Journal of Chromatography B, 910 (2012) 114–121

Contents lists available at SciVerse ScienceDirect

Journal of Chromatography B

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

Potentially antioxidant compounds indicated from *Mallotus* and *Phyllanthus* species fingerprints☆

S. Thiangthum[a,b], B. Dejaegher[a], M. Goodarzi[a], C. Tistaert[a], A.Y. Gordien[c], N. Nguyen Hoai[d], M. Chau Van[d], J. Quetin-Leclercq[c], L. Suntornsuk[b], Y. Vander Heyden[a,*]

[a] *Department of Analytical Chemistry and Pharmaceutical Technology, Center for Pharmaceutical Research (CePhaR), Vrije Universiteit Brussel (VUB), Laarbeeklaan 103, B-1090 Brussels, Belgium*
[b] *Department of Pharmaceutical Chemistry, Faculty of Pharmacy, Mahidol University, 447 Sri-Ayudhaya Road, Rajathevee, Bangkok 10400, Thailand*
[c] *Analytical Chemistry, Drug Analysis and Pharmacognosy Unit, Louvain Drug Research Institute, Université Catholique de Louvain (UCL), Avenue E. Mounier 72, B-1200 Brussels, Belgium*
[d] *Institute of Marine Biochemistry, Vietnam Academy of Science and Technology (VAST), 18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam*

ARTICLE INFO

Article history:
Received 18 February 2012
Accepted 21 June 2012
Available online 28 June 2012

Keywords:
Chromatographic fingerprints
Antioxidant activity
Multivariate calibration
Mallotus species
Phyllanthus species

ABSTRACT

The genera of *Mallotus* and *Phyllanthus* contain several species that are commonly used as traditional medicines in oriental countries. Some species show interesting pharmaceutical activities, such as an antioxidant activity. To produce clinically useful medicines or food supplements (nutraceuticals) from these herbs, the species should be identified and a thorough quality control should be implemented. Nowadays, the integration of chromatographic and chemometric approaches allows a high-throughput identification and activity prediction of medicinal plants. In this study, Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA) were applied and compared to distinguish *Mallotus* and *Phyllanthus* species. Moreover, peaks from their chromatographic fingerprints, which were responsible for their antioxidant activity were assigned. For the latter purpose, the relevant information was extracted from the chromatographic fingerprints using linear multivariate calibration techniques, i.e., Partial Least Squares (PLS) and Orthogonal Projections to Latent Structures (O-PLS). Results reveal that exploratory analysis using PCA shows somewhat diverging clustering tendencies between *Mallotus* and *Phyllanthus* samples than HCA. However, both approaches mainly confirm each other. Concerning the multivariate calibration techniques, both PLS and O-PLS models demonstrate good predictive abilities. By comparing the regression coefficients of the models with the chromatographic fingerprints, the peaks that are potentially responsible for the antioxidant activity of the extracts could be confirmed.

1. Introduction

Herbal medicines have been used by many people around the world for thousands of years. Unfortunately, their quality, safety, and efficacy are not always sufficiently evaluated [1,2]. Because of their complex composition, the development of a suitable analytical procedure to separate all or as many compounds as possible from such herbal sample is a challenging task. Furthermore, the concentrations of the herbal components can vary significantly depending on the cultivation conditions of the plant, the drying process, and the harvest season [2]. Some researchers only use information from one or a limited number of compounds, the so-called markers, to evaluate the quality of herbal medicines. However, it seems doubtful that only focusing on some compounds will describe and evaluate the complexity of the herbal sample properly.

The World Health Organization (WHO) has accepted chromatographic fingerprint analysis as a strategy for the assessment of herbal medicines [3]. A chromatographic fingerprint can be obtained by, for example, reversed-phase high-performance liquid chromatography (RPLC), and typifies the complete composition of a herbal medicine. A fingerprint represents a chromatographic profile in which the detectable chemical constituents are separated as much as possible. The obtained fingerprints can be used as a unique identification tool to evaluate the authenticity of a herbal sample, the quality and assurance of the consistency, and the stability of a herbal medicine. Nowadays, the combination of (hyphenated) chromatographic instruments and chemometrical approaches for data (pre-) treatment allows a fast investigation of herbal samples [4–11]. Moreover, chemometric treatment of

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

* Corresponding author. 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.2012.06.025

the chromatographic fingerprints also allows modeling and predicting pharmacological activities (e.g. antioxidant and cytotoxic activities) and/or indicating peaks potentially responsible for the modeled activities [2,4,12–16].

The *Mallotus* and *Phyllanthus* genera, belonging to the family of the *Euphorbiaceae*, are widely distributed in most tropical and sub-tropical countries. They have been extensively used in folk medicine in India, China, Vietnam and other countries for thousands of years for the treatment of a broad spectrum of diseases. For instance, the genus *Mallotus* has been used for the treatment of chronic hepatitis and enteritis [17,18], while the genus *Phyllanthus* has been used for kidney, urinary bladder, and intestinal infections, and for diabetes [19,20]. Recently, many studies were performed concerning the chemical components of *Mallotus* and *Phyllanthus* species and several pharmacologically active constituents were determined [12,21–25]. Cytotoxic (*Mallotus apelta* [22]), antimicrobial (*Mallotus peltalus* [22], *Phyllanthus emblica* [24]), anti-inflammatory (*Mallotus peltalus* [22], *Mallotus spodocarpys* [22]), and antioxidant activities (*Mallotus metcalfianus* [23], *Phyllanthus emblica* [21], *Phyllanthus niruri* [25]) have been reported for both genera.

In a parallel study [26], RPLC fingerprints of 36 samples, i.e., 10 *Mallotus* and 26 *Phyllanthus* samples, were developed. Then, unsupervised and supervised classification methods were used to classify the *Mallotus* and *Phyllanthus* samples according to genera (*Mallotus* and *Phyllanthus*) and species (*Mallotus apelta*, *Mallotus paniculatus, Phyllanthus emblica*, *Phyllanthus reticulatus*, *Phyllanthus urinaria* L., *Phyllanthus amarus*). As unsupervised techniques, Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA) were evaluated, while as supervised classification techniques, Linear Discriminant Analysis, Quadratic Discriminant Analysis, and Classification and Regression Trees, were compared. The methods were applied to classify the samples in two classes (i.e. the two genera) or in 6 classes (i.e. the 6 species). Results showed the applicability of both unsupervised and supervised methods to discriminate between the samples.

