Liquid Chromatography for Plant Metabolite Profiling in the Field of Drug Discovery

Luiz Carlos Klein-Júnior, Johan Viaene, Amorn Slosse, and Yvan Vander Heyden

1 Introduction

"A new golden age of natural products discovery" (Shen 2015), "The re-emergence of natural products for drug discovery in the genomics era" (Harvey et al. 2015), and "The impending renaissance in discovery and development of natural products" (Pawar et al. 2017) are some review titles, used in recent publications, which highlight the importance of natural products in drug discovery nowadays.

The historic significance of natural products is very well described in the literature (Atanasov et al. 2015; Li and Weng 2017). Probably, one of the most important marks is the *Ebers Papyrus*, a scroll dating back to 1500 BC. It contains more than 800 medicinal preparations (mostly plant based) used in the ancient Egyptian medicine, stored since 1873 at the University of Leipzig (Atanasov et al. 2015; Universitätsbibliothek Leipzig 2016; Li and Weng 2017).

The use of complex mixtures for the treatment of pathological conditions lasted until the beginning of the nineteenth century. In 1817, the pharmacist Friedrich Wilhelm Adam Sertürner reported the isolation and evaluation of the *principium somniferum* from the opium poppy (*Papaver somniferum* L.), nowadays known as morphine (Atanasov et al. 2015). This boosted the isolation of many other important natural products, such as quinine, caffeine, and atropine, creating the foundations of the Western medicine.

Although history serves as a proof of concept for the importance of natural products in drug discovery, the Big Pharma companies abandoned most of their research

L. C. Klein-Júnior (✉)
Chemical-Pharmaceutical Investigation Nucleus, University do Vale do Itajaí – UNIVALI, Itajaí, SC, Brazil
e-mail: lcklein@univali.br

J. Viaene · A. Slosse · Y. Vander Heyden
Department of Analytical Chemistry, Applied Chemometrics and Molecular Modelling, Vrije Universiteit Brussel – VUB, Brussels, Belgium

V. Cechinel Filho (ed.), *Natural Products as Source of Molecules with Therapeutic Potential*, https://doi.org/10.1007/978-3-030-00545-0_3

programs in this field during the last decades (David et al. 2015; Harvey et al. 2015; Shen 2015). Constant taxonomic modifications, low yield, ecological and legal aspects, accessibility issues, and variation in quality of plant material are some of the reasons for the decline in natural product interest (Atanasov et al. 2015; Harvey et al. 2015).

Nevertheless, one of the most important challenges for pharmaceutical companies in natural product researching is also the main reason for the renewed interest in the area. In the late twentieth century, Big Pharma centered their studies in combinatorial synthesis and high-throughput screening (HTS). Unfortunately, plant extracts are not compatible with HTS campaigns, mainly due to their high viscosity, aggregation and/or precipitation, non-specific binding, and presence of fluorescent and/or quenching compounds, mostly requiring adaptation, purification, and additional steps for proper evaluation, hampering this methodology for drug discovery (Atanasov et al. 2015; Shen 2015). However, the number of new drugs approved from the use of combinatorial synthesis and HTS is considerably low, demonstrating that this approach was quite frustrating (Atanasov et al. 2015; Harvey et al. 2015).

To the poor results obtained in the last decades, the recent awarding of the 2015 Nobel Prize in Physiology or Medicine to Youyou Tu was added. Prof. Tu discovered the plant natural product artemisinin, used for the treatment of malaria. Together with recent analytical and chem-bioinformatical technological advances, the renewal of interest in natural products seems to be undeniable (Atanasov et al. 2015; Shen 2015).

Traditionally, bioactivity-guided fractionation is used in the study of natural products. In this approach, pharmacological assays are performed in order to drive the isolation of active compounds, theoretically avoiding the isolation of non-active compounds (Hubert et al. 2017). Usually, this workflow is very time- and money-consuming. In addition, it is not uncommon to isolate already known compounds with already known pharmacological properties (Atanasov et al. 2015; Hubert et al. 2017).

In this sense, new technologies have been introduced in order to avoid this very laborious methodology. An emerging approach is the metabolic profiling (Atanasov et al. 2015). Using chemometric tools, it is possible to correlate a chemical profile and a bioactivity of plant extracts, giving valuable information regarding the most active compounds. In addition, using hyphenated techniques, the early dereplication is also possible, avoiding the isolation of already known bioactive compounds (Allard et al. 2017; Begou et al. 2017).

Several analytical techniques have been used for metabolic profiling, such as direct nuclear magnetic resonance (NMR) and mass spectrometry (MS) analysis (Gemperline et al. 2016; Kumar 2016). However, because of the complexity of extracts, a preliminary separation, such as by liquid chromatography (LC), is often applied (Allard et al. 2017; La Barbera et al. 2017).

In this chapter, focus is given to the workflow in LC plant metabolic profiling for drug discovery, which includes sample preparation, development of the metabolite profile and pharmacological-assay considerations, as well as the overall correlation of metabolite profiles and bioactivity using chemometric tools.

2 Sample Preparation

Sample preparation is the first major step to be considered in the analysis of secondary metabolites, in the context of drug discovery, because it has a great impact on the metabolite contents and consequently the obtained results (Rates 2001; Huie 2002; Kim and Verpoorte 2010; Vuckovic 2012; Wu et al. 2013; La Barbera et al. 2017). However, this step is often still being done manually causing high costs. Therefore, to handle a large number of samples simultaneously (high throughput) and to minimize possible degradation of the metabolites, sample preparation should be simple and fast. Results of the chemical analysis generally depend on sampling and extraction within sample preparation. Usually sampling is correctly being performed with appropriate quenching (methods that prevent or minimize the enzymatic or biochemical processes of the plants because these could result in metabolic profile modification) to assure experimental reproducibility. However, much attention has to be paid to the improvement of the extraction step, as researchers use own experience-based protocols omitting adequate extraction evaluation (Kim and Verpoorte 2010; Mushtaq et al. 2014; Wen et al. 2014; Klein-Júnior et al. 2016a). Extraction, removal of possible interfering compounds, cleanup, enrichment of the metabolites of interest, and, if necessary, the transformation of analytes in a more suitable form that is compatible with, for instance, LC-MS analysis should be done in such a way that the composition of the components remains more or less constant.

Extraction is the main step in sample preparation; thus these methods have to be considered in regard to their suitability to meet the aims of a study (Kim and Verpoorte 2010; Sasidharan et al. 2011; Gupta et al. 2012). This important procedure, executed before chromatographic analysis, differs depending on the choice of the extraction solvent, which, among other things, is determined by the used analytical method and the chemical characteristics of the considered compounds. Moreover, different aspects have to be taken into account when extraction is carried out, e.g., solvent properties, the solvent-sample ratio, and the extraction time, pressure, and temperature, as these determine its success (Kim and Verpoorte 2010; Choi and Verpoorte 2014; Azmir et al. 2013). There is however no single solvent which dissolves all compounds in a sample, so multiple solvent extractions are needed to get a total view of the metabolome (Azmir et al. 2013; Martin et al. 2014; Mushtaq et al. 2014; La Barbera et al. 2017). Hence comprehensive methods, using a gradient of different solvents with increasing polarity, allow to efficiently extract a wide range of metabolites in one run. Moreover, these methods provide the potential to extract metabolites without degradation of any kind (De Monte et al. 2014; Yuliana et al. 2011; Mushtaq et al. 2014; Hill and Roessner 2015).

Extraction can be split in an actual extraction and the cleanup/enrichment of metabolites (fractionation of crude extracts of metabolite groups of interest) as a consequence of their low concentration in complex matrices. The conventional or classical extraction methods, such as maceration, percolation, and Soxhlet extraction, are the methods at first applied, making use of suitable solvents. The main

drawbacks of these methods are a possibly protracted extraction time, consumption of large amounts of unhealthy and polluting solvents, high costs, low selectivity, and the potential degradation of metabolites (Sasidharan et al. 2011; Gupta et al. 2012; Azmir et al. 2013; Mushtaq et al. 2014; Brusotti et al. 2014; Azwanida 2015; Raks et al. 2018). Consequently alternative techniques, which give minimal sample degradation, are less hazardous solvent consuming, are less energy demanding, are time saving, have more environment-friendly properties, and provide better extraction efficiency and selectivity, were and are still being developed (Sasidharan et al. 2011; Gupta et al. 2012). Frequently used techniques are, for instance, microwave-assisted extraction (MAE), ultrasound-assisted extraction (UAE), supercritical fluid extraction (SFE), and pressurized solvent extraction (PSE). Such variety of techniques, however, requires careful evaluation as each technique may considerably influence the composition and hence the biological activity of the extract (Azwanida 2015; Atanasov et al. 2015).

UAE is the most commonly used technique; is very simple, fast, and cheap without solvent-choice limitation; and is environment-friendly. It is based on an increased surface contact between solvents and samples by using ultrasound waves, facilitating solvent transport in the plant matrix. High quantities of metabolites are obtained using small amounts of solvents. The application of ultrasound energy may have a possible negative impact on the active metabolites (Kim and Verpoorte 2010; De Monte et al. 2014; Azwanida 2015).

MAE, consisting of closed- or open-vessel MAE, makes use of microwave energy to facilitate the distribution of analytes from the plant matrix into the solvent. Advantages of this technique are lower solvent consumption, its speed and efficiency compared to Soxhlet and UAE, and its extraction quality and yield against traditional methods. MAE is less time-consuming but more expensive than UAE and cheaper and less environment-friendly than SFE. However, its technique cannot be used for temperature-labile components (De Monte et al. 2014; Raks et al. 2018).

SFE as extraction technique mainly uses CO_2 as supercritical extraction solvent. Thanks to the low critical temperature of CO_2 (31.1 °C), thermolabile constituents may be extracted without degradation or denaturation. CO_2 (greenhouse gas) is cheap, safe, and easily available. It is also less toxic, nonflammable, and easily removable when ending the extraction. It provides high-quality and reliable extracts. One of the main drawbacks is the high cost of the equipment.

PSE is based on the principle that at increased pressure the solvent remains in the liquid phase after heating, which allows fast and little solvent-consuming extractions. Simultaneously higher yields are obtained. H_2O is often used as a green solvent. This technique can be hyphenated with UAE to gain efficiency (De Monte et al. 2014; Gupta et al. 2012; Mushtaq et al. 2014; Klein-Júnior et al. 2016c; Raks et al. 2018).

For the cleanup and concentration step, solid-phase extraction (SPE) is extensively used because of its possible automation and the availability of a wide range of sorbents, usually applied in cartridges (Tu et al. 2010; Wen et al. 2014). It is based on the adsorption or partition of the extracted compounds on a solid phase, which initially retains the interested group of analytes. After washing of the unwanted

components, the metabolites of interest are desorbed (eluted) with an appropriate solvent. Basically SPE is a fractionation rather than an extraction technique. Liquid-liquid extraction (LLE) was and still is a frequently used cleanup technique, but because it is too time- and solvent consuming, it is more and more replaced by SPE. A related SPE technique that was developed is the solid-phase micro-extraction. It is similar to SPE but miniaturized, faster, and greener (Mushtaq et al. 2014; Klein-Júnior et al. 2016c).

The actual extraction methods do not result in collecting the entire plant metabolome; thus compromises are to be made (Vuckovic 2012; Brusotti et al. 2014; Wen et al. 2014; Klein-Júnior et al. 2016a). Hence, the enormous number of plant metabolites (± 30,000) with considerable differences in polarity, stability, and chemical diversity, as well as the extraction reproducibility, makes the design of efficient protocols mandatory (Kaiser et al. 2009; t'Kindt et al. 2009; Choi and Verpoorte 2014; Mushtaq et al. 2014). Metabolite profiling studies must be designed in such a way that a maximum of metabolites are detected, in other words, i.e., that different metabolite groups are extracted (Sasidharan et al. 2011; Gupta et al. 2012).

An arising issue is the use of chemometrics and experimental-design approaches in optimizing the extraction and, occasionally, the fractionation processes (Klein-Júnior et al. 2016a). Several studies have been dealing with this optimization, where different factors, such as solvent concentrations, temperature, extraction time, sample-solvent ratio, particle size, and pH, are studied in order to obtain maximal responses, for, for instance, yield, amount of a specific class of metabolites, and biological activities (Souza et al. 2007; Das et al. 2013; Zhu and Liu 2013; Martin et al. 2014; Wang et al. 2014; Hammi et al. 2015; Wu et al. 2015; Izadiyan and Hemmateenejad 2016; Ćujić et al. 2016; de O Silva et al. 2017; Dary et al. 2017).

As a rule, when performing metabolome analysis, the number of extracted compounds reflects the best metabolomics conspectus. However, in case of targeted metabolite profiling, in order to obtain an accurate overview of the chemical diversity of the metabolite group of interest, other responses/approaches are being elaborated.

In this regard, Klein-Júnior et al. (2016a) extensively studied the optimization of indole alkaloid extraction and fractionation based on UPLC-DAD metabolite profiling with the aim to develop a less time- and solvent-consuming method, as well as to represent maximally the entire chemical composition of the plant being extracted. In other words, how can an efficient extraction method be developed to obtain a maximal number of metabolites? In a first step in the optimization of an UAE method, a fractional factorial screening design was executed to determine the significant effects of the selected factors. The evaluation of this design was done in two steps. Firstly, the response yield (obtained with LLE fractions), number of peaks, and sum of peak areas showed to have little meaning in relation to chemical diversity. In order to obtain a comprehensive picture of the metabolic profile, a new approach was presented to overcome the inadequate information of the abovementioned responses. Euclidean distance measurements between the metabolic profiles of the alkaloid fractions and the blank injection were calculated. This distance is an indication of the extract chemical diversity. In other words, the higher the distance,

the higher the chemical diversity of the extract, i.e., the higher the metabolite content compared to the blank signal.

