Journal of Chromatography A, 1218 (2011) 6522–6530

Contents lists available at ScienceDirect

Journal of Chromatography A

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

iMatch: A retention index tool for analysis of gas chromatography–mass spectrometry data

Jun Zhang[a], Aiqin Fang[a], Bing Wang[a], Seong Ho Kim[b], Bogdan Bogdanov[a], Zhanxiang Zhou[c], Craig McClain[c,d,e], Xiang Zhang[a,*]

[a] *Department of Chemistry, University of Louisville, Louisville, KY 40202, USA*
[b] *Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40202, USA*
[c] *Department of Medicine, University of Louisville, Louisville, KY 40202, USA*
[d] *Department of Pharmacology & Toxicology, University of Louisville, Louisville, KY 40202, USA*
[e] *Louisville VAMC, Louisville, KY 40202, USA*

ARTICLE INFO

Article history:
Received 29 January 2011
Received in revised form 28 June 2011
Accepted 10 July 2011
Available online 23 July 2011

Keywords:
Retention index
Empirical distribution function
Identification
GC–MS

ABSTRACT

A method was developed to employ National Institute of Standards and Technology (NIST) 2008 retention index database information for molecular retention matching via constructing a set of empirical distribution functions (DFs) of the absolute retention index deviation to its mean value. The effects of different experimental parameters on the molecules' retention indices were first assessed. The column class, the column type, and the data type have significant effects on the retention index values acquired on capillary columns. However, the normal alkane retention index (I_{norm}) with the ramp condition is similar to the linear retention index (I^T), while the I_{norm} with the isothermal condition is similar to the Kováts retention index (I). As for the I_{norm} with the complex condition, these data should be treated as an additional group, because the mean I_{norm} value of the polar column is significantly different from the I^T. Based on this analysis, nine DFs were generated from the grouped retention index data. The DF information was further implemented into a software program called *iMatch*. The performance of *iMatch* was evaluated using experimental data of a mixture of standards and metabolite extract of rat plasma with spiked-in standards. About 19% of the molecules identified by ChromaTOF were filtered out by *iMatch* from the identification list of electron ionization (EI) mass spectral matching, while all of the spiked-in standards were preserved. The analysis results demonstrate that using the retention index values, via constructing a set of DFs, can improve the spectral matching-based identifications by reducing a significant portion of false-positives.

1. Introduction

Gas chromatography coupled to mass spectrometry (GC–MS) is one of the most widely used analytical techniques for analysis of small molecules such as metabolites in metabolomics, where analytes are first separated on a GC column and then subjected to MS measurement. The mass spectrometer in GC–MS is usually equipped with an electron ionization (EI) ion source. The EI process fragments the analyte's molecular ions resulting in mass spectrum. For molecular identification using the EI mass spectrum, several software packages have been developed by calculating the mass spectral similarity between the experimental mass spectrum and the mass spectrum recorded in a reference database [1–3]. However, the mass spectrum represents only partial information of a molecular structure. Identifying molecules based on spectrum matching only, therefore, may introduce false-positive identifications, especially for the analysis of isomers. Additional molecular information may be employed to increase the identification confidence. One approach is to combine molecular separation information with the mass spectrum matching.

The chromatographic separation information in GC–MS is the analyte's retention time, which varies from experiment to experiment. Kováts retention index (I) [4] and linear retention index (I^T) [5] were proposed to reduce the dependency of retention time values on the experimental conditions. However, the retention index value is still affected by several experimental conditions. For example, the retention index value of an analyte measured using different stationary phases of GC columns can be significantly different [6]. Several approaches have been proposed to use retention index value to assist molecular identification: Smith et al. suggested a constant retention index deviation window [7]. Zenkevich employed the average retention index value and standard deviation of reference retention indices calculated from the whole set of reference indices for identification [8].

* Corresponding author. Tel.: +1 502 852 8878; fax: +1 502 852 8149.
E-mail address: xiang.zhang@louisville.edu (X. Zhang).

0021-9673/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2011.07.039

Even though several retention index databases have been developed [9–14], the application of using retention index databases to aid molecular identification is not widely employed yet. Two main reasons prohibit the wide usage of the retention index values recorded in the current databases. One is that the retention index values recorded in the databases may not be reliable. The National Institute of Standards and Technology (NIST) retention index database [12] is currently the largest database. In spite of the fact that some erroneous or suspicious retention index data were removed from its 2008 version (NIST08), the retention index values of some molecules still exhibit a relatively large deviation, of which molecular misidentification in the literature is one of the main causes [15]. Second, compared to the mass spectral database, a relatively small number of retention time data are available. For example, only 21,847 molecules have retention index values in the NIST08 database while 192,108 molecules have mass spectra. One approach to increase the volume of retention index data is to employ quantitative structure-(chromatographic) retention relationships (QSRRs) to predict the chromatographic relationship from the numerical descriptors of each molecule [16–19]. However, the reliability of the QSRR models depends on a set of more reliable retention index data collection, which is used as input data of the QSRR model [20].

The objective of this work is to develop a method that uses the retention index data recorded in the NIST08 retention index database to increase the probability of correct molecular identification in GC–MS. The distribution of retention index values was analyzed to find the experimental parameters that do not significantly influence the retention index values, and then all the retention index values acquired under these experimental parameters were grouped together. If a database recorded experimental parameter has a strong effect on the retention index value, the retention index data were divided into different groups according to the values of this experimental parameter. After grouping all the retention index data based on their retention index deviations, the empirical distribution function (DF) of each grouped retention index data set was constructed, from which an appropriate retention index deviation window of each grouped retention index data set can be calculated by setting a statistical confidence interval. The results of this analysis were further implemented into a bioinformatics tool named *iMatch* using MATLAB 2008b to assist the molecular identification of mass spectrum similarity matching. The effectiveness of *iMatch* software was tested using experimental data of a mixture of 116 standards and a rat plasma metabolite extract spiked with 6 standards.

The following notations will be used throughout the article. Each retention index value recorded in the NIST08 retention index database is associated with experimental conditions including column type (capillary and packed), column class (standard non-polar, semi non-polar and standard polar), data type (Kováts retention index I, linear retention index I^T, Lee retention index I_{Lee} and normal alkane retention index I_{norm}), program type (ramp, isothermal and complex), and others (active phase, column length, carrier gas, substrate, column diameter, phase thickness, start temperature, end temperature, heat rate, start time and end time). The column type, column class, data type and program type are notated as experimental parameters, and further the information listed in the parenthesis of each experimental parameter is notated as the values of the corresponding experimental parameters.

