Unsupervised methods in LC-MS data treatment

Turova, Polina; Styles, Iain; Timashev, Vladimir; Kravets, Konstantin ; Grechnikov, Alexander ; Lyskov, Dmitry ; Samigullin, Tahir ; Podolskiy, Ilya ; Shpigun, Oleg ; Stavrianidi, Andrey

DOI:
10.1016/j.jpba.2021.114382

Citation for published version (Harvard):
Turova, P, Styles, I, Timashev, V, Kravets, K, Grechnikov, A, Lyskov, D, Samigullin, T, Podolskiy, I, Shpigun, O & Stavrianidi, A 2021, 'Unsupervised methods in LC-MS data treatment: application for potential chemotaxonomic markers search', *Journal of Pharmaceutical and Biomedical Analysis*, vol. 206, 114382. https://doi.org/10.1016/j.jpba.2021.114382

Unsupervised methods in LC-MS data treatment: application for potential chemotaxonomic markers search

Polina Turova*[a], Iain Styles[b], Vladimir Timashev[a], Konstantin Kravets[c], Alexander Grechnikov[c], Dmitry Lyskov[d], Tahir Samigullin[e], Ilya Podolskiy[f], Oleg Shpigun[a], Andrey Stavrianidi[a]

[a]M.V. Lomonosov Moscow State University, Faculty of Chemistry, 1-3 Leninskie Gory, Moscow, 119991, Russia
[b]University of Birmingham, School of Computer Science, Edgbaston, Birmingham B15 2TT, United Kingdom
[c]Vernadsky Institute of Geochemistry and Analytical Chemistry of the Russian Academy of Sciences, Kosygina 19, Moscow, 119991, Russia
[d]M.V. Lomonosov Moscow State University, Faculty of Biology, 1-12 Leninskie Gory, Moscow, 119234, Russia
[e]M.V. Lomonosov Moscow State University, Belozersky Institute of Physico-Chemical Biology, 1–40 Leninskie Gory, 119234, Moscow, Russia
[f]Bruker Ltd., Pyatnitskaya 50/2 build. 1, Moscow, 119017, Russia

*Corresponding author. Address: Lomonosov Moscow State University, Faculty of Chemistry, 1-3 Leninskie Gory, Moscow, 119991, Russia; email: turova.polina@gmail.com; phone: +79854722433

Abstract

The combination of Liquid Chromatography and Mass Spectrometry (LC-MS) is commonly used to determine and characterize biologically active compounds because of its high resolution and sensitivity. In this work we explore the interpretation of LC-MS data using multivariate statistical analysis algorithms to extract useful chemical information and identify clusters of similar samples. Samples of leaves from 19 plants belonging to the Apiaceae family were analyzed in unified LC conditions by high- and low-resolution mass spectrometry in a wide range scan mode. LC-MS data preprocessing was performed followed by statistical analysis using tensor decomposition in the form of Parallel Factor Analysis (PARAFAC); matrix factorization following tensor unfolding with principal component analysis (PCA), independent component analysis (ICA), non-negative matrix factorization (NMF); or unsupervised feature selection (UFS). The optimal number of components for each of these methods were found and results were compared using four different metrics: silhouette score, Davies-Bouldin index, computational time, number of noisy components. It was found that PCA, ICA and UFS give the best results across the majority of the criteria for both low- and high-resolution data. An algorithm for biomarker signal selection is suggested and 23 potential chemotaxonomic markers were tentatively identified using MS^2 data. Dendrograms constructed by the methods were compared to the molecular phylogenic tree by calculating pixel-wise mean square error (MSE). Therefore, the suggested approach can support chemotaxonomic studies and yield valuable chemical information for biomarker discovery.

Keywords

Liquid chromatography, mass spectrometry, machine learning, Apiaceae, multi-way data

1. Introduction

In recent years many approaches for the investigation of plants' taxonomy have been developed. These includes morphological, anatomical and chemotaxonomic classification. Chemotaxonomy is used for the classification of plants on the basis of their chemical composition[1]. The main task of this approach is to search for primary and secondary metabolites and on the basis of their presence or concentration create new classifications and reveal their relation to the molecular phylogeny classification. In previous works it was shown that for the Rutaceae family such markers are coumarins[2]. Coumarins are secondary metabolites which are considered as a chemical defense against predators and their content depends heavily on the growing conditions. In previous works[3] it was shown that some coumarins can appear or disappear from the chemical composition depending on the variety of conditions: geographical origin of the plants; environmental conditions (climate, pollution, light irradiation, etc); physiological variations (stage of development of the plant organ, plant part used, etc.); sample storage conditions and many others. Plants from the Apiaceae family are also rich sources of biologically active compounds such as coumarins and they are useful as food and flavoring and possess diverse pharmacological activities[4]. For the chemotaxonomic markers search it is necessary to use highly precise analytical methods as chromatography, mass spectrometry, nuclear magnetic resonance and complex statistical algorithms.

Liquid chromatography coupled with mass spectrometry (LC-MS) provides rich information about biological samples and is widely used in plant extract analysis. One of the major difficulties in the LC-MS method is that raw data, which is naturally structured as a 3D array, is difficult to interpret manually and automated analysis methods are needed to extract the most important information. In popular metabolomics approaches, “peak picking” software (e.g. MZmine, XCMS) and peak alignment algorithms [5,6] are widely used to reduce the 3D dataset to a set of peaks determined, by some means, to be the most informative. The first general problem in this approach is that some information will inevitably be lost because many peaks are discarded. The second problem is that other methods of analysis which may be more informative have not been fully investigated. They typically involve decompositions of the data into a set of factors which may be more easily interpretable.

Data decomposition methods typically belong to two classes[7]. In the first, the 3d array of LC-MS data is treated “as is” and tensor decomposition methods are applied. The most widely used 3D tensor decompositions are Parallel factor analysis (PARAFAC) and Tucker decomposition, both of which decompose a tensor into a set of matrices. These methods have previously been applied to different types of mass spectrometry data [8]. The alternative approach is to unfold the 3D data into a 2D array by reshaping a tensor of size X×Y×Z into a X×N matrix (where N = Y×Z) which can then be factorized using a wide range of techniques for matrix factorization. In LC-MS the dimensions that are combined in reshaping are retention time and m/z values. Tensor unfolding for LC-MS data is not widely described in the literature, but it has been successfully used for mass spectrometry imaging (MSI) data[9]. Unfolding data in this way opens up a much wider range of potential factorization methods, but it has the disadvantage of combining two orthogonal dimensions, which may remove some of the data’s structure and information content.

A second choice that must be made is whether subsequent chemometric analysis is supervised or unsupervised. In general, unsupervised techniques are applied (following any necessary preprocessing) when there is no or little prior knowledge about samples; or unobvious patterns are expected to be revealed; or when the goal is to identify which intrinsic (latent) factors are responsible for the greatest variability in the data. Results of unsupervised approaches are therefore typically most suitable for the discovery of markers present in significant concentrations. Lower abundance analytes can most reliably be identified using a supervised knowledge-based approach or by informed selection of specific areas/windows of the dataset[10]. A wide range of unsupervised approaches have been considered in the literature for applications including dimensionality reduction; resolution of samples; biomarker discovery; outlier identification; interference identification[9]. Among the most common unsupervised methods applied to mass spectrometry data are principal component analysis (PCA), independent component analysis (ICA), and non-negative matrix factorization (NMF) [11].