In a second step, the entire metabolic profiles were used as responses to determine detailed information of the factor effects on these profiles. Effect plots or effect fingerprints were calculated for each factor and graphically evaluated. From these plots thermolabile compounds could be indicated, and peaks (compounds) as well as the important factors were selected for further optimization. This was done performing a central composite design with temperature and extraction time as optimization factors. Here, the heights of the selected peaks were determined as responses. After response modeling (and avoiding degradation), the optimal combination of time and temperature was determined. For the optimization of the alkaloid fractionation, SPE was applied as alternative technique of LLE. Applying a Box-Behnken design, three factors were studied, and as response the sum of the peak areas of six metabolites in the profiles was used. Sample concentration, percentage of acetonitrile, and eluting volume were set to obtain the best fractionation conditions. It can be concluded that the Euclidean distance approach and the entire metabolic profiles are useful as responses for extraction optimization of specific component groups. This study allowed developing a time- and solvent-saving method as well as a reliable extraction and fractionation method of indole alkaloids, without component degradation.

3 Development of the Metabolite Profile

Herbal samples have a complex composition. This is related to the multitude of metabolic pathways involved in transforming the nutrients and compounds taken up by the plant through its root system and aerial parts, into compounds required by the plant at a given stage in its life cycle. Additionally, these processes highly depend on light exposure, rainfall, soil type, and numerous other external factors, resulting in a high variability in chemical composition between different samples of a same plant species (Wagner and Ulrich-Merzenich 2009; Lu et al. 2005; Li et al. 2010; Liu 2011). Further, when the sample extraction is set up to isolate compounds of a specific group (for instance, alkaloids), the compounds in the mixture may show a high degree of similarity for several physicochemical properties. Because in drug discovery, the aim is to find and fully characterize potential new drug candidates, it is essential that the chemical analysis allows separating the various compounds in the (cleaned-up) extracts. A wide range of chemical separation techniques is available to obtain this goal (Liu 2011).

Spectral techniques are also very popular tools in chemical analysis. They can be applied to characterize a compound in terms of its ultraviolet, visible or infrared radiation absorbance, fluorescence, mass spectrum, nuclear magnetic resonance spectrum, etc. (Gauglitz and Vo-Dinh 2003; Gunzler and Williams 2001). Although this spectral information has its value in various contexts, it is important to know that not all of these techniques do separate the information of the different

compounds. As a result, the UV spectrum, for instance, of a mixture of compounds will contain information from all UV-absorbing compounds present. However, information from individual compounds is required in the context of drug discovery because of the need to find active compounds, which may develop as potential new lead compounds (Liu 2011). To assure that the information provided by spectroscopic techniques is specific for given compounds in a mixture, these compounds need to be separated prior to the spectral detection. Various separation techniques, including chromatography (Tistaert et al. 2011) and electrophoresis (Gunzler and Williams 2001), are available.

In this chapter, the discussion is restricted to chromatographic separation techniques. These separate compounds in a mixture based on their different interaction behavior in a two-phase system, called the stationary and the mobile phase, which are (relatively to one another) moving in opposite directions. Various forms exist, which will be briefly overviewed in this section, focusing on their use in the context of drug discovery.

3.1 Thin-Layer Chromatography

The simplest technique is thin-layer chromatography (TLC) (Fig. 1) (Sherma and Fried 2003; Tang et al. 2014). In TLC, a layer of the stationary phase, often bare silica or chemically modified silica (see further), is attached to a plate. Small volumes of sample solution are spotted on one side of the plate in little spots or bands. The analysis is started by placing the plate in a recipient, the development chamber, containing the mobile phase, which is usually a mixture of organic solvents, as shown in Fig. 1. The mobile phase then starts moving through the stationary phase by capillary forces, dragging also the sample compounds with it. Compounds with a relatively high affinity for the stationary phase will migrate slowly, while those with a higher affinity for the mobile phase will migrate faster. This results in a separation of the compounds according to their affinity differences for stationary and mobile phases. When the solvent has moved a given distance on the plate – when the separation is maximal at the conditions applied – the plate is removed from the

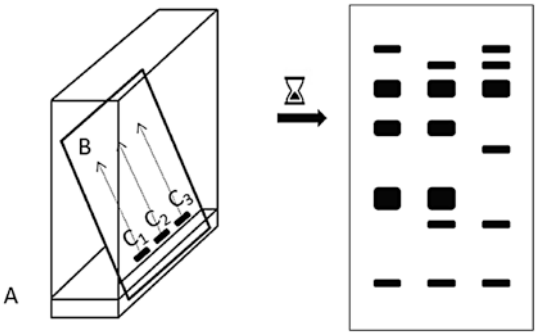

Fig. 1 Thin-layer chromatography. Left side: A, development chamber; B, TLC plate; C1–3, spots of samples or standard solutions. Right side: TLC plate after separation and revelation: pattern of spots is visualized

development chamber and dried and the compounds are revealed through the application of spray agents or by UV light radiation which causes quenching of fluorescence on pretreated plates or by densitometric measurements (Tang et al. 2014), which reveal a pattern of spots for each sample. TLC is often used in traditional drug discovery because it allows the simultaneous analysis of several samples, it is relatively cheap, and a wide range of stationary and mobile phases can be used. Since the plates are single use, applying aggressive solvents is not even an issue. A major drawback of the technique, however, is low efficiency and resolution, which means that compounds with similar properties are likely to show overlapping spots. Therefore, TLC is rather applied to separate groups of compounds, after which their spots can be individually scraped off the plate, redissolved in an appropriate solvent, and subjected to column chromatography (see next section), since on modern chromatography columns, compounds with similar properties can be better separated (Sherma and Fried 2003; Tang et al. 2014). However, given the latter, very often the TLC step may be skipped, and after sample preparation, immediately column chromatography is applied.

3.2 *Column Chromatography*

In column chromatography (Fig. 2), a small amount of sample, containing compounds that need to be separated, is injected in a constantly flowing mobile phase, which can be a liquid, a gas, or a supercritical fluid, called, respectively, liquid (LC), gas (GC), and supercritical fluid chromatography (SFC). Case studies applying GC and SFC can be found in refs (Liu et al. 2016; Li et al. 2013). However, in this chapter, we will further focus on LC. Although in the early days of LC, gravitation force

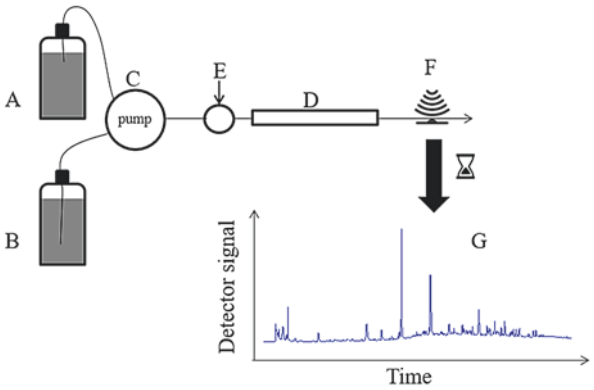

Fig. 2 Schematic representation of HPLC. A and B: solvents used as mobile phase components, C: pump moving the solvents at a constant flow rate, E: injector where the sample is introduced in the mobile phase flow, D: analytical column, F: detector, G: graphical output of the detector: chromatogram

was used as a motor for the separation process, nowadays, the mobile phase is pumped through the column at a tunable flow rate. This process is called high-pressure/high-performance liquid chromatography (HPLC) (Lough and Wainer 1996; Dong 2006; Waksmundzka-Hajnos and Sherma 2011). HPLC nowadays has become a benchmark technique in analytical chemistry and is a common tool in drug discovery. The mobile phase is thus pumped through a column containing the stationary phase, which is usually composed of particles that are mechanically immobilized in the column. Classically the particles are silica based, which has polar properties. Apolar and intermediately polar stationary phases have been created by chemically binding apolar or intermediately polar functional groups (for instance, C18 or C_3-CN chains) to the silanol groups on the surface of the silica particles. As a result, continuous developments have resulted in the availability of a large choice of stationary phases (Lough and Wainer 1996; Dong 2006; Waksmundzka-Hajnos and Sherma 2011).

A stationary phase should be selected in such a way that it has chemical properties that differ in some way (for instance, polarity) from the mobile phase. Mobile and stationary phases with various properties can be selected. In liquid chromatography, two modes of chromatography are classically defined, normal and reversed phase. In normal-phase chromatography, the mobile phase has an apolar nature, for instance, hexane-based, and the stationary phase a polar, for instance, silica. In reversed-phase chromatography, the opposite is true with polar mobile phases, i.e., water-based, to which methanol or acetonitrile (or mixtures) may be added, and apolar stationary phases, e.g., C18. A number of variations on these modes exist, including ion-pair and micellar chromatography (Lough and Wainer 1996; Dong 2006; Waksmundzka-Hajnos and Sherma 2011).

In modern drug discovery, reversed-phase LC is extensively used because of a more limited use of organic solvents and its compatibility with mass spectrometry (Watson and Sparkman 2007) (see further), which is progressively more frequently used as a detector in comprehensive metabolomics profiling.

The principle behind the chromatographic separation is that different compounds interact differently with the mobile and stationary phases. Compounds with a relatively higher affinity for the mobile phase will go rather fast through the column, while compounds with more affinity for the stationary phase will require more time to travel the same distance. As a result, the compounds are separated and elute from the column at different times. After the separation, a detector (see further) then continuously measures specific information, thus detecting what elutes from the column. Elution of only mobile phase results in a baseline signal. A deviation (peak) from the baseline signal indicates the presence of a compound (see Fig. 2). The graphic representation of the detector signal, which is called a chromatogram, shows a peak for each detected compound. Each peak has a maximum, which occurs at a time that is called the retention time of the corresponding compound.

In herbal analysis the composition of the mobile phase is often changed as a function of time, which is called gradient elution. As a result, each compound detaches from the stationary phase when the mobile phase reaches a composition necessary to overpower the compound's interaction with the stationary phase

(Lough and Wainer 1996; Dong 2006; Waksmundzka-Hajnos and Sherma 2011). The result is that mixtures of polar and apolar compounds can be separated and determined in one run. Optimizing chromatographic methods essentially means finding a set of conditions (mobile phase composition, stationary phase, gradient conditions, etc.) allowing an acceptable separation and a reasonable analysis time (Dejaegher et al. 2010; Alaerts et al. 2007). In drug discovery, in our opinion, priority usually is given to the quality of the separation, to allow a maximal separation of the compounds and their specific characterization.

Compounds may be separated by HPLC, when their interaction with mobile and stationary phases is sufficiently different. When the compounds are known, and when the sample is not too complex, the separation can be optimized by screening different stationary phase-mobile phase combinations to find the combination resulting in the best separation. However, when the compounds in a sample are not a priori known or very numerous (as is the case in drug discovery), even after separation optimization, it is often not possible to find mobile and stationary phase conditions where all compounds are completely baseline separated. Then the peaks of given compounds show (partial) overlap. The quality of separation between two consecutively eluting compounds can be quantified as the resolution (Lough and Wainer 1996). This parameter quantifies the ratio of the difference between the retention times and the sum of the peak widths of two consecutively eluting compounds. For a baseline separation of two peaks of similar height, a role of thumb specifies that resolution should be above 1.5.

The mobile and stationary phases; the mobile phase flow rate; the instrument itself; the column chemistry, dimensions, and brand; the size and shape of the stationary phase particles; and the column temperature all may influence the retention times of the compounds in a given mixture. The peak widths are also influenced by multiple factors, including column length, flow rate of the mobile phase, stationary and mobile phase chemistry, dimensions of the tubings, analysis temperature, and size of the particles. However, on a given system with a fixed set of conditions, the column properties play a key role in the observed peak width and thus the quality of the separation that can be achieved. A column's separation power is quantified as the efficiency and is expressed as the number of theoretical plates. The higher this number, the better the column is expected to perform, i.e., the more complex mixtures can be separated. The number of theoretical plates on a column increases when the theoretically defined height equivalent to a theoretical plate (HETP) decreases, and with increasing column length (Lough and Wainer 1996). Because of the complexity of herbal samples, the efficiency of a column is decisive for the quality of the fingerprint that is developed. Strategies to increase column efficiency are thus based on two principles: increasing the column length on the one hand and decreasing HETP on the other (Lough and Wainer 1996; Dong 2006; Waksmundzka-Hajnos and Sherma 2011).

The first option, however, is limited by another phenomenon, the back pressure that is generated by the column. Classical HPLC instruments and columns are built to withstand back pressures of up to 400 bar. Since increasing the column length fastly increases back pressure, this option is limited. However, a solution was provided with the development of monolithic columns. These columns have a rather

low efficiency, but they have larger pores and thus a low back pressure. Consequently, they can be serially coupled, which in the end results in more efficient separations (Dejaegher et al. 2010; Alaerts et al. 2007).

