.8	0.037	
Pyrazosulfuron Ethyl	93697-74-6		2989.2	0.385[a]	
Butachlor	23184-66-9		2171.3	0.100	
Propanil	709-98-8		1555.4	0.027	
Fenobucarb	3766-81-2	Insecticides	1562.4	0.029	
Cartap hydrochloride	15263-52-2		1753.8	0.058	
lambda-cyhalothrin	91465-08-6		3129.0	0.291[a]	
Deltamethrin	52918-63-5		3168.3	0.303[a]	
Buprofezin	69327-76-0		2135.8	0.084	
Isoprocarb	2631-40-5		1467.6	0.025	
Alphamethrin	67375-30-8		2905.8	0.229[a]	
Fipronil	120068-37-3		3014.9	0.262[a]	
Etofenprox	80844-07-1		2753.7	0.198[a]	

[a] Molecules with leverage value above the warning leverage ($h^* = 0.130$).

4. Conclusions

The retention indices of volatile organic compounds detected by the DVB-CAR-PDMS fiber in the headspace of rice are described and predicted by a conformation-independent QSPR model obtained in this work. The use of the Replacement Method allows for the selection of an optimal subset of three topological and constitutional Dragon descriptors. This model is based on a reduced number of molecular descriptors and could be useful for chromatographers working on the aromatic profile of rice, as well as its quality control based on the GC technique. Moreover, when modeling gas-chromatographic retention indices, the use of the conformation-independent QSPR approach represents an efficient alternative to develop models based on topological and constitutional molecular aspects of chemicals.

Acknowledgments

Cristian Rojas is grateful for his PhD Fellowship from the National Secretary of Higher Education, Science, Technology and Innovation from the Republic of Ecuador. Pablo R. Duchowicz wishes to thank the National Scientific and Technical Research Council of Argentina (CONICET) for the project grant PIP11220130100311, and the Minister of Science, Technology and Productive Innovation for the use of the electronic library facilities. Pablo R. Duchowicz and Reinaldo Pis Diez are members of the Scientific Researcher Career of CONICET.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jcs.2017.11.004

References

Berg, H., 2001. Pesticide use in rice and rice–fish farms in the Mekong Delta, Vietnam. Crop Prot. 20, 897–905.
Berthold, M., Cebron, N., Dill, F., Gabriel, T., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B., 2008. KNIME: the konstanz information miner. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (Eds.), Data Analysis, Machine Learning and Applications. Springer Berlin, Heidelberg, pp. 319–326.
Bryant, R., McClung, A., 2011. Volatile profiles of aromatic and non-aromatic rice