There have been only a few attempts to directly compare unsupervised treatment of mass spectrometry data. Different approaches to chemometric analysis of LC-MS data have been compared by classification accuracy, computational time and F1 score [12], but this work used a specific preprocessing protocol in which only the data points with the highest intensity within each peak were retained. In other work[9] different unsupervised treatment methods applied to MALDI imaging MS data were compared. It was shown that NMF and ICA produced components which mapped the spatial distribution of molecules and for which the associated spectra featured lower noise.

The aim of the present study was to compare the possibilities of different unsupervised factorization approaches in LC-MS data treatment, and to discover potential chemotaxonomic markers for 19 plant species from Apiaceae family. Data was acquired from the samples on two instruments and two types of data were investigated: LC-MS with low resolution MS (LRMS) and with high resolution MS (HRMS). In both cases raw LC-MS data was recorded in tensor form with three dimensions corresponding to samples, retention time, and mass to charge ratio. Different data factorization techniques were applied to the data using two general approaches: direct decomposition of the 3d tensor, and decomposition of the unfolded tensor

(Fig. 1). For direct tensor factorization we used non-negative PARAFAC decomposition. On unfolded tensors, we applied a range of dimensionality reduction and feature selection methods. For dimensionality reduction PCA, ICA and NMF were used; for feature selection variance-based feature filtering was employed.

For this research we used a dataset consisting of 57 samples from 19 plants belonging to the Apiaceae family and representing 7 genuses: Prangos, Ferulago, Cachrys, Bilacunaria, Diplotaenia, Azilia and Seseli (Table 1). Application of unsupervised methods to such a diverse dataset can reveal the most variable chemotaxonomic markers of this family.

2. Experimental

2.1. Instrumentation

The LC-LRMS apparatus consisted of a HPLC Thermo Scientific Dionex Ultimate 3000 (MA, USA) system with a binary analytical pump, an automatic sample injector coupled on-line with AB Sciex Qtrap 3200 (ON, Canada) mass spectrometer with an electrospray ionization interface. The column effluent was analyzed by ESI-MS in positive ion mode and the mass spectra were acquired and processed using the Analyst software (version 1.5) provided by AB Sciex. For the MS, the following conditions were used: ion spray voltage: 5500V; ion source heater temperature: 350°C; entrance potential: 10 V; declustering potential: 40 V; mass range 100-1200 Da.

The LC-HRMS apparatus consisted of a Thermo Scientific Accela HPLC system (CA, USA) coupled on-line with Orbitrap Exactive mass spectrometer (Dreieich, Germany). The column effluent was analyzed by HESI-MS in positive ion mode and the mass spectra were acquired and processed using the Xcalibur™ Software (version 2.2) provided by Thermo Scientific™. For the MS, the following conditions were used: spray voltage: 3.90 kV, capillary temperature: 300 °C, capillary voltage: 50.0 V, tube lens voltage: 100.0 V, skimmer voltage: 20 V, heater temperature: 350 °C, resolution 35 000, mass range 100-1200 Da.

In both LC-LRMS and LC-HRMS experiments the HPLC separation was conducted on a C18 column (Acclaim RSLC 2.1×150 mm, 2.2 μm) at a flow rate of 0.35 mL/min and oven temperature 35 °C. Two solvents were used: (A) 0.5% HCOOH aqueous solution and (B) MeCN. The gradient was as follows: 0 – 3 min 10 % B; 3 – 20 min linear gradient from 10 to 95 % B; 20 – 22 min 95 % B; 22 – 22.2 min linear gradient from 95 to 10 %; 22.2 – 27 min 10 % B.

Biomarker identification was performed on a Bruker Elute LC system coupled on-line with a Bruker Impact II high-resolution Quadrupole Time-of-Flight Instrument. HPLC separation was conducted on a C18 column (Intensity Solo 1.8 2.1× 100 mm) at a gradient flow rate (from 0.200 to 0.480 mL/min). Two solvents were used: (A) 5 mM Ammonium Formate and 0.01 % FA in MeOH:H_2O 1:99 mixture with and (B) 5 mM Ammonium Formate and 0.01 % FA in MeOH. The gradient was as follows: 0 – 0.1 min 4 % B; 0.1 – 1 min linear gradient from 4 to 18.3 %; 1 – 2.5 min linear gradient from 18.3 to 50 % B; 2.5 – 14 min linear gradient from 50 to 99.9 % B; 14 – 16 min 99.9 % B; 16 – 16.1 min linear gradient from 99.9 to 4 % B; 16.1 – 20 min 4 % B.

2.2. Materials and reagents

Deionized water was from a Milli-Q system from Millipore (MA, USA); HPLC-grade acetonitrile was purchased from Panreac (Barcelona, Spain) and >99.8% pure ethanol was from Sigma-Aldrich (Steinheim, Germany); Formic acid >99.9% purity was purchased from Acros (Geel, Belgium); MeOH >99.9% purity was purchased from Burdick & Jackson (Seelze, Germany) and Ammonium Formate ≥99.0% purity was purchased from Sigma-Aldrich

(Steinheim, Germany). Plant material was collected by botanists from Lomonosov Moscow State University.

2.3. Sample preparation

Plant material was collected in Iran, Portugal, Kyrgyzstan, and Uzbekistan in 2013 – 2019 and housed in Moscow University Herbarium (MW) or in the private collection of Dmitry Lyskov (information about herbarium specimens is available at https://plant.depo.msu.ru/). All plant specimens used for the analysis are listed in Table 1. Material was dried; extracts were prepared by weighting 0.01 g of a plant sample, adding 1 mL of methanol:water (3:1, v/v) mixture and extracting in an ultrasonic bath for 30 minutes, all extracts were prepared in three replicates. Extracts were centrifuged and diluted by a factor of ten with 10% aqueous acetonitrile. 2 mg/mL solution of eleutheroside B was used as an internal standard (IS) and 20 µL of this solution was added to each sample. For most of the plants, leaf samples were used, but for some samples leaves were missing and stems were used instead.

2.4. Software and packages

All LC-MS files were converted into mzXML format using MSConvert from ProteoWizard Tools. Data analysis was performed in Python 3 using the following modules: pymzML for mzML data files parsing[13]; scipy.signal for signal smoothing; pandas for arrays pretreatment; tensorly for PARAFAC decomposition; scikit-learn for PCA, ICA, NMF, UFS algorithms and performance metrics; matplotlib for data visualization, biopython for hierarchical and molecular phylogenetic trees visualization. Corcondia criteria and explained variance for PARAFAC models were calculated in MATLAB using the N-way toolbox. For dendrogram construction unweighted pair group method with arithmetic mean was used to cluster objects. Minkowski distance was used as a metric to evaluate object similarity. For phylogenetic tree ‘identity’ model for distance calculation was employed. All files from LC-MS analysis in mzML format and implemented algorithms are available at the github repository (https://github.com/turovapolina/unsupervised-LC-MS-data-treatment).