A second option is modifying the column properties to decrease HETP. The Van Deemter equation (HETP = A + B/u + Cu) is very helpful to rationalize this process (Dong 2006). It expresses HETP as the sum of three processes causing dispersion or band broadening: Eddy diffusion, due to different paths followed by molecules through the porous particles and represented by the A-term; longitudinal diffusion, leading to diffusion along the axis of the column and expressed by the B-term; and a C-term, reflecting the mass transfer of the analyte between the mobile and stationary phases. The Eddy diffusion is independent from the linear velocity, u, while the longitudinal is inversely correlated and the mass transfer term directly. These coefficients can be affected either by reducing the permeable zone in the particles (which has led to the development of superficially porous or core-shell particles) (Hayes et al. 2014; Guiochon and Gritti 2011) or by decreasing the size of the particles. Especially the latter option causes again an increase in the back pressure, which resulted in the development of ultrahigh-pressure liquid chromatography (UHPLC) instruments and columns, which can be used at pressures till 1000 bar (Waksmundzka-Hajnos and Sherma 2011). These latter techniques show improved efficiencies and have the advantage that in the same time more compounds can be separated, either a similar separation can be obtained in a shorter analysis time. Therefore, UHPLC has made its way into herbal fingerprint analysis, including drug discovery.

3.3 Detectors

3.3.1 Spectroscopic Detection

Both classical HPLC and its core-shell, monolithic, and ultrahigh-pressure variants are applied in drug discovery. The amount of information that is collected depends very much on the detector used (Lough and Wainer 1996). UV detectors are very popular. The simplest can measure only the absorbance at one tunable wavelength. The output generated is a chromatogram, showing a two-dimensional peak pattern for each sample along a time axis and an absorbance axis.

The chromatographic fingerprints are of interest in drug discovery and were applied, for instance, by Ben Ahmed et al. (2017) to indicate the potentially antioxidant compounds in *Pistacia atlantica* leaf extracts. Nevertheless, the information obtained is too limited to allow identification of the compounds. Additionally, when not all compounds are fully separated, it is difficult to estimate the contribution of each individual compound. Technological developments have resulted in the development of a more advanced type of detectors that can register a spectrum at each time point. These detectors are called diode-array detectors (DAD), and they generate a three-dimensional output, with a time, a wavelength, and an absorbance axis (Lough and Wainer 1996).

Even though DAD offers a UV spectrum for each peak, which can be used for identification purposes, UV spectra do not always allow the unambiguous identification of compounds. It often allows identifying the group of compounds, but since UV spectra of closely related compounds tend to be very similar, in essence, they lack the required specificity to truly identify the compounds (Lough and Wainer 1996). Techniques that provide highly specific identification are those that can distinguish between compounds with small structural differences. Mass spectrometry (Watson and Sparkman 2007) and nuclear magnetic resonance (Garrido and Beckmann 2013; Qin et al. 2009) are nowadays progressively used in drug discovery since they provide characteristic structural information. Fluorescence (He et al. 2013), evaporative light scattering (Alaerts et al. 2007), and electrochemical (He et al. 2013) detectors are also occasionally used when developing fingerprint profiles.

3.3.2 Mass Spectrometric Detection

Mass spectrometry allows measuring chemically charged species in the gas phase. A number of processes are necessary to obtain and analyze the charged species: ionization, mass analysis (including fragmentation possibilities), and detection.

Ionization

Since the output of the HPLC system is a liquid, an interface is required to bring the compounds from this liquid to a charged gas phase. In mass spectrometry various ionizers exist to make this transfer. In modern instruments, a popular choice is to use electrospray ionization. Its principle is to lead the LC eluent through a capillary with a nanoscale diameter, which can be electrically charged. Additionally, a temperature increase and a gas flow are applied. This leads to the formation of progressively smaller and smaller droplets of liquid that finally explode due to an overload in charges of the same polarity. As a result, the compounds in the sample get a positive or negative charge and are brought in the gas phase. Once these charged species are formed, they can be manipulated by changing the electric charge (or magnetic fields) in various parts of the mass spectrometer. In order to avoid interferences in these processes and to result in the highest signal, these manipulations need to be conducted in a vacuum environment, making a vacuum pump an essential part of any mass spectrometer.

Mass Analysis

Another crucial process in mass spectrometry is separating the charged species according to their mass/charge ratio. This is done in the mass analyzer. Again, various mass analyzers are available. Most can either be applied in a mode allowing to

scan charged compounds in a given m/z range or in a mode focusing on compounds with a specific m/z value. Quadrupole mass analyzers are extensively used, mainly to assay a number of known (targeted) compounds. These quadrupole analyzers contain four cylindrical rods on which charges are applied that are changed in polarity at a very high frequency. Charged species are attracted to the rods, followed by a repulsion when the polarity changes. For each m/z value, a given set of charges on the rods results in the charged compound passing through to the next zone in the mass spectrometer, which can be a collision cell, a second analyzer, or a detector. Other compounds with other m/z values either are attracted or repulsed by the charges and do not make it to the end of the analyzer. When such a quadrupole analyzer is applied in scanning mode, the charges on the rods are changed as a function of time to consecutively fulfill the requirements for all m/z values in a given m/z range. Some analyzers (so-called trap analyzers) are built in such a way that they even allow to immobilize selected compounds and eject them in a very precise way according to their m/z range. In principle, they also manipulate charged species through application of changes in electric fields in the mass analyzer, which is similar to the simple quadrupole analyzer.

An analyzer that does not allow the selection of ions with a specific m/z is the time-of-flight mass analyzer. In essence, it is a tube with a fixed distance. Since the m/z of an ion determines the time required to complete this distance, the m/z of the ion is determined from its travel time.

Fragmentation

When a compound is ionized through electrospray ionization, it gets either a positive or a negative charge, but it does not fall apart into fragments. From the obtained information, the molecular mass of small molecules can be determined. However, when fragmenting the ionized compound, a number of fragment ions are created, which are highly characteristic for a given compound. These fragment ions can then be used to elucidate the chemical structure of the compound. To create such fragments, energy is required. This energy is usually provided through a gas stream, which is kept at high energy in a collision cell. Through the invention of tandem mass spectrometry, nowadays, it is possible to select in a first analyzer a specific charged compound (called a mother or precursor ion), which can lead to the collision cell, where it is fragmented, followed by a separation of the fragment ions (also called product or daughter ions) in a second mass analyzer. The first mass analyzer is often a quadrupole; for the second several choices are available: another quadrupole analyzer or, for instance, a time-of-flight analyzer. Alternatively, in trap analyzers, the selection of the precursor ion and fragmentation and selection of the fragment ions can consecutively happen in one single space.

From MS Spectra to Compound Identification

Mass analyzers are usually grouped in low-resolution and high-resolution instruments. The difference is very important in drug discovery, since it determines how appropriate the instrument is for identification of unknown compounds. Low-resolution mass analyzers, for instance, quadrupole analyzers, can only determine the m/z value to unit value. For identification, this is often too limited, due to the occurrence of compounds that have the same mass at unit resolution but differ in the numbers behind the comma. Therefore, quadrupoles alone are usually not sufficient when the aim is to elucidate a compound's structure. This requires a high-resolution mass analyzer (for instance, time-of-flight or Orbitrap analyzer), which can determine a compound's mass, called the accurate mass, with a very good accuracy. Identification is based on the comparison of this measured accurate mass with the accurate masses of a range of compounds with some restrictions, for instance, on the number of carbon, hydrogen, and oxygen atoms. The differences of the measured accurate mass with the theoretical masses (also called the mass differences) are then determined and used as an identification criterion. Another aspect that is usually considered in compound identification is the correspondence of its isotopic pattern with that of candidate molecules. The most likely molecular formula is then obtained as the one with the smallest mass difference and the best correspondence in isotopic pattern (Watson and Sparkman 2007).

Commercial MS processing software packages (for instance, Masslynx and Excalibur) provided by instrument suppliers can be used in drug discovery. LC-MS data are processed in several steps. In a first step, the precursor ion data are processed in order to select characteristic marker compounds. In commercial software, the precursor ion LC-MS data of a sample is often presented as a list of intensities with given retention times and m/z values. Exploratory multivariate analysis techniques (for instance, principal component analysis) are used to visualize similarities and differences in the intensities at the observed retention time-m/z pairs in different samples. In a next step, multivariate discrimination models can be built, linking the intensity information of the samples to a given property. This property can be the plant species, plant part, growing region, harvesting season, or any other property that might relate to the chemical composition. The model's information is then used to determine m/z retention time pairs that are characteristic for each property, resulting in a list of characteristic marker compounds, for instance, to distinguish between two considered herbal species. The compounds are then tentatively identified based on the m/z values of their precursor ion.

Confirmation of the tentative identification of the compounds is done using the information obtained after fragmentation. This is done by comparison of the observed fragmentation pattern with the fragmentation pattern of a limited number of compounds with a good fit (corresponding isotopic pattern and low mass defect) for the precursor ion. This is done for each marker compound and often results in the identification of already known compounds. However, sometimes, potentially new compounds are detected, which can then be further characterized to confirm their chemical structure by complementary structural elucidation techniques, like NMR (Roessner and Dias 2013).

4 Pharmacological Evaluation

To use the metabolic profiling approach, the chemical analysis of a sample is linked to a biological activity. Although in vivo assays using rodents have already been used for this purpose (Cardoso-Taketa et al. 2008), it is not recommended. In vivo assays usually are less reproducible (Reardon 2016), a problem which will have impact on the mathematical model established to indicate the bioactive compounds. In fact, lack of reproducibility is a recurring concern in scientific publications (Plant et al. 2014; Munafò et al. 2017). In addition, in vivo assays go against the spirit of 3R (reduction, refinement, and replacement of animal experimental testing) that should be used in laboratories (Doke and Dhawale 2015). In this sense, in vitro assays can properly replace in vivo approaches in the screening of natural products.

Usually, in vitro assays are based on simple chemical reactions, such as antioxidant activity determination using 2,2-diphenyl-1-picrylhydrazyl (DPPH) (Ben Ahmed et al. 2017) or on a purified protein and molecule interaction, such as the inhibition of enzymes, e.g., of acetylcholinesterase (Klein-Júnior et al. 2016a). In addition, cell cultures can also be used as an alternative, such as in antiproliferative assays (Li et al. 2017).

As already highlighted in this chapter (see Sample Preparation), an important step in the metabolic profiling approach is the extraction procedure. This will also have impact on the pharmacological assays. Traditionally, in vitro assays are performed using organic-solvent extracts, such as ethanol or methanol, and/or organic-solvent mixture extracts, such as dichloromethane/methanol (1:1), and/or aqueous extracts (McCloud 2010). However, because of the extract chemical diversity, some problems are often faced, mostly related to pan-assay interference compounds (PAINS). These compounds are promiscuous artifacts, such as catechols, quinones, phenolic Mannich bases, and hydroxyphenylhydrazones (Baell 2016), that interact in a non-drug-like manner with the target (Baell and Walters 2014). In natural product research, the most usual problems are (i) tannins interaction with proteins, such as enzymes, leading to false-positive results; (ii) antagonism due to compounds with opposing pharmacological activities; (iii) active compound dilution in non-active compounds, not reaching a high enough concentration to elicit their activity in the extract; (iv) presence of naturally fluorescent and quenching products in the extract, which might affect fluorescence-based in vitro assays; (v) presence of colored compounds in the extract, which might affect colorimetric-based in vitro assays; (vi) micelle formation, due to amphipathic compounds, such as saponins, which might lead to cell death; and (vii) chelation of metals essential for biochemical assays, such as those estimating the activity of metalloproteases (Henrich and Beutler 2013).

To overcome these difficulties, some strategies can be applied. The most frequent is prefractionation, when complex extracts are semi-purified usually based on their polarity to obtain a simpler sample (Henrich and Beutler 2013). This can be achieved using liquid-liquid (LLE) or solid-phase extraction (SPE), as well as chromatographic methods. For a metabolic profiling approach, prefractionation by chromato-

graphic methods is less used since the extract composition could be oversimplified, losing some valuable information, especially related to natural product synergy. However, using LLE or SPE, it is possible to retain most of the extract complexity, while removing interfering compounds. Klein-Júnior et al. (2016a, b) used SPE cartridges to obtain an alkaloid fraction from *Psychotria nemorosa* which was able to inhibit both butyrylcholinesterase and monoamine oxidase B.

However, some compounds that initially can be considered as interfering, such as phenols, may also be relevant in some studies. Ben Ahmed et al. (2016) showed that phenolic compounds, such as galloylquinic acid and gallic acid, were the ones responsible for the antioxidant activity of extracts of *Pistacia atlantica.* Thus, prefractionation may then also not be a good approach. To avoid PAINS, some modifications to the in vitro methodology can be proposed, such as multiple concentration testing (also essential for IC_{50} estimation), choice of the method based on the extracts' fluorescent or colorimetric characteristics, and the use of additional agents to reduce aggregation and other non-specific bindings (Henrich and Beutler 2013). Butler et al. (2014) suggested the use of generic inactive extracts, obtained by the degradation of randomly selected specimen extracts, as negative controls. To predict the matrix effect, this inactive extract could be spiked with the positive control, giving additional information regarding PAINS or PAINS-like compounds in the extract.

In vitro assays have also their own limitations, which include the lack of bioavailability and metabolism information. In this context, in vivo experiments using alternative organisms could additionally be performed; *Danio rerio* (zebra fish), *Drosophila melanogaster* (fruit fly), and *Caenorhabditis elegans* (a nematode) are some alternative species that can be used in medium- to high-throughput approaches. However these organisms are not that frequently used for the metabolic profiling approach (Atanasov et al. 2015; Doke and Dhawale 2015).