cultivars using SPME/GC-MS. Food Chem. 124, 501–513.
Draper, N.R., Smith, H., 1981. Applied Regression Analysis. New York.
Duchowicz, P.R., Castro, E.A., Fernández, F.M., 2006. Alternative algorithm for the search of an optimal set of descriptors in QSAR-QSPR studies. MATCH Commun. Math. Comput. Chem. 55, 179–192.
Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T., McDowell, R.M., Gramatica, P., 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ. Health Perspect. 111, 1361–1375.
Escarrone, A., Caldas, S., Furlong, E., Meneghetti, V., Fagundes, C., Arias, J., Primel, E., 2014. Polycyclic aromatic hydrocarbons in rice grain dried by different processes: evaluation of a quick, easy, cheap, effective, rugged and safe extraction method. Food Chem. 146, 597–602.
Fatemi, M.H., Malekzadeh, H., 2014. CORAL: predictions of retention indices of volatiles in cooking rice using representation of the molecular structure obtained by combination of SMILES and graph approaches. J. Iran. Chem. Soc. 12, 405–412.
Fukuda, T., Takeda, T., Yoshida, S., 2014. Comparison of volatiles in cooked rice with various amylose contents. Food Sci. Technol. Res. 20, 1251–1259.
Golbraikh, A., Tropsha, A., 2002. Beware of q2! J. Mol. Graph. Modell. 20, 269–276.
Gramatica, P., 2007. Principles of QSAR models validation: internal and external. QSAR Comb. Sci. 26, 694–701.
Grimm, C.C., Champagne, E.T., Lloyd, S.W., Easson, M., Condon, B., McClung, A., 2011. Analysis of 2-acetyl-1-pyrroline in rice by HSSE/GC/MS. Cereal Chem. 88, 271–277.
Grimm, C.C., Champagne, E.T., Ohtsubo, K.i., 2002. Analysis of volatile compounds in the headspace of rice using SPME/GC/MS. In: Marsili, R. (Ed.), Flavor, Fragrance, and Odor Analysis. Marcel Dekker, Inc., pp. 229–248
Hoffmann, R., Minkin, V.I., Carpenter, B.K., 1996. Ockham's razor and chemistry. Bull. Société Chim. Fr. 133, 117–130.
Kaliszan, R., 2007. QSRR: quantitative structure-(chromatographic) retention relationships. Chem. Rev. 107, 3212–3246.
Kode srl, 2016. Dragon (Version 7). Software for Molecular Descriptor Calculation. https://chm.kode-solutions.net.
Liu, X., Korenaga, T., 2001. Dynamics analysis for the distribution of polycyclic aromatic hydrocarbons in rice. J. health Sci. 47, 446–451.
Lohninger, H., 1993. Evaluation of neural networks based on radial basis functions and their application to the prediction of boiling points from structural parameters. J. Chem. Inf. Comput. Sci. 33, 736–744.
Moran, P.A., 1950. Notes on continuous stochastic phenomena. Biometrika 37, 17–23.
O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., Hutchison, G.R., 2011. Open Babel: an open chemical toolbox. J. cheminformatics 3, 1–14.
Organisation for Economic Co-operation and Development, 2007. Guidance Document on the Validation of (Quantitative)Structure-activity Relationships [(Q)SAR] Models. OECD Publishing, Paris.
Riahi, S., Ganjali, M.R., Pourbasheer, E., Norouzi, P., 2008. QSRR study of GC retention indices of essential-oil compounds by multiple linear regression with a genetic algorithm. Chromatographia 67, 917–922.
Rojas, C., Duchowicz, P.R., Tripaldi, P., Pis Diez, R., 2015a. QSPR analysis for the retention index of flavors and fragrances on a OV-101 column. Chemom. Intelligent Laboratory Syst. 140, 126–132.
Rojas, C., Duchowicz, P.R., Tripaldi, P., Pis Diez, R., 2015b. Quantitative structure–property relationship analysis for the retention index of fragrance-like compounds on a polar stationary phase. J. Chromatogr. A 1422, 277–288.
Rücker, C., Rücker, G., Meringer, M., 2007. Y-Randomization and its variants in QSPR/QSAR. J. Chem. Inf. Model 47, 2345–2357.
Stewart, J.J.P., 2016. Stewart Computational Chemistry, USA. MOPAC2016. http://OpenMOPAC.net.
Tao, S., Jiao, X., Chen, S., Liu, W., Coveney, R., Zhu, L., Luo, Y., 2006. Accumulation and distribution of polycyclic aromatic hydrocarbons in rice (Oryza sativa). Environ. Pollut. 140, 406–415.
The MathWorks Inc., MatLab, http://www.mathworks.com.
Todeschini, R., Consonni, V., 2009. Molecular Descriptors for Chemoinformatics. WILEY-VCH, Weinheim.
Yan, J., Huang, J.-H., He, M., Lu, H.-B., Yang, R., Kong, B., Xu, Q.-S., Liang, Y.-Z., 2013. Prediction of retention indices for frequently reported compounds of plant essential oils using multiple linear regression, partial least squares, and support vector machine. J. Sep. Sci. 36, 2464–2471.