2. Experimental

2.1. Mixture of standards

A mixture of 76 compounds (8270 MegaMix, Restek Corp., Bellefonte, PA) and C_7–C_{40} *n*-alkanes (Sigma–Aldrich Corp., St. Louis, MO) were spiked with a deuterated six components semi-volatiles internal standard (ISTD) mixture (Restek Corp., Bellefonte, PA) at a concentration of 2.5 μg/mL prior to comprehensive gas chromatography time-of-flight mass spectrometry (GC × GC/TOF-MS) analysis.

2.2. Rat plasma sample

A 200 μL rat plasma sample was mixed with 800 μL of an organic solvent mixture (chloroform:methanol:water = 2:5:2) to both precipitate proteins and extract metabolites from the sample. After sitting at room temperature for 1.0 h and being sonicated for 10 min, the sample was centrifuged at 15,000 × *g*. Supernatants from the mixture were collected and evaporated to dryness with a SpeedVac and then redissolved in 100 μL of pyridine [21]. 50 μL of the metabolite extract was treated with 100 μL of 50 mg/mL ethoxyamine hydrochloride pyridine solution for 30 min at 60 °C. Subsequently, the spiked extracts were derivatized with 100 μL of N-(tert-butyldimethylsilyl)-N-methyltrifluoroacetamide (MTBSTFA) for 1 h at 60 °C. After derivatization, 250 μL of the derivatized sample was spiked with the ISTD mixture at a concentration of 2.5 μg/mL prior to GC × GC/TOF-MS analysis.

2.3. GC × GC/TOF-MS analysis

All GC × GC/TOF-MS analyses were performed on a LECO Pegasus® 4D time-of-flight mass spectrometer (TOF-MS) (LECO Corporation, St. Joseph, MI) equipped with a Gerstel MPS2 auto-sampler (GERSTEL Inc, Linthicum, MD). The Pegasus 4D GC × GC/TOF-MS instrument was equipped with an Agilent 7890 gas chromatograph featuring a LECO two-stages cryogenic modulator and a secondary oven. A 30 m × 0.25 mm ^1d_c × 0.25 μm ^1d_f, Rxi-5 ms GC capillary column (5% diphenyl/95% dimethyl polysiloxane, Restek Corp., Bellefonte, PA) was used as the primary column for the GC × GC/TOF-MS analysis. A second column of 1.2 m × 0.10 mm ^2d_c × 0.10 μm ^2d_f, BPX-50 (50% phenyl polysilphenylene-siloxane, SGE Incorporated, Austin, TX) was placed inside the secondary oven after the thermal modulator. The helium carrier gas flow rate was set to 1.0 mL/min at a corrected constant flow via pressure ramps. A 1.0 μL liquid sample was injected into the liner using the splitless mode with the injection port temperature set at 260 °C. The primary column temperature was programmed with an initial temperature of 60 °C for 0.5 min and then ramped at a temperature gradient of 7 °C/min to 315 °C. The secondary column temperature program was set to an initial temperature of 65 °C for 0.5 min and then also ramped at the same temperature gradient employed in the first column to 320 °C accordingly. The thermal modulator was set to +20 °C relative to the primary oven and a modulation time of 5 s was used. The MS mass range was m/z = 10–750 with an acquisition rate of 150 spectra per second. The ion source chamber was set at 230 °C with the MS transfer line temperature set to 260 °C and the detector voltage was 1800 V with an electron energy of 70 eV.

2.4. Data reduction

LECO's ChromaTOF software package (version 4.21) equipped with the National Institute of Standards and Technology (NIST) MS database (NIST MS Search 2.0, NIST/EPA/NIH Mass Spectral Library; NIST 2002) was used for instrument control, spectrum deconvolution, and metabolite identification. The manufacturer's recommended parameters for ChromaTOF were used to reduce the raw instrument data into a metabolite peak list. These parameters are: baseline offset = 0.5; smoothing = auto; peak width in first dimension = 6 s; peak width in the second dimension = 0.1 s; signal-to-noise ratio (S/N) = 100.0; match required to combine

peaks = 500; R.T. shift = 0.08 s; minimum forward similarity match before name is assigned = 600. The peak true spectrum was also exported as part of the information for each peak in absolute format of intensity values.

3. Theoretical basis

3.1. The retention index

Four types of retention measurements are recorded in the NIST08 database: Kováts retention index (I), linear retention index (I^T), normal alkane retention index (I_{norm}) and Lee retention index (I_{Lee}) [22]. I, I^T and I_{norm} use the homologous n-alkane series as the references. The I is measured under isothermal conditions and the I^T is measured under temperature-programmed conditions (referred as ramp conditions in the NIST08 retention index database). Some n-alkane retention index data were categorized as normal alkane retention index in the NIST08 retention index database because the retention index calculation equation cannot be determined from the original literature. The I and the I^T are calculated as follows:

$$I = 100z + 100\left(\frac{\log(t'_{R(s)}) - \log(t'_{R(z)})}{\log(t'_{R(z+1)}) - \log(t'_{R(z)})}\right) \quad (1)$$

$$I^T = 100z + 100\left(\frac{t_{R(s)} - t_{R(z)}}{t_{R(z+1)} - t_{R(z)}}\right) \quad (2)$$

where I and I^T are the Kováts and linear retention index, respectively, t'_R is the adjusted retention time and t_R is retention time [23], s refers to the target compound that elutes off the GC column between two adjacent n-alkane reference compounds with carbon numbers z and $z+1$, respectively, z refers to the n-alkane with z carbon atoms and $z+1$ represents the n-alkane with $z+1$ carbon atoms.