3. Results and discussion

3.1. Data acquisition and preparation

The model dataset consisted of 57 samples which are three replicates of 19 plant species which represents 7 genera from Apiaceae family. As a preliminary step, four extraction systems of methanol, water and dichloromethane mixtures were tested to maximize the signal[14]. The methanol:water (75:25, v/v) systems provided maximum peak capacity and the highest intensities in the chromatograms of all samples. Composition of the mobile phase was varied in a wide range during the gradient program in order to elute both polar and non-polar compounds and resolve as many distinct peaks as possible. MS data was collected in scan mode in the range 100–1200 m/z. All samples were analyzed in the same chromatographic conditions by LC-LRMS and LC-HRMS.

LC-LRMS data treatment. For mass chromatogram smoothing continuous wavelet transform, Baseline Estimation and Denoising With Sparsity (BEADS) approach and Savitzki-Golay filter were assessed with different parameters[15]. The goal was to choose a smoothing algorithm and associated parameters which will work successfully, i.e. smooth as much noise as possible but preserve peak shapes, on all mass chromatograms. In particular, the percentage of acetonitrile in the mobile phase was found to increase noise, and the smoothing algorithm should be capable of removing this noise. The optimal method was found to be a Savitzki-Golay filter with window size 13 and polynomic order 1, results of its implementation for both noisy and informative chromatograms (with or without distinct peaks) are shown in Fig. 2 (A,B). The step between time points varies across samples between 0.02 and 0.04 min and the time axis was linearly interpolated with a step size of 0.05 min in order to unify the time axis. A final time scale with a range from 2 to 22 minutes was chosen in order to disregard unretained compounds at the beginning of the chromatogram, very noisy signals at high percentages of acetonitrile and reequilibration time at the end of the chromatogram. For the mass axis unification, intensities for signals with residual masses in the range from -0,35 to +0,65 were summed and assigned to a cell with the corresponding integer m/z value. Data from all samples were combined into one tensor with dimensions 57 × 380 × 1200 corresponding to number of samples, number of retention time points and number of m/z values respectively.

LC-HRMS data treatment. A significant challenge when dealing with HRMS data is that the instrument is able to separate ion peaks with an m/z difference of 0.00001 which means that theoretically the spectrum may contain up to 100000*1200 (mass range) components. In reality, each time point of each sample had about 15000 ion peaks in the spectrum, but only a small portion of them had significant intensities. Thus, to reduce computational costs only signals with intensities higher than 5 % of the highest peak in the spectrum were selected (shown in Fig. 2 (C,D)). After the elimination of weak and noisy signals 40-60 important peaks were left in each spectrum. A dataframe containing the first timepoint of the first sample was created and was filled sequentially by all subsequent time points from all samples. When an m/z was found that had not been seen in previous time points, a new column was created and filled with zeros for all preceding rows. This procedure constructs a unified mass scale across all time points and all samples. To assess both environment-dependent and instrument-dependent fluctuations in measured masses and retention times, an internal standard (IS) was added to each sample. The mean absolute error (MAE) of the IS measured mass (m/z 395.13180) across all samples calculated for inter-day measurements was less than 0.005 Da. The MAE is greater for bigger masses, therefore it was decided to set the m/z window size equal to 0.01 Da. At the next step intensities of m/z signals in the dataframe which have mass differences lower than 0.01 Da were considered to results from for one m/z and summed. Cells with missing m/z signals were replaced by zero values. Finally, due to low reproducibility of retention times probably caused by the unstable performance of the LC pump, time periods of length 0.5 minutes were used instead of a continuous time axis. Unlike the LRMS dataset where the m/z scale interval is constant, the size of the final dataframe for HRMS data will depend on the number of unique m/z values observed in the particular dataset. However, the approach adopted here can be applied to data produced by any HRMS system. The final array was reshaped into a tensor with dimensions 51 × 45 × 2580 with the same axes as the LRMS data tensor.

3.2. Chemometric analysis

The obtained tensors were either directly subjected to PARAFAC decomposition or unfolded into a 2D array. The unfolding procedure takes a tensor of dimensions I×J×K and rearranges it in such a way that the number of samples I remains unchanged and two other dimensions (m/z and retention time (RT)) are combined into a single new dimension with size J×K. Therefore, the new feature space consists of the concatenation of the mass spectra and retention time pairs for each sample. PCA, ICA, NMF and UFS methods were applied on data organised by this approach and compared with the direct tensor decomposition. For PARAFAC, PCA, ICA and NMF a critical parameter is the number of components, which was chosen based on statistical analysis of each method without using any prior information about dataset.

For the PARAFAC one- to fifteen-component PARAFAC models with non-negative constrains were fitted to each dataset; the explained variance, corcondia criteria, error and number of iterations for all models were compared and finally the optimal number of components was selected to find the best balance across the criteria as shown in Fig. S1 (A, B)[16]. The choice of the number of components was validated by half-split analysis (Fig. S1 C,D). Results of hierarchical clustering analysis following PARAFAC and all other methods are presented in the Supplementary Information (Fig. S3 – S12). For PCA we selected the number of components that was sufficient to explain 95% of the variance in the data, which was 13 components (for LRMS) and 10 components (for HRMS). For the determination of the number of ICA components, ICA-by-blocks method was used[17]. In Fig. S2 (A,B) signal-correlation plots for LRMS and HRMS datasets are presented. It can be seen that for both LRMS and HRMS data after extracting more than 4 components, the curves decrease progressively which means that the correlations between the components of the different blocks are much lower. Thus, the optimal number of components in those datasets is 4. To identify the optimal number of components (rank of the matrix factors) for NMF, the residual sum of squares (RSS) was calculated and its correlation with the number of components was visualized, as shown in Fig. S2 (C,D). The optimal number was decided using a previously suggested method[18] of identifying where the graph of RSS against the number of components shows an inflection point (8 for LRMS and 9 for HRMS). Among different feature selection methods variance-based UFS was chosen as the most suitable approach. It eliminates features with variances below a predefined threshold which in this case was the mean of all variances[19]. Using this threshold 97 % and 99 % of features from LRMS and HRMS datasets respectively were excluded.

At the next step all five methods in the optimized conditions were applied to the obtained datasets.

The results of the applied algorithms were compared using multiple criteria: computational time, number of noisy components, silhouette score, and Davies-Bouldin index. All results are presented in Table 2.

The Silhouette score was also used to understand how close the sample is to its parent cluster compared with its neighboring cluster[20]. Silhouette coefficients close to +1 suggest the sample is near to its true parent but distant from the neighboring clusters. A value of 0 means that the sample is between two adjacent clusters on or very close to the decision boundary and negative values indicate incorrect cluster assignment for that sample. Values for all samples are calculated and average among all of them is considered as silhouette score. PCA showed the best performance with 0.71 and 0.48 scores for LRMS and HRMS data respectively.

