Journal of Chromatography B, 910 (2012) 61–70

Contents lists available at SciVerse ScienceDirect

Journal of Chromatography B

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

Similarity analyses of chromatographic fingerprints as tools for identification and quality control of green tea☆

G. Alaerts[a], J. Van Erps[b], S. Pieters[a], M. Dumarey[a], A.M. van Nederkassel[a], M. Goodarzi[a], J. Smeyers-Verbeke[a], Y. Vander Heyden[a,*]

[a] Department of Analytical Chemistry and Pharmaceutical Technology (FABI), Center for Pharmaceutical Research (CePhaR), Vrije Universiteit Brussel (VUB), Laarbeeklaan 103, 1090 Brussel, Belgium

[b] Department of Applied Physics and Photonics (TONA-FirW), Brussels Photonics Team (B-Phot), Vrije Universiteit Brussel (VUB), Pleinlaan 2, 1050 Brussel, Belgium

ARTICLE INFO

Article history:
Received 28 December 2011
Accepted 27 April 2012
Available online 18 May 2012

Keywords:
Correlation and distance matrix
Fingerprint chromatography
Green tea
Quality control
Sample identification
Similarity analysis

ABSTRACT

Similarity assessment of complex chromatographic profiles of herbal medicinal products is important as a potential tool for their identification. Mathematical similarity parameters have the advantage to be more reliable than visual similarity evaluations of often subtle differences between the fingerprint profiles. In this paper, different similarity analysis (SA) parameters are applied on green-tea chromatographic fingerprint profiles in order to test their ability to identify (dis)similar tea samples. These parameters are either based on correlation or distance measurements. They are visualised in colour maps and evaluation plots. Correlation (r) and congruence (c) coefficients are shown to provide the same information about the similarity of samples. The standardised Euclidean distance (ds) reveals less information than the Euclidean distance (de), while Mahalanobis distances (dm) are unsuitable for the similarity assessment of chromatographic fingerprints. The adapted similarity score (ss^*) combines the advantages of r (or c) and de. Similarity analysis based on correlation is useful if concentration differences between samples are not important, whereas SA based on distances also detects concentration differences well. The evaluation plots including statistical confidence limits for the plotted parameter are found suitable for the evaluation of new suspected samples during quality assurance. The ss^* colour maps and evaluation plots are found to be the best tools (in comparison to the other studied parameters) for the distinction between deviating and genuine fingerprints. For all studied data sets it is confirmed that adequate data pre-treatment, such as aligning the chromatograms, prior to the similarity assessment, is essential. Furthermore, green-tea samples chromatographed on two dissimilar High-Performance Liquid Chromatography (HPLC) columns provided the same similarity assessment. Combining these complementary fingerprints did not improve the similarity analysis of the studied data set.

1. Introduction

In herbal samples, the variability of active compounds and their concentrations is well known. They vary with the species and with factors such as the cultivating region, the climate (temperature, humidity, light, wind) and the harvest time. Differences are also caused by the method of drying, washing, crushing and pulverising plants, as well as storage and conservation [1,2]. Proper identification and quality control is required in the crusade against the commercialisation of low-quality 'lookalikes', containing lower concentrations of active compounds or higher concentrations of contaminants (like pesticides) [3,4]. (Un)conscious fraud might also be caused by language confusions or by a lower harvest quality due to climate conditions [4–6]. Therefore, identification, as part of the quality control of herbal medicines or nutraceuticals, is essential for the user's safety.

Regulatory instances provide monographs and guidelines to ensure the quality of medicines. In monographs of, for instance, The European Pharmacopoeia [7], The United States Pharmacopeia [8] and The Pharmacopoeia of the People's Republic of China [9], besides macroscopic and microscopic identification, markers are often specified for the identification and quality control of bulk herbal material. Because of the highly complex and unknown composition of herbs and the lack of unique markers, this approach is not always appropriate for the identification and global quality control of a herb [10,11]. Identification based on a limited number of

☆ This paper belongs to the Special Issue Chemometrics in Chromatography, Edited by Pedro Araujo and Bjørn Grung.

* Corresponding author at: Department of Analytical Chemistry and Pharmaceutical Technology (FABI), Vrije Universiteit Brussel (VUB), Laarbeeklaan 10, B-1090 Brussel, Belgium. Tel.: +32 2 477 47 34; fax: +32 2 477 47 35.
E-mail address: yvanvdh@vub.ac.be (Y. Vander Heyden).

1570-0232/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jchromb.2012.04.031

markers is thus not always sufficient and could be replaced by the information originating from the entire fingerprint, *i.e.* a characteristic profile of the herb [12,13]. Fingerprints can be obtained by spectroscopic or separation (mainly chromatographic) techniques [10,14–19]. Regulatory agencies, such as the European Medicines Agency (EMA) [20], the American Food and Drug Administration (FDA) [21], the Chinese State Food and Drug Administration [22], the World Health Organisation (WHO) [23] and the abovementioned Pharmacopeia commissions accept in monographs the use of fingerprints, besides macro- and microscopic identification. An overview of existing regulations and guidelines about the quality control of herbal medicines is presented in Ref. [24].

A proper identification should confirm that a sample is originating from the expected herb and exclude that it is from another. The fingerprint of a sample is commonly compared with that of a reference standard extract. Since chromatographic fingerprints of complex samples, like herbal extracts, may contain large numbers of low concentrated compounds, a visual evaluation cannot always discriminate between the profiles [25,26]. Therefore mathematical data handling techniques are recommended.

To evaluate (dis)similarities, two types of mathematical data handling approaches can be used, *i.e.* ‘similarity analysis’ and ‘exploratory data analysis’. Exploratory data analysis techniques visualise trends within large groups of samples, characterised by many variables. New samples are positioned relative to the abovementioned groups of samples. Principal Component Analysis (PCA) or Hierarchical Clustering Analysis (HCA) is frequently used technique [1,14,18,27–33]. An overview of these and other techniques, illustrated with examples, can be found in Ref. [10].

The second approach, *i.e.* similarity analysis, compares the samples two-by-two. SA parameters, *e.g.* correlation coefficients (r), are also widely used to evaluate (dis)similarities between herbal fingerprints [2,27–31,34–40]. Correlation coefficients evaluation of HPLC fingerprints has been used to distinguish between substitutes and adulterants [36]. Inter- and intra-manufacturer batch-to-batch consistency may be another objective of SA [34,35]. SA is occasionally based on a number of selected peaks [27,41]. In our opinion, SA in quality control is more informative when the entire profile is used, as dissimilarities in the non-selected peaks can be important as well.

Besides correlations, also measures of distance can be used for SA. However, distance calculations, *e.g.* Euclidean and Mahalanobis distances, are mostly performed in combination with an exploratory data analysis [27,33,42,43]. The choice for either a correlation or a distance parameter requires a consideration of the objective goal [44,45] and is a part of our study. In the literature [10], it is noticed that the choice of a good reference chromatogram is critical to obtain representative similarity values for the samples to be evaluated. Similarities are occasionally determined after comparison with a genuine sample, identified as that with the highest similarity to all others [2]. Often the mean or median fingerprint of the samples is taken as the reference when standard extracts of the herb are unavailable [27–31,46]. According to [47], the mean fingerprint should be used if no outlying fingerprints are present, otherwise the median can act as reference. Similarity values for samples are preferably to be determined relative to a group of genuine fingerprints. Comparison with a range of similarity values from a number of genuine samples is therefore also used, for instance, in Ref. [34]. This approach was also applied in this study.

The main goal of this paper is to compare different correlation and distance measures to evaluate their suitability for similarity analysis of chromatographic fingerprint profiles as a tool for identification and quality control of herbal samples. Three data sets of green-tea fingerprints are used as case studies. A second goal of this paper is to evaluate the usefulness of dissimilar chromatographic fingerprints, *i.e.* chromatograms obtained on dissimilar chromatographic systems. It is investigated whether or not the combination of such fingerprints reveals more information about the (dis)similarities between samples.

2. Theory

Correlation and distance measures can be used for similarity analysis of herbal chromatographic fingerprints.