The I_{Lee} system employs polycyclic aromatic hydrocarbons (PAHs): naphthalene, phenanthrene, chrysene and picene, i.e., compounds consisting of two, three, four and five fused benzene rings, respectively, as retention markers for gas chromatography of polyaromatic hydrocarbons and derivatives [12]. The value of the I_{Lee} can be calculated using Eq. (1) or (2) depending on the experimental conditions. All Lee indices were categorized as complex, isothermal or ramp index in the NIST08 retention index database. There are 572, 239 and 3447 values for the complex, isothermal and ramp Lee retention indices, respectively. For the purpose of comparison, all I_{Lee} isothermal values were converted into I values as follows [24,25]:

$$I = (194.4 - 0.201T) + L(4.48 + 3.72 \times 10^{-3}T) + L^2(4.21 \times 10^{-6}T - 1.16 \times 10^{-5}) \quad (3)$$

where T is temperature in °C, L the isothermal Lee retention index, and I the converted Kováts retention index. The complex and ramp Lee indices were converted into I values as follows [26]:

$$I = 127.7 + 4.5269 \times L + 2.6193 \times 10^{-3} \times L^2 + 5.00 \times 10^{-7} \times L^3 \quad (4)$$

where L is the complex or ramp Lee retention index and I the converted Kováts retention index. It can be expected that the converted Kováts retention indices may have large variation because Eqs. (3) and (4) are empirical. Furthermore, the Lee retention index is approximately six times smaller than the Kováts retention index. Most of the Lee retention indices were rounded to integers in the NIST08 retention index database, which also contributes to large variation in the converted values.

3.2. Column class

The column class refers to the stationary phase type. The column with similar stationary phase made by different manufacturers is divided into the same column class. All columns are classified into three column classes in the NIST08 retention index database: standard non-polar, semi non-polar and standard polar column. The typical standard non-polar column is DB-1 (100% dimethylpolysiloxane), semi non-polar column is DB-5 ((5%-phenyl)-methylpolysiloxane, 95% dimethyl) and standard polar column is DB-WAX (polyethylene glycol (PEG)).

3.3. Statistical analysis methods

Statistical methods can be employed to evaluate the effect of an experimental parameter on the retention index values. If different values of an experimental parameter significantly affect the retention index values of the majority of the molecules, the retention index data should be split into multiple groups according to the values of this experimental parameter. For example, the experimental parameter "column class" has three category values: standard non-polar, semi non-polar, and standard polar. If the column class does not significantly affect the retention time value, each molecule should have similar retention index values regardless whether it was analyzed on a standard non-polar, a semi non-polar, or a standard polar column. Otherwise, the retention index values of the same molecule should be statistically different according to each value of the experimental parameter.

The analsysis of variance (ANOVA) [27] is a statistical method to test whether the means of several groups of data are all equal. However, ANOVA assumes normal distribution of the test data. Heberger [28] found that the distribution of some molecules' retention index data does not follow normal distribution even though the experiments were conducted in the same lab. Kolmogorov–Smirnov test [29] was employed to check the distribution of all types of retention index values, i.e., I, I^T, I_{norm} and I_{Lee} were recorded in the NIST08 retention index database. It was concluded that these data acquired using some experimental parameters do not follow the normal distribution (data not shown). For this reason, a non-parametric alternative, the Kruskal–Wallis test [30], was used to determine the equality of the retention index values of the same molecule measured under different values of each experimental parameter since the Kruskal–Wallis test does not rely on the assumption of normal distribution. The Kruskal–Wallis test statistic is defined as follows:

$$\text{Null hypothesis } H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_k \quad (5)$$

$$\text{Alternative hypothesis } H_1: \quad \mu_i \neq \mu_j \quad (6)$$

$$H = \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(n+1) \quad (7)$$

where n is the total sample size, n_i ($i = 1, 2, \ldots, k$) represents the sample size of the ith group, R_i is the sum of the ranks for the ith group, and H is the Kruskal–Wallis statistic. The statistic approximates a chi-square distribution with $k-1$ degrees of freedom, if the null hypothesis of equal populations is true (H_0), otherwise, the H_0 will be rejected. In this study, the Kruskal–Wallis test was performed at an error level of 0.05.

For the molecules having multiple retention index values acquired under the same experiment conditions, two outlier detection algorithms were used to remove the outlier retention index values of each molecule before the analysis. The Grubbs's test was used for the molecules with more than 6 retention index values [31], and the Q-test was employed for retention index values smaller than 6 but larger than or equal to 3.

The retention index deviation of the same molecule to its mean value is given as follows:

$$dev_p^i = I_p^i - \frac{1}{N}\sum_{i=1}^{N} I_p^i \quad (8)$$

$$dev_p^{a,i} = \left| I_p^i - \frac{1}{N}\sum_{i=1}^{N} I_p^i \right| \quad (9)$$

where I_p^i is the ith retention index value of molecule p recorded in the NIST08 retention index database under an experiment condition of interest, N is the number of retention index values, dev_p^i is the deviation of the ith retention index value of p to its mean retention index value, and $dev_p^{a,i}$ is the absolute deviation of the ith retention index value of molecule p.

After grouping the retention index data according to the results of the Kruskal–Wallis test, the absolute mean difference of the retention index values of each molecule measured in two groups is defined as follows:

$$diff_p^g = \bar{I}_p^{g1} - \bar{I}_p^{g2} \quad (10)$$

where $diff_p^g$ is the retention index mean difference of molecule p measured when the value of experiment parameter g was set as $g1$ and $g2$, respectively; $\bar{I}_p^{g1}$ is the mean retention index value at $g1$; $\bar{I}_p^{g2}$ the mean retention index value at $g2$.

Every retention index value has its deviation in each group and the deviation values of all molecules in one group can form a deviation distribution. The empirical distribution function (DF) of absolute deviation can be created from this distribution. The DF is a function that assigns probability $1/n$ to each of n retention index database values. Its graph has a stair-step appearance. If a sample comes from a distribution in a parametric family such as a normal distribution, its empirical DF is likely to resemble the parametric distribution. If not, its empirical distribution still gives an estimate of the DF for the distribution that generated the data. From the DF curve, the size of retention index deviation window and its confidence level can be determined.

4. Results and discussion

In order to use retention index value to aid molecular identification, the ideal situation is that every molecule of interest has a reference retention index value, a variation window, and a statistical confidence interval under certain experimental conditions. This, however, is not true because of the very limited retention index information is recorded in the current retention index databases. The majority of molecules have a single retention index value and therefore, the size the retention index variation window cannot be statistically derived. In order to estimate the retention index variation window for a molecule that do not have enough reference retention index values in the database under certain experimental conditions, the effect of each experimental parameter on the retention index values was analyzed. If an experiment parameter does not have a significant effect on the retention index value, the retention index values of different molecules can be grouped together and the variance of this group of molecules is used to estimate the variance of the molecules that do not have enough retention index values in the database. During the grouping analysis, molecules that have at least 4 retention index database values in each group were used.