Another metric used for clustering performance evaluation was Davies-Bouldin index[20]. This criterion is based on an averaged ratio “within-cluster” and “between-cluster” distances. If two clusters are close together and have a large spread then this ratio will be large, indicating that clusters are not very distinct. PCA and UFS produced the best results for LRMS and HRMS respectively.

Computational costs for PCA, ICA, UFS were relatively the same and less than for NMF and PARAFAC. It was shown that PCA and ICA produced fewer noisy components.

Based on the discussed criteria PCA, ICA and UFS methods demonstrated similar performance in LRMS and HRMS data treatment. Therefore, the next stage was to compare them by ability to discover biomarkers and by closeness of their clustering results to biological molecular phylogenetic tree. However, all of these unsupervised techniques allow the most variable markers in the composition of investigated samples to be identified.

3.3. Biomarker identification

The final stage of the data analysis was identification of the markers which were the most important for clustering. In the metabolomic approach each feature ultimately represents a single compound, because redundancy related to isotopic peaks and adduct ions is removed, and only one time point of the peak vertex is taken into account for each feature. In our study extra information about isotopologues and peak shapes is preserved, however the related signals from one peak should be regarded as one compound for the purpose of biomarker discovery. To extract such signals from LC-LRMS data treated by either PCA or ICA, a retention time window of 0.4 min was established to group signals with the highest weights in each component as well as m/z values of signals attributable to one isotopic pattern (A, A+1, A+2). Each group of signals, therefore, could be regarded as one compound. Although the number of such components was different for these three methods (see section 3.2), it was decided to extract 50 compounds for each method by evenly extracting them from all components. In the same manner signals corresponding to the first 50 most significant compounds were selected after the UFS procedure. Further, an intersection of all these lists of signals was obtained. Approximately 50 % of signals from each method's list were captured in the intersection list. Among them 23 compounds were interpreted and remaining signals known to be noisy (from high retention time) were not considered.

For LC-HRMS data same strategy was employed. The results of three methods (PCA, ICA, and UFS) were intersected and the same potential chemotaxonomic markers were observed. They correlate with most of the features from the intersection list generated using LC-LRMS data.

Finally, it was tried to perform dereplication of these compounds based on the literature data and available databases. Representative samples which contain compounds of interest were reanalyzed on the qTOF instrument in auto-MS^2 mode. The results of the annotation are presented in Table 3. Spectra for all compounds are presented in Supplementary (Fig. S14 – S62)

Compound 1 possessed a molecular weight of 328 deduced from the protonated molecule ([M + H]$^+$) peak at m/z 329.1596 ($C_{16}O_7H_{25}$, eluted at 4.7 min), which produced predominant fragment ion m/z 167.1061 corresponding to the cleavage of glucose molecule. The exact position of the substituent could not be assigned and this compound was tentatively assigned as verbenone glycoside or one of its isomers previously isolated from Prangos species along with γ-pyrone glucosides[21], which are structurally similar to **compound 2.**

Compound 2 had a molecular weight 432 deduced from the sodium adduct ([M + Na]$^+$) signal at m/z 455.1162 and [M + H]$^+$ ion peak at m/z 433.1342 ($C_{18}H_{25}O_{52}$, eluted at 6.2). An [M + H]$^+$ precursor ion produced the predominant fragment ion m/z 127.0387, which allowed to suspect a structure similar to maltol glucoside. The fragment ion at m/z 329.0839 observed in the ESI/MS^2 spectrum of the sodium adduct may be interpreted as a fragment of hydroxy-3-methylglutaric acid (HMG) substituted glucopyranosyl side chain. Thus, **compound 2** can be tentatively identified as previously reported licoagroside B or its isomer[22].

Compound 3 was detected as the precursor ion [M+H]$^+$ at m/z 425.1451 ($C_{20}O_{10}H_{25}$, eluted at 8.0 min). The observed fragment ions at m/z 263 and m/z 245 in its MS^2 spectrum can be produced by loss of sugar moiety with a successive loss of a neutral fragment (H_2O). Presence of the most intensive ion peak at m/z 191 corresponding to the additional loss of C_4H_6 allowed preliminary identification of this compound as rutarin or its positional isomer. Other candidates were rejected after manual comparison with the spectra from GNPS library.

Compound 4 showed a protonated molecule ([M+H]$^+$) ion peak at m/z 479.0821 ($C_{21}H_{18}O_{13}$, eluted at 8.3 min) and a predominant fragment at m/z 303.0501 in its MS^2 spectrum, which should be attributed to the mild elimination of a glucuronic acid as a neutral loss. The ions produced by the m/z 303.0501 precursor are in accordance with typical fragmentation pattern of quercetin[23]. Therefore, this compound was tentatively identified as quercetin glucuronide.

Compounds 5 and 6 are a pair of isomers with molecular weight of 217 determined by the protonated molecule ([M+H]$^+$) peak at m/z 217.0495 ($C_{12}H_8O_4$, eluted at 12.3 and 13.1 min). In the ESI/MS^2 spectra these precursor ions displayed the ion peak at m/z 202.0261 ($C_{11}H_5O_4$) corresponding to demethylation together with a signal at m/z 174 produced via additional drop of CO. The ion peaks detected at m/z 189 and 161 resulted from the successive losses of two CO molecules. Although these compounds could not be distinguished by their ESI/MS spectra, the comparison of their elution order on a RP-C18 column with reported in literature[23] allows a tentative identification of compound **5** as xanthotoxin and thus compound **6** as bergapten.

Compound 7 possessed a molecular weight of 246 deduced from the protonated molecule ($[M+H]^+$) ion peak at m/z 247.0603 ($C_{13}H_{11}O_5$, eluted at 13.1 min). The molecular weight of compound 7 is 30 Da larger than xanthotoxin (**5**), which corresponds to the additional -OCH_3 substituent. A similar fragmentation pattern to other linear furanocoumarins (Table 3) allows tentative identification of **compound 7** as isopimpinellin[24].

Compound 8 had a molecular weight 260 determined by the protonated molecule ($[M+H]^+$) signal at m/z 261.1123 ($C_{15}H_{16}O_4$, eluted at 13.3 min). Among the observed peaks of its isomers, this one is the most retained. Moreover, the ion peak corresponding to the loss of H_2O was not observed in its ESI/MS spectra, while the predominant fragment ion was detected at 189 m/z. Therefore, compound **8** was tentatively characterized as isomeranzin[25].

Compounds 9 and 10 both had a molecular weight 286 Da determined by the presence of a protonated molecule peak at m/z 287.0918 ($C_{16}H_{15}O_5$, eluted at 13.4 and 14.1 min). These two compounds showed the typical fragmentation patterns of monosubstituted furanocoumarins, with the presence of m/z 203, 175, 159 and 145 in their ESI/MS^2 spectra (Table 3). With respect to the presence of the fragment ion at m/z 269 $[M + H - 18]^+$ in the MS^2 spectrum of **compound 9**, it was tentatively identified as pabulenol, and thus **compound 10** would be oxypeucedanin[24].