2.1. Similarity analysis based on correlation

The correlation parameters used in the literature can be reduced to the (Pearson product-moment) correlation coefficient r and the congruence coefficient c (Eqs. (1)–(3)). Both r and c are calculated between each pair of fingerprints, $\mathbf{x}_i$, with $i = 1, 2, \ldots, p$, and where each fingerprint is composed of measurements at $j = 1, 2, \ldots, q$ time points.

$$r(\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathrm{cov}(\mathbf{x}_1\mathbf{x}_2)}{s_{x1}s_{x2}} = \frac{\sum_{j=1}^{q}(x_{1j}-\bar{x}_1)(x_{2j}-\bar{x}_2)}{\sqrt{\sum_{j=1}^{q}(x_{1j}-\bar{x}_1)^2\sum_{j=1}^{q}(x_{2j}-\bar{x}_2)^2}} = \frac{(\mathbf{x}_1-\bar{x}_1)(\mathbf{x}_2-\bar{x}_2)}{||\mathbf{x}_1-\bar{x}_1||\,||\mathbf{x}_2-\bar{x}_2||} \qquad (1)$$

$$c(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{j=1}^{q}x_{1j}x_{2j}}{\sqrt{\sum_{j=1}^{q}x_{1j}^2\sum_{j=1}^{q}x_{2j}^2}} = \frac{(\mathbf{x}_1)(\mathbf{x}_2)}{||\mathbf{x}_1||\,||\mathbf{x}_2||} \qquad (2)$$

with $\mathbf{x}_1$ and $\mathbf{x}_2$ the fingerprints considered, x_{1j} and x_{2j} the absorbances measured at the jth time point, $\bar{x}_1$ and $\bar{x}_2$ the respective means of the absorbances, cov the covariance of the fingerprints, s_{xi} the standard deviation, and $||\mathbf{x}_i||$ the norm of the fingerprint, *i.e.* the length of the corresponding vector $\mathbf{x}_i$, given by:

$$norm = ||\mathbf{x}_i|| = \sqrt{\sum_{j=1}^{q}x_{ij}^2} \qquad (3)$$

To evaluate whether chromatographic fingerprints are similar or not, the correlation coefficient r (Eq. (1)) is most frequently used [48]. Correlation coefficient calculations are used in a variety of applications [2,27–31,34–39]. As the correlation coefficient between two fingerprints is by definition equal to the scalar product of the normed mean-centred fingerprints, it is the ratio of the covariance of two fingerprints to the product of their standard deviations [49]. The more r is approaching 1, the more linear the relation between both fingerprints is and the more similar they are. This parameter r is integrated in the ‘Similarity evaluation system for chromatographic fingerprints of Traditional Chinese Medicines (Chinese Pharmacopoeia Committee, 2004)’ software [50]. Liang's group [38] developed a software package, Computer Aided Similarity Evaluation (CASE), for processing fingerprint data, in which the correlation coefficient is called linear correlation coefficient (*LCC*).

The congruence coefficient c (Eq. (2)) [51] is a correlation calculated with respect to the origin (as opposed to the correlation coefficient, which is calculated with respect to the mean). The congruence coefficient is also called the reflective correlation or the angular separation [45]. In the CASE software [38], this parameter is named the ‘Similarity Index’ and is expressed as the cosine of the angle between the fingerprints as such, *i.e.* not the mean-centred fingerprints as in r. Most often r is used, but depending on the objective, c can sometimes better explain the correlation between samples [45,52]. Therefore, it is interesting to check whether this is also the case in the similarity analysis of green-tea fingerprints.

2.2. Similarity analysis based on distance

Distance measurements between two fingerprints are another approach in similarity analysis. The larger the distance between two fingerprints, the more dissimilar they are.

When comparing with a reference extract, a large distance implies that the quality of the studied extract may be unacceptable [53]. If the difference between the fingerprints is acceptable, the sample can be accepted as having the same nutritional or pharmaceutical properties as the reference.

The distance between two fingerprints can be calculated in different ways. Most commonly used is the Euclidean distance *de* (Eq. (4)). In other cases, the standardised Euclidean distance *ds* (Eq. (5)) can be used. Another distance measure is the Mahalanobis distance *dm* (Eq. (6)) [44] which takes into account a covariance matrix **C**, *i.e.* the measure of the degree to which both variables are correlated. Therefore, it corrects for their correlation [54]. If no correlation exists between two variables, *i.e.* if $\mathbf{C}^{-1}$ equals the unit matrix **I**, the Mahalanobis distance equals the Euclidean distance. An important restriction is that **C** has to be regular to calculate $\mathbf{C}^{-1}$.

$$de(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{q}(x_{1j} - x_{2j})^2} = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T(\mathbf{x}_1 - \mathbf{x}_2)} \quad (4)$$

$$ds(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{q}[(x_{1j} - x_{2j})/s_j]^2} \quad \text{with} \quad s_j = \sqrt{\frac{1}{p}\sum_{i=1}^{p}(x_{ij} - \bar{x}_j)^2} \quad (5)$$

$$dm(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T\mathbf{C}^{-1}(\mathbf{x}_1 - \mathbf{x}_2)} \quad (6)$$

with $\bar{x}_j$ the mean and s_j the standard deviation of the values in the jth time point, and **C** the covariance matrix. The other symbols are explained higher.

Most distance measures are less obvious to evaluate than correlations because their numerical values can take different orders of magnitude. To determine whether a distance is small or large, it would be easier to use values between 0 and 1, as for correlation coefficients, representing large and small differences, respectively. The similarity score *ss*, given in Eq. (7) [55], is such an adaptation of a distance measure obtained by dividing the Euclidian distance by the sum of all absorbance values of fingerprint $\mathbf{x}_1$ and by subtracting that ratio from 1. The *ss* is close to 1 if the fingerprints, and thus the samples, are similar. This *ss* is not influenced by the scale of the absorbance values. Nevertheless, in the similarity score, the value of the comparison of fingerprint 1 with fingerprint 2 will be different from that of the comparison of fingerprint 2 with 1. We therefore propose a small adaptation by dividing the Euclidean distance by the maximum of the sums of the absorbances of fingerprints $\mathbf{x}_1$ and $\mathbf{x}_2$. This adapted similarity score, given in Eq. (8), is represented by *ss**.

$$ss(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{\sqrt{\sum_{j=1}^{q}(x_{1j} - x_{2j})^2}}{\sum_{j=1}^{q}x_{1j}} \quad (7)$$

$$ss*(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{\sqrt{\sum_{j=1}^{q}(x_{1j} - x_{2j})^2}}{\max\left(\sum_{j=1}^{q}x_{1j}, \sum_{j=1}^{q}x_{2j}\right)} \quad (8)$$

3. Data sets

The evaluation of similarity analysis parameters is performed on three data sets consisting of HPLC fingerprints of green-tea extracts [43,56]. These fingerprints were used to model the anti-oxidant capacity of the samples. The green-tea samples chromatographed in both studies were different.

One data matrix consists of the chromatographic fingerprints of p herbal samples. The rows represent the p samples and the columns the time points at which absorbances were measured. A value x_{ij} in the matrix represents the absorbance (mAU) in the ith sample (extract) at the jth time data point (Ultra-Violet (UV) detection was performed).

In the study of van Nederkassel et al. [43] the green-tea samples were chromatographed on a Chromolith SpeedROD column coupled to a Chromolith Performance column (total column length 150 mm) resulting in a chromatographic profile of 11 min. The data set consists of 55 fingerprints, gathered in data matrix **Xt** (55 × 3100). In Ref. [43], the first 52 fingerprints were marked as genuine green-tea samples, and fingerprints 53, 54 and 55 as outliers.

In the study of Dumarey et al. [56] two dissimilar columns were used, a Chromolith Performance and a Waters Xterra column, both of 100 mm length. The fingerprints of 63 green-tea samples were gathered in the data sets **Xa** (63 × 2701) and **Xb** (63 × 2952), respectively. In Ref. [56], the first 60 fingerprints were marked as genuine green-tea samples and fingerprints 61, 62 and 63 as outliers.