5 Multivariate Data Handling of Chromatographic Fingerprints

After chemical analysis (fingerprint development) and pharmacological evaluation of the samples, both datasets must be linked in order to indicate the bioactive constituents in the extract. In this sense, the chromatographic fingerprints are subjected to a multivariate data handling procedure. However, raw fingerprint data usually need to be *pretreated* appropriately. *Peak alignment* is often necessary. It is imperative that peaks corresponding to the same molecule occur at the same retention time in different samples. However, in practice, retention time shifts are observed between analyses. These are due to several minor variations in experimental conditions, such as temperature, mobile phase composition, and flow rate (Alaerts et al. 2010a; Korifi et al. 2014).

Typically, *warping* is the most appropriate alignment methodology for chromatograms obtained from the same species. In warping, peaks are shifted, stretched, and/or compressed along the *x*-axis (time), without changing their sequence, aiming to enhance similarity between profiles (Bloemberg et al. 2013). In general, the profile to be aligned (*P*) is fitted to a target chromatogram (*T*) (Bloemberg et al. 2013). Different techniques can be used for warping, such as dynamic time warping (DTW), parametric time warping (PTW), and fuzzy warping (FW) (Alaerts et al. 2010a; Bloemberg et al. 2013).

One of the most used methods is correlation optimized warping (COW). COW was created by Vest Nielsen et al. (1998) and can be applied to either single or multitrace profiles. By this approach, both *P* and *T* are equally segmented, allowing each segment in *P* to be compressed or stretched to better resemble *T*. Two parameters must be optimized: the segment length *m* and the number of points *per* segment that a segment might be compressed or elongated, also called slack *t* (Bloemberg et al. 2013; Korifi et al. 2014). Klein-Júnior et al. (2016a), for instance, used this approach to warp UPLC-DAD data. In Fig. 3a, untreated data are observed, where shifts are quite visual in the chromatograms. This is reinforced by the correlation graphic (bottom). After warping (Fig. 3b), it is evident that the correlation was improved, highlighted by both the chromatograms and the plot.

Other preprocessing methods, with other goals, might also be used. In fact, *autoscaling* is often applied. It aims to make all metabolites equally important using standard variation as scaling factor. It can be calculated as follows:

$$\tilde{x} = \frac{x_{ij} - \overline{x}_i}{s_i}$$

where x_{ij} is the measured variable, $\overline{x}_i$ is the column mean, and s_i is the column standard deviation (Van den Berg et al. 2006).

Column centering is another pretreatment frequently used, since it removes the vertical offset from the data. It removes from each column the column mean, as follows:

$$\tilde{x}_{ij} = x_{ij} - \overline{x}_i$$

where x_{ij} is the variable and $\overline{x}_i$ is column mean (van den Berg et al. 2006).

Some other preprocessing techniques can also be applied, such as other scaling approaches, normalizations, and standard normal variate (SNV) (Van den Berg et al. 2006; Zeaiter and Rutledge 2009; Alaerts et al. 2010a). However, there is no recipe which method is best used, since each data has its own particularities. Moreover, a combination of preprocessing steps can also be applied, and its outcome should be evaluated to be able to select the best combination (Bloemberg et al. 2013; Gerretzen et al. 2015).

As a second step in the data treatment, *unsupervised data analysis* is performed to explore the data structure. It only takes into account matrix **X**, containing *p* fin-

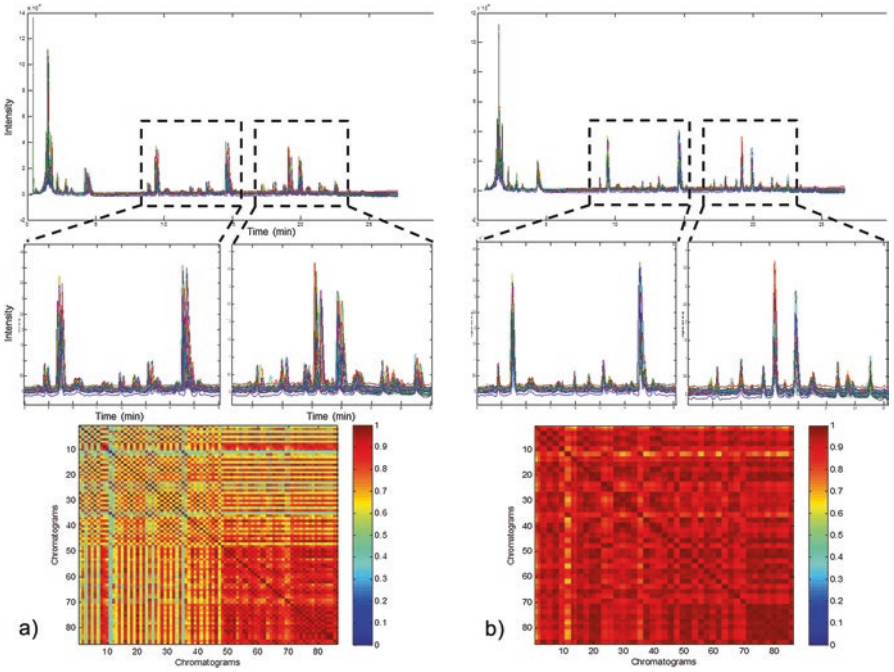

Fig. 3 Chromatographic fingerprints and correlation graphics (at the bottom) before (**a**) and after (**b**) warping, data obtained from the UPLC-DAD analysis of alkaloid fractions of *Psychotria nemorosa* (Klein-Júnior et al. 2016a)

gerprints, consisting of *n* measurement points. It aims to detect cluster formation tendency, giving a general idea of the data structure (Alaerts et al. 2010a; Goodarzi et al. 2013; Ren et al. 2015). For that purpose, two main techniques are used: *principal component analysis* (PCA) and *hierarchical cluster analysis* (HCA).

PCA is a technique that allows variable reduction, making it easier to visualize the data. It involves computation of new variables, known as principal components (PCs), orthogonal to each other, that retain most of the remaining variation information in the original or pretreated matrix **X**. The first PC explains most of the data variation, and each following PC contains less information regarding the variance in the dataset. The objects are then projected on (the first) two or three PCs and visualized in a score plot, used to observe cluster tendency in the samples (Fig. 4a), while the contribution of each original variable to the PC score can be visualized in a loading plot (Fig. 4b), used to evaluate the influence of each variable on the clustering (Esbensen and Geladi 2009; Goodarzi et al. 2013; Ren et al. 2015).

Another technique is HCA. It creates a dendrogram, which represents the cluster formation in the matrix **X**, organized as a hierarchical tree. Usually, the most similar objects are merged firstly. In an iterative process, objects are merged by similarity until all objects form only one cluster containing the entire dataset. This similarity is measured either as a distance, such as Euclidean distance and Mahalanobis dis-

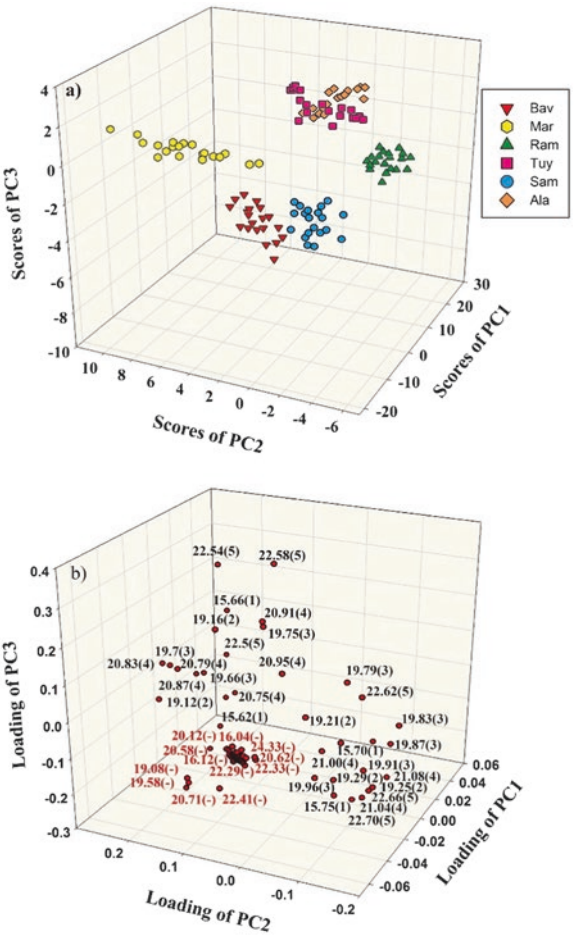

Fig. 4 Principal component analysis (PCA) score plot (**a**) and loading plot (**b**) for the fatty acid fingerprints of walnut samples from six regions in Iran (Esteki et al. 2017)

tance, or as a correlation, such as Pearson correlation coefficients. Another item that influences the dendrogram is the linkage function, which gives the similarity metric for pairs of groups (Fig. 5). Usually applied linkage functions include, among others, single linkage, unweighted average linkage, and centroid (Lee and Yang 2009; Goodarzi et al. 2013; Ren et al. 2015). HCA and PCA are commonly performed together to highlight cluster tendency in the dataset.

As a final step in the data handling for metabolic profile approach, *supervised data analysis techniques* are used. Specifically, *multivariate calibration* approaches are often applied. Then, both matrix **X** information (fingerprints) and the response vector **y** (biological activity, usually given as continuous values – e.g., IC_{50}) are used. Their main purpose is to build a regression model between the response (**y**) and the predictors (**X**). The regression coefficients of the model enable to indicate peaks in the fingerprint associated to the considered activity (Goodarzi et al. 2013; Ren et al. 2015). The model is described as:

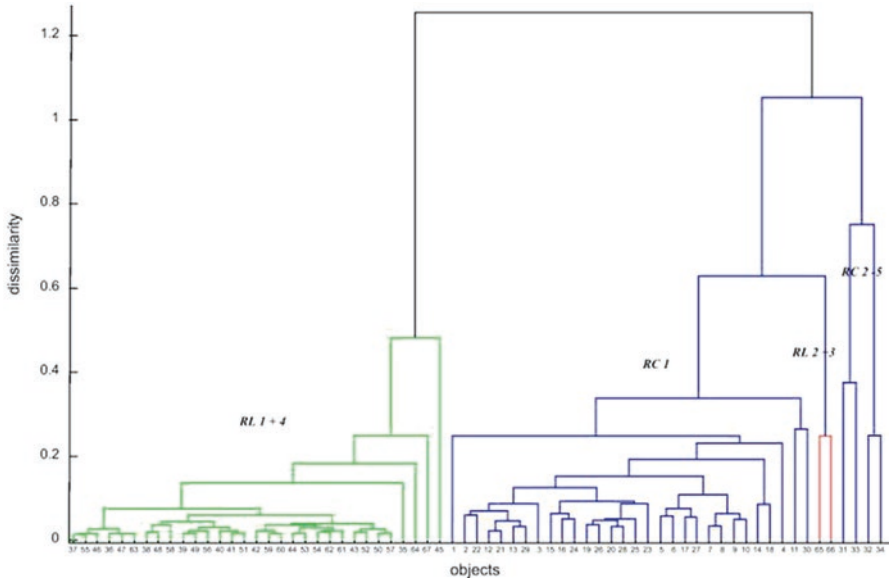

Fig. 5 Hierarchical cluster analysis (HCA) dendrogram for the fingerprints of two Chinese herbs: rhizome *Chuanxiong* (RC) and rhizome *Ligustici* (RL) (Alaerts et al. 2010b)

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{e}$$

where $\boldsymbol{b}$ is a $p \times 1$ vector of regression coefficients and $\boldsymbol{e}$ an $n \times 1$ residual vector (Tistaert et al. 2009).

Usually, a training set and a test set are applied, when enough samples are available. The training set is used to build a model and the test set to evaluate the predictive properties of the model. However, sometimes, not enough samples are available to establish a calibration and test set, which are both sufficiently representative. Therefore, in such case, almost the whole matrix **X** is used to build the model, and its predictive properties are evaluated using cross-validation (CV), in which the root-mean-squared error of cross-validation (RMSECV) is determined, often from one sample at the time that is left out (leave-one-out cross-validation). The model with the lowest RMSECV, found for different models constructed, usually is selected as the best (Alaerts et al. 2010a; Ren et al. 2015). RMSECV is calculated as follows:

$$RMSECV = \sqrt{\sum_{i=1}^{N} \frac{\left(\hat{y}_{CV,i} - y_i\right)2}{N}}$$

where N is the number of calibration samples, y_i the experimentally obtained response of the ith sample, and $\hat{y}_{CV,i}$ the response predicted by the model for the ith sample.

Another feature that is used to evaluate the model is the simplicity, which is directly correlated to the lower number of components. Finally, interpretability, related to the regression coefficients, is also important, since it must be easy to determine the contribution of the variables to the model (Alaerts et al. 2010a). Therefore, the number of (latent) variables in the model is a compromise between a model complexity and a low RMSECV.

Different techniques can be used to build the calibration models, such as stepwise multiple linear regression (stepMLR), principal component regression (PCR), partial least squares (PLS), and orthogonal projection to latent structures (OPLS) (Alaerts et al. 2010a). One of the most used methods is PLS. A PLS model can be written as:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\boldsymbol{T}} + \boldsymbol{E}$$

$$\boldsymbol{y} = \boldsymbol{T}\boldsymbol{P}^{\boldsymbol{T}}\boldsymbol{b} + \boldsymbol{f} = \boldsymbol{T}\boldsymbol{q} + \boldsymbol{f}$$

$$\boldsymbol{b} = \boldsymbol{P}\boldsymbol{q}$$

where $\boldsymbol{T}$ gives the score matrix of $\mathbf{X}$ and $\mathbf{y}$, $\boldsymbol{P}$ the loading matrix of $\mathbf{X}$ on $\boldsymbol{T}$ and $\boldsymbol{P}^{\boldsymbol{T}}$ its transposal, $\boldsymbol{E}$ the residual matrix, $\boldsymbol{b}$ the regression coefficients, $\boldsymbol{q}$ the loading vector of $\mathbf{y}$ on $\boldsymbol{T}$, and $\boldsymbol{f}$ the residual vector of $\mathbf{y}$ (Nguyen Hoai et al. 2009).