Compound 11 was detected by the presence of a the protonated molecule ($[M+H]^+$) ion peak at m/z 323.0679 ($C_{16}H_{16}O_5Cl$, eluted at 14.6 min) with a specific isotopic distribution corresponding to a monochlorinated compound. In the MS^2 spectrum this precursor produced the predominant fragment ions corresponding to the loss of HCl and side chain cleavage at m/z 287 and 203, respectively. Other observed fragment ions were the same as for other monosubstituted fumarocoumarins. Therefore, this compound was tentatively identified as saxalin[21].

Compounds 12 and 22 both had a molecular weight 270 Da determined by the presence of protonated molecule ($[M+H]^+$) signal at m/z 271.0961 ($C_{16}H_{15}O_4$, eluted at 15.8 and 16.8 min) in their ESI/MS spectra. These precursors have shown ion peaks at m/z 203, 175, and 159 common for monosubstituted furanocoumarins (Table 3). However, efforts to distinguish the paired isomers **12** and **22** by ESI/MS^2 analysis were unsuccessful, and these compounds were differentiated by comparison of their elution order on a RP-C18 column with reported in literature[24]. Thus, **compounds 12** and **22** were tentatively identified as imperatorin and isoimperatorin, respectively.

Compounds 13 had a molecular weight 316 Da determined by the presence of protonated molecule ($[M+H]^+$) ion peaks at m/z 317.1384 ($C_{18}H_{21}O_5$, eluted at 16.0 min), it exhibited the same fragmentation pattern as compounds **14**, **17**, **18**, **20**. Thus, **compound 13** was tentatively assigned as linear isomer of cnidiadin[26]. It should be noted, that its pyrano-analogue might also be found in some of the samples at almost the same retention time.

Compounds 16 and 23 were a pair of isomers with molecular weight of 244 determined by the protonated molecule ($[M+H]^+$) ion peak at m/z 245.1177 ($C_{16}H_{15}O_4$, eluted at 16.5 and 17.2 min), but their ESI/MS fragmentation patterns were quite different. Compound **23** showed the presence of characteristic fragment ion at m/z 187 $[245\text{-}C_4H_{10}]^+$, while compound **16** exhibits the predominant ion at m/z 189 $[245\text{-}C_4H_8]^+$, which was probably caused by different π–π conjugation extensions. Accordingly, compounds **16** and **23** were assigned as osthol[23] and suberosin[21], which could be also confirmed by their retention in RP HPLC[27].

For **compounds 14,15,17-21**, all showed protonated molecule ($[M+H]^+$) peak at m/z 329,1387 ($C_{19}H_{20}O_5$, eluted between 16.3 and 17.0 min). These isomers may belong to the classes of furanocoumarins and pyranocoumarins. Their ESI/MS^2 spectra demonstrated two distinguishing patterns. One of them includes predominant ion peaks at m/z 229, 247 and 213, while the second exhibits the most intensive ion peaks at m/z 229, 187 and 159. It was found from literature that linear monosubstituted furanocoumarins and pyranocoumarins exhibit the first fragmentation pattern while the angular ones show predominant ion peak at m/z 187[28]. Tentative assignments of angular and linear structures could be also confirmed by comparison of their relative retention time in a RP HPLC column. It is known that angular coumarins are more strongly retained compared to their linear isomers[29], and angelate isomer is eluted after its senecioic acid analogue[30]. Moreover, pyranocoumarins tend to be eluted before furanocoumarins[31]. Therefore compound **14** and **17** were tentatively identified as decursin and decursinol angelate[32], and thus **compounds 18** and **20** would be prantschimgin and deltoin (syn. sprengelianin)[33]. Similarly, compound **15** was tentatively assigned as jatamansin, thus compounds **19** and **21** would be libanorin and columbianadin[34].

Many more chromatographic peaks of structurally similar compounds were observed in the chromatograms. However, it is nearly impossible to differentiate all of them, because the corresponding mass spectra are sometimes missing or not well described in the available literature. Thus, however, the application of the suggested data treatment techniques allowed identification of the most plausible chemotaxonomic marker candidates. Moreover, these markers are expected to be significant due to the fact that they were found in the high-ranking components of all three selected methods.

3.4. Application to chemotaxonomic purposes

Classification of plants on the basis of their secondary metabolites and their biosynthetic pathways is called chemotaxonomy[1]. The main purposes of chemotaxonomy are to improve the existing system of plant differentiation and to incorporate the modern knowledge of the natural relationship of plants. One example of compounds which might be used as chemotaxonomic markers are coumarins[2,4]. In the work [4] coumarin-containg species, namely, *Angelica Sinensis, Angelica Dahurica, Angelica Decursiva, Peucedanum Praeruptorium, Peucedanum Pubescens* were analyzed by direct injection MS in positive multiple ion monitoring mode and the results showed that only several sample classes could be separated from the main cluster in the PCA score plot. The variables responsible for this classification were structurally described as angular-type pyranocoumarins, linear-type pyranocoumarins, angular-type furanocoumarins, and ligustilide derivatives. In the present work coumarins profiles were for the first time compared in the range of genera and species. The distribution of biomarkers identified in this work is shown in Fig. 3. After careful consideration of the identified biomarkers, it was concluded that there are no unique compounds for any of the genera. In order to find out which markers depend on the growing conditions and what are markers of each genus, a more extensive research with larger number of biological replicates of each species and more representatives of each genus should be conducted in future. We also note that for this particular task supervised techniques might show significantly improved classification performance.

Another way to visualize the results of unsupervised learning is hierarchical tree construction (Fig. S3-S12). Comparison by closeness of dendrograms created by each method to the molecular phylogenetic tree (Fig. S13) was done. It should be noted that trees generated from LC-MS data show differences in chemical composition which is not correlated with plant molecular phylogenetic analysis results. For the evaluation of these differences, an approach that involves computing pairwise distances between all data items and showing the distances in a matrix form was employed[35]. As a quantitative characteristic pixel-wise mean square error (MSE) can be calculated by Eq. (1) (where $I^1_{i,j}$ and $I^2_{i,j}$ are the i,j elements of the first and second distance matrix respectively), in the form where instead of pixel values the original distance values in the matrices are considered.

$$MSE\ (I^1, I^2) = \frac{1}{n}\sum_{i,j}(I^1_{i,j} - I^2_{i,j})^2 \qquad (1)$$

Errors calculated by this method were compared and the lowest value were obtained by the UFS method for both LRMS and HRMS data: 0.105 and 0.144 respectively. Although plant tissue chemical composition is highly variable, it may be beneficial to use the combination of LC-MS-based methods and unsupervised machine learning algorithms along with molecular phylogeny data in chemotaxonomic studies.

4. Conclusions

Two types of data analysis were considered for LC-LRMS and -HRMS data: tensor decomposition by PARAFAC and decomposition following tensor unfolding into two dimensions. For unfolded tensors, four approaches to data reduction and factorization were considered: PCA, ICA, NMF and UFS. Results obtained by these methods from both datasets were compared by several criteria. Applied to LC-LRMS and LC-HRMS data treated by suggested approaches, PCA showed the best results according to silhouette coefficient, Davies-Bouldin index, computational time and number of noisy components. However, PCA, ICA and UFS demonstrated comparable performance and similar lists of biomarkers were revealed from their results. A list of 23 compounds, most of which belong to the coumarin class were extracted from the intersection of the results from all employed methods. These compounds were tentatively identified by comparing their ESI/MS spectra with published data. The distribution of these biomarkers in different species from the Apiaceae family was shown. The identified compounds can potentially serve as chemotaxonomic markers because they were chosen by the algorithms as features with highest dispersion across the samples.