In both studies, each green-tea sample was chromatographed twice. The chromatograms were first aligned using Correlation Optimised Warping [57] to correct for retention time shifts. Then, for each sample the average chromatogram was calculated and included in the data sets **Xt**, **Xa** and **Xb**. The fingerprints of the three data sets are plotted in Fig. 1.

To test the effect of the alignment of chromatographic fingerprints on the similarity analysis, results from the original data sets, *e.g.* **XOa** (126 × 2701), were compared with those on the aligned (also called warped) data sets, *e.g.* **XWa** (126 × 2701). Notice that the latter matrices contain 126 fingerprints, the double of **Xa**, where duplicated fingerprints were averaged. However, averaging can only be done meaningfully if the fingerprints are well aligned.

To investigate the usefulness of dissimilar fingerprints for identification and quality control, data set **Xab** (63 × 5653) was created by concatenating for each sample the fingerprint of **Xa** by those of **Xb**.

All calculations and figures are made in Matlab 7.1 (The Mathworks, Natick, MA) on a computer with an Intel® Core™ i5-2400 CPU clocked at 3.10 GHz, and with 8.00 GB of RAM.

4. Results and discussion

To evaluate which similarity analyses are best to identify genuine and false (outlying) green-tea samples, those based on correlations (r and c) are evaluated in a first section. SA based on distance parameters (*de*, *ds*, *dm* and *ss**) is discussed in Section 4.2. A comparison with the correlation parameters is also made. In a preliminary study, the influence of aligning the fingerprints on the SA results was checked. Although it would be much simpler and

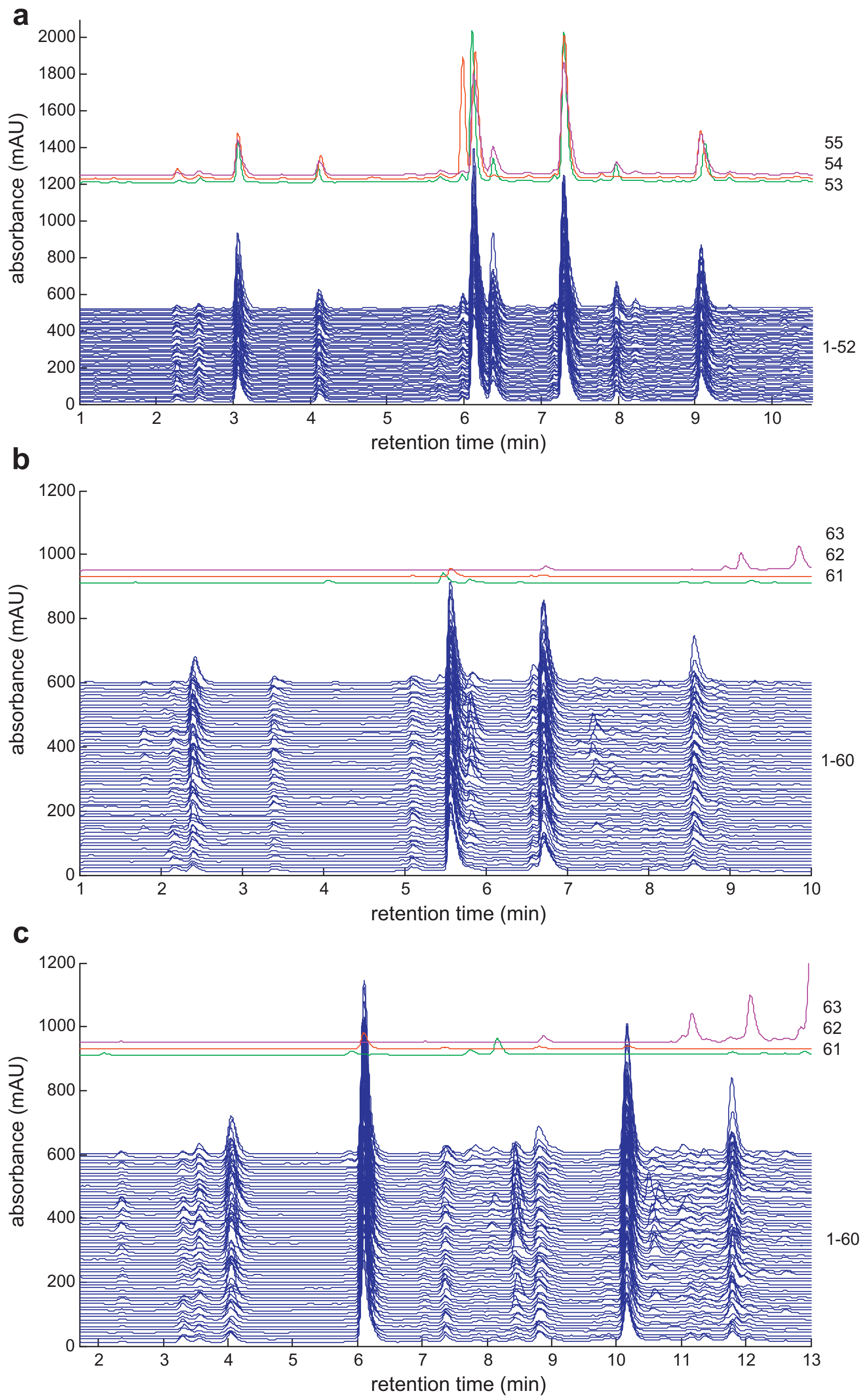

Fig. 1. Fingerprints of the green-tea samples from data sets (a) **Xt**, (b) **Xa** and (c) **Xb**. The suspected fingerprints are plotted above the regular.

less time consuming if no alignment was necessary for SA on chromatographic fingerprints, it turned out to be an absolutely essential step. An example is provided in Section 4.3.

Finally, the usefulness of the combination of dissimilar chromatographic fingerprints in SA is evaluated in Section 4.4.

4.1. Similarity analysis based on correlation

The correlation parameters are calculated pairwise between the p fingerprints in data sets **Xt**, **Xa** and **Xb**. The correlation (r) and congruence (c) coefficients (Eqs. (1) and (2)) of the data sets **Xt** and **Xa** form $p \times p$ correlation matrices [49] and are plotted in the colour maps of Fig. 2a and b, respectively. For both parameters, the higher the correlation is (dark red), the more similar the fingerprints and thus the green-tea samples are. The correlation (Fig. 2a$_1$ and b$_1$) and congruence (Fig. 2a$_2$ and b$_2$) coefficients-based colour maps are comparable for a given data set. Visually distinguishable samples, show low r and c values (blue) *versus* the majority of the fingerprints (red), *i.e. versus* the genuine samples. This is, for example, clearly the case for samples 53 and 54 in Fig. 2a, and 61 and 63 in Fig. 2b.