The retention indices of molecules with chemical abstract service (CAS) numbers were selected for analysis. A total 242,116 retention index values were extracted from the NIST08 retention index database for 14,878 molecules. Many experimental conditions categorized as “Others” in the NIST08 retention index database are not available for most of the molecules. For this reason, the scope of this study was further limited to investigate the effect of column type, column class, data type, and program type on the retention index value. Table 1 summarizes the number of molecules and their corresponding retention index values extracted from the NIST08 retention index database. Most of the retention index data in the database are either I or I^T values, while a small number of the I_{Lee} data was recorded in the NIST08 database. The bulk of retention index data recorded in the NIST08 retention index database were obtained on capillary columns since a capillary column provides much better GC separation efficiency than a packed column [32].

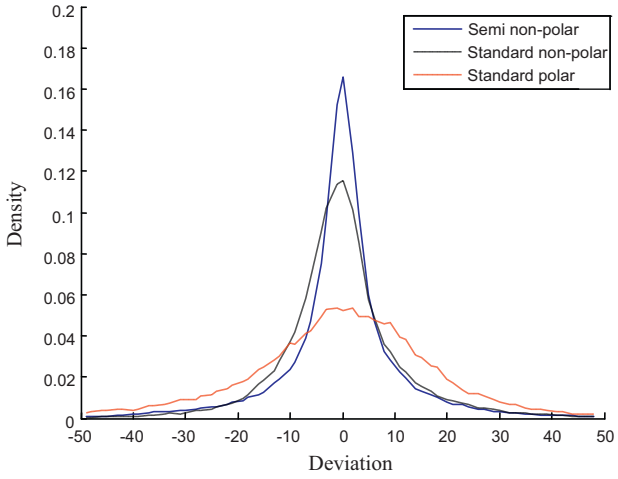

Fig. 1. The distribution curve of retention index deviation grouped only by column class. The retention index data extracted from NIST08 retention index database were divided into three groups according to the column class: semi non-polar, standard non-polar and standard polar column. The abscissa is retention index deviation defined by Eq. (8) and ordinate is the density of the deviation.

4.1. The effect of the column class

It has been reported that the column class has significant influence on retention index [6]. In this analysis, all database extracted retention index data were first divided into three groups according to the values of column class: standard non-polar, semi non-polar, and standard polar, defined by the NIST08 retention index database. A pairwise Kruskal–Wallis test was performed to study the effect of the column class values on the retention index value. There are 1749 molecules that each has at least four retention index values measured on both the standard non-polar column and the standard polar column. 1742 (99.6%) molecules have significantly different retention index values on the standard polar column compared to the standard non-polar column. This means that these two column classes have a significant effect on the retention index. Similar results were found between the standard polar column and the semi non-polar column, where 1505 molecules have at least four retention index values measured on the standard polar column and also on the semi non-polar column. Of the 1505 molecules, 1501 molecules (99.7%) have different retention index values and just 4 molecules have similar mean retention index values. As for the standard non-polar column and the semi non-polar column, 59.8% of molecules (1315 out of 2198) have statistically different retention index values between these two column classes. This indicates that the values of the column class can significantly affect the retention index and therefore, the retention index values acquired under different values of column class cannot be merged into one group.

Fig. 1 is the distribution of the retention index deviation grouped only by column class. The retention index deviation was calculated according to Eq. (8). The deviation distribution of the retention

Table 1
Summary of the retention index data extracted from the NIST08 retention index database (retention index values/molecules).

Column class		Data type		Program type		Column type	
Semi non-polar	91,365/9470	I	58,546/7144	Isothermal	45,105/6639	Capillary	221,704/13,001
Standard non-polar	79,766/10,645	I^T	72,551/7219	Ramp	157,138/11,619	Packed	19,293/5483
Standard polar	70,940/5689	I_{norm}	106,696/9917	Complex	39,808/5366		
		I_{Lee}	4258/1080				

index values measured on the standard polar column is much wider than the distributions of the other two column classes, indicating that the retention index deviation of standard polar columns is larger than the other two column classes. It can be concluded that the chromatographic reproducibility of standard polar columns reported in the literature is lower than the other two types of columns.

4.2. The effect of the data type

A total of four data types were recorded in the NIST08 retention index database: I, I^T, I_{norm}, and I_{Lee}. The most popular retention index types are I and I^T. In order to study the influence of the data type on the retention index system, the retention index data must be grouped not only by the data type, but also by the column class since the influence of the column class to the retention index system is statistically significant and cannot be ignored.

The I_{norm} recorded in the NIST08 retention index database is a data type for which data treatment was not clearly stated in the original literature but the alkane scaling was applied. There are three temperature program types recorded in the NIST08 retention index database for the I_{norm}: isothermal, ramp, and complex condition. For this reason, the I_{norm} values were dissected based on the temperature programmed type for comparative analysis. As for the I_{Lee} values, Eqs. (3) and (4) were used to convert them into the corresponding I values, respectively. Because the number of molecules that have both the I_{Lee} and the other type of retention index values is limited, the I_{Lee} was compared with all other types of retention index values.

Table 2 shows the analysis results based on a total of 17 comparative analyses using the pairwise Kruskal–Wallis test between different data types. For every pairwise comparison, the retention index values selected for a test must have the same column class. To a pair of two types of retention index listed in the first column, $N_{similar}$ represents the number of molecules with similar mean retention index values between the two types of retention index, while $N_{dissimilar}$ represents the number of molecules with different mean retention index values. Δ_{mean} is the mean difference of retention index between two types of retention index. There is not enough data to draw a sound statistical conclusion about the effect between the I versus the I_{norm} under isothermal conditions, and the same as I_{Lee} versus other retention index types. The mean value of the retention index difference defined in Eq. (10) between the I_{norm} under the ramp condition and the I^T is smaller than 3 i.u. (retention index units). The mean value of the retention index difference between the I_{norm} using the complex condition and the I^T is close to zero except for the polar column type. Therefore, these retention index data can be merged as one group with limited variations introduced. However, the I_{norm} values acquired under the complex condition has an obvious difference with the I values, and the percentage of molecules with significant different values ranges from 33.8% to 46.2%. There is also a significant difference between the I^T and the I, with more than 32.0% of the molecules having different values. Because lack of data, the big difference between the I_{Lee} and other retention index types demonstrates that the conversion equations of the I_{Lee} are not accurate. This indicates that the data type affects the retention index values, and the retention index values of the same molecule measured under these data types cannot be merged.