Although the methodology allowed successful separation of each sample with its replicates from the rest of the dataset, it has demonstrated some limitations in application to biological classification. It was shown that dendrograms constructed by the employed methods differ from the molecular phylogenetic tree, which may be caused by changes in chemical composition of the studied extracts related to different environmental factors. Due to the high chemical diversity of coumarins and other plant constituents, future studies should use a larger number of biological replicates for each species.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Process of plant material collection for this research was supported by the Russian Foundation for Basic Research (project no. 19-04-00496 a). All other parts of the reported study including experiment and data analysis were funded by RFBR, project number 19-33-90036. HPLC-MS-TOF analysis was performed using the equipment of the demo laboratory of Bruker Ltd., Moscow, Russia.

References

[1] R. Singh, Chemotaxonomy: A Tool for Plant Classification, Journal of Medicinal Plants Studies. 4 (2016) 90–93.

[2] A. Dugrand-Judek, A. Olry, A. Hehn, G. Costantino, P. Ollitrault, Y. Froelicher, F. Bourgaud, The distribution of coumarins and furanocoumarins in Citrus species closely matches Citrus phylogeny and reflects the organization of biosynthetic pathways, PLoS ONE. 10 (2015). https://doi.org/10.1371/journal.pone.0142757.

[3] A. Forycka, W. Buchwald, Variability of composition of essential oil and coumarin compounds of Angelica archangelica L., Herba Polonica. 65 (2019) 62–75. https://doi.org/10.2478/hepo-2019-0027.

[4] X. Xu, W. Li, T. Li, K. Zhang, Q. Song, L. Liu, P. Tu, Y. Wang, Y. Song, J. Li, Direct Infusion-Three-Dimensional-Mass Spectrometry Enables Rapid Chemome Comparison among Herbal Medicines, Analytical Chemistry. 92 (2020). https://doi.org/10.1021/acs.analchem.0c00483.

[5] K. Kumar, Introducing an integral optimised warping (IOW) approach for achieving swift alignment of drifted chromatographic peaks: an optimisation of the correlation optimised warping (COW) technique, Analytical Methods. 10 (2018). https://doi.org/10.1039/C8AY00963E.

[6] K. Kumar, Optimizing the process of reference selection for correlation optimised warping (COW) and interval correlation shifting (icoshift) analysis: automating the chromatographic alignment procedure, Analytical Methods. 10 (2018). https://doi.org/10.1039/C7AY02340E.

[7] W. Sun, R.D. Braatz, Opportunities in tensorial data analytics for chemical and biological manufacturing processes, Computers & Chemical Engineering. 143 (2020) 107099. https://doi.org/10.1016/j.compchemeng.2020.107099.

[8] P. Turova, I. Rodin, O. Shpigun, A. Stavrianidi, A new PARAFAC-based algorithm for HPLC–MS data treatment: herbal extracts identification, Phytochemical Analysis. 31 (2020) 948–956. https://doi.org/10.1002/pca.2967.

[9] P.W. Siy, R.A. Moffitt, R.M. Parry, Y. Chen, Y. Liu, M.C. Sullards, A.H. Merrill, M.D. Wang, Matrix factorization techniques for analysis of imaging mass spectrometry data, in: 2008 8th IEEE International Conference on BioInformatics and BioEngineering, IEEE, 2008: pp. 1–6. https://doi.org/10.1109/BIBE.2008.4696797.

[10] N. Kuhnert, R. Jaiswal, P. Eravuchira, R.M. El-Abassy, B. von der Kammer, A. Materny, Scope and limitations of principal component analysis of high resolution LC-TOF-MS data: The analysis of the chlorogenic acid fraction in green coffee beans as a case study, Analytical Methods. 3 (2011) 144–155. https://doi.org/10.1039/c0ay00512f.

[11] Y. Gut, M. Boiret, L. Bultel, T. Renaud, A. Chetouani, A. Hafiane, Y.M. Ginot, R. Jennane, Application of chemometric algorithms to MALDI mass spectrometry imaging of pharmaceutical tablets, Journal of Pharmaceutical and Biomedical Analysis. 105 (2015) 91–100. https://doi.org/10.1016/j.jpba.2014.11.047.

[12] P. Kharyuk, D. Nazarenko, I. Oseledets, I. Rodin, O. Shpigun, A. Tsitsilin, M. Lavrentyev, Employing fingerprinting of medicinal plants by means of LC-MS and machine learning for species identification task, Scientific Reports. 8 (2018). https://doi.org/10.1038/s41598-018-35399-z.

[13] T. Bald, J. Barth, A. Niehues, M. Specht, M. Hippler, C. Fufezan, pymzML--Python module for high-throughput bioinformatics on mass spectrometry data, Bioinformatics. 28 (2012) 1052–1053. https://doi.org/10.1093/bioinformatics/bts066.

[14] M.B. Gholivand, Y. Yamini, M. Dayeni, Y. Shokoohinia, The influence of the extraction mode on three coumarin compounds yield from Prangos ferulacea (L.) Lindl roots, Journal of the Iranian Chemical Society. 12 (2015) 707–714. https://doi.org/10.1007/s13738-014-0529-0.

[15] K. Kumar, Standardising the chromatographic denoising procedure, Analytical Methods. 10 (2018). https://doi.org/10.1039/C8AY01606B.

[16] R. Bro, H.A.L. Kiers, A new efficient method for determining the number of components in PARAFAC models, Journal of Chemometrics. 17 (2003) 274–286. https://doi.org/10.1002/cem.801.

[17] D. Jouan-Rimbaud Bouveresse, A. Moya-González, F. Ammari, D.N. Rutledge, Two novel methods for the determination of the number of components in independent components analysis models, Chemometrics and Intelligent Laboratory Systems. 112 (2012) 24–32. https://doi.org/10.1016/j.chemolab.2011.12.005.

[18] L.N. Hutchins, S.M. Murphy, P. Singh, J.H. Graber, Position-dependent motif characterization using non-negative matrix factorization, Bioinformatics. 24 (2008) 2684–2690. https://doi.org/10.1093/bioinformatics/btn526.

[19] D. Weigt, D.A. Sammour, T. Ulrich, B. Munteanu, C. Hopf, Automated analysis of lipid drug-response markers by combined fast and high-resolution whole cell MALDI mass spectrometry biotyping, Scientific Reports. 8 (2018). https://doi.org/10.1038/s41598-018-29677-z.

[20] L. Vendramin, R.J.G.B. Campello, E.R. Hruschka, Relative clustering validity criteria: A comparative overview, Statistical Analysis and Data Mining: The ASA Data Science Journal. 3 (2010). https://doi.org/10.1002/sam.10080.