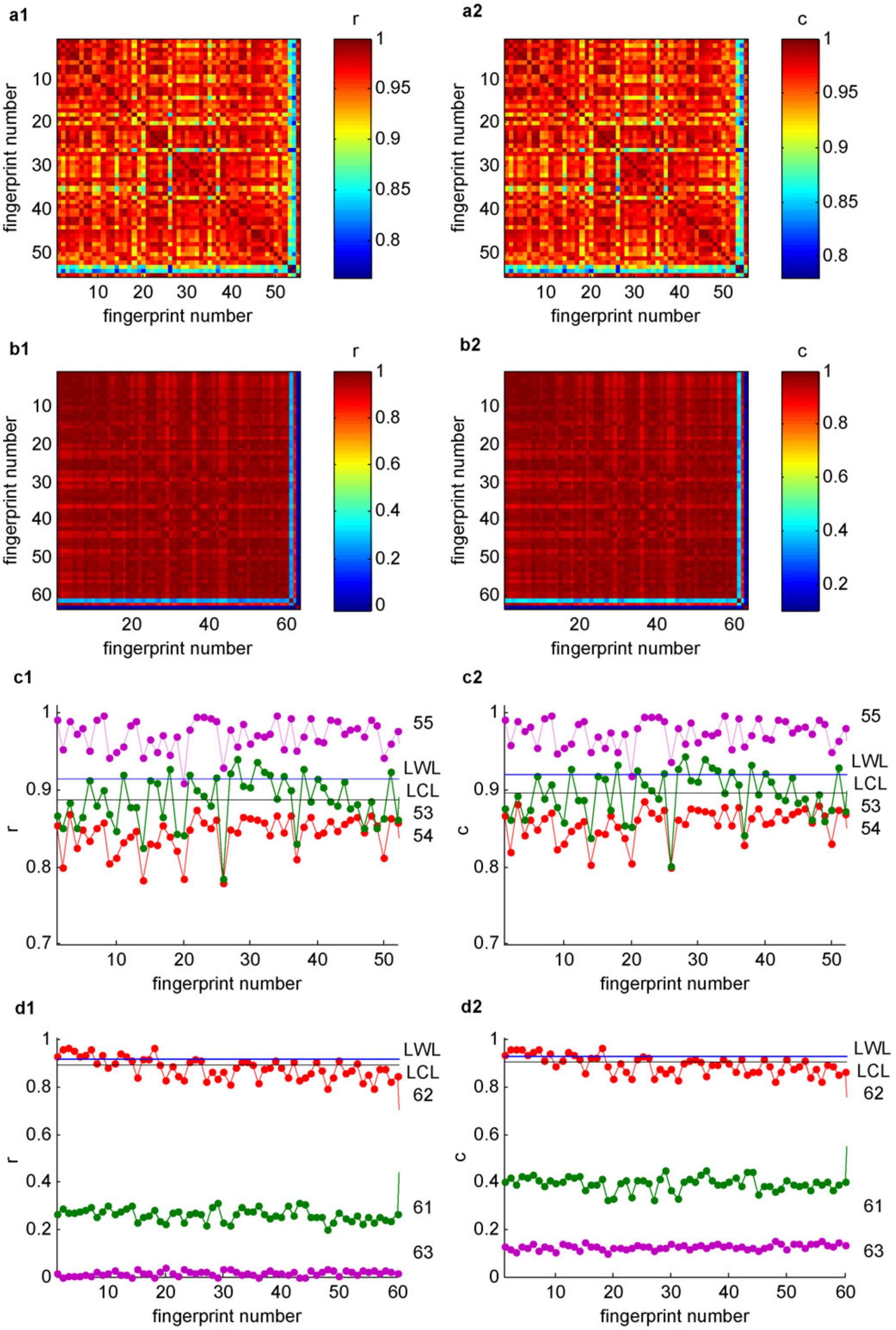

Fig. 2. (a and b) Colour maps of the correlation coefficients r (a_1, b_1) and congruence coefficients c (a_2, b_2) of data sets **Xt** (a) and **Xa** (b). (c and d) Evaluation plots for the correlation coefficients (c_1, d_1) and congruence coefficients (c_2, d_2) of the suspected fingerprints 53 (green), 54 (red) and 55 (magenta) from **Xt** (c), and 61 (green), 62 (red) and 63 (magenta) from **Xa** (d). Horizontal lines: lower warning and control limits (LWL and LCL) for the correlation and congruence coefficients, derived from the genuine green-tea fingerprints. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

To evaluate suspected samples and to distinguish between deviating and genuine ones, so-called 'evaluation plots' were also created. In dataset **Xt**, the first 52 fingerprints were defined as genuine samples and for data sets **Xa** and **Xb** the first 60 were. The 'evaluation plots' of the suspected samples, *i.e.* 53, 54 and 55 for data set **Xt**, and 61, 62 and 63 for **Xa**, show their correlation values with the genuine samples in Fig. 2c and d, respectively. For these evaluation plots, a reference range of similarity values is determined from the genuine samples. Representative critical values are important. In this study, the reference range of similarity values is determined from the genuine samples as the statistical confidence limits used in quality control charts [49,58]. In a normal distribution, 95% of, for instance, the correlation coefficients of genuine samples is situated in the interval $\bar{r} \pm 1.96 s_r$, with $\bar{r}$ the average correlation coefficient and s_r the standard deviation of the r-values. In control charts, these limits are called 'warning limits', or more specifically the lower warning limit (LWL), $\bar{r} - 1.96s_r$ and upper warning limit (UWL), $\bar{r} + 1.96s_r$. Thus 97.5% of the correlation or congruence coefficients between genuine samples lay above the lower warning limit (LWL), *i.e.* $\bar{r} - 1.96s_r$ or $\bar{c} - 1.96s_c$, respectively. Stricter limits are the 'control limits'. 99.7% of the correlation coefficients between genuine samples is situated between the 'control limits' ($\bar{r} \pm 3.09s_r$). Above the lower control limit (LCL), *i.e.* $\bar{r} - 3.09s_r$ or $\bar{c} - 3.09s_c$, lays 99.85% of the regular correlation or congruence coefficient values, respectively. The LWL and LCL values are indicated on the evaluation plots.

Notice that for a given data set the evaluation plots based on the correlation (Fig. 2c$_1$ and d$_1$) and congruence (Fig. 2c$_2$ and d$_2$) coefficients also show comparable results. The interpretation of the similarity values of the suspected samples is made as follows: a suspected sample is considered not genuine if more than 2.5% (two of the 52 and 60 results for **Xt** and **Xa**, respectively), of, for instance, its correlation coefficients with the genuine samples is below the LWL and if more than 0.15% (one correlation value of the suspected sample) is below the LCL.

The proposed evaluation plots with statistical confidence limits for suspected samples could be useful during quality control. During the development of a reference data set (preferably more than 20 samples used for quality control charts [49]) to determine the statistical warning and control limits for these evaluation plots, the presence of suspected samples is often not known in advance. This may affect the limits on the evaluation plots. However, suspected samples may be indicated from these evaluation plots and the colour maps visually reveal a global idea of similarities between all fingerprints in the data set. In this study, both colour maps and evaluation plots are therefore used to evaluate whether the applied similarity parameters are able to distinguish between deviating and genuine samples.

On the correlation- and congruence-based colour maps of data set **Xt** (Fig. 2a) the most dissimilar fingerprint is 54. This is also clear on the evaluation plots (Fig. 2c). All correlation values of fingerprint 54 are below the LCL. This fingerprint has indeed a clearly different pattern, *i.e.* an extra peak around 6 min and no peak around 8 min (Fig. 1a). In [43] only this fingerprint was considered as outlier based on robust PCA. Fingerprints 53 and 55 on the other hand, were excluded from the data set based on deviating Trolox Equivalent Antioxidant Capacity (TEAC)-values. Sample 53 was considered an outlier because of a very high TEAC-value [56]. This sample also shows a slightly different pattern *versus* the majority of fingerprints (Fig. 1a), resulting in lower r and c values (Fig. 2a). Indeed, on the evaluation plots in Fig. 2c, considerably more than 0.15% or 2.5% of the values is found below the LCL and LWL, respectively. Fingerprint 55 on the other hand, visually seems similar to the genuine fingerprints in the colour maps. It was considered as outlier because of experimental errors during the TEAC-assay, not based on its fingerprint profile. The SA (Fig. 2) confirms that fingerprint 55 should not be excluded as outlier and can be identified as a genuine green-tea sample.

Fig. 2b and d shows the colour maps and evaluation plots for data set **Xa**. Fingerprints 61 and 63 clearly have a low correlation with the other samples (Fig. 2b). The evaluation plots shown in Fig. 2d confirm the observations about these suspected fingerprints, *i.e.* they are clearly rejected as all their correlation values are below the LCL. In Ref. [56], these fingerprints were considered as outliers based on a visual evaluation of their profile and low TEAC-values. Fingerprint 62, also excluded in Ref. [56], is not clearly rejected based on the colour maps (Fig. 2b). However, from the evaluation plots, it is detected as an outlying sample. The fingerprint profile hardly shows any peak (Fig. 1b and c) and is thus logically rejected because it can be considered as a sample of bad quality.