4.3. The effect of the column type

To show the effect of column type on the retention index value, the retention index data were pairwise compared according to the column type. Table 3 shows the comparison results of the Kruskal–Wallis test between different column types. All retention index values selected for each pairwise comparison have the same data type and column class. Since the molecules that have at least four I^T values acquired from the same column class on both the capillary and the packed column is limited, no valuable statistical results can be obtained from these data. To a fixed data type and column class, $N_{similar}$ represents the number of molecules with similar mean of retention index values measured on capillary and packed columns, while $N_{dissimilar}$ represents the number of molecules with different mean of retention index values. Δ_{mean} is the mean difference of retention index between the retention index values measured on the capillary and packed columns. As for the I values, the mean difference between the capillary column and the packed column ranges from 3 to 13 i.u., while the standard deviation ranges from 12 to 27 i.u. The results show that the effect of the column types on the I data is significant and cannot be ignored.

4.4. Grouping the retention index data

According to the analysis results presented above, the column class, column type, and data type all have an effect on the retention index value. However, the I_{norm} with the ramp condition can be merged with the I^T, while the I_{norm} with isothermal condition can be merged with the I values. As for the I_{norm} with complex condition, because the mean value of the standard polar column is significantly different from the I^T, these retention index data should be treated as an additional group. All retention index data acquired on the packed column are excluded for further analysis due to the limited data volume. By doing so, all the extracted retention index data of the molecules that have at least four retention index values acquired on capillary columns are categorized into 9 groups.

Fig. 2 shows the empirical distribution function (DF) of the 9 groups based on the absolute deviation of retention index values recorded in the NIST08 retention index database. The probability in each DF curve increases with the increase of the absolute deviation, and all of the DF curves level off approaching a value of 1.0 before the absolute deviation reaches 50 i.u. However, the retention index data acquired on the semi non-polar capillary columns have the best quality followed by the standard non-polar capillary columns. The standard polar capillary column has the worst performance. For example, when the cumulative probability is set as 0.95, the absolute deviations of the I^T on the semi non-polar, standard non-polar, and standard polar capillary columns are 18, 18, and 35 i.u., respectively.

To study the relation between the DFs and the number of retention index values measured for each molecule, the molecules with retention index data larger than 30, 60, and 100 records were chosen and the corresponding absolute deviation were used to create

Table 2
The pairwise comparison results of retention indices grouped by column class and data type.

Data type	Column class	$N_{similar}$	$N_{dissimilar}$	Δ_{mean}
I_{norm} with ramp condition vs. I^T	Standard non-polar	573	162	3
	Standard polar	548	154	1
	Semi non-polar	600	95	1
I_{norm} with isothermal condition vs. I	Standard non-polar	6	0	/
	Standard polar	0	1	/
	Semi non-polar	4	0	/
I_{norm} with complex condition vs. I	Standard non-polar	185	114	5
	Standard polar	135	116	8
	Semi non-polar	213	69	6
I_{norm} with complex condition vs. I^T	Standard non-polar	212	81	0
	Standard polar	197	217	11
	Semi non-polar	378	62	1
I^T vs. I	Standard non-polar	425	290	6
	Standard polar	259	122	5
	Semi non-polar	266	171	6
I_{Lee} vs. others	Standard non-polar	6	11	/
	Standard polar	0	0	/
	Semi non-polar	3	78	/

the DF (Fig. 3). Compared to Fig. 2, Fig. 3 shows that the absolute deviation of retention index values is slightly decreasing with the increase of the number of retention index data except for the I on the semi non-polar capillary columns. Further study shows that this was induced by one molecule, benzene (CAS: 71-43-2), which has 552 database recorded I values measured on semi non-polar columns. The histogram of these I values shows a very broad bimodal distribution (Fig. 4). Many factors such as false molecular identifications and inaccurate column classification may contribute to such a broad retention index distribution. Therefore, the influence of this molecule to the whole distribution will increase with the decrease of the number of molecules. In our study, the retention indices of such molecules were not removed because there is no clear evidence showing which fraction of the retention indices are the true positives.

4.5. Implementation of DF functions to aid compound identification

Fig. 5 shows the mean I^T values versus their corresponding standard deviations of 1506 molecules on the standard non-polar capillary columns. The Spearman's rank-order correlation coefficient between the standard deviations and the mean retention index values is only 0.319. The other columns show similar results (data are not shown). This suggests that using a relative retention index deviation window [33] to aid in molecular identification may not be an ideal approach. In this work, a maximum absolute retention index deviation window ΔI was set as the threshold of retention index value matching as follows:

$$\left|I_{\text{exp}} - I_{ref}\right| \leq \Delta I \tag{11}$$

where I_{exp} and I_{ref} are the retention index values of the experiment and reference values, respectively. The value of ΔI can be determined from the DF curve at a preferred confidence interval decided by the user. If the experimental retention index value satisfies this equation, the identification may be correct. Otherwise, the identification result of the mass spectrum matching is questionable and further validation is needed. For example, if the accumulative probability (confidence level) was set to 0.95, the threshold of the retention index window ΔI for molecules of interested analyzed on a semi non-polar capillary column in the temperature gradient mode will be 18 i.u. (Fig. 2a).

A software package entitled *iMatch* was developed to aid molecular identifications using the DF curves. *iMatch* uses the ChromaTOF results as its input and generates two lists, a preserved list and a filtered list. The preserved list contains all identified molecules whose retention index values equal to or less than ΔI and molecules that do not have retention index information in the NIST08 retention index database. The filtered list contains all molecules with experimental retention index values larger than ΔI, and these molecules are considered as false-positive identifications.

It should be noted that the retention index of some molecules has larger deviations and therefore, does not follow the DF distribution of the rest of molecules in that group. To detect these molecules, the mean standard deviation ($\overline{STD}$) of the retention index values in each of the 9 groups was calculated. If the STD of the retention index values of a molecule is larger than $2 \times \overline{STD}$, the CAS number of that molecule is kept in a separate list in *iMatch* software. The experimental retention index values of these molecules will not be evaluated, e.g. the mass spectrum identification results of these molecules will not be filtered regardless of the value of ΔI. A total of 549 of such molecules were detected. The information of these molecules is listed as S-Table 1 of Supplementary Material.