[21] Y. Shikishima, Y. Takaishi, G. Honda, M. Ito, Y. Takeda, O.K. Kodzhimatov, O. Ashurmetov, Terpenoids and γ-pyrone derivatives from Prangos tschimganica, Phytochemistry. 57 (2001) 135–141. https://doi.org/10.1016/S0031-9422(00)00407-6.

[22] S. Kozachok, Ł. Pecio, J. Kolodziejczyk-Czepas, S. Marchyshyn, P. Nowak, J. Mołdoch, W. Oleszek, γ-Pyrone compounds: flavonoids and maltol glucoside derivatives from Herniaria glabra L. collected in the Ternopil region of the Ukraine, Phytochemistry. 152 (2018) 213–222. https://doi.org/10.1016/j.phytochem.2018.05.009.

[23] X. Zheng, X. Zhang, X. Sheng, Z. Yuan, W. Yang, Q. Wang, L. Zhang, Simultaneous characterization and quantitation of 11 coumarins in Radix Angelicae Dahuricae by high performance liquid chromatography with electrospray tandem mass spectrometry, Journal of Pharmaceutical and Biomedical Analysis. 51 (2010) 599–605. https://doi.org/10.1016/j.jpba.2009.09.030.

[24] B. Li, X. Zhang, J. Wang, L. Zhang, B. Gao, S. Shi, X. Wang, J. Li, P. Tu, Simultaneous characterisation of fifty coumarins from the roots of angelica dahurica by off-line two-dimensional high-performance liquid chromatography coupled with electrospray ionisation tandem mass spectrometry, Phytochemical Analysis. 25 (2014) 229–240. https://doi.org/10.1002/pca.2496.

[25] L. Duan, L. Guo, K. Liu, E.H. Liu, P. Li, Characterization and classification of seven Citrus herbs by liquid chromatography-quadrupole time-of-flight mass spectrometry and genetic algorithm optimized support vector machines, Journal of Chromatography A. 1339 (2014) 118–127. https://doi.org/10.1016/j.chroma.2014.02.091.

[26] Y. Chen, G. Fan, Q. Zhang, H. Wu, Y. Wu, Fingerprint analysis of the fruits of Cnidium monnieri extract by high-performance liquid chromatography-diode array detection-electrospray ionization tandem mass spectrometry, Journal of Pharmaceutical and Biomedical Analysis. 43 (2007) 926–936. https://doi.org/10.1016/j.jpba.2006.09.015.

[27] M. Figueroa, I. Rivero-Cruz, B. Rivero-Cruz, R. Bye, A. Navarrete, R. Mata, Constituents, biological activities and quality control parameters of the crude extract and essential oil from Arracacia tolucensis var. multifida, Journal of Ethnopharmacology. 113 (2007) 125–131. https://doi.org/10.1016/j.jep.2007.05.015.

[28] V. Shukla, P. Singh, D. kumar, R. Konwar, B. Singh, B. Kumar, Phytochemical analysis of high value medicinal plant Valeriana jatamansi using LC-MS and it's in-vitro anti-proliferative screening, Phytomedicine Plus. 1 (2021) 100025. https://doi.org/10.1016/j.phyplu.2021.100025.

[29] G.F. Spencer, L.W. Tjarks, R.G. Powell, Analysis of linear and angular furanocoumarins by dual-column high-performance liquid chromatography, Journal of Agricultural and Food Chemistry. 35 (1987) 803–805. https://doi.org/10.1021/jf00077a040.

[30] J. Zhang, L. Li, T.W. Hale, W. Chee, C. Xing, C. Jiang, J. Lü, Single oral dose pharmacokinetics of decursin and decursinol angelate in healthy adult men and women, PLoS ONE. 10 (2015). https://doi.org/10.1371/journal.pone.0114992.

[31] S.Y. Kang, K.Y. Lee, S.H. Sung, M.J. Park, Y.C. Kim, Coumarins isolated from Angelica gigas inhibit acetylcholinesterase: Structure-activity relationships, Journal of Natural Products. 64 (2001) 683–685. https://doi.org/10.1021/np000441w.

[32] M.J. Ahn, M.K. Lee, Y.C. Kim, S.H. Sung, The simultaneous determination of coumarins in Angelica gigas root by high performance liquid chromatography-diode array detector coupled with electrospray ionization/mass spectrometry, Journal of Pharmaceutical and Biomedical Analysis. 46 (2008) 258–266. https://doi.org/10.1016/j.jpba.2007.09.020.

[33] Y. Xu, H. Cai, G. Cao, Y. Duan, K. Pei, S. Tu, J. Zhou, L. Xie, D. Sun, J. Zhao, J. Liu, X. Wang, L. Shen, Profiling and analysis of multiple constituents in Baizhu Shaoyao San before and after processing by stir-frying using UHPLC/Q-TOF-MS/MS coupled with multivariate statistical analysis, Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences. 1083 (2018) 110–123. https://doi.org/10.1016/j.jchromb.2018.03.003.

[34] B. Wang, X. Liu, A. Zhou, M. Meng, Q. Li, Simultaneous analysis of coumarin derivatives in extracts of Radix Angelicae pubescentis (Duhuo) by HPLC-DAD-ESI-MSntechnique, Analytical Methods. 6 (2014) 7996–8002. https://doi.org/10.1039/c4ay01468e.

[35] J. Wang, X. Liu, H.W. Shen, High-dimensional data analysis with subspace comparison using matrix visualization, Information Visualization. 18 (2019) 94–109. https://doi.org/10.1177/1473871617733996.

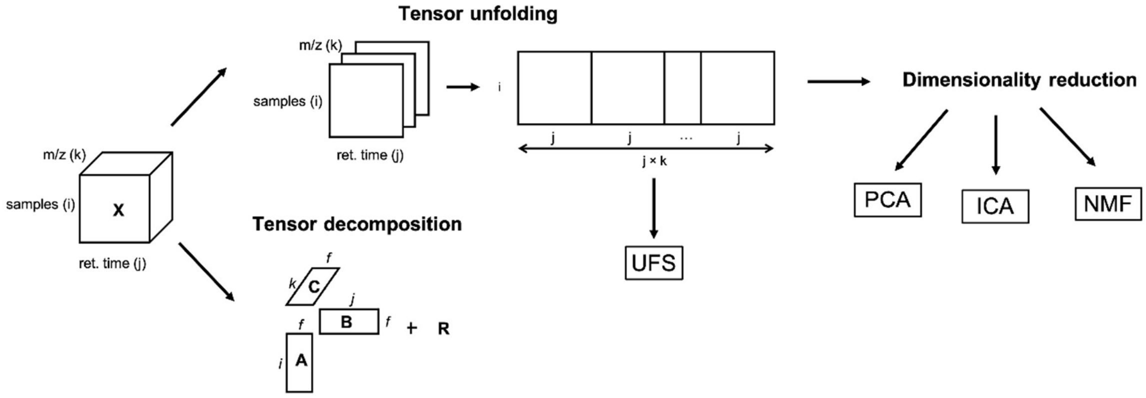

Fig. 1. Data organization and chemometric treatment workflow.