We can conclude that both the correlation and congruence coefficients provide similar information and are valuable similarity analysis parameters. The colour maps and evaluation plots overcome the need for selection of a reference fingerprint. They allow evaluating the similarity of a fingerprint *versus* a set of reference (genuine) fingerprints. On the evaluation plots, all non-genuine samples were successfully detected, including the low quality sample 62 of dataset **Xa**. Sample 55 of dataset **Xt** was correctly detected as a genuine sample. The correlation-based evaluation plots with the statistical warning and control limits are thus efficient for the distinction between genuine and deviating samples.

4.2. Similarity analysis based on distance

Similarity analysis can also be performed by estimating distance parameters. The usefulness of these parameters in distinguishing between deviating samples and genuine ones, is studied in this section.

First the Euclidean (*de*) and standardised Euclidean (*ds*) distances (Eqs. (4) and (5)) are evaluated. The *de* and *ds* distances between each pair of fingerprints are calculated and form $p \times p$ distance matrices. In Fig. 3a and b the Euclidean (a_1, b_1) and standardised Euclidean (a_2, b_2) distance matrices are plotted for the fingerprints of data sets **Xt** and **Xa**, respectively. For *de* and *ds*, the blue colour represents low distance values, corresponding to high similarities, while the red colour represents low similarity. Generally, the colour maps of *de* (Fig. 3a$_1$ and b$_1$) and *ds* (Fig. 3a$_2$ and b$_2$) show different patterns for the two data sets. Notice that the values of *ds* are lower than *de*, as they are divided by their standard deviation (Eq. (5)).

The evaluation plots of the suspected samples 53, 54 and 55 for data set **Xt** (Fig. 3c) and 61, 62 and 63 for **Xa** (Fig. 3d) are constructed as described higher. The UWL is defined as mean distance +1.96s_d and the upper control limit (UCL) as mean distance +3.09s_d. A suspected sample is identified as not genuine in a given data set if more than 2.5% of its distance values with the genuine samples are above the UWL or more than 0.15% are above the UCL. For both data sets, the evaluation plots of *de* (Fig. 3c$_1$ and d$_1$) and *ds* (Fig. 3c$_2$ and d$_2$) show again different patterns.

In the *de* colour map, fingerprint 54 from data set **Xt** (Fig. 3a$_1$), is detected as deviating (plotted in red). This was also the case using the correlation measures. Fingerprint 53 is somehow suspicious in the *de* colour map. However, this was more obvious in the colour map of the correlations (Fig. 2a). Notice that some of the genuine fingerprints also show a lower similarity, which makes the interpretation of the *de* colour map in general less obvious than that of the correlations. In the *ds* colour map (Fig. 3a$_2$), many genuine samples show low similarities, which renders these colour maps less useful for similarity analysis. In the *de* evaluation plots (Fig. 3c$_1$), fingerprints 53 and 54 in data set **Xt** are detected as deviating samples. However, as for the *ds* colour maps, the *ds* evaluation plots (Fig. 3c$_2$) give less obvious results. Just 2.5% of the *ds* results of fingerprint 53 are above the UWL, but none above the UCL. Fingerprint 55 is correctly not detected as deviating profile in both the colour maps (Fig. 3a) and evaluation plots (Fig. 3c).

From the *de* colour map (Fig. 3b$_1$) one concludes for data set **Xa** that fingerprints 61, 62 and 63 are dissimilar to the genuine green-tea fingerprints. Fingerprint 62 is here clearly detected as deviating, which was not the case on the r or c colour maps. As this sample has small peaks and thus is of lower quality, we can conclude that the Euclidian distance detects this type of lower quality sample better. The *de* evaluation plot (Fig. 3d$_1$) confirms the colour maps, *i.e.* fingerprints 61, 62 and 63 are clearly rejected. The *de* better distinguishes low quality green-tea samples, caused by low concentrations, than the correlation measurements do. On the *ds*

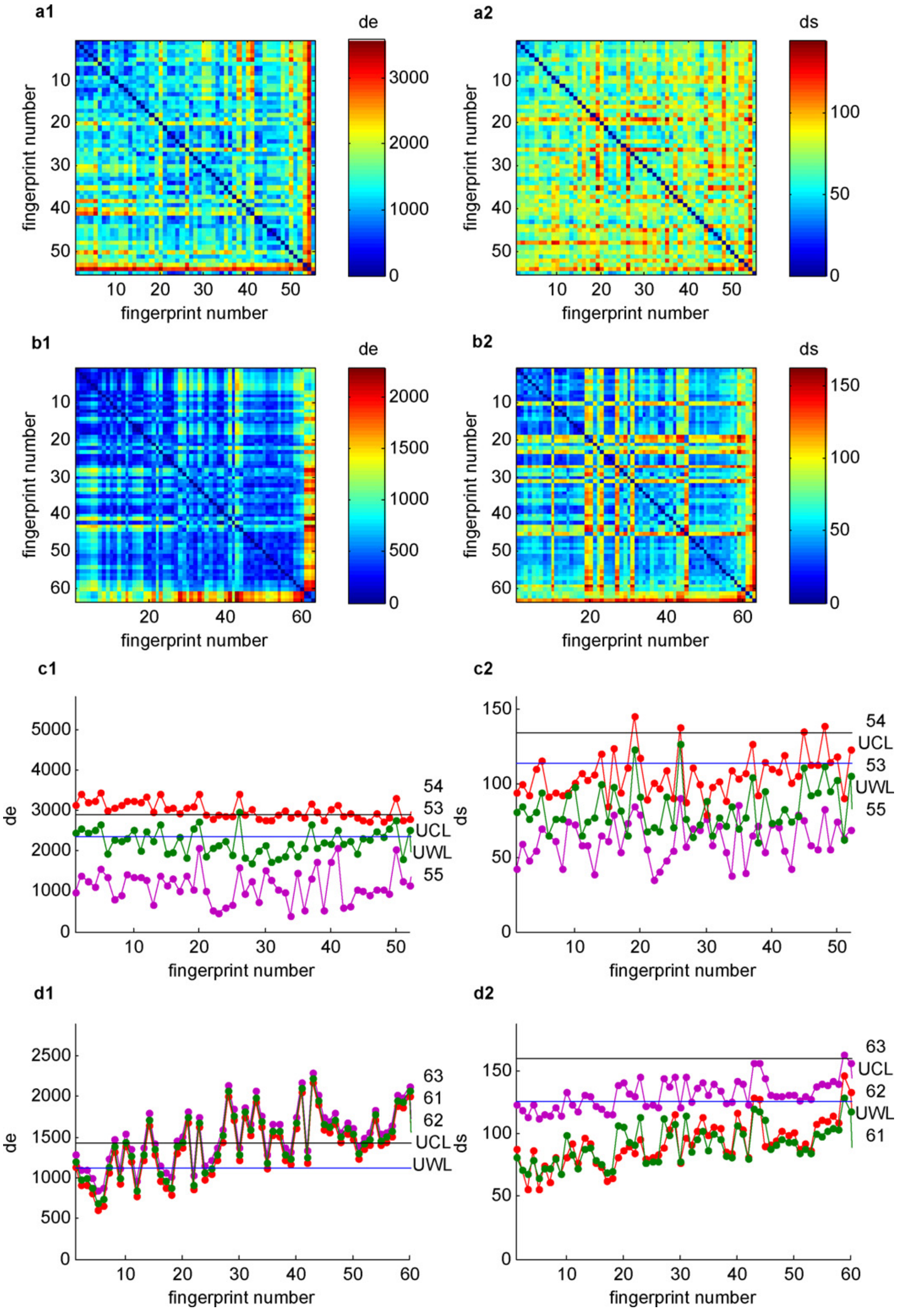

Fig. 3. (a and b) Colour maps of the Euclidean distances *de* (a_1, b_1) and the standardised Euclidean distances *ds* (a_2, b_2) of data sets **Xt** (a) and **Xa** (b). (c and d) Evaluation plots for the Euclidean distances *de* (c_1, d_1) and the standardised Euclidean distance *ds* (c_2, c_2) of the suspected fingerprints 53 (green), 54 (red) and 55 (magenta) from **Xt** (c), and 61 (green), 62 (red) and 63 (magenta) from **Xa** (d). Horizontal lines: upper warning and control limits (UWL and UCL) for the (standardised) Euclidean distances, derived from the genuine green-tea fingerprints. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

colour map (Fig. 3b_2) many genuine samples show a high distance and thus would indicate low similarities. The *ds* evaluation plot gives less obvious and non-consistent results. Just 2.5% of the *ds* results of fingerprint 62 are above the UWL, but none above the UCL, while fingerprint 61 was even not detected at all as deviating sample. This confirms that *ds* does not seem to be a valuable SA parameter to distinguish deviating profiles from a data set.

