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.

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

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

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:

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

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