Table 3
The pairwise comparison results of retention indices group by column class, data type and column type.

Column type	Data type	Column class	$N_{similar}$	$N_{dissimilar}$	Δ_{mean}
Capillary vs. packed	I	Standard non-polar	124	104	4
		Standard polar	38	28	13
		Semi non-polar	170	73	3
	I^T	Standard non-polar	40	21	/
		Standard polar	3	1	/
		Semi non-polar	0	1	/

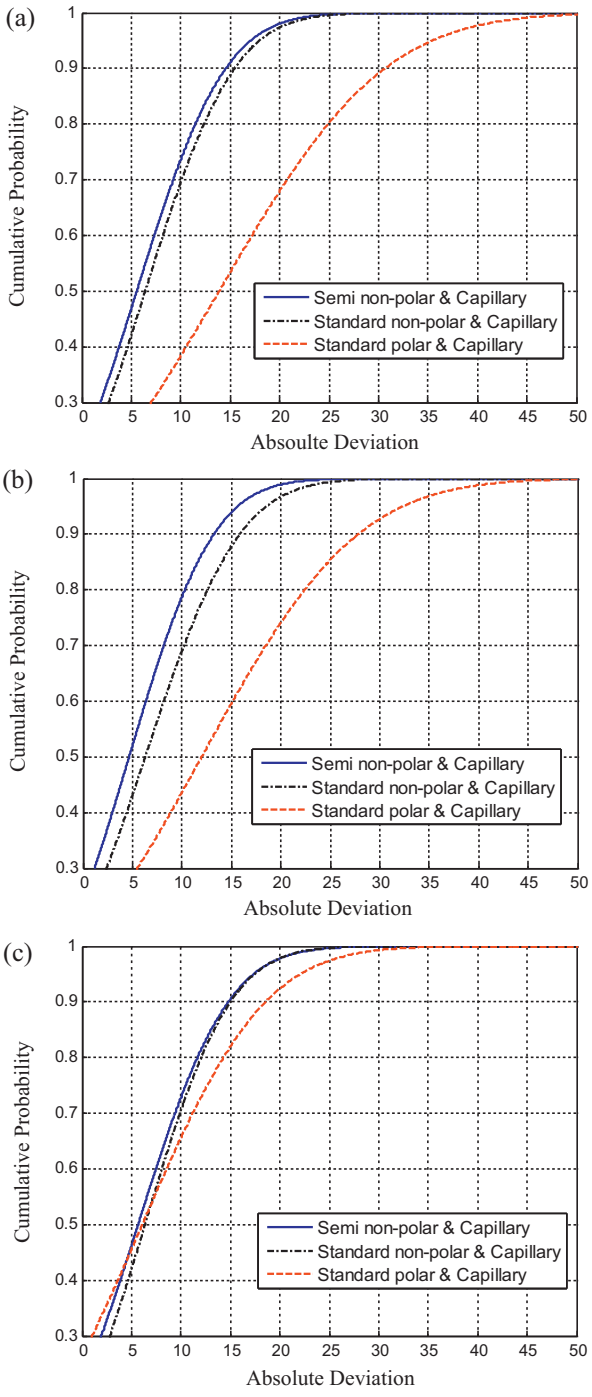

Fig. 2. The empirical distribution function (DF) of the absolute deviation on capillary column. The absolute deviation is defined in Eq. (9). (a) The DF of the absolute deviation on I^T, (b) the DF of the absolute deviation on I, and (c) the DF of the absolute deviation on normal alkane retention index with complex condition, all figures, the blue line is the semi non-polar column, the black line is the standard non-polar column, and the red line is the standard polar column. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.6. Analysis of the experimental data

The mixture of standards consists of 116 molecules, of which 34 are *n*-alkanes, 76 are from the MegaMix, and 6 are from the ISTD. A total of 26 alkanes (C_8–C_{34}) were detected in this study. The retention times of these detected *n*-alkanes were used to calculate the I^T values of the remaining standards. The information of the molecular identification by mass spectrum matching, I^T information obtained in this work, and those recorded in the NIST08 retention index database are provided as the Supplementary Material as S-Table 2.

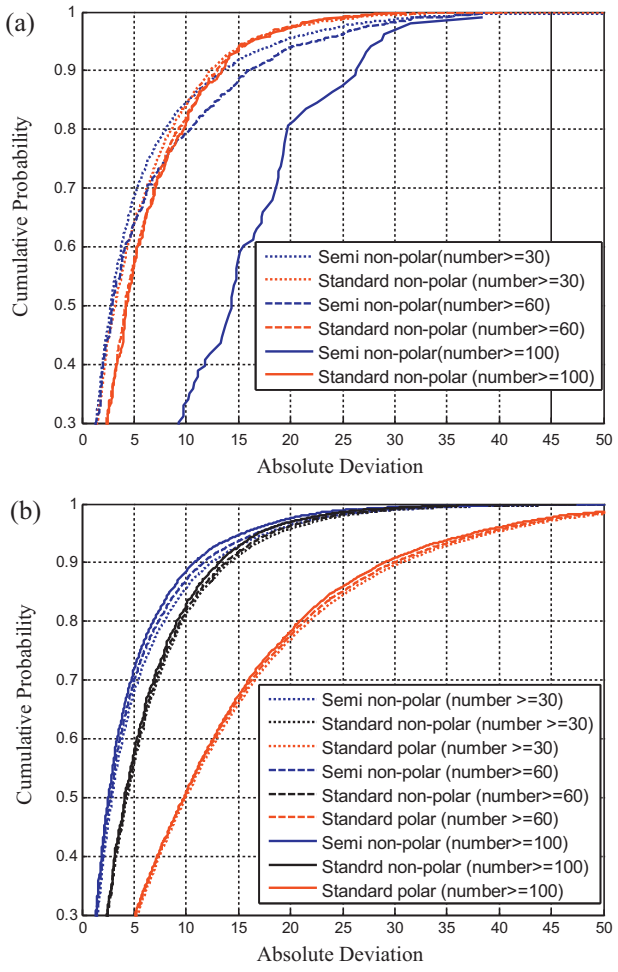

Fig. 3. The empirical distribution function (DF) of absolute deviation on capillary column. The retention index data of all molecules that have more than 30, 60 and 100 retention index values within one group were used to create the DFs. (a) The DF of absolute deviation on the I and (b) the DF of absolute deviation on the I^T.