The third distance measure evaluated as similarity parameter is the Mahalanobis distance (*dm*) (Eq. (6)). The more similar the fingerprints are, the smaller the Mahalanobis distance is expected to be. From the colour map pattern no interpretation can be made (Fig. 4). The covariance matrices of the data sets are close to singular which makes the calculation of *dm* unreliable. The calculation time of the Mahalanobis distances was high, *e.g.* about 70 times higher

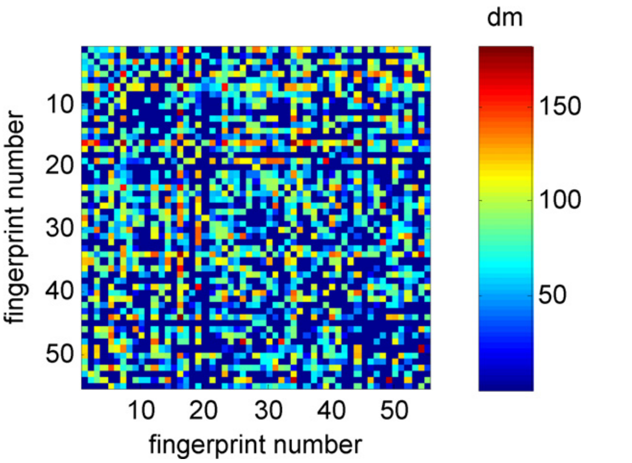

Fig. 4. Colour map of the Mahalanobis distances *dm* of data set **Xt**. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

compared to the case where this parameter was excluded from the SA calculation. The Mahalanobis distance is thus not useful as SA parameter for chromatographic fingerprints.

As a fourth distance measure, the adapted similarity score (*ss**), defined in Eq. (8), is introduced. Using distances like the Euclidean distance *de*, results in a wide range of values, *e.g.* the *de* ranges for the genuine green-tea fingerprints of data sets **Xt**, **Xa** and **Xb** are 362–3238, 60–1612 and 135–2863, respectively. Distance values between 0 and 1 are obtained using the adapted similarity score *ss**. It is a scaled distance, which is subtracted from 1. Consequently a similar scale as for the correlation parameters is obtained. A good similarity between fingerprints will thus result in a high *ss** (close to 1), a low similarity in a low (close to 0).

In Fig. 5a and b, the *ss** are plotted for the fingerprints of data sets **Xt** and **Xa**, respectively. The higher *ss** is (dark red), the more similar the fingerprints and thus the green-tea samples are. The blue colour represents a low similarity. In Fig. 5c and d the evaluation plots of the suspected samples 53, 54 and 55 for data set **Xt**, and 61, 62 and 63 for **Xa**, respectively, are shown. They are interpreted in a similar manner as the correlation-based evaluation plots, as described in Section 4.1.

The *ss** colour map from **Xt** (Fig. 5a), yields similar results as for the correlations, *i.e.* fingerprints 53 and 54 are detected as deviating and 55 is not. Nevertheless, the similarities and dissimilarities are visually clearer in the *r* and *c* colour maps (Fig. 2a). The *ss** evaluation plot (Fig. 5c) again provides similar information about the suspected samples: 53 and 54 are considered dissimilar, whereas 55 is identified as a genuine sample. All *ss** values of fingerprint 55 are above the LWL. This matches the conclusions drawn in Section 4.1 from the fingerprint patterns in Fig. 1a.

The *ss** colour map from **Xa** (Fig. 5b) provides some extra information compared to the *r* and *c* colour maps (Fig. 2b), which was also noticed using the Euclidean distance. The three suspected samples 61, 62 and 63 are clearly detected as deviating fingerprints. Fingerprint 62, which is not distinguished as low quality sample based on the *r* and *c* colour maps, is correctly detected from the *ss** colour maps. Therefore, the *ss** colour maps seem better than those of the correlation parameters to visually distinguish between deviating and genuine samples. However, from the *r* and *c* evaluation plots (Fig. 2d), this sample was also correctly detected. The *ss** evaluation plot (Fig. 5d) confirms the observations from the colour map, *i.e.* 61, 62 and 63 are clearly rejected as genuine samples. For this data set, the *ss** evaluation plot was even clearer than the one based on the Euclidean distance.

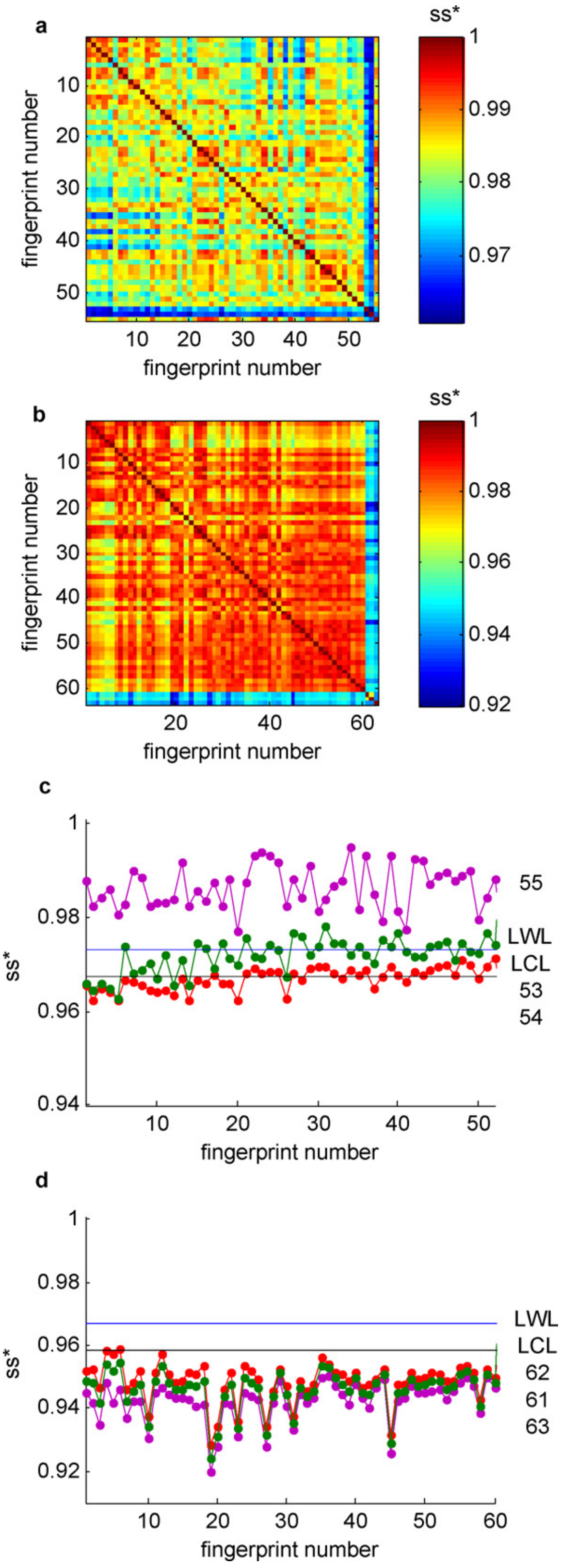

Fig. 5. (a and b) Colour maps of the adapted similarity score *ss** of data sets **Xt** (a) and **Xa** (b). (c and d) Evaluation plots for the similarity score of the suspect fingerprints 53 (green), 54 (red) and 55 (magenta) from **Xt** (c), and 61 (green), 62 (red) and 63 (magenta) from **Xa** (d). Horizontal lines: lower warning and control limits (LWL and LCL) for the adapted similarity scores, derived from the genuine green-tea fingerprints. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

In summary, we observed that standardised Euclidean distances reveal less interpretable information than the Euclidean distances, and that the Mahalanobis distances are not interesting for similarity analysis. Based on *de* or *ss** colour maps and evaluation plots, the same conclusions are drawn. The adapted similarity score combines the advantages of the correlation and the Euclidean distance matrices. First, distances like *de* and *ss** detected quality differences based on concentration differences. The low-quality sample 62 in data set **Xa** was detected on the *r* or *c* evaluation plots, whereas on the *r* or *c* colour maps it was not. Secondly, the *ss** parameter gives easily interpretable similarity results between 0 and 1. For the evaluation of new data sets, the *ss** colour maps and evaluation plots turns out to be valuable tools to distinguish between deviating and genuine samples.