The goal of this study was to model the antioxidant activity of the 36 *Mallotus* and *Phyllantus* samples, originating from different genera, species, origins and/or collection times, as a function of their chromatographic fingerprints. The goal of this modeling is not to use the model to predict the activity of future samples but to indicate peaks potentially responsible for the antioxidant activity [16]. The antioxidant activity of the herbal extracts was determined with a 1,1-diphenyl-2-picrylhydrazyl (DPPH) radical scavenging activity test and expressed in term of the median inhibition concentration (IC_{50}). First, an unsupervised data analysis using PCA and HCA was performed to verify whether the antioxidant samples could be distinguished from the less or non-active samples. Then, the antioxidant activity was modeled as a function of the earlier developed HPLC fingerprints [26] using the multivariate calibration techniques Partial Least Squares (PLS) [27] and Orthogonal Projections to Latent Structures (O-PLS) [28]. The regression coefficients of the resulting models were evaluated to indicate the peaks possibly responsible for the antioxidant activity.

2. Theory

2.1. Data preprocessing

Prior to data analysis, the chromatographic fingerprints are organized in an $n \times p$ data matrix **X**, where the n objects (herbal samples) constitute the rows and the p variables (time points) the columns. In each cell of the matrix, the detector signal intensity at a given time point is presented.

The results of the data analysis are influenced by the applied data preprocessing method. In this study, different methods to pretreat the data, i.e. column centering, normalization followed by column centering, and standard normal variate (SNV) followed by column centering, were applied and compared [4,27,29,30].

Column centering removes the column mean from each corresponding column element. Normalization of the chromatograms scales the rows to a constant total by dividing each row by its corresponding norm. SNV corresponds to row centering, followed by row scaling, where row centering removes the row mean from each corresponding row element and row scaling divides each row element by its corresponding row standard deviation.

2.2. Unsupervised exploratory data analysis

Unsupervised data analysis only uses information contained in the data matrix **X**, i.e. the fingerprints, and does not use the information contained in the response vector **y**.

2.2.1. Principal Component Analysis

Principal Component Analysis (PCA) reduces the number of variables and visualizes the information included in the $n \times p$ data matrix **X** [27,30]. PCA makes linear combinations of the original variables, thus creating the so-called latent variables or principal components (PCs), in such a way that the latter describe the largest possible remaining variation in **X** and are orthogonal. The projection of an object on a PC is called a score on this PC, while the projection of each original variable to the PC is called a loading. A score plot represents the scores on two PC's and shows information regarding the (dis)similarity of the objects, while a loading plot provides information on the contribution of the original variables to the considered PC's.

2.2.2. Hierarchical Cluster Analysis

Hierarchical Cluster Analysis (HCA) is a clustering method applied to reveal the underlying structure of objects through an iterative process that associates (agglomerative method) or dissociates (divisive method) the data set object by object, and that is stopped when all objects have been processed [30,31].

A divisive method starts with all objects in one cluster and divides them into subsets, continuously making smaller clusters until all objects are individually in a cluster [30,32]. An agglomerative procedure, on the other hand, starts with each object in a separate cluster and combines the clusters sequentially, reducing the number of clusters at each step until all objects belong to only one cluster [30]. The hierarchical clustering process can be represented as a tree or dendrogram, where each step in the clustering process is illustrated by a joint of clusters.

In this study, an agglomerative HCA was selected to visualize the data contained in the fingerprint matrix **X** and give insight in the clustering tendency of the data. Several (dis)similarity measures to cluster the objects can be used, e.g. Euclidean distance, Mahalanobis distance, Pearson correlation distance, and Spearman's rank correlation coefficient [30].

In this work, the Euclidean distance and the Pearson correlation distance, which is calculated as "1 − Pearson correlation coefficient r", were evaluated as distance measures [30]. Consider $\mathbf{x}$ ($x_1, x_2, \ldots, x_n$) and $\mathbf{z}$ ($z_1, z_2, \ldots, z_n$) as two sets of n measurements/variables (fingerprints) with means (averages) $\bar{x}$ and $\bar{z}$. The Euclidean distance (ED) between $\mathbf{x}$ and $\mathbf{z}$ is calculated as follows

$$\mathrm{ED} = \sqrt{\sum_{i=1}^{n} (x_i - z_i)^2} \quad (1)$$

where x_i and z_i are the ith elements of $\mathbf{x}$ and $\mathbf{z}$.

The Pearson correlation coefficient r is estimated as

$$r(x,z) = \frac{cov(x,z)}{s_x s_z} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(z_i - \bar{z})^2}} \quad (2)$$

where cov(x,z) is the covariance of the variables **x** and **z**, and s_x and s_z are the standard deviations of **x** and **z**.

There are several linkage methods using different criteria to decide which individual objects should be merged, i.e. single linkage, complete linkage, average linkage, weighted average linkage, centroid's method, median's method, and Ward's method. In this study, all linkage methods were applied and evaluated. In all methods, in a first step, the two objects, o_1 and o_2, representing fingerprints, that are most similar (with the smallest distance between them, $D_{o_1o_2}$) are clustered to a new combined object, o^*, which means that the two objects with the smallest distance between them, $D_{o_1o_2}$, need to be found. The distance of the new combined object o^* to the remaining objects i, D_{io*}, or to another cluster in a further stage of the iterative procedure, D_{i*o*}, then can be obtained in different ways.

Single linkage, also called nearest neighbor linkage, then uses the smallest distance between the two clustered objects, o_1 and o_2, and the other remaining objects i, i.e. $D_{io*} = \min(D_{io_1}, D_{io_2})$. In case of another cluster i^*, the distance is considered to be equal to the smallest distance between two individual objects, of which one belongs to cluster o^* and one to cluster i^*.

Complete linkage follows the opposite approach of single linkage. It uses the largest distance between the two clustered objects, o_1 and o_2, and the other remaining objects i, i.e. $D_{io*} = \max(D_{io_1}, D_{io_2})$. In case of another cluster i^*, the distance is considered to be equal to the largest distance between two individual objects, of which one belongs to cluster o^* and one to cluster i^*.

In average linkage, the average distance between the two clustered objects, o_1 and o_2, and the other remaining objects i, i.e. $D_{io*} = (D_{io_1} + D_{io_2})/2$, is then used. In case of another cluster i^*, the distance is considered to be equal to the average distance between both clusters. Both unweighted and weighted average linkage procedures exist [30].

The centroid's and the median's methods use the distance between the centroids and weighted centroids of the two clustered objects, o_1 and o_2, and the other remaining objects i or cluster i^*, respectively.

Another approach is Ward's method, which uses an a posteriori heterogeneity criterion. This criterion is defined as the sum of the squared distances of each member of a cluster to the centroid of that cluster. Elements or clusters are joined with as criterion that the sum of heterogeneities of all clusters should increase as little as possible.

2.3. Supervised data analysis: multivariate calibration techniques

Linear multivariate calibration techniques try to relate the information contained in the $n \times p$ data matrix **X** to an $n \times 1$ response vector **y**, with **y** being a continuous response. For this purpose, several techniques have been proposed [28,30,33].