The overlay plot of the regression coefficients with the original chromatograms is important for peak indication. In this sense, if using IC_{50} as response, negative regression coefficient values indicate active peaks, since lower IC_{50} indicates higher activity. Based on this prediction, peaks that match with "negative peaks" on the regression coefficient plot indicate active compounds (Figs. 9 and 10). In this sense, hyphenated techniques are important for the early dereplication of the active extract. LC-MS and LC-NMR techniques avoid the tedious and costly isolation of already known bioactive compounds. However, if the compound had never been assayed for a given activity, its isolation is mandatory to confirm the model prediction. Usually, modern techniques, such as medium-pressure liquid chromatography (MPLC) and/or preparative HPLC, are used for this purpose. Finally, the chemical structure can be determined and its activity determined.

6 Applications of Metabolite Profile

6.1 Linked to In Vitro Reaction Assays

Kvalheim et al. (2011) applied PLS regression to model the ferric reducing antioxidant power (FRAP) assay value (as a response) as a function of the obtained fingerprints (gradient HPLC with UV detection at one selected wavelength) for 60 mixtures, prepared according to an experimental design approach (no extraction involved). The mixtures were composed of 12 compounds with antioxidant

properties, varying in compound-dependent concentration and composition. Forty mixtures were used as training set to build the model and the remaining 20 as test samples to validate the model. Except for two outlying mixtures in the training set, the prediction results obtained by PLS showed very good correspondence with the measured FRAP measurements, both for training and test set (not shown here; for the figure the reader is referred to Kvalheim et al. 2011).

The FRAP measurements of the pure compounds indicated that they could be sorted according to their decreasing antioxidant capacity, the most active compounds having a high FRAP value and vice versa. A typical fingerprint is shown in Fig. 6a. Three approaches were compared to retrieve the same sequence of the antioxidant potential. In a first approach, the regression coefficients of the PLS model were evaluated. Based on the regression coefficient plot, the most important peaks were indicated. The regression coefficient plot is shown in Fig. 6b. The size of the regression coefficients depended largely on the size of the peaks, making it impossible to rank the compounds according to their antioxidant capacity in the same sequence as obtained through the FRAP assay of the pure compounds. In the second approach (not shown here; for the figure the reader is referred to Kvalheim et al. 2011), target projection loadings (Kvalheim et al. 2011) were calculated and visualized in a profile. The size of the peaks resulted in less concentration-dependent conclusions; however it did not yet fully correspond to the sequence of the antioxidant capacity determined by FRAP. The third alternative multiplied the target projection loadings with the selectivity ratio (Kvalheim et al. 2011) and is shown in Fig. 6c. The size of the peaks in this profile showed the same trend as the FRAP results of the pure compounds. The third method showed the best correspondence with the results of the FRAP assay performed on pure standards.

As a result, the latter method is presented as a good basic procedure when the compounds are unknown, as is the case in the drug discovery context. Since in this study the compounds were all a priori known, UV spectra were sufficient for unambiguous identification. However, when the compounds would not be known in advance, techniques providing structural information (MS, NMR) would be necessary to identify the compounds after isolation or separation of the compounds, for instance, by means of liquid chromatography.

Xu et al. (2015) developed a data-driven method to determine the bioactive components in HPLC profiles of a set of Radix *Puerariae lobatae* samples and in a synthetic sample set. Like Kvalheim et al. (2011), they also focused on the antioxidant activity, measured by the FRAP assay. Fingerprint profiles were registered by gradient HPLC, and one detection wavelength was selected per sample set. Pretreatment of the HPLC fingerprint data was done using asymmetric least squares (background correction) and COW (retention time shift correction). They applied a variant on classic PLS modeling, called sure independence screening interval PLS (SIS-iPLS). This algorithm eliminates the variables that are only weakly correlated to the modeled response. The rationale for this method is that uninformative variables included in the PLS model, as well as the absence of important variables, will decrease the model's predictive properties (while the best model is more complex). The algorithm initially selects a number of m inter-

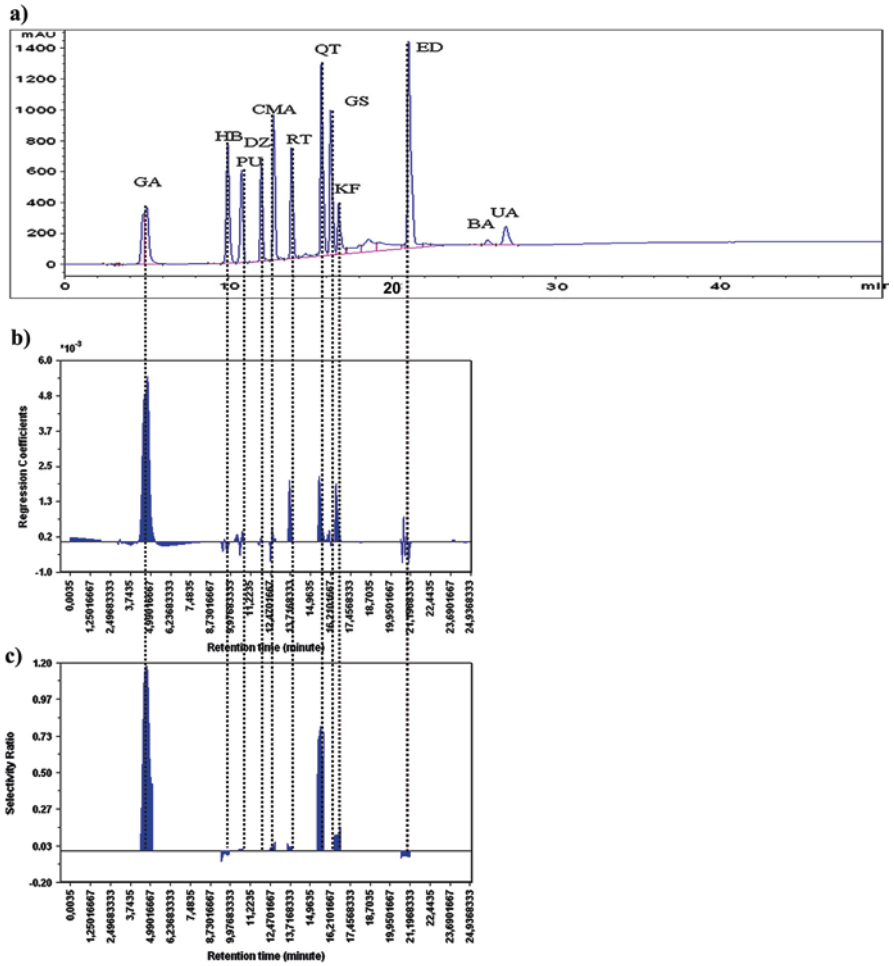

Fig. 6 (**a**) Typical fingerprint of the 12-compound mixture, (**b**–**c**) indication of important time zones in the PLS model by evaluation of the regression coefficients (**b**) and the multiplication of the selectivity ratio and the target projection loadings (**c**). Fine gray vertical lines highlight retention times of the 12 compounds. (Adapted from Kvalheim et al. 2011)

vals (coinciding with peaks in the chromatograms) which have the highest correlation with the FRAP value (see Fig. 7a).

The initially selected intervals are iteratively eliminated one by one (according to decreasing correlation with y), and their predictive value is assessed through the calculation of the RMSEP (root-mean-squared error of prediction) and plotted as a function of the eliminated intervals. Which variables are finally retained (Fig. 7c) is decided evaluating the minimum in this graph (Fig. 7b).

Next, PLS models were made leaving out one of the included components (intervals) and were compared to the performance of the PLS model with all components,

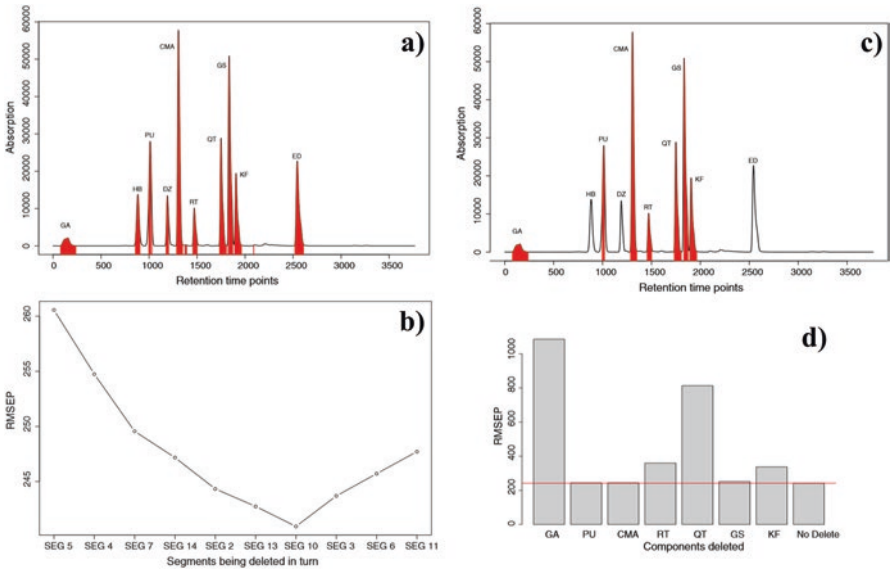

Fig. 7 SIS-iPLS procedure to link fingerprint information to the antioxidant activity. Initially selected m intervals (**a**): red peaks, RMSEP as a function of the eliminated intervals (**b**), the intervals maintained after iterative backward variable selection (**c**): red peaks, RMSEP increase for PLS models created after deletion of the specific intervals: *GA* gallic acid, *PU* puerarin, *CMA* coumaric acid, *RT* rutin, *QT* quercetin, *GS* genistein, *KF* kaempferol, No delete: model with all seven intervals (=reference level, indicated with the red line), (**d**) (Xu et al. 2015)

providing the reference RMSEP. The importance of each individual interval is assessed by the influence of eliminating it from the PLS regression model. The most important variables lead to the highest increase in RMSEP relative to the reference level (Fig. 7d).

For the synthetic sample set, the findings were checked against and confirmed the wet chemistry measurements. For the real herbal samples, the method's validity was studied differently. The intervals were compared to those obtained by other variable selection algorithms. The quality of all these models was compared by studying the distributions of the selected variables for the samples with the 30% lowest and 30% highest similarity values with the average profile and overlaying these with the FRAP distribution of the samples. Consequently, Xu et al. (2015) concluded that their method included more relevant variables than the other models. For the real plant samples, the study did not include identification of the compounds. This could be done by LC-MS or NMR, which provide more detailed structural information.

Ben Ahmed et al. (2016) studied the antioxidant compounds in *Pistacia atlantica* leaf extracts. The antioxidant activity of 28 samples was determined by two in vitro methods. The first assay used 2,2-diphenyl-1-picrylhydrazyl (DPPH), the other potassium ferricyanide (PFC). Both assays are complimentary, since they measure different aspects of antioxidant properties. Additionally, gradient HPLC profiles,

using two serially coupled monolithic columns, were developed for the samples. The raw HPLC profiles showed retention time differences, which were corrected with correlation optimized warping. This resulted in well-aligned HPLC profiles (see Fig. 8a). As a result, a matrix **X**, containing the (aligned) HPLC profiles of the 28 samples, and two column vectors, **y**, containing the antioxidant activities of the samples determined by DPPH and PFC, respectively, were obtained. Multivariate calibration models (PLS and OPLS) were built based on column-centered, normalized plus column-centered, and standard-normal-variate transformed plus column-centered fingerprint data and used to predict the antioxidant activities of the samples from their fingerprint data. The predictive properties of the models were assessed based on the correspondence of the predicted and the measured antioxidant activities for the samples and expressed as the root-mean-squared error of cross-validation (RMSECV). The best models are those with a low RMSECV value.

The models were mainly used to evaluate the contribution of sample compounds to the measured activities. The antioxidant activity of a sample can be calculated as the summed products of the measured detector signal at a given time point in a sample's fingerprint (depending on the sample and the time point) and the regression coefficient characteristic for that time point. The regression coefficients of the different time points also constitute a profile (see Fig. 8b–d for column-centered, normalized plus column-centered, and standard-normal-variate transformed plus column-centered fingerprints, respectively). The peaks in this profile will be high (in absolute value) for compounds that have a large influence on the antioxidant activity and are small or absent for those that are not important for the antioxidant activity. Depending on the direction (positive or negative peaks), the compound's increase relates to an increase or decrease of the antioxidant activity. Comparing the regression coefficient profiles of the two assays with the fingerprints allowed identifying compounds influencing mainly the PFC measurement, others mainly influencing the DPPH measurement, and a third group influencing both. The figures of the regression coefficients for the DPPH and PFC models, overlaid with the fingerprints, are shown in Fig. 8a, b, respectively.

Identification of the indicated compounds was done by LC-ESI-QToF-MS in negative ionization mode, registering MS data in MS^E mode. In this mode, low-energy precursor ion information is registered, next to fragment ion information. Combination of both precursor and fragment ion information allowed the authors to tentatively identify 12 of the 13 compounds, yielding one unidentified compound. Future research should focus on the isolation and structural elucidation of this compound using structural elucidation approaches.