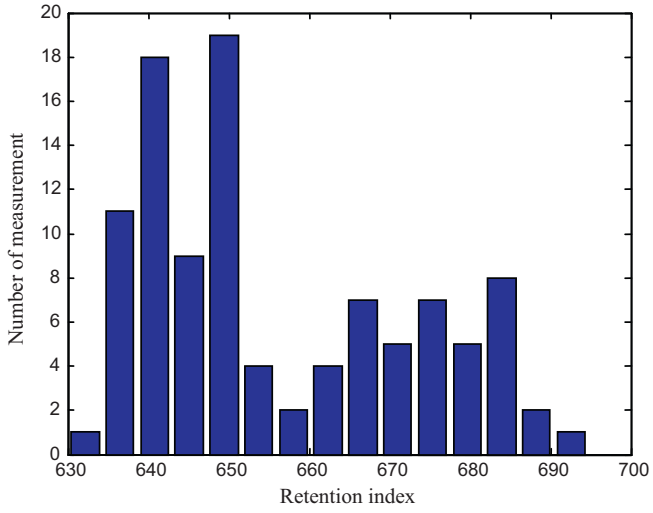

Fig. 4. The histogram of the Kováts retention index values of benzene [CAS number = 71-43-2] on the capillary semi non-polar column.

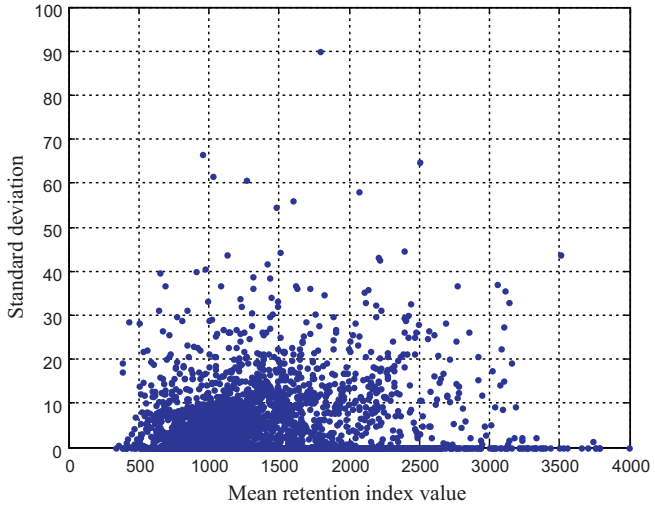

Fig. 5. The relationship between the mean retention index value and the corresponding standard deviation. The retention index data are the I^T values acquired on the standard non-polar capillary column. Most of the standard deviation is smaller than 40 i.u. and standard deviation does not increase with the increase of the mean I^T value.

After removing the molecules identified in the blank/solvent sample, ChromaTOF identified 162, 185, 154 molecules from the experimental data of three replicate injections of the mixture of standards, even though the mixture actually contains only 82 standards from MegaMix and ISTD. Of the 82 standards, 65, 63 and 63 were identified by ChromaTOF with a similarity score ranging from 602 to 957. The ISTD mixture consists of 6 molecules and all of them were identified by ChromaTOF via EI mass spectrum matching. Molecules 1,4-dichlorobenzene-D_4 and perylene-D_{12} do not have I^T information in the NIST08 retention index database. The I^T deviation of naphthalene-D_8, phenathrene-D_{10}, acenaphthene-D_{10}, and chrysense-D_{12} are 10, 30, 25 and 80 i.u., respectively.

iMatch uses the ChromaTOF results as its input and generates two lists, a preserved list and a filtered list. For the ISTD standards, four (naphthalene-D_8, acenaphthene-D_{10}, chrysene-D_{12} and phenanthrene-D_{10}) passed the I^T matching and two (1,4-dichlorobenzene-D_4, perylene-D_{12}) were preserved because of no reference I^T value in the database. This means that all the 6 ISTD standards were identified by EI mass spectrum matching and all of the identified standards were kept after *iMatch* analysis. Of the identified MegaMix standards, 10 fall into the list with large I^T deviations and 5 do not have I^T values in the NIST08 retention index database. All of these 15 molecules were kept in the preserved list by *iMatch*. For the rest of the standards, 40, 41 and 39 passed the I^T filtering when the cumulative probability was set to 0.999. *iMatch* removed 25, 25 and 22 molecules that were identified by EI mass spectrum matching, because of the large deviation between the experimental I^T values and the database values. These removed molecules are actually not present in the mixture and therefore, are false-positive identifications. *iMatch* also rejected 4, 1 and 3 standards that present in the sample and identified by EI mass spectrum matching. These molecules are considered as the false-negatives generated by *iMatch* analysis. However, it should be noted that only one of these false-negatives was identified in all of the three replicate injections. It is possible that the rest of the false-negatives are actually random matches, e.g., most likely they are false-positive identifications of EI mass spectrum matching. Comparing the number of false-positives and the number of false-negatives rejected by *iMatch*, it can be concluded that using retention index as a filtering method can identify and remove a major portion of false-positive identifications of EI mass spectrum matching.

MTBSTFA derivatized metabolites extracted from rat plasma with spiked-in ISTD were analyzed five times using GC × GC/TOF-MS. After processing the instrumental data using ChromaTOF, five peak lists were generated. Each of them consisted of 1176, 1155, 1163, 1116 and 1202 EI mass spectrum matching identified molecules. *iMatch* software was then employed to process each of these peak lists for I^T matching. By setting the confidence level to 0.999 in *iMatch*, the ISTD standards were first manually checked in the output files of *iMatch*. All the six molecules (naphthalene-D_8, acenaphthene-D_{10}, phenanthrene-D_{10}, 1,4-dichlorobenzene-D_4, perylene-D_{12} and chrysene D_{12}) were preserved. This is consistent to the results obtained from the mixture of standards, which means that the sample complexity does not significantly affect the identification of the ISTD standards.

After *iMatch* analysis, a total of 970, 941, 898, 918 and 978 molecules were preserved in the identification lists of the five replication injections. Of these preserved molecules, 129, 134, 122, 134 and 139 passed the I^T filtering criteria in the five peak lists, respectively, while the rest were persevered because of no I^T information in the NIST08 retention index database. A total of 206, 214, 265, 198 and 224 identified molecules were rejected by *iMatch* because of large retention index deviations, respectively. This represents an average rejection ratio of 19%. Even though it is impossible to assess the rate of false-positives and false-negatives removed by *iMatch* because of the sample complexity, the observation of ISTD standards strongly suggests that the methods proposed in this study can remove a significant portion of false-positive identifications.