Generally, the relationship between **X** and **y** can be described as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \quad (3)$$

where **b** represents a $p \times 1$ vector of regression coefficients and **e** an $n \times 1$ residual vector. In this study, two linear multivariate calibration techniques, i.e. Partial Least Squares (PLS) and Orthogonal Projections to Latent Structures (O-PLS), are used for modeling and their regression coefficients were studied in order to indicate peaks possibly responsible for the antioxidant activity.

2.3.1. Partial Least Squares

Partial Least Squares (PLS) [27,30] is a latent-variable technique that maximizes the covariance between **X** and **y**. The PLS model can be presented as follows

$$\mathbf{X} = \mathbf{TP}^{\mathrm{T}} + \mathbf{E} \quad (4)$$

$$\mathbf{y} = \mathbf{TP}^{\mathrm{T}}\mathbf{b} + \mathbf{f} = \mathbf{Tq} + \mathbf{f} \quad (5)$$

$$\mathbf{b} = \mathbf{Pq} \quad (6)$$

where **T** represents the $n \times n$ score matrix for **X** and **y**, **P** the $p \times n$ loading matrix of **X** on **T**, **E** the $n \times p$ residual matrix of **X**, **b** the $p \times 1$ vector of regression coefficients, **q** the $n \times 1$ loading vector of **y** on **T**, and **f** the $n \times 1$ residual vector of **y**. The regression coefficients **b** can be used to evaluate the contribution of the original variables to the final model [2,4,13–16].

In our study, the optimal model complexity was determined by a leave-one-out cross-validation procedure (LOO-CV). During LOO-CV each object is left out once and the model is built using the remaining objects. The root mean squared error of cross-validation (RMSECV) (Eq. (7)) is then calculated for models with different complexities [30].

$$\mathrm{RMSECV}(f) = \sqrt{\sum_{i=1}^{N} \frac{(\hat{y}_{cv,i} - y_i)^2}{N}} \quad (7)$$

where f is the model complexity, N the number of calibration samples, y_i the measured response of the ith sample, and $\hat{y}_{cv,i}$ the response for the ith sample predicted from the calibration model obtained without the ith sample. The optimal model complexity corresponds to the number of latent factors resulting in the (nearly) lowest RMSECV.

2.3.2. Orthogonal Projections to Latent Structures

Orthogonal Projections to Latent Structures (O-PLS) [28] remove the variation in **X** that is not correlated to **y**. This is done by subtracting PLS components, orthogonal to **y**, from the original data matrix **X**. Thus, the original data is split into two data sets, i.e. one that contains the information relevant to **y** and another with the information orthogonal to **y**.

An O-PLS model can be written as follows

$$\mathbf{X} = \mathbf{TP}^{\mathrm{T}} + \mathbf{T}_{\mathrm{orth}}\mathbf{P}_{\mathrm{orth}}^{\mathrm{T}} + \mathbf{E} \quad (8)$$

where **T** represents the orthonormal $n \times n$ score matrix for **X** and **y**, **P** the orthonormal $p \times n$ loading matrix representing the regression coefficients of **X** on **T**, $\mathbf{T}_{\mathrm{orth}}$ the orthogonal $n \times n$ score matrix for **X** and **y**, $\mathbf{P}_{\mathrm{orth}}$ its corresponding orthogonal $p \times n$ loading matrix, and **E** the $n \times p$ residual matrix of **X**. **y** and **b** are calculated according to Eqs. (5) and (6), respectively. Again, the regression coefficients **b** (Eq. (6)) can be used to evaluate the contribution of the original variables to the final model [2,4,12,13,16].

Removing the orthogonal information from the original data set leads to a reduction of the number of PLS components in the O-PLS model, i.e. only one PLS component is used. This allows an improved interpretability of the regression coefficients [28].

3. Experimental

3.1. Herbs and preparation of the extracts

Ten *Mallotus* and twenty-six *Phyllanthus* samples, from 6 different species, were collected in different Vietnamese regions (Table 1). To protect the forests, only the leaves were collected. The samples were authenticated by Professor Nguyen Nghia Thin (Hanoi National University, Vietnam).

Table 1
The *Mallotus* and *Phyllanthus* samples with their species name, collection time, origin, and the IC_{50}. / = not specified and CI = confidence interval. The highly antioxidant samples are marked in bold.

Sample	Species	Collection time	Origin	Average IC_{50} (95%CI)	
1	*Mallotus apelta*	August, 2009	Van Ban-Lao Cai	72.66	(51.36–101.20)
2	***Phyllanthus emblica***	**August, 2009**	**Van Ban-Lao Cai**	**9.90**	**(7.61–13.89)**
3	***Phyllanthus emblica***	**November, 2009**	**Dong Dang-Lang Son**	**9.55**	**(7.54–12.58)**
4	*Mallotus apelta*	November, 2009	Tam Dao-Vinh Phuc	74.59	(54.91–102.40)
5	*Mallotus apelta*	November, 2009	Dong Dang-Lang Son	98.42	(65.89–137.60)
6	*Mallotus paniculatus*	December, 2009	Dong Van-Ha Giang	37.31	(27.84–55.08)
7	***Phyllanthus emblica***	**December, 2009**	**Dong Van-Ha Giang**	**8.57**	**(6.95–10.57)**
8	*Mallotus apelta*	December, 2009	Ham Yen-Tuyen Quang	110.00	(/)
9	***Phyllanthus reticulatus***	**February, 2010**	**Nghia Trai-Hung Yen**	**28.06**	**(19.25–38.18)**
10	***Phyllanthus urinaria* L.**	**February, 2010**	**Nghia Trai-Hung Yen**	**9.75**	**(7.36–12.43)**
11	*Phyllanthus amarus*	February, 2010	Nghia Trai-Hung Yen	99.32	(29.23–142.50)
12	***Phyllanthus amarus***	**March, 2010**	**Van Dien-Hanoi**	**13.74**	**(9.64–19.08)**
13	*Mallotus paniculatus*	March, 2010	Huong Hoa-Quang Tri	38.87	(31.07–52.68)
14	***Phyllanthus reticulatus***	**March, 2010**	**Van Dien-Hanoi**	**29.70**	**(21.74–45.66)**
15	***Phyllanthus emblica***	**March, 2010**	**Huong Hoa-Quang Tri**	**10.92**	**(7.61–14.79)**
16	***Phyllanthus urinaria* L.**	**March, 2010**	**Van Dien-Hanoi**	**8.45**	**(6.03–11.46)**
17	*Mallotus paniculatus*	April, 2010	Me Linh-Vinh Phuc	35.76	(27.30–48.59)
18	***Phyllanthus reticulatus***	**April, 2010**	**Me Linh-Vinh Phuc**	**13.42**	**(10.51–17.28)**
19	***Phyllanthus amarus***	**April, 2010**	**Me Linh-Vinh Phuc**	**13.12**	**(10.77–15.87)**
20	***Phyllanthus reticulatus***	**April, 2010**	**Lan Ong-Hanoi**	**22.18**	**(13.47–33.60)**
21	***Phyllanthus emblica***	**April, 2010**	**Me Linh-Vinh Phuc**	**8.75**	**(6.99–10.85)**
22	***Phyllanthus amarus***	**April, 2010**	**Lan Ong-Hanoi**	**14.33**	**(9.90–20.69)**
23	***Phyllanthus urinaria* L.**	**April, 2010**	**Me Linh-Vinh Phuc**	**8.83**	**(7.25–10.70)**
24	***Phyllanthus urinaria* L.**	**April, 2010**	**Lan Ong-Hanoi**	**8.56**	**(6.95–10.49**
25	***Phyllanthus reticulatus***	**May, 2010**	**Ninh Hiep-Hanoi**	**23.49**	**(15.23–33.48)**
26	***Phyllanthus reticulatus***	**May, 2010**	**Dong Anh-Hanoi**	**27.31**	**(13.42–42.45)**
27	***Phyllanthus amarus***	**May, 2010**	**Que Vo-Bac Ninh**	**15.51**	**(11.14–21.22)**
28	***Phyllanthus emblica***	**May, 2010**	**Dong Anh-Hanoi**	**10.26**	**(8.32–12.73)**
29	***Phyllanthus amarus***	**May, 2010**	**Dong Anh-Hanoi**	**10.15**	**(6.91–14.95)**
30	***Phyllanthus urinaria* L.**	**May, 2010**	**Que Vo-Bac Ninh**	**17.99**	**(11.95–19.88)**
31	***Phyllanthus urinaria* L.**	**May, 2010**	**Ninh Hiep-Hanoi**	**21.13**	**(13.71–28.78)**
32	***Phyllanthus amarus***	**May, 2010**	**Ninh Hiep-Hanoi**	**16.12**	**(11.62–21.39)**
33	***Phyllanthus urinaria* L.**	**May, 2010**	**Dong Anh-Hanoi**	**13.60**	**(8.35–19.86)**
34	*Mallotus apelta*	June, 2010	Pa Co-Hoa Binh	105.56	(70.08–151.70)
35	*Mallotus paniculatus*	June, 2010	VQG-Pumat	31.47	(23.72–42.63)
36	*Mallotus paniculatus*	July, 2010	Cuc phuong-Ninh Binh	36.70	(28.85–49.01)