6.2 Linked to In Vitro Enzymatic Assays

To indicate the peaks responsible for the inhibition of butyrylcholinesterase and monoamine oxidase-A in an alkaloid fraction of the leaves of *Psychotria nemorosa*, Klein-Júnior et al. (2016a) used OPLS models. These enzymes are related to the

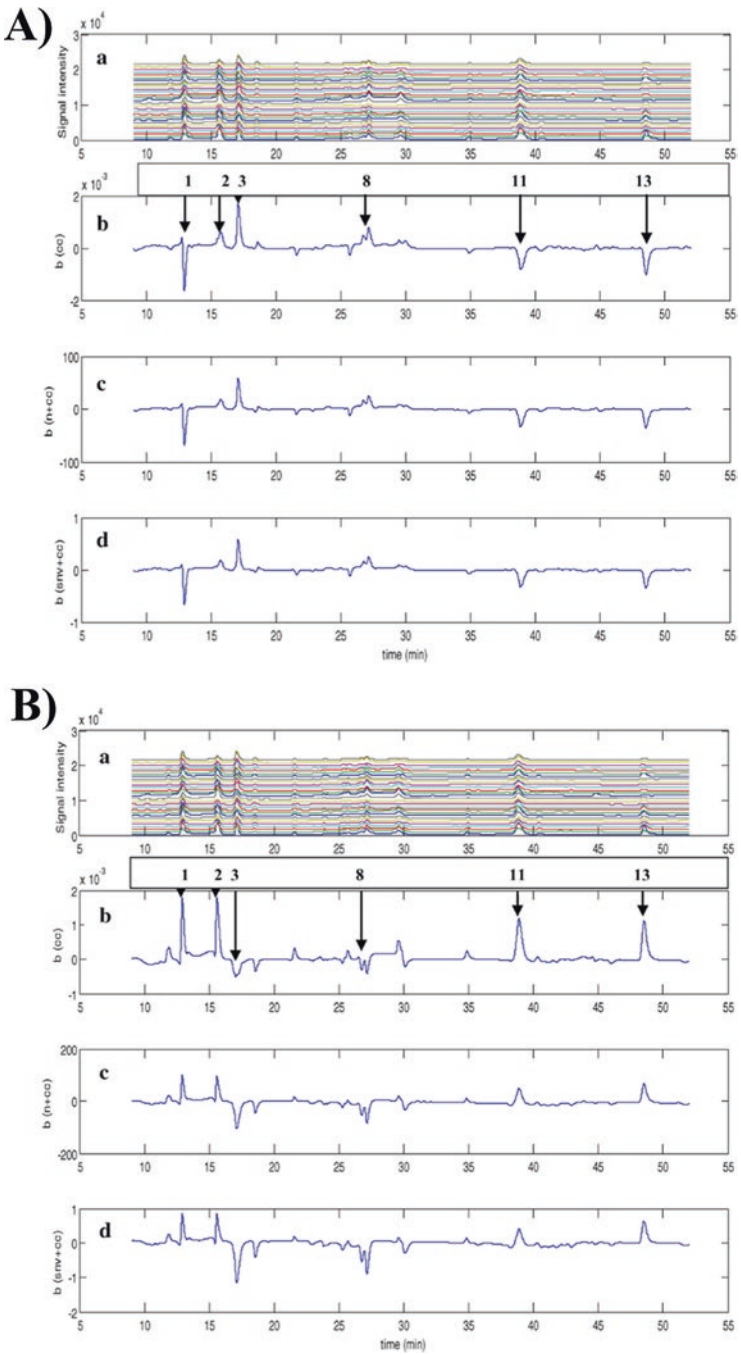

Fig. 8 OPLS regression modeling of the antioxidant activity determined using the DPPH (**a**) and PFC (**b**) methods. Preprocessed *Pistacia atlantica* fingerprints (a), OPLS regression coefficients obtained with column-centered (b), normalized plus column-centered (c), and standard-normal-variate transformed plus column-centered (d) fingerprint data. Numbers 1, 2, 3, 8, 11, and 13 refer to the peaks identified with LC-MS (Ben Ahmed et al. 2016)

remediation of the symptoms of neurodegenerative diseases, such as Alzheimer's and Parkinson's disease. Forty-three samples were collected from five different locations and were extracted using methanol-assisted micro-extraction by an ultrasonic bath. These extracts were fractionated by solid-phase extraction to elute alkaloid fractions. These fractions were evaluated for their modulation of enzymatic activity, and their chemical profiles were determined by ULPC-DAD, detected at 280 nm. The matrix **X**, consisting of 43 rows (samples) and 36,001 columns (time points), was submitted to COW alignment, followed by SNV and column centering. To model the activity as a function of the fingerprints, Klein-Júnior et al. evaluated different techniques: PLS and OPLS (Fig. 9). For both enzymes, the best results were obtained by OPLS, since it showed less noisy regression coefficients than PLS. OPLS removes the variation in matrix **X** that is not correlated to the response **y**. Four compounds were indicated as multifunctional, meaning able to inhibit both butyrylcholinesterase and monoamine oxidase-A (Fig. 10). Although these compounds were not isolated yet, their indole nature was confirmed by LC-SPE-NMR, and a fraction, enriched in these compounds (as well as others), was able to significantly inhibit the enzymatic activity.

Kang et al. (2013) applied the metabolic profiling approach to study the skin whitening effect of *Morus alba*. Since tyrosinase plays a pivotal role in the production of melanin pigment, its inhibition was measured to evaluate the in vitro effect of the *M. alba* at 10 μg/mL. The extracts were obtained from six different samples of the root bark using pressurized liquid extraction. Methanol, methanol/water (8:2, 5:5, 2:8,

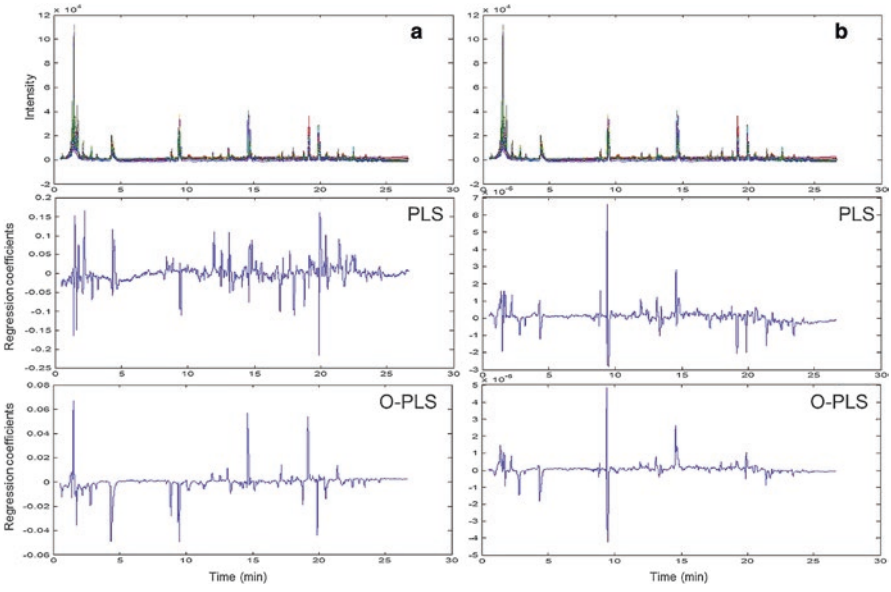

Fig. 9 Chromatographic fingerprints (top figure) and the regression coefficients from PLS and O-PLS models for butyrylcholinesterase inhibitory activity (**a**) and for monoamine oxidase-A inhibitory activity (**b**) (Klein-Júnior et al. 2016a)

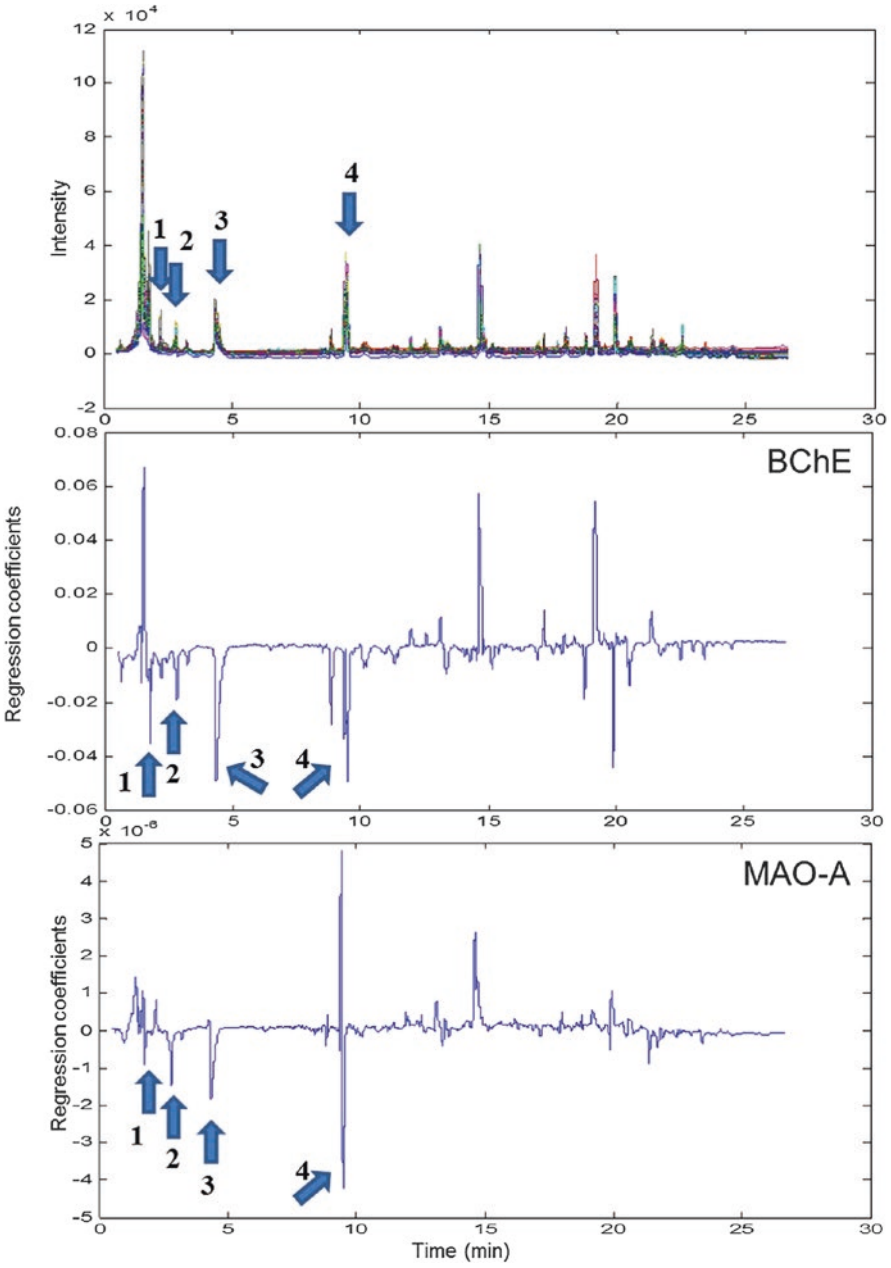

Fig. 10 Chromatographic fingerprints and the regression coefficients from O-PLS models for butyrylcholinesterase (BChE) inhibitory activity and for monoamine oxidase-A (MAO-A) inhibitory activity. The arrows indicate potentially multi-target compounds (Klein-Júnior et al. 2016a)

v/v), methanol/ethyl acetate (9:1, 7:3, v/v), and methanol/ethyl acetate/water (7:1,5:1,5, v/v) were used as solvents, totalizing 42 samples. The chemical analysis was performed by HPLC-DAD, giving chromatograms with 14,400 time points at 254 nm. Before linking fingerprint and bioactivity, the chromatograms were aligned using COW. Finally, using PLS modeling, it was possible to correlate the tyrosinase inhibitory effect of each extract to the chromatographic profile. To validate the model, two approaches were used: one where the samples were divided according to their origin (six groups) and the other one where the samples were divided according to the extraction solvent (seven groups). Then, a PLS model was created from a calibration set in which one of the groups was removed. The removed group, called the prediction set, was used to verify the prediction ability of the model. It was observed that, regardless of the source of difference between the samples, PLS was able to satisfactorily predict the activity of the extracts. Therefore, comparing the regression coefficient plot of the model with the chromatogram, it was possible to indicate that oxyresveratrol and its mono- and diglycosides, mulberrofuran G, kuwanon G, kuwanon H, and morusin exhibited positive regression coefficients, in different intensities, which are in line with previous studies, according to the authors.

Ben Ahmed et al. (2018) studied the leaves of *Pistacia atlantica*, an Algerian medicinal plant (previously studied for its antioxidant activity; Ben Ahmed et al. 2016), used for its antidiabetic and antihypertensive effects. The group evaluated the α-amylase, α-glucosidase, and angiotensin I-converting enzyme inhibitory effects of phenolic compound-enriched fractions. The plant material, collected in different periods and regions, was extracted with acetone/water (7:3) by maceration. After removal of acetone, the aqueous fraction was defatted, and ethyl acetate was used to obtain the evaluated fractions. The chromatographic fingerprints were obtained by HPLC-DAD, with a detection wavelength set at 254 nm. Prior to multivariate calibration, fingerprints were aligned by COW, and the matrix **X** was further pretreated by SNV followed by column centering. Twenty-eight samples were used to build a PLS model, using the inhibitory activity, expressed as $IC_{50,}$ as response **y**. Through evaluation of the regression coefficient plots and using LC-MS-based identification, glucogallin, quinic acid, and galloylquinic acid were indicated as α-amylase inhibitors; methyl gallate and tetragalloylglucoside as α-glucosidase inhibitors; gallic acid, gentisic acid, and digalloylquinic acid as α-amylase and α-glucosidase inhibitors; and glucogallin, gallic acid, galloylshikimic acid, methyl gallate, digalloylquinic acid, digallic acid, trigalloylglucose, and tetragalloylquinic acid as angiotensin I-converting enzyme inhibitors.