4.3. Evaluation of fingerprint alignment on the similarity analysis

The influence of aligning the chromatographic fingerprints on the SA results was evaluated in a preliminary study. Should alignment not affect the SA results, then this data pre-treatment step could be removed and the approach would then be faster.

To study the alignment effect, we compared the results from data sets consisting of the original (not aligned) fingerprints, *e.g.* **XOa**, with their aligned fingerprints, *e.g.* **XWa**. In general, a very high influence on the SA results is seen. As an example, the colour map of *ss** from **XOa** (Fig. 6a) is plotted against its **XWa** (Fig. 6b) counterpart. It is obvious that warping increases the similarity parameter *ss** between genuine or similar samples. After correction of small retention time shifts, consequence of the warping, all genuine samples show a high similarity (Fig. 6b). Although it would be simpler and faster if no alignment was necessary, it turns out to be an absolutely essential step in similarity analysis of chromatographic fingerprints.

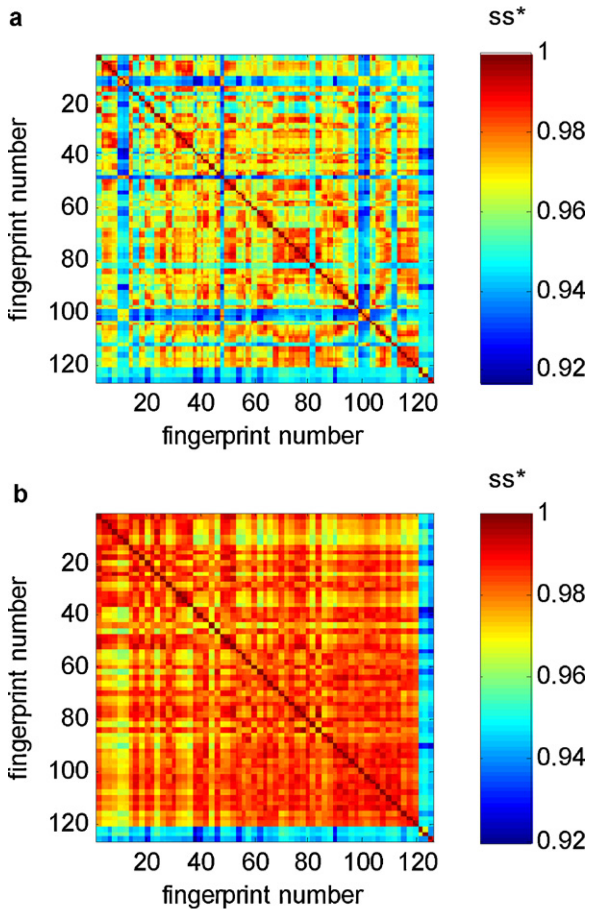

Fig. 6. (a and b) Colour maps of the adapted similarity scores *ss** of the original data set **XOa** (a) and its warped matrix **XWa** (b). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

4.4. Evaluation of the benefit of dissimilar fingerprints

As data sets **Xa** and **Xb** are built up by the fingerprints obtained on two dissimilar chromatographic systems, it allows to check whether such fingerprints provide complementary information about sample similarity. Differences in parameter ranges are observed between the genuine green-tea fingerprints in **Xa** and **Xb**. For instance, *r* ranges between 0.877 and 0.999 for **Xa** and between 0.845 and 0.999 for **Xb**. Differences are also seen for *de* or *ss**. Despite those range differences, the SA parameters *r*, *c*, *de*, *ds*, *dm* and *ss** resulted in similar colour maps and evaluation plots for both data sets. The fingerprints of a sample, obtained at both dissimilar chromatographic systems, provide an equivalent similarity evaluation in this study.

Furthermore, it is evaluated whether the inclusion of more chromatographic information in the SA, *i.e.* concatenating both fingerprints of **Xa** and **Xb** into a data set **Xab**, provides better identification or quality control results. If a fingerprint contains more useful information, one expects the result of the similarity analysis to be more decisive. However, the colour maps and evaluation plots from **Xab** were nearly the same as those from the individual data sets **Xa** or **Xb** and thus did not provide extra information. In addition, the calculation time for **Xab** was higher. However, these observations about the use of combined dissimilar fingerprints should be confirmed from other case studies.

5. Conclusion

A visual evaluation can discriminate between chromatographic fingerprint profiles when differences are large and the number of samples is limited. However, if differences are more subtle, mathematical approaches are necessary. This study compared different SA parameters as a tool for identification and quality control of herbal samples.

If no official reference material of the herbal product is available, the use of data sets with a sufficient number of genuine samples is important in SA. The use of such libraries, instead of one reference sample, showed its usefulness. For a proper data interpretation, alignment of the chromatograms is a required data pre-treatment step.

The colour maps and evaluation plots showed to be useful tools to distinguish between deviating and genuine samples. The evaluation plots with the statistical confidence limits are efficient tools to evaluate suspected samples during quality assurance. They can be constructed for different parameters.

The correlation *r* and congruence *c* coefficients provided very similar information. The standardised Euclidean distances *ds* provided less information than the Euclidean distances *de*; while the Mahalanobis distance *dm* was not useful for SA. The adapted similarity score *ss** combines the advantages of the correlation coefficients and the Euclidean distances. SA based on correlation is useful if concentration differences between samples are not important, whereas similarity analysis based on distances also detects concentration differences well. The *ss** colour maps and evaluation plots are therefore valuable tools to distinguish between deviating and genuine fingerprints.

Finally, the combination of the dissimilar chromatograms of each sample did not reveal extra information concerning the (dis)similarity of the considered sample.