Extracts were prepared by weighing 10.0 g of sample which is then extracted with three times 100 mL methanol in an ultrasonic bath (Branson Ultrasonic Corporation, Connecticut, USA) at a temperature between 30 and 45 °C during 1 h. Afterwards, the extract was filtered through a 240 nm pore size filter paper (Whatman, Hanoi, Vietnam) and evaporated at reduced pressure (60 Pa) and elevated temperature (40 °C). The obtained crude extract was divided into 3 fractions, i.e. one for the DPPH radical scavenging assay, one for the HPLC analysis, and one as a library sample for reference purposes in Vietnam.

3.2. HPLC

3.2.1. Equipment, chemicals and reagents

The experiments were performed on a Shimadzu Prominence HPLC system (Shimadzu, Tokyo, Japan), equipped with an autosampler, a vacuum degasser, a quaternary pump, a column oven and a photo diode array detector. All data were acquired and processed by LC solution software (Shimadzu). Two coupled Chromolith™ Performance RP-18e columns (each 100 mm × 4.6 mm I.D.) with a Chromolith™ Performance RP-18e guard column (5 mm × 4.6 mm I.D.) were used as stationary phase. HPLC grade acetonitrile (ACN) (Fisher Scientific, Leicestershire, UK), trifluoroacetic acid (TFA) (Sigma–Aldrich, Steinheim, Germany) and ultra pure water, obtained from the Arium® pro UV Ultrapure Water system (Sartorius Stedim Biotech, Aubagne, France), were used to prepare the mobile phases. All solvents were degassed for 15 min on an Ultrasonic bath (Branson Ultrasonic Corporation, Connecticut, USA) prior to HPLC analysis.

3.2.2. Sample preparation

The development of the sample preparation method was based on the methodology described in [34]. To prepare the samples for HPLC analysis, 50.0 mg crude extract was weighed and diluted to volume with methanol in a 2.0 mL volumetric flask. Then, the solution was mixed on a shaking bath (Edmund Bühler, Hechingen, Germany) during 15 min at 400 rpm. This was followed by filtration through a filter (Schleicher & Schuell, Dassel, Germany) with a diameter of 125 mm and a pore size smaller than 2 μm and consecutively by filtration through a 25 mm syringe filter having a polypropylene membrane with 0.2 μm pore size (VWR International, Leuven, Belgium).

3.2.3. Chromatographic conditions

The development of the chromatographic fingerprints was based on the methodology described in [34]. The mobile phase consisted of (A) 0.05%TFA in ACN, and (B) 0.05%TFA in ultra pure water. Gradient elution was applied for 60 min. The gradient was 5–20% A in 0–20 min, 20–95% A in 20–50 min and 95% A during 50–60 min. The column temperature was 25 °C, the flow rate 1.0 mL min^{-1}, the injection volume 10 μL, and the detection wavelength 254 nm.

3.3. DPPH radical scavenging assay

The antioxidant activity of the samples was determined using the DPPH radical scavenging assay. The DPPH radical scavenging assay [35] measures the capacity of a compound or a sample to scavenge the stable 1,1-diphenyl-2-picrylhydrazyl radical (DPPH•). DPPH• has an absorption band at 515 nm, which disappears upon reduction by an antiradical compound. The assay was carried out