6.3 Linked to In Vivo Assays

Li et al. (2017) studied how in various fruit parts of *G. xanthochymus* differences in fingerprints obtained by UHPLC-QToF-MS could correlate with bioactivities measured with zebra fish models, in order to determine biflavonoids with anti-angiogenic activity. To prepare crude extracts, the pericarp, aril (= sheath that encloses the

seed), and seed of the fruits were separately extracted in a 70% methanol/H_2O solvent and then with a subsequent ultrasonic extraction and finally a combination of these extracted parts were put together. After redissolution in methanol, nine samples (three different collections) were analyzed using UHPLC. Zebra fish embryos were subdivided in six groups of twenty embryos that were treated with water, either containing extracts or not. Growth of the subintestinal vessels (SIVs) of the embryos were then microscopically investigated. Pericarp and aril extracts at different low concentrations show considerable length decreases in vessel growth. However, seed extracts have no significant influence. These results indicated that in pericarp and aril extracts, comparable bioactive components are expected to occur, different from these in seed. From the raw chromatographic data, retention time, exact mass, and ion intensity of the compounds in the fingerprint were used as variables for the chemometric study. Prior to this analysis, fingerprints were preprocessed, making use of peak identification, peak integration, and peak alignment. The obtained data was then further normalized.

The principal component analysis (PCA) score plot showed clustering of aril and pericarp samples, separated from the seed extracts. Calculation of goodness of fit and predictability of the PCA values stated that there were similar metabolic components in aril and pericarp. The PCA score plot clearly indicated that the chemical composition of the seeds is considerably different. This trend was also observed in the total ion chromatograms and the PCA loading plot, where differences could be seen in the variables. In the latter plot, marker ions from seed can be discriminated from those seen in the aril-pericarp. To compare the inactive seed group and the active aril-pericarp group, orthogonal partial least square-discriminant analysis (OPLS-DA) with scatterplot (S-plot) was applied.

Important marker ions that differentiate the two groups were determined. This way 13 markers, at the top of the S-plot, were indicated as candidates that potentially have anti-angiogenic activity. From these 13 components, two biflavonoids, xanthochymol and amentoflavone, were further studied for their potential bioactivity in zebra fish embryos. No effect of xanthochymol on the growth of SIV was seen. However, amentoflavone importantly decreased the vessel growth at certain concentrations. Moreover, both compounds showed downregulation of the expressions of certain genes. The study results show that amentoflavone has anti-angiogenic effects. Further study of seven other biflavonoids for their bioactivity-structure relation in the in vivo method displayed that only fukugetin inhibited the growth of SIVs, in other words had anti-angiogenic activity.

Wen et al. (2018) studied the targeted isolation and identification of compounds with potential antihyperlipidemic activity in crabapple using UHPLC-DAD-MS-SPE/NMR fingerprints, mice experiments, and PLS-DA. Twelve crude extracts from three crabapple species, four extracts of each variety, were prepared, making use of different solvents and solvent concentrations. The in vivo mice model was set up, consisting of eighteen groups (eight animals per group), three control groups and fifteen test groups each containing randomly chosen mice, in which obesity was induced. Twelve of the 15 groups were treated with (12 groups) or without (three groups) the 12 different extracts in order to investigate their possible

cholesterol-reducing effect. Only six of the 12 extracts showed antihyperlipidemic activity, so the extracts were classified in an active and an inactive group. In order to determine which compounds in the extracts correlate with the results of the mice experiments, the 12 samples were analyzed by LC-MS of which the raw data were preprocessed through alignment and autoscaling before chemometric analysis.

PCA, PLS-DA, and independent sample t-test of the UHPLC data were carried out. In order to generate appropriate classification models, the fingerprints were measured with + and − electrospray ionization. PCA positive and negative score plots clearly showed the two distinct groups. No outliers were observed; therefore all data were used for PLS-DA. Performing the latter analysis, significant values for goodness of fit and predictability were calculated, indicating a well-defined model to distinguish between active and inactive samples. After evaluation of the variables through Variable Importance in Projection, the Pearson correlation coefficient, and the p-value in the independent samples t-test, 22 differentiating variables were selected.

After automated MS-guided SPE trapping, ten enriched target compounds with potential cholesterol-reducing activity were obtained, eluted with deuterated methanol and introduced in an NMR system. By means of the obtained NMR and MS/MS spectra, seven compounds were identified by comparing with existing data from a local database, while the remaining three constituents went through the usual chemistry research channels for identification. Out of these ten constituents, six were already studied earlier and reported to have cholesterol-reducing capacities.

7 Conclusions

The study of plant extracts, aiming the discovery of new chemical entities with therapeutic potential, involves a complex process, merging areas, such as chemistry, pharmacy, biology, and pharmacology. To rationalize this intricate procedure, given methodologies have been defined, avoiding the trial-and-error approach. One methodology drawing the attention is the metabolic profiling approach, where the chemical fingerprint is correlated chemometrically with a specific activity, affording valuable information regarding the active compounds in a given extract.

As highlighted in this book chapter, although this methodology aims guiding (and theoretically, simplifying) the identification of the active compounds, several steps are important and highly influenced by different factors, demanding a well-designed experimental procedure. From sample preparation till data analysis, each step has a given number of possible approaches, making it challenging to choose the most appropriate combination.

The first important step is the plant-material extraction. It will directly impact the final result. Too complex extracts may contain several artifacts that may trouble the LC analysis, as well as impair the in vitro evaluation, mainly because of the presence of PAINS. In addition, greener and faster extraction techniques, such as supercritical fluid extraction, are preferred over the classical methods, e.g., maceration.

Therefore, in order to prevent future problems, the extraction procedure and, eventually, the cleanup methodology should be optimized, avoiding meaningless information to be obtained.

In the chromatographic analysis, usually reversed-phase LC is used. Even so, a huge number of possibilities can be screened, such as different stationary and mobile phases, flow rates, and detectors. This will be limited, in the end, mainly by the resources of the laboratory. However, detectors that can give structural information, e.g., mass spectrometry, may be recommended to indicate potentially active compounds, thus avoiding the re-isolation of already known (bioactive) secondary metabolites.

In parallel, the pharmacological evaluation methodology also plays an important role in the output. Classical in vivo experiments may not be a good choice, since they are less reproducible, weakening the mathematical model. Last but not least, ethical aspects must also be taken into account here. Therefore, in vitro assays are preferred, giving responses with lower error and providing a high-throughput possibility. However, because of its own limitations, e.g., lack of pharmacokinetic information, alternative organisms for in vivo assays have been proposed, such as zebra fish.

As a final step, chemometric tools are employed to link chemical and pharmacological information. In this sense, multivariate calibration techniques are used. However, most of the time, data pretreatment is a must. Peak alignment procedures, such as COW, are often applied. Then, unsupervised data analysis techniques, e.g., PCA and HCA, are used to detect outliers and to observe trends in the (pretreated) matrix **X**. Finally, in order to correlate the (pretreated) matrix **X** with the response **y**, calibration methodologies are applied. Usually PLS is used, often giving good prediction models. However, in some cases, variations of this method, such as OPLS, are used. Regardless of the method, the regression coefficient plot comparison to the LC fingerprints can be used to indicate the potentially active compounds. These might be identified by hyphenated techniques, e.g., LC-MS and LC-NMR. However, the isolation of the bioactive metabolites, if not already described in the literature, is mandatory to validate the prediction, as well as for identification purposes in case unknown compounds are involved.

In the end, the metabolic profiling approach works as a sieve. It retains (and eliminates) unimportant information (inactive compounds) and lets the important data (active compounds) pass over the modeled mesh. However, one must keep in mind that the mesh selectivity is based on earlier knowledge. Therefore, a poor matrix **X** will cause deformations in the sieve, retaining either more or less information than it should and allowing to pass either more or less important data. Ultimately, inactive compounds thus can be indicated, and valuable compounds could be missed. When it happens, and when it is recognized what is not evident, there is no other way than to “sift” everything again.

References

Alaerts G, Matthijs N, Smeyers-Verbeke J et al (2007) Chromatographic fingerprint development for herbal extracts: a screening and optimization methodology on monolithic columns. J Chromatogr A 1172:1–8

Alaerts G, Dejaegher B, Smeyers-Verbeke J et al (2010a) Recent developments in chromatographic fingerprints from herbal products: set-up and data analysis. Comb Chem High Throughput Screen 13:900–922

Alaerts G, Merino-Arévalo M, Dumarey M et al (2010b) Exploratory analysis of chromatographic fingerprints to distinguish rhizoma chuanxiong and rhizoma ligustici. J Chromatogr A 1217:7706–7716

Allard PM, Genta-Jouve G, Wolfender JL (2017) Deep metabolome annotation in natural products research: towards a virtuous cycle in metabolite identification. Curr Opinion Chem Biol 36:40–49

Atanasov AG, Waltenberger B, Pferschy-Wenzing EM et al (2015) Discovery and resupply of pharmacologically active plant-derived natural products: a review. Biotechnol Adv 33:1582–1614

Azmir J, Zaidul ISM, Rahman MM et al (2013) Techniques for extraction of bioactive compounds from plant materials: a review. J Food Eng 117:426–436

Azwanida N (2015) A review on the extraction methods used in medicinal plants, principle, strength and limitation. Med Aromat Plants 4(2–6):196

Baell JB (2016) Feeling nature's PAINS: natural products, natural product drugs, and pan assay interference compounds (PAINS). J Nat Prod 79:616–628

Baell JB, Walters MA (2014) Chemical con artists foil drug discovery. Nature 513:481–483

Begou O, Gika HG, Wilson ID et al (2017) Hyphenated MS-based targeted approaches in metabolomics. Analyst 142:3079–3100

Ben Ahmed Z, Yousfi M, Viaene J et al (2016) Antioxidant activities of *Pistacia atlantica* extracts modeled as a function of chromatographic fingerprints in order to identify antioxidant markers. Microchem J 128:208–217

Ben Ahmed Z, Yousfi M, Viaene J et al (2017) Seasonal, gender and regional variations in total phenolic, flavonoid, and condensed tannins contents and in antioxidant properties from *Pistacia atlantica* ssp. leaves. Pharm Biol 55:1185–1194

Ben Ahmed Z, Yousfi M, Viaene J et al (2018) Potentially antidiabetic and antihypertensive compounds identified from *Pistacia atlantica* leaf extracts by LC fingerprinting. J Pharm Biomed Anal 149:547–556

Bloemberg TG, Gerretzen J, Lunshof A et al (2013) Warping methods for spectroscopic and chromatographic signal alignment: a tutorial. Anal Chim Acta 781:14–32

Brusotti G, Cesari I, Dentamaro A (2014) Isolation and characterization of bioactive compounds from plant resources: the role of analysis in the ethnopharmacological approach. J Pharm Biomed Anal 87:218–228

Butler MS, Fontaine F, Cooper MA (2014) Natural product libraries: assembly, maintenance, and screening. Planta Med 80:1161–1170

Cardoso-Taketa AT, Pereda-Miranda R, Choi YH et al (2008) Metabolic profiling of the Mexican anxiolytic and sedative plant *Galphimia glauca* using nuclear magnetic resonance spectroscopy and multivariate data analysis. Planta Med 74:1295–1301

Choi YH, Verpoorte R (2014) Metabolomics: what you see is what you extract. Phytochem Anal 25:289–290

Ćujić N, Šavikin K, Janković T et al (2016) Optimization of polyphenols extraction from dried chokeberry using maceration as traditional technique. Food Chem 194:135–142

de O Silva E, Borges LL, da Conceição EC et al (2017) Box–Behnken experimental design for extraction of artemisinin from *Artemisia annua* and validation of the assay method. Rev Bras Farmacogn 27:519–524