5. Conclusions

A software entitled *iMatch* was developed to aid molecular identification using the retention index information recorded in National Institute of Standards and Technology (NIST) 2008 retention index database. Kruskal–Wallis test was used to assess the effect of various experimental parameters to the retention index values. The columns class, the column type and data type affect the retention index values. However, the normal alkane retention index I_{norm} with ramp condition, i.e., temperature-programmed condition, can be merged with the linear retention index I^T, while the I_{norm} with isothermal condition can be merged with the Kováts retention index I. As for the I_{norm} with complex condition, because the mean value of the polar column is significantly different from the I^T, these retention index data should be treated as an additional group. According to these analysis results, all retention index values extracted from the NIST08 retention index database were categorized into nine groups. An empirical distribution function (DF) was generated from the absolute deviation of retention index for each group, from which retention index variation window can be obtained at a specified confidence level. The DF information is further incorporated in the *iMatch* software, where the user can specify the confidence level. The performance of *iMatch* was evaluated using experimental data of a mixture of standards and metabolite extract of rat plasma extract with spiked-in standards. About 19% of the molecules identified by ChromaTOF were filtered out by *iMatch* from the EI mass spectrum matching identification results of plasma data, while all of the spiked-in standards were preserved. These analysis results demonstrate that using retention index can improve the spectral similarity-based identifications by reducing a significant portion of false-positive identifications.

Acknowledgements

The authors thank Drs. Steve Stein and Edward White of the National Institute of Standards and Technology (NIST) for their help with the interpretation of the NIST08 retention index database. This

work was supported by National Institute of Health (NIH) grant 1RO1GM087735 through the National Institute of General Medical Sciences (NIGMS) and 1RC2AA019385 through National Institute on Alcohol Abuse and Alcoholism, respectively.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.chroma.2011.07.039.

References

[1] S.E. Stein, J. Am. Soc. Mass Spectrom. 10 (1999) 770.
[2] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, Anal. Chem. 78 (2006) 779.
[3] J.M. Halket, A. Przyborowska, S.E. Stein, W.G. Mallard, S. Down, R.A. Chalmers, Rapid Commun. Mass Spectrom. 13 (1999) 279.
[4] E. Kováts, Helv. Chim. Acta 41 (1958) 1915.
[5] H. van Den Dool, P. Dec. Kratz, J. Chromatogr. 11 (1963) 463.
[6] C.F. Poole, T.O. Kollie, S.K. Poole, Chromatographia 34 (1992) 281.
[7] D.H. Smith, M. Achenbach, W.J. Yeager, P.J. Anderson, W.L. Fitch, T.C. Rindfleisch, Anal. Chem. 49 (1977) 1623.
[8] I.G. Zenkevich, J. Ecol. Chem. 3 (1994) 111.
[9] V.I. Babushok, P.J. Linstrom, J.J. Reed, I.G. Zenkevich, R.L. Brown, W.G. Mallard, S.E. Stein, J. Chromatogr. A 1157 (2007) 414.
[10] R. Richmond, J. Chromatogr. A 758 (1997) 319.
[11] T. Kind, G. Wohlgemuth, D.Y. Lee, Y. Lu, M. Palazoglu, S. Shahbaz, O. Fiehn, Anal. Chem. 81 (2009) 10038.
[12] S.E. Stein, Retention Indices in NIST Chemistry WebBook. NIST Standard Reference Database Number 69, versions 2005 and 2008, 2008.
[13] The Sadtler Standard Gas Chromatography Retention Index Library Sadtler-Heyden, Philadelphia, PA, 1986.
[14] V. Pacakova, L. Feltl, Chromatographic Retention Indices and Aid to Identification of Organic Compounds , Ellies Horwood, New York, 1992.
[15] I.G. Zenkevich, V.I. Babushok, P.J. Linstrom, E. White, S.E. Stein, J. Chromatogr. A 1216 (2009) 6651.
[16] T. Hancock, R. Put, D. Coomans, Y. Vander Heyden, Y. Everingham, Chemom. Intell. Lab. Syst. 76 (2005) 185.
[17] Z. Garkani-Nejad, M. Karlovits, W. Demuth, T. Stimpfl, W. Vycudilik, M. Jalali-Heravi, K. Varmuza, J. Chromatogr. A 1028 (2004) 287.
[18] V.V. Mihaleva, H.A. Verhoeven, R.C.H. de Vos, R.D. Hall, R.C.H.J. van Ham, Bioinformatics 25 (2009) 787.
[19] K. Heberger, J. Chromatogr. A 1158 (2007) 273.
[20] R. Kaliszan, Chem. Rev. 107 (2007) 3212.
[21] X.D. Huang, F.E. Regnier, Anal. Chem. 80 (2008) 107.
[22] M.L. Lee, D.L. Vassilaros, C.M. White, M. Novotny, Anal. Chem. 51 (1979) 768.
[23] G. Castello, P. Moretti, S. Vezzani, J. Chromatogr. A 1216 (2009) 1607.
[24] V.I. Babushok, P.J. Linstrom, Chromatographia 60 (2004) 725.
[25] V.I. Babushok, P.J. Linstrom, Abstr. Pap. Am. Chem. S. 228 (2004) U124.
[26] S.E. Stein, V.I. Babushok, R.L. Brown, P.J. Linstrom, J. Chem. Inf. Model. 47 (2007) 975.
[27] R.V. Hogg, J. Ledolter, Engineering Statistics , MacMillan, New York, 1987.
[28] K. Heberger, Anal. Chim. Acta 223 (1989) 161.
[29] M.A. Stephens, J. Am. Stat. Assoc. 69 (1974) 730.
[30] K.A. Wallis, J. Am. Stat. Assoc. 47 (1952) 583.
[31] F.E. Grubbs, Technometrics 11 (1969) 1.
[32] C.A. Cramers, H.G. Janssen, M.M. van Deursen, P.A. Leclercq, J. Chromatogr. A 856 (1999) 315.
[33] N. Strehmel, J. Hummel, A. Erban, K. Strassburg, J. Kopka, J. Chromatogr. B 871 (2008) 182.