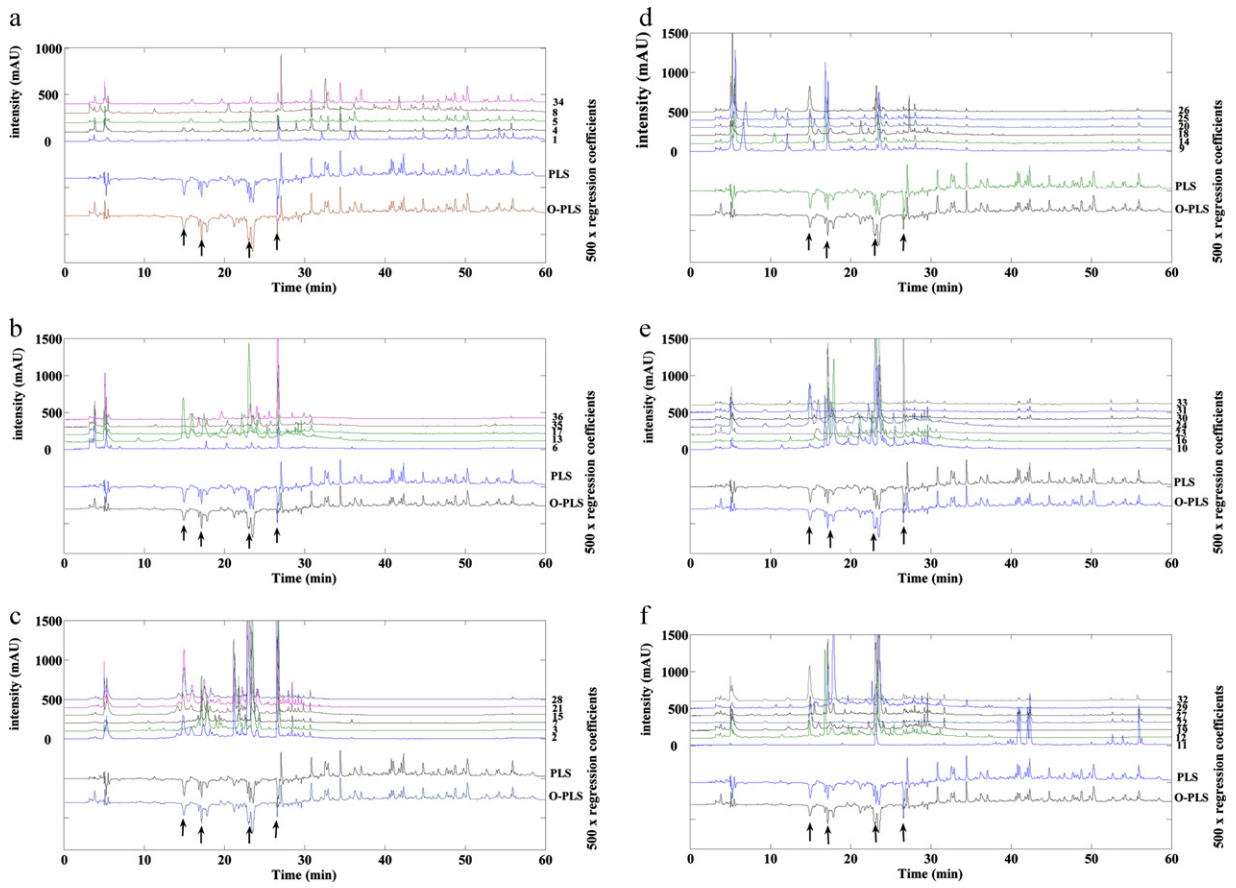

Fig. 1. 60 min chromatographic fingerprints of (a) *Mallotus apelta* samples (1, 4, 5, 8, 34), (b) *Mallotus paniculatus* samples (6, 13, 17, 35, 36), (c) *Phyllanthus emblica* samples (2, 3, 7, 15, 21, 28), (d) *Phyllanthus reticulatus* samples (9, 14, 18, 20, 25, 26), (e) *Phyllanthus urinaria* L. samples (10, 16, 23, 24, 30, 31, 33) and (f) *Phyllanthus amarus* samples (11, 12, 19, 22, 27, 29, 32). The regression coefficients from PLS and O-PLS models, preprocessed with normalization and column centering, are also given.

as follows. Stock solutions of samples were prepared at a concentration of 2 mg/mL in HPLC grade methanol (Prolabo, Paris, France). Polystyrene flat-bottomed 96-well plates (Nunc MicroWell, Thermo Scientific, USA) were filled with 100 μL methanol. 100 μL sample stock solution was added to the first wells, serial 2-fold dilutions were performed and 100 μL was discarded from the last wells. (±)-α-tocopherol (Sigma, Munich, Germany) was used as an antiradical control. Blank controls (methanol only) were also introduced. 100 μL freshly prepared solution of DPPH in methanol (50 μg/mL) was added to each well giving a final DPPH concentration of 25 μg/mL (63.4 μM) and final concentrations of samples ranging from 1 to 500 μg/mL. Plates were kept in the dark at room temperature for 20 min. After 20 min, absorbances were read at 515 nm using a SpectraMax 190 Microplate Reader equipped with the SoftMax Pro software (Molecular Devices, Ismaning, Germany). A DPPH calibration curve was measured between 0 and 50 μg/mL. Percentages of remaining DPPH, $\%DPPH_{remaining}$, were calculated as follows:

$$\%DPPH_{remaining} = \frac{A_{sample}}{A_{MeOH\ blank}} \times 100 \quad (9)$$

with A_{sample} being the absorbance of sample extract and $A_{MeOH\ blank}$ being the absorbance of blank control.

Two independent assays were carried out in duplicate to assess intra and inter-day repeatability. For each assay, the percentages of remaining DPPH are reported as the mean ± standard deviations of 2 duplicate values. Using the software GraphPad Prism 4 (GraphPad software, San Diego, CA, USA), the percentages of remaining DPPH were plotted against the $\log_{10}$ of the corresponding concentration. A sigmoidal curve was then fitted on the experimental points allowing determination of the IC_{50}, i.e. the sample concentration (in μg/mL) that is needed to scavenge half of the DPPH initially introduced. IC_{50}'s for each assay are reported with their 95% confidence intervals. The final IC_{50} values are reported as mean and 95% confidence intervals of the two IC_{50} values obtained for each independent assay.

3.4. Data analysis

Computations were performed on a PC with an Intel 2.8 GHz Pentium-IV processor, 512 MB RAM and running on Microsoft Windows XP and Matlab™ 7.1 (The Mathworks, Natick, MA). All data (pre)processing methods were performed using m-files written for Matlab™ 7.1.

4. Results and discussion

4.1. DPPH radical scavenging test

Table 1 presents the 36 *Mallotus* and *Phyllanthus* samples with their species name, collection time, origin, and the DPPH scavenging activity results given as average IC_{50} values with their 95% confidence interval. The lower the average IC_{50} value is, the higher the antioxidant activity of the sample.

Except for *P. amarus* (sample 11), all *Phyllanthus* samples (2, 3, 7, 9, 10, 12, 14, 15, 16, 18–33) are considered to have a high antioxidant activity (average $IC_{50} < 30$ μg/mL) with the range of IC_{50} situated between 8.45 and 29.70 μg/mL. The five *M. paniculatus* samples (6, 13, 17, 35, 36) are considered to have an intermediate

Table 2
Number of components and RMSECV for the calibration models built with the 60 min fingerprints and three preprocessing approaches.

Calibration technique	Preprocessing	# of components	RMSECV
PLS	Column centering	4	20.1
	Normalization and column centering	2	16.5
	SNV and column centering	2	16.8
O-PLS	Column centering	1 (3[a])	19.7
	Normalization and column centering	1 (1[a])	16.1
	SNV and column centering	1 (1[a])	16.4

[a] For O-PLS, the number of removed orthogonal components is given between brackets.

antioxidant activity (30 μg/mL < average IC_{50} < 50 μg/mL) with the IC_{50} range situated between 31.47 and 38.87 μg/mL. The five *M. apelta* samples (1, 4, 5, 8, 34), on the other hand, do not possess any antioxidant activity (average IC_{50} > 70 μg/mL) with an IC_{50} range between 72.66 and 110.00 μg/mL. The IC_{50} values of the *Mallotus* and *Phyllanthus* samples range between 31.47–110.00 μg/mL and 8.45–99.32 μg/mL, respectively.