Dary C, Baghdikian B, Kim S et al (2017) Optimization of ultrasound-assisted extraction of bioactive alkaloids from *Stephania cambodica* using response surface methodology. C R Chim 20:996–1005
Das AK, Mandal V, Mandal SC (2013) A brief understanding of process optimisation in microwave-assisted extraction of botanical materials: options and opportunities with chemometric tools. Phytochem Anal 25:1–12
David B, Wolfender JL, Dias DA (2015) The pharmaceutical industry and natural products: historical status and new trends. Phytochem Rev 14:299–315
De Monte C, Carradori S, Granese A et al (2014) Modern extraction techniques and their impact on the pharmacological profile of *Serenoa repens* extracts for the treatment of lower urinary tract symptoms. BMC Urol 14(63):1–11
Dejaegher B, Alaerts G, Matthijs N (2010) Methodology to develop liquid chromatographic fingerprints for the quality control of herbal medicines. Acta Chromatogr 22:237–258
Doke SK, Dhawale SC (2015) Alternatives to animal testing: a review. Saudi Pharm J 23:223–229
Dong MW (2006) Modern HPLC for practicing scientists. Wiley, Hoboken, 304pp
Esbensen KH, Geladi P (2009) Principal component analysis: concept, geometrical interpretation, mathematical background, algorithms, history, practice (Chapter 2.13). In: Brown SD, Tauler R, Walczak B (eds) Comprehensive chemometrics, vol 2. Elsevier, Amsterdam, pp 211–226
Esteki M, Farajmand B, Amanifar S et al (2017) Classification and authentication of Iranian walnuts according to their geographical origin based on gas chromatographic fatty acid fingerprint analysis using pattern recognition methods. Chem Intell Lab Syst 1217:7706–7716
Garrido L, Beckmann N (eds) (2013) New applications of NMR in drug discovery and development. RSC Publishing, Cambridge, UK, 530pp
Gauglitz G, Vo-Dinh T (eds) (2003) Handbook of spectroscopy. Volume 1. Wiley-VCH, Weinheim, 538pp
Gemperline E, Keller C, Li L (2016) Mass spectrometry in plant-omics. Anal Chem 88:3422–3434
Gerretzen J, Szymańska E, Jansen JJ et al (2015) Simple and effective way for data preprocessing selection based on design of experiments. Anal Chem 87:12096–12103
Goodarzi M, Russell PJ, Vander Heyden Y (2013) Similarity analyses of chromatographic herbal fingerprints: a review. Anal Chimica Acta 804:16–28
Guiochon G, Gritti F (2011) Shell particles, trials, tribulations and triumphs. J Chromatogr A 1218:1915–1938
Gunzler H, Williams A (eds) (2001) Handbook of analytical techniques. Wiley-VCH, Weinheim, 1182pp
Gupta A, Naraniwal M, Kothari V (2012) Modern extraction methods for preparation of bioactive plant extracts. Int J Appl Nat Sci 1:8–26
Hammi KM, Jdey A, Abdelly C et al (2015) Optimization of ultrasound-assisted extraction of antioxidant compounds from Tunisian *Zizyphus lotus* fruits using response surface methodology. Food Chem 184:80–89
Harvey L, Edrada-Ebel R, Quinn RJ (2015) The re-emergence of natural products for drug discovery in the genomics era. Nat Rev Drug Discov 14:111–129
Hayes R, Ahmed A, Edge T et al (2014) Core-shell particles: preparation, fundamentals and applications in high performance liquid chromatography. J Chromatogr A 1357:36–52
He Y, Wu Q, Hansen SH et al (2013) Differentiation of tannin-containing herbal drugs by HPLC fingerprints. Pharmazie 68:155–159
Henrich CJ, Beutler JA (2013) Matching the power of high throughput screening to the chemical diversity of natural products. Nat Prod Rep 30:1284–1298
Hill CB, Roessner U (2015) Advances in high-throughput LC-MS analysis for plant metabolomics. In: de Vooght-Johnson R (ed) Advanced LC-MS applications for metabolomics. Future Science Ltd, London, pp 58–71
Hubert J, Nuzillard JM, Renault JH (2017) Dereplication strategies in natural product research: how many tools and methodologies behind the same concept? Phytochem Rev 16:55–95
Huie CW (2002) A review of modern sample-preparation techniques for the extraction and analysis of medicinal plants. Anal Bioanal Chem 373:23–30

Izadiyan P, Hemmateenejad B (2016) Multi-response optimization of factors affecting ultrasonic assisted extraction from Iranian basil using central composite design. Food Chem 190:864–870

Kaiser KA, Barding GA Jr, Larive CK (2009) A comparison of metabolite extraction strategies for ^{1}H-NMR-based metabolic profiling using mature leaf tissue from the model plant *Arabidopsis thaliana*. Magn Reson Chem 47:S147–S156

Kang KB, Lee DY, Kim TB et al (2013) Prediction of tyrosinase inhibitory activities of *Morus alba* root bark extracts from HPLC fingerprints. Microchem J 110:731–738

Kim HK, Verpoorte R (2010) Sample preparation for plant metabolomics. Phytochem Anal 21:4–13

Klein-Júnior LC, Viaene J, Salton J et al (2016a) The use of chemometrics to study multifunctional indole alkaloids from *Psychotria nemorosa* (*Palicourea comb. nov.*). Part I: extraction and fractionation optimization based on metabolic profiling. J Chromatogr A 1463:60–70

Klein-Júnior LC, Viaene J, Tuenter E et al (2016b) The use of chemometrics to study multifunctional indole alkaloids from *Psychotria nemorosa* (*Palicourea comb. nov.*). Part II: indication of peaks related to the inhibition of butyrylcholinesterase and monoamine oxidase-A. J Chromatogr A 1463:71–80

Klein-Júnior LC, Vander Heyden Y, Henriques A (2016c) Enlarging the bottleneck in the analysis of alkaloids: a review on sample preparation in herbal matrices. Trends Anal Chem 80:66–82

Korifi R, Le Dréau Y, Dupuy N (2014) Comparative study of the alignment method on experimental and simulated chromatographic data. J Sep Sci 37:3276–3291

Kumar D (2016) Nuclear magnetic resonance (NMR) spectroscopy for metabolic profiling of medicinal plants and their products. Crit Rev Anal Chem 46:400–412

Kvalheim OM, Chan HY, Benzie IFF et al (2011) Chromatographic profiling and multivariate analysis for screening and quantifying the contributions from individual components to the bioactive signature in natural products. Chemom Intell Lab Syst 107:98–105

La Barbera G, Capriotti AL, Cavaliere C et al (2017) Liquid chromatography – high resolution mass spectrometry for the analysis of phytochemicals in vegetal-derived food and beverages. Food Res Int 100:28–52

Lee I, Yang J (2009) Common clustering algorithms (Chapter 2.27). In: Brown SD, Tauler R, Walczak B (eds) Comprehensive chemometrics, vol 2. Elsevier, Amsterdam, pp 557–618

Li FS, Weng JK (2017) Demystifying traditional herbal medicine with modern approaches. Nat Plants 3(17109):1–7

Li Y, Wu T, Zhu J et al (2010) Combinative method using HPLC fingerprint and quantitative analyses for quality consistency evaluation of a herbal medicinal preparation produced by different manufacturers. J Pharm Biomed Anal 52:597–602

Li JR, Li M, Xia B et al (2013) Efficient optimization of ultrahigh-performance supercritical fluid chromatographic separation of *Rosa sericea* by response surface methodology. J Sep Sci 36:2114–2120

Li P, Yue GGL, Kwok HF et al (2017) Using ultra-performance liquid chromatography quadrupole time of flight mass spectrometry-based chemometrics for the identification of anti-angiogenic biflavonoids from edible *Garcinia* species. J Agric Food Chem 65:8348–8355

Liu WJH (ed) (2011) Traditional herbal medicine research methods. Identification, analysis, bioassay, and pharmaceutical and clinical studies. Wiley, Hoboken, 466pp

Liu CT, Zhang M, Yan P et al (2016) Qualitative and quantitative analysis of volatile components of Zhengtian pills using gas chromatography mass spectrometry and ultra-high performance liquid chromatography. J Anal Methods Chem 2016:1–8

Lough WJ, Wainer IW (1996) High performance liquid chromatography: fundamental principles and practice, 1st edn. Blackie Academic and Professional, Glasgow, 288pp

Lu GH, Chan K, Liang YZ et al (2005) Development of high-performance liquid chromatographic fingerprints for distinguishing Chinese Angelica from related umbelliferae herbs. J Chromatogr A 1073:383–392

Martin AC, Pawlus AD, Jewett EM et al (2014) Evaluating solvent extraction systems using metabolomics approaches. RSD Adv 4:26325–26334

McCloud TG (2010) High throughput extraction of plant, marine and fungal specimens for preservation of biologically active molecules. Molecules 15:4526–4563
Munafò MR, Nosek BA, Bishop DVM et al (2017) A manifesto for reproducible science. Nat Human Behav 1:0021
Mushtaq MY, Choi YH, Verpoorte R et al (2014) Extraction for metabolomics: access to the metabolome. Phytochem Anal 25:291–306
Nguyen Hoai N, Dejaegher B, Tistaert C et al (2009) Development of HPLC fingerprints for *Mallotus* species extracts and evaluation of the peaks responsible for their antioxidant activity. J Pharm Biomed Anal 50:753–763
Pawar SV, Ho JC, Yadav GD et al (2017) The impending renaissance in discovery & development of natural products. Curr Top Med Chem 17:251–267
Plant AL, Locascio LE, May WE et al (2014) Improved reproducibility by assuring confidence in measurements in biomedical research. Nat Method 11:895–898
Qin HL, Deng AJ, Du GH et al (2009) Fingerprinting analysis of *Rhizoma chuanxiong* of commercial types using 1H nuclear magnetic resonance spectroscopy and high performance liquid chromatography method. J Integr Plant Biol 51:537–544
Raks V, Al-Suod H, Buszewski B (2018) Isolation, separation, and preconcentration of biologically active compounds from plant matrices by extraction techniques. Chromatographia 81:189–202
Rates SMK (2001) Plants as source of drugs. Toxicon 39:603–613
Reardon S (2016) A mouse's house may ruin studies. Nature 530:264
Ren S, Hinzman AA, Kang EL et al (2015) Computational and statistical analysis of metabolomics data. Metabolomics 11:1492–1513
Roessner U, Dias DA (eds) (2013) Metabolomics tools for natural product discovery: methods and protocols. Humana Press, New York, 311pp
Sasidharan S, Chen Y, Saravanan D et al (2011) Extraction, isolation and characterization of bioactive compounds from plants' extracts. Afr J Tradit Complement Altern Med 8:1–10
Shen B (2015) A new golden age of natural products discovery. Cell 163:1297–1300
Sherma J, Fried B (eds) (2003) Handbook of thin-layer chromatography, 3rd edn. Marcel Dekker, New York, 1331pp
Souza CRF, Bott RF, Oliviera WP (2007) Optimization of the extraction of flavonoid compounds from herbal material using experimental design and multi-response analysis. Lat Am J Pharm 26:682–690
t'Kindt R, Morreel K, Deforce D et al (2009) Joint GC-MS and LC-MS platforms for comprehensive plant metabolomics: repeatability and sample pre-treatment. J Chrom B 877:3572–3580
Tang TX, Guo WY, Xu Y et al (2014) Thin-layer chromatographic identification of Chinese propolis using chemometric fingerprinting. Phytochem Anal 2:266–272
Tistaert C, Dejaegher B, Nguyen Hoai N et al (2009) Potential antioxidant compounds in *Mallotus* species fingerprints. Part I: Indication, using linear multivariate calibration techniques. Anal Chim Acta 652:189–197
Tistaert C, Dejaegher B, Vander Heyden Y (2011) Chromatographic separation techniques and data handling methods for herbal fingerprints: a review. Anal Chim Acta 690:148–161
Tu Y, Jeffries C, Ruan H et al (2010) Automated high-throughput system to fractionate plant natural products for drug discovery. J Nat Prod 73:751–754
Universitätsbibliothek Leipzig (2016) Available at https://papyrusebers.de. Accessed Dec 2017
Van den Berg RA, Hoefsloot HCJ, Westerhuis JA et al (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 7:142
Vest Nielsen N-P, Carstensen JM, Smedsgaard J (1998) Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. J Chromatogr A 805:17–35
Vuckovic D (2012) Current trends and challenges in sample preparation for global metabolomics using liquid chromatography-mass spectrometry. Anal Bioanal Chem 403:1523–1548
Wagner H, Ulrich-Merzenich G (2009) Synergy research: approaching a new generation of phytopharmaceuticals. Phytomedicine 16:97–110

Waksmundzka-Hajnos M, Sherma J (eds) (2011) High performance liquid chromatography in phytochemical analysis. CRC Press, Boca Raton, 996pp

Wang D, Wang S, Du Q et al (2014) Optimization of extraction and enrichment of steroidal alkaloids from bulbs of cultivated *Fritillaria cirrhosa*. Biomed Res Int 2014:1–15

Watson JT, Sparkman OD (2007) Introduction to mass spectrometry: instrumentation, applications, and strategies for data interpretation, 4th edn. Wiley, Chichester, 864pp

Wen Y, Chen L, Li J et al (2014) Recent advances in solid-phase sorbents for sample preparation prior to chromatographic analysis. Trends Anal Chem 59:26–41

Wen C, Wang D, Li X et al (2018) Targeted isolation and identification of bioactive compounds lowering cholesterol in the crude extracts of crabapples using UPLC-DAD-MS-SPE/NMR based on pharmacology-guided PLS-DA. J Pharm and Biomed Anal 150:144–151

Wu H, Guo J, Chen S et al (2013) Recent developments in qualitative and quantitative analysis of phytochemical constituents and their metabolites using liquid chromatography-mass spectrometry. J Pharm Biomed Anal 72:267–291

Wu J, Yu D, Sun H et al (2015) Optimizing the extraction of anti-tumor alkaloids from the stem of *Berberis amurensis* by response surface methodology. Ind Crop Prod 69:68–75

Xu J, Xu QS, Chan CO et al (2015) Identifying bioactive components in natural products through chromatographic fingerprint. Anal Chim Acta 870:45–55

Yuliana ND, Khatib A, Verpoorte R et al (2011) Comprehensive extraction method integrated with NMR metabolomics: a new bioactivity screening method for plants, adenosine a1 receptor binding compounds in *Orthosiphon stamineus* Benth. Anal Chem 83:6902–6906

Zeaiter M, Rutledge D (2009) Preprocessing methods (Chapter 3.04). In: Brown SD, Tauler R, Walczak B (eds) Comprehensive chemometrics, vol 3. Elsevier, Amsterdam, pp 121–231

Zhu C, Liu X (2013) Optimization of extraction process of crude polysaccharides from pomegranate peel by response surface methodology. Carbohydr Polym 92:1197–1202