Acknowledgements

S. Pieters and J. Van Erps thank the Fund for Scientific Research (FWO, Vlaanderen, Belgium) for financial support.

References

[1] P. Zou, Y. Hong, H.L. Koh, J. Pharm. Biomed. Anal. 38 (2005) 514.
[2] L.-W. Yang, D.-H. Wu, X. Tang, W. Peng, X.-R. Wang, Y. Ma, W.-W. Su, J. Chromatogr. A 1070 (2005) 35.
[3] C. Aschwanden, Bull. World Health Org. 79 (2001) 691.
[4] D. Bensky, S. Clavey, E. Stöger, Chinese Herbal Medicine I: Materia Medica, 3rd ed., Eastland Press, Seattle, USA, 2004, pp. 158–160, 178–181.
[5] J. Park, H.-J. Park, H.-J. Lee, E. Ernst, Am. J. Chin. Med. 30 (2002) 419.
[6] P. Drasar, J. Moravcova, J. Chromatogr. B 812 (2004) 3.
[7] The European Pharmacopoeia, 7th ed., Council of Europe, Strasbourg, France, 2011, http://online.pheur.org/entry.htm (accessed 27.12.11).
[8] The United States Pharmacopeia 28, The National Formulary 23, The United States Pharmacopeial Convention, Rockville, MD, 2005.
[9] Pharmacopoeia of the People's Republic of China, vol. 1, Chinese Pharmacopoeia Commission, Peoples Medical Publishing House, China, 2005.
[10] G. Alaerts, B. Dejaegher, J. Smeyers-Verbeke, Y. Vander Heyden, Comb. Chem. High Throughput Screen. 13 (2010) 900.
[11] G.-H. Lu, K. Chan, Y.-Z. Liang, K. Leung, C.-L. Chan, Z.-H. Jiang, Z.-Z. Zhao, J. Chromatogr. A 1073 (2005) 383.
[12] Y.-B. Ji, Q.-S. Xu, Y.-Z. Hu, Y. Vander Heyden, J. Chromatogr. A 1066 (2005) 97.
[13] L. Zhao, C. Huang, Z. Shan, B. Xiang, L. Mei, J. Chromatogr. B 821 (2005) 67.
[14] Y. Liang, P. Xie, F. Chau, J. Sep. Sci. 33 (2010) 410.
[15] G. Alaerts, N. Matthijs, J. Smeyers-Verbeke, Y. Vander Heyden, J. Chromatogr. A 1172 (2007) 1.
[16] B. Dejaegher, G. Alaerts, N. Matthijs, Acta Chromatogr. 22 (2010) 237.
[17] S. Pieters, C. Tistaert, G. Alaerts, K. Bodzioch, D. Mangelings, B. Dejaegher, C. Rivière, N. Nguyen Hoai, M. Chau Van, J. Quetin-Leclerq, Y. Vander Heyden, Talanta 83 (2011) 1188.
[18] G. Alaerts, M. Merino-Arévalo, M. Dumarey, B. Dejaegher, N. Noppe, N. Matthijs, J. Smeyers-Verbeke, Y. Vander Heyden, J. Chromatogr. A 1217 (2010) 7706.
[19] Y.-B. Ji, G. Alaerts, C.-J. Xu, Y.-Z. Hu, Y. Vander Heyden, J. Chromatogr. A 1128 (2006) 273.
[20] Guideline on specifications: test procedures and acceptance criteria for herbal substances, herbal preparations and herbal medicinal products/traditional herbal medicinal products, Committee for medicinal products for human use (CHMP), European Medicines Agency Inspections, 30 March 2006, CPMP/QWP/2820/00 Rev 1, EMEA/CVMP/815/00 Rev 1. http://www.emea.europa.eu/pdfs/human/qwp/282000en.pdf (accessed 27.12.11).
[21] Guidance for Industry: Botanical Drug Products, U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), 2004, June, p. 10. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm070491.pdf (accessed 27.12.11).
[22] Status Quo of Drug Supervision in China, Information Office of the State Council of the People's Republic of China, State Food and Drug Administration, Beijing, China, 2008, http://www.gov.cn/english/2008-07/18/content_1049011.htm (accessed 27.12.11).
[23] WHO Traditional Medicine Strategy 2002–2005, WHO/EDM/TRM/2002.1, World Health Organization, Geneva, Switzerland, 2002, http://whqlibdoc.who.int/hq/2002/WHO_EDM_TRM_2002.1.pdf (accessed 27.12.11).
[24] C. Tistaert, B. Dejaegher, Y. Vander Heyden, Anal. Chim. Acta 690 (2011) 148.
[25] S.-K. Wong, S.-K. Tsui, S.-Y. Kwan, X.-L. Su, R.-C. Lin, L.-M. Tang, J.-X. Chen, J. Food Drug Anal. 12 (2004) 110.
[26] S.-L. Li, J.-Z. Song, F.F.K. Choi, C.-F. Qiao, Y. Zhou, Q.-B. Han, H.-X. Xu, J. Pharm. Biomed. Anal. 49 (2009) 253.
[27] W.-J. Kong, Y.-L. Zhao, X.-H. Xiao, J.-B. Wang, H.-B. Li, Z.-L. Li, C. Jin, Y. Liu, Anal. Chim. Acta 634 (2009) 279.
[28] W.-J. Kong, Y.-L. Zhao, X.-H. Xiao, C. Jin, Z.-L. Li, Phytomedicine 16 (2009) 950.
[29] L.-Z. Yi, D.-L. Yuan, Y.-Z. Liang, P.-S. Xie, Y. Zhao, Anal. Chim. Acta 588 (2007) 207.
[30] Y.-Y. Zhao, Y. Zhang, R.-C. Lin, W.-J. Sun, Fitoterapia 80 (2009) 333.
[31] Y. Chen, S.-B. Zhu, M.-Y. Xie, S.-P. Nie, W. Liu, C. Li, X.-F. Gong, Y.-X. Wang, Anal. Chim. Acta 623 (2008) 146.
[32] A. Gledhill, The Column 4 (2008) 19.
[33] Z.-D. Zeng, Y.-Z. Liang, T. Zhang, F.-T. Chau, Y.-L. Wang, Anal. Bioanal. Chem. 385 (2006) 392.
[34] Y. Li, T. Wu, J. Zhu, L. Wan, Q. Yu, X. Li, Z. Cheng, C. Guo, J. Pharm. Biomed. Anal. 52 (2010) 597.
[35] L. Chen, Y.P. Tang, M.J. Chen, H.S. Deng, X.P. Yan, D.K. Wu, Phytomedicine 17 (2010) 100.
[36] G.X. Xie, M.F. Qiu, A.H. Zhao, W. Jia, Chromatographia 64 (2006) 739.
[37] X.-H. Fan, Y.-Y. Cheng, Z.-L. Ye, R.-C. Lin, Z.-Z. Qian, Anal. Chim. Acta 555 (2006) 217.
[38] F. Gong, B.-T. Wang, F.-T. Chau, Y.Z. Liang, Anal. Lett. 38 (2005) 2475.
[39] Y. Chen, W. Bicker, J.Y. Wu, M.Y. Xie, W. Lindner, J. Chromatogr. A 1217 (2010) 1255.
[40] H. Wei, L. Sun, Z. Tai, S. Gao, W. Xu, W. Chen, Anal. Chim. Acta 662 (2010) 97.
[41] F.-Q. Guo, Y.-Z. Liang, C.-J. Xu, L.-F. Huang, X.-N. Li, J. Chromatogr. A 1054 (2004) 73.
[42] H.L. Zhai, F.D. Hu, X.Y. Huang, J.H. Chen, Anal. Chim. Acta 657 (2010) 131.
[43] A.M. van Nederkassel, M. Daszykowski, D.L. Massart, Y. Vander Heyden, J. Chromatogr. A 1096 (2005) 177.
[44] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics B, Data Handling in Science and Technology 20B, Elsevier, Amsterdam, 1998.
[45] I. Borg, P.J.F. Groenen, Modern Multidimensional Scaling Theory and Applications, 2nd ed., Springer Science + Business Media, New York, 2005, pp. 111–133, 389–409, 429–447.
[46] Y. Xie, Z.-H. Jiang, H. Zhou, X. Cai, Y.-F. Wong, Z.-Q. Liu, Z.-X. Bian, H.-X. Xu, L. Liu, J. Pharm. Biomed. Anal. 43 (2007) 204.
[47] V.M. Arlt, M. Stiborova, H.H. Schmeiser, Mutagenesis 17 (2002) 265.
[48] K.-T. Fang, Y.-Z. Liang, X.-L. Yin, K. Chan, G.-H. Lu, Chemom. Intell. Lab. Syst. 82 (2006) 236.
[49] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics A, Data Handling in Science and Technology 20A, Elsevier, Amsterdam, 1997.
[50] M. Ganzera, Planta Med. 75 (2009) 776.
[51] Y.-Z. Liang, P. Xie, K. Chan, J. Chromatogr. B 812 (2004) 53.
[52] Manual Praat program doing phonetics by computer, Phonetic Sciences, University of Amsterdam, updated 22 March 2011. http://www.fon.hum.uva.nl/praat/ (accessed 27.12.11).
[53] A.M. van Nederkassel, V. Vijverman, D.L. Massart, Y. Vander Heyden, J. Chromatogr. A 1085 (2005) 230.
[54] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, Chemom. Intell. Lab. Syst. 50 (2000) 1.
[55] T. Neely, B. Walsh-Mason, P. Russell, A. Van Der Horst, S. O'Hagan, P. Lahorkar, Toxicol. Int. 18 (2011) 20.
[56] M. Dumarey, I. Smets, Y. Vander Heyden, J. Chromatogr. B 878 (2010) 2733.
[57] N.-P. Vest Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A 805 (1998) 17.
[58] M. Thompson, R. Wood, Pure Appl. Chem. 67 (1995) 649.