4.2. HPLC fingerprints

In an earlier study [26], 60 min HPLC fingerprints have been developed. The data consists of two genera comprising a total of six species. Visual examination of the fingerprints (Fig. 1) reveals highly similar chromatographic profiles within each species except for *P. amarus*, where sample 11 seems to have a different profile than all other samples of that species (Fig. 1f). Additionally, the profiles between the species are rather different, thus the chemical constituents differ considerably between the species. Therefore, different pharmaceutical activities can possibly be attributed to different species (as seen in Table 1).

4.3. Evaluation of antioxidant activity versus fingerprint

4.3.1. Data preprocessing

Prior to the modeling, the data was preprocessed. Different preprocessings were evaluated, i.e. column centering, normalization followed by column centering, and standard normal variate (SNV) transformation followed by column centering. For the PLS and O-PLS techniques, normalization followed by column centering was best for this data set (see further, Table 2). For PCA and HCA, SNV followed by column centering was required (see further, Figs. 2 and 3). All results discussed further are from data preprocessed as indicated above.

Aligning the peaks is usually recommended, because of retention time shifts between chromatograms due to experimental error. For this purpose, many so-called warping techniques can be applied. In this study, the fingerprints were not aligned, because of the large diversity in species, which makes it challenging to align the peaks properly, and because the goal of the study is not prediction of antioxidant activity for new samples but indication of peaks possibly responsible for the antioxidant activity. This problem is studied and discussed in detail in [16]. Moreover, it is rather difficult to identify corresponding peaks that should be aligned, since no mass spectrometry (MS) data was available. Thus, aligning peaks is not evident, because peaks representing different components can be forced to align. The problem (or its absence because anyway the proper peaks are indicated) of non-alignment of peaks in such situations was discussed in [16]. Therefore, in this study, it was also decided to treat, model, and interpret the non-aligned fingerprints.

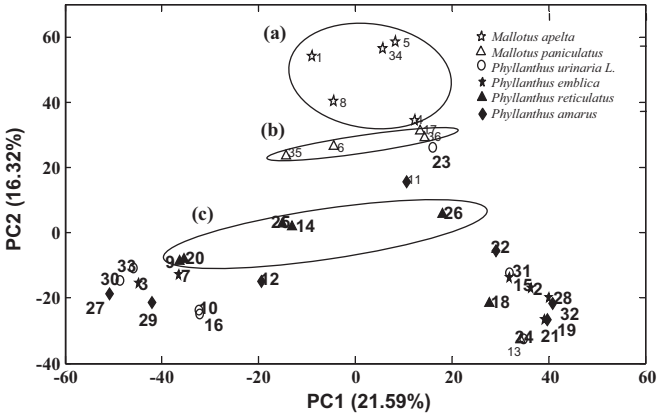

Fig. 2. PC1–PC2 score plot for the 60 min fingerprints of the 36 *Mallotus* and *Phyllanthus* samples (pretreatment: standard normal variate and column centering). Three groups of species are distinguished, i.e. containing (a) samples 1, 4, 5, 8, 34, (b) samples 6, 17, 35, 36, and (c) samples 9, 14, 20, 25, 26. The highly active antioxidant samples are marked in bold.

4.3.2. Exploratory analysis

Principal Component Analysis (PCA) has been applied to verify whether groups of samples could be distinguished, occasionally according to their antioxidant activity or species. The score plot (Fig. 2) of PC1 (21.59% explained variation) versus PC2 (16.32% explained variation) was drawn after the above-mentioned preprocessing. Standard normal variate followed by column centering was considered as the best preprocessing. All samples with highly antioxidant activity (marked in bold) are clustered, i.e. can be distinguished along PC2. Combining the proximity of samples on the score plot and previous knowledge of their fingerprint profiles and species, results in the distinction of the two groups of genera. The samples belonging to the genus *Mallotus* (samples 1, 4, 6, 8, 17, 34, 36) are distinguished from those from the genus *Phyllanthus* (samples 2, 3, 7, 9, 12, 14, 16, 18, 22, 24, 33) except for *M. paniculatus* (sample 13). The above-mentioned distinction between antioxidant and non-antioxidant samples along PC2 is the same as that between *Mallotus* and *Phyllanthus* samples. Three groups of species, i.e. *M. apelta* (samples 1, 4, 5, 8, 34), *M. paniculatus* (samples 6, 17, 35, 36) and *P. reticulatus* (samples 9, 14, 20, 25, 26) are clustered nicely,

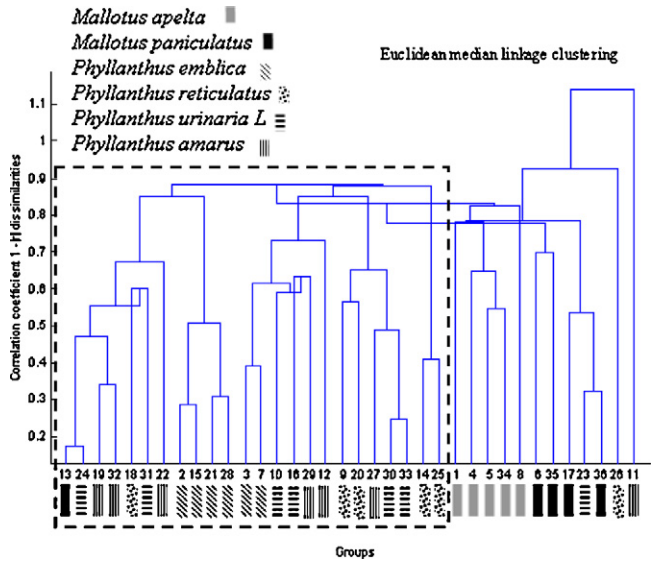

Fig. 3. HCA dendrogram obtained using Euclidean distance as distance measure and median's method as linkage method (pretreatment: standard normal variate and column centering).

i.e. they are situated in each other proximity. Obviously, sample 13 seems to be outlying. Its (dis)similarity to the other *M. paniculatus* samples or to fingerprints from other species can be evaluated in a similarity analysis [36]. Sample 18 (*P. reticulatus*) is also not clearly clustered with the other *P. reticulatus* samples. However, given the variability within the fingerprints of *P. reticulatus* and the arbitrariness of defining the groups on the PCA plot, sample 18 could have been included in the *P. reticulatus* group. The fact that the different species are not separated in largely distinct groups is both an indication and a consequence of the fact that the fingerprints are not extremely different.

In order to confirm the clustering results of PCA, the Hierarchical Cluster Analysis (HCA) technique also has been utilized. Several distance measures (i.e. Euclidean distance and the Pearson correlation coefficient) and linkage methods (i.e. single linkage, complete linkage, average linkage, weighted average linkage, centroid's method, median's method and Ward's method) were evaluated and compared. When using the Euclidean distance measure and the median linkage method, after SNV and column centering as preprocessing, the dendrogram given in Fig. 3 was generated. This dendrogram also revealed relationships amongst the studied samples. It is clear that the samples with antioxidant activity can be distinguished, except for samples 23 (*Mallotus urinaria* L.) and 26 (*P. reticulatus*). Secondly, sample 13, which is a medium active *M. paniculatus* sample, is clustered with the highly active *Phyllanthus* samples. These results are already seen with PCA. The *Mallotus* and *Phyllanthus* samples can again more or less be distinguished, except for *M. paniculatus* (sample 13), *P. urinaria* L. (sample 23), *P. amarus* (sample 11), and *P. reticulatus* (sample 26). For samples 13 and 23, this was also seen from the PC1–PC2 score plot of Fig. 2. Additionally, on the PC1–PC2 score plot, sample 26 is situated further away from the majority of *P. reticulatus* samples and sample 11 is situated at the border between the *Mallotus* and *Phyllanthus* species. Moreover, sample 11 has a low antioxidant activity compared to all other *Phyllanthus* samples. The fact that sample 26 was included in the group of antioxidative *P. reticulatus* samples in the PCA plot and not in the cluster analysis is not contradictory, because in PCA a subjectively determined cluster is drawn based on some labeling information, while in the dendrogram splits are determined by correlation coefficients. Both plots provide complementary information.

4.3.3. Linear multivariate calibration

Multivariate calibration models for the antioxidant activity were established using two techniques i.e. PLS and O-PLS, on the data matrix **X** consisting of the 36 fingerprints and the response vector **y**, representing the IC_{50} test results. No division of the data into a calibration and a test set was made because the number of samples is rather small and prediction of the antioxidant activity of new samples is not the primary goal of this study, but the indication of peaks potentially responsible for the antioxidant activity is. For PLS, the optimal model complexity (A_{opt}) was chosen from a LOO-CV procedure and the simplest model with (nearly) the lowest RMSECV was selected. For O-PLS, a one-component PLS model with $A_{opt} - 1$ orthogonal components was built. In Table 2, the number of model components and the RMSECV are presented for both calibration techniques, applying three preprocessing approaches. The O-PLS models after normalization or SNV followed by column centering were found to be most simple, i.e. they contain the lowest number of components. For both PLS and O-PLS, normalization followed to column centering lead to the model with the best predictive ability, i.e. the lowest RMSECV for the whole data set (16.5 and 16.1 μg/mL, respectively).

To evaluate the model's ability to predict the antioxidant activity of the samples, the prediction of their antioxidant activity was taken into account. Table 3 shows the results of the 25 highly antioxidant samples (IC_{50} < 30 μg/mL) for the DPPH radical scavenging assay and the predictions from the models after application of normalization and column centering as preprocessing approach. For these highly antioxidant samples, none of the two models predicted any as being non-active (IC_{50} > 70 μg/mL). The PLS model predicted 24 out of 25 highly active samples correctly as being highly antioxidant. Only sample 22 was predicted as being intermediately active (30 μg/mL < IC_{50} < 50 μg/mL). The mean bias of these highly antioxidant samples was found to be 4.49 μg/mL. The O-PLS model predicted 21 highly active samples correctly and four (22, 23, 25, and 26) incorrectly as intermediately active. For O-PLS, the mean bias of these highly antioxidant samples was found to be 5.95 μg/mL. None of the models predicted inactive samples to have an intermediate or highly antioxidant activity.

Table 3
Results from the DPPH radical scavenging assay (IC_{50}) and predictions from the models built with the 60 min fingerprints. Preprocessing: normalization and column centering.

Sample no.	IC_{50} (μg/mL)	Predicted IC_{50} (μg/mL)	
		PLS	O-PLS
2	9.90	6.09	11.48
3	9.55	6.94	7.55
7	8.57	8.16	9.91
9	28.06	26.53	26.05
10	9.75	12.73	8.53
12	13.74	15.82	16.14
14	29.70	25.82	29.26
15	10.92	7.82	13.86
16	8.45	9.23	4.31
18	13.42	15.39	20.45
19	13.12	12.02	12.94
20	22.18	26.76	28.03
21	8.75	6.92	7.11
22	14.33	45.36	45.67
23	8.83	18.56	43.20
24	8.56	15.94	12.58
25	23.49	25.35	31.27
26	27.31	24.45	35.49
27	15.51	13.64	9.54
28	10.26	6.52	9.55
29	10.15	14.35	9.52
30	17.99	19.83	17.29
31	21.13	23.66	28.96
32	16.12	17.88	18.93
33	13.60	26.34	25.21
Mean bias		4.49	5.95

4.3.4. Regression coefficients: potentially antioxidant compounds

In order to indicate in the fingerprints those peaks potentially responsible for the antioxidant activity of the measured samples, the regression coefficients of the models are examined. Chromatographic fingerprint peaks of potentially antioxidant compounds correspond to negative regression coefficient peaks (indicated with an arrow in Fig. 1) as the IC_{50} test result decreases with increasing activity. In this study, the fingerprints of each species were compared with the obtained regression coefficients (Fig. 1) from both models i.e., PLS and O-PLS. For this case study, the regression plots from the two models showed similar coefficient profiles.

The major negative coefficient peaks corresponding to the potentially antioxidant compounds are found in four peaks (at retention times of about 15.0, 17.0, 23.0 (splitted coefficient peaks), and 27.0 min). The antioxidant samples have no major peaks present at retention times corresponding to the positive coefficient peaks in the regression plots. In Fig. 1b (*M. paniculatus* samples), none of the negative coefficient peaks match major compounds present in these fingerprints, except for sample 13. This sample has an antioxidant activity and was already found atypical from the PCA score plot and the HCA dendrogram. In Fig. 1c–f, all samples have a high antioxidant activity and the major negative

coefficients correspond to substance peaks seen in the fingerprints. These peaks are probably responsible for most antioxidant activity. In a next step, the indicated relevant peaks should be isolated, identified, and investigated using structure-elucidation techniques. More information related to these steps can be found in [16].

5. Conclusions

HPLC fingerprints of *Mallotus* and *Phyllanthus* samples were combined with data-handling techniques in order to model the antioxidant activity and indicate peaks possibly responsible for this activity.

In a first step, exploratory analysis using PCA and HCA was performed to verify the data structure, e.g. to see whether the antioxidant samples could be distinguished from the less or non-active samples, or whether atypical fingerprints occur. The PCA score plot showed some diverging clustering tendency compared to HCA. However, the dendrogram mainly confirmed what was seen on the PCA score plot.

Then, the antioxidant activities of the samples were modeled as a function of the fingerprints using PLS and O-PLS. The peaks potentially responsible for the antioxidant activity of the samples were indicated studying the regression coefficients of the models. In this study, the regression coefficients plot of both models not only showed similar coefficient profiles but also the coefficient peaks linked to the compounds probably responsible for the antioxidant activity can be indicated at the same positions. In a next step, the substances from these relevant peaks should be isolated, identified, and further examined.

Acknowledgements

The financial support from the Bilateral Project (BL/03/V21) between Vietnam and Belgium was greatly appreciated. Bieke Dejaegher is a post-doctoral fellow of the Research Foundation Flanders, Belgium (FWO).

References

[1] L. Giri, H.C. Andola, V. Kant Purohit, M.S.M. Rawat, R.S. Rawal, I.D. Bhatt, Res. J. Phytochem. 4 (2010) 234.
[2] M. Dumarey, A.M. van Nederkassel, E. Deconinck, Y. Vander Heyden, J. Chromatogr. A 1192 (2008) 81.
[3] Guidelines, for the Assessment of Herbal Medicines, World Health Organization, Geneva, Switzerland, 1991, http://whqlibdoc.who.int/HQ/1991/WHO_TRM_91.4.pdf (accessed 25.05.11).
[4] C. Tistaert, B. Dejaegher, Y. Vander Heyden, Anal. Chim. Acta 690 (2011) 148.
[5] L. Regina Rocha Martins, E. Rodrigues Pereira-Filho, Q. Bezerra Cass, Anal. Bioanal. Chem. 400 (2011) 469.
[6] Y. Li, T. Zhang, X. Zhang, H. Xu, C. Liu, J. Sep. Sci. 33 (2010) 3347.
[7] M. Costas-Rodríguez, I. Lavilla, C. Bendicho, Anal. Chim. Acta 664 (2010) 121.
[8] M. Vallejo, S. Angulo, D. García-Martínez, A. García, C. Barbas, J. Chromatogr. A 1187 (2008) 267.
[9] 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.
[10] J. Yang, L.-H. Chen, Q. Zhang, M.-X. Lai, Q. Wang, J. Sep. Sci. 30 (2007) 1276.
[11] F. Gong, B.-T. Wang, Y.-Z. Liang, F.-T. Chau, Y.-S. Fung, Anal. Chim. Acta 572 (2006) 265.
[12] C. Tistaert, B. Dejaegher, G. Chataigné, C. Van Minh, J. Quetin-Leclercq, Y. Vander Heyden, Talanta 83 (2011) 1198.
[13] C. Tistaert, B. Dejaegher, N. Nguyen Hoai, G. Chataigné, C. Rivière, V. Nguyen Thi Hong, M. Chau Van, J. Quetin-Leclercq, Y. Vander Heyden, Anal. Chim. Acta 652 (2009) 189.
[14] N. Nguyen Hoai, B. Dejaegher, C. Tistaert, V. Nguyen Thi Hong, C. Rivière, G. Chataigné, K. Phan Van, M. Chau Van, J. Quetin-Leclercq, Y. Vander Heyden, J. Pharm. Biomed. Anal. 50 (2009) 753.
[15] A.M. van Nederkassel, M. Daszykowski, D.L. Massart, Y. Vander Heyden, J. Chromatogr. A 1096 (2005) 177.
[16] C. Tistaert, B. Dejaegher, G. Chataigné, C. Rivière, M. Chau Van, J. Quetin-Leclerq, Y. Vander Heyden, Anal. Chim. Acta 721 (2012) 35.
[17] P. Van Kiem, N. Hai Dang, H. Viet Bao, H. Thanh Huong, C. Van Minh, L. Mai Huong, J. Joon Lee, Y. Ho Kim, Arch. Pharm. Res. 28 (2005) 1131.
[18] T.-Y. An, L.-H. Hu, X.-F. Cheng, Z.-L. Chen, Phytochemistry 57 (2001) 273.
[19] A. Kumaran, R. Joel Karunakaran, LWT – Food Sci. Technol. 40 (2007) 344.
[20] M. Xu, Z.-J. Zha, X.-L. Qin, X.-L. Zhang, C.-R. Yang, Y.-J. Zhang, Chem. Biodivers. 4 (2007) 2246.
[21] W. Luo, M. Zhao, B. Yang, J. Ren, G. Shen, G. Rao, Food Chem. 126 (2011) 277.
[22] C. Rivière, V. Nguyen Thi Hong, Q. Tran Hong, G. Chataigné, N. Nguyen Hoai, B. Dejaegher, C. Tistaert, T. Nguyen Thi Kim, Y. Vander Heyden, M. Chau Van, J. Quetin-Leclercq, Phytochem. Rev. 9 (2010) 217.
[23] C. Rivière, V. Nguyen Thi Hong, L. Pieters, B. Dejaegher, Y. Vander Heyden, M. Chau Van, J. Quetin-Leclercq, Phytochemistry 70 (2009) 86.
[24] X. Liu, M. Zhao, W. Luo, B. Yang, Y. Jiang, J. Med. Food 12 (2009) 423.
[25] M.K. Sarkar, M. Kinter, B. Mazumder, P.C. Sil, Food Chem. 114 (2009) 1405.
[26] M. Goodarzi, B. Dejaegher, A. Hoang Le Tuan, N. Nguyen Hoai, M. Chau Van, J. Quetin-Leclercq, Y. Vander Heyden, in preparation.
[27] H. Marten, M. Martens, Multivariate Analysis of Quality: An Introduction, Wiley, Chichester, 2001.
[28] J. Trygg, S. Wold, J. Chemometr. 16 (2002) 119.
[29] 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. Part A, Elsevier, Amsterdam, 1997.
[30] 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. Part B, Elsevier, Amsterdam, 1998.
[31] M. Daszykowski, B. Walczak, D.L. Massart, Anal. Bioanal. Chem. 380 (2004) 370.
[32] N. Bratchell, Chemometr. Intell. Lab. Syst. 6 (1989) 105.
[33] O.E. de Noord, Anal. Chem. 68 (1996) 3851.
[34] B. Dejaegher, G. Alaerts, N. Matthijs, Acta Chromatogr. 22 (2010) 237.
[35] W. Brand-Williams, M.E. Cuvelier, C. Berset, Lebensm. Wiss. U. Technol. 28 (1995) 25.
[36] G. Alaerts, J. Van Erps, S. Pieters, M. Dumarey, A.M. van Nederkassel, M. Goodarzi, J. Smeyers-Verbeke, Y. Vander Heyden, Journal of Chromatography B (2012), http://dx.doi.org/10.1016/j.jchromb.2012.04.031, in press.