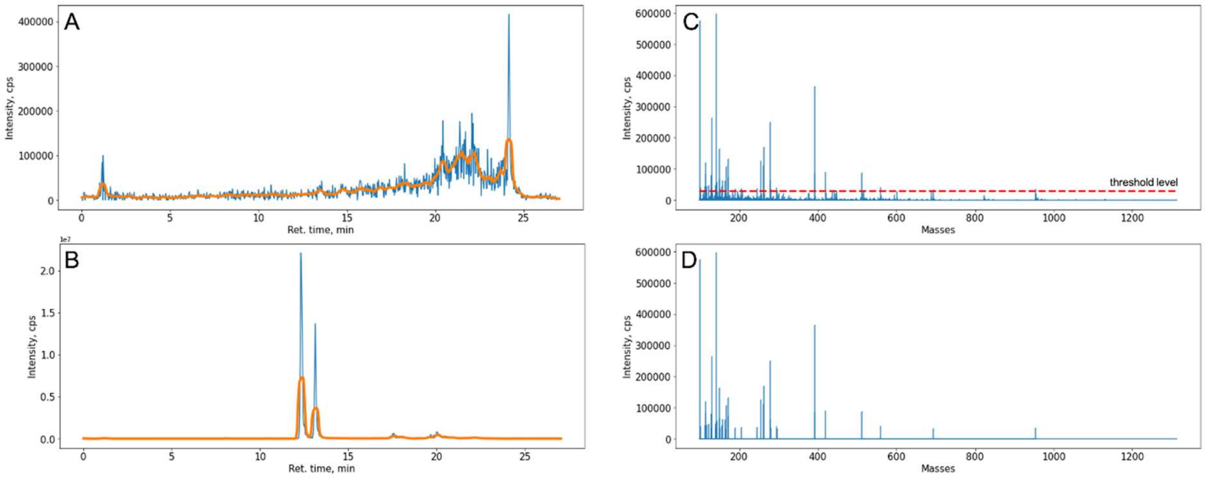

Fig. 2. An example of noisy raw LC-LRMS mass chromatogram smoothing by Savitzki-Golay filter (A). An example of informative raw LC-LRMS mass chromatogram smoothing by Savitzki-Golay filter (B). A representative raw mass spectrum from LC-HRMS before (C) and after (D) noise subtraction below the threshold line.

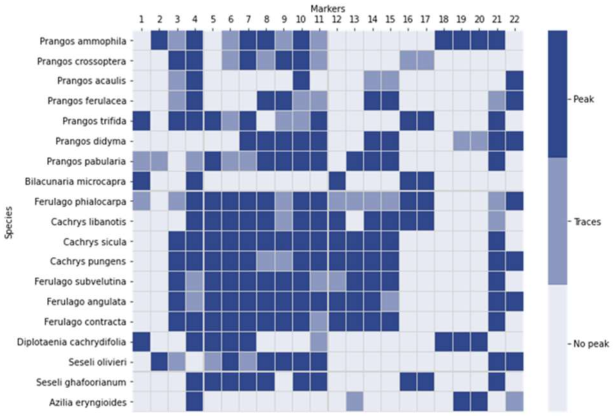

Fig. 3. Distribution of revealed biomarkers in studied species.

Table 1. List of specimens used in the experiments

#	Plant species	Part	Specimen's voucher
1	*Prangos pabularia*	Leaves	MW0858238
2	*Cachrys libanotis*	Leaves	MW0798144
3	*Prangos acaulis*	Leaves	MW0744005
4	*Prangos ferulacea*	Stems	MW0751912
5	*Prangos didyma*	Stems	MW0857912
6	*Ferulago subvelutina*	Leaves	098-IR-19
7	*Prangos ammophila*	Leaves	MW0857867
8	*Prangos trifida*	Leaves	MW0798580
9	*Ferulago angulata*	Leaves	085-IR-19
10	*Cachrys sicula*	Leaves	MW0798143
11	*Ferulago contracta*	Leaves	053-IR-19
12	*Cachrys pungens*	Leaves	MW0784701
13	*Diplotaenia cachrydifolia*	Leaves	164-IR-19
14	*Ferulago phialocarpa*	Leaves	169-IR-19
15	*Azilia eryngioides*	Leaves	167-IR-19
16	*Seseli olivieri*	Leaves	173-IR-19
17	*Prangos crossoptera*	Leaves	MW0753036
18	*Bilacunaria microcapra*	Leaves	028-IR-19
19	*Seseli ghafoorianum*	Leaves	124-IR-19

Table 2. Comparison of data treatment techniques.

Method	**Davies-Bouldin index**	**Silhouette score**	**Computational time, sec**	**Noisy components**
LRMS data				
PCA	0.33	0.71	4.26	1
ICA	0.48	0.64	5.69	1
NMF	0.50	0.59	108.87	3
PARAFAC	0.43	0.47	140.59	3
UFS	0.52	0.53	1.26	–
HRMS data				
PCA	0.83	0.48	2.48	0
ICA	1.25	0.44	1.80	0
NMF	1.90	0.25	78.42	1
PARAFAC	1.05	0.38	122.86	1
UFS	0.75	0.40	0.26	–

Table 3. Chromatographic and mass-spectral data for compounds defined as biomarkers.

Number	RT (min), m/z	Components	$[M+H]^+$, m/z (formula, error (ppm))	Adduct ions, m/z	Key MS/MS fragmentation	Identity	Reference
1	4.7 125	LRMS: PCA 2, ICA 2, NMF 2, FS	329.1595 ($C_{16}H_{24}O_7$, -0.1)	351.1419 $[M+Na]^+$	329-167 = Glc 167-125 = C_2H_2O	Verbenone glycoside	[21]
2	6.2 127	LRMS: PCA 4, ICA 4, NMF 5, FS HRMS: PARAFAC 9	433.1342 $C_{18}H_{24}O_{12}$, -0.4)	455.1162 $[M+Na]^+$	433-127 = $C_{12}H_{18}O_9$ 127-85 = C_2H_2O 455-329 = $C_6H_6O_3$	Licoagroside B	[22]
3	8.0 263	LRMS: PCA 9, ICA 3, NMF 3, PARAFAC 9 FS HRMS: PCA 4, ICA 1, NMF 3, PARAFAC 3	425.1451 ($C_{20}H_{24}O_{10}$, -1.9)	442.1716 $[M+NH_4]^+$ 447.1271 $[M+Na]^+$	263-245 = H_2O 245-191 = C_4H_6	Rutarin	GNPS library
4	8.3 303	LRMS: PCA 8, ICA 2, NMF 1 PARAFAC 9, FS HRMS: PARAFAC 3	479,0821 ($C_{21}H_{18}O_{13}$, -0.3)	501.0639 $[M+Na]^+$	479-303 = GluA 303-229 = $C_2H_2O_3$ 303-153 = $C_8H_6O_3$ 303-137 = $C_8H_6O_4$	Quercetin glucuronide	[23]
5	12.3 217	LRMS: PCA 1, ICA 3, NMF 3, PARAFAC 2, FS HRMS:	217.0495 ($C_{12}H_8O_4$, 0.1)	234.0760 $[M+NH_4]^+$	217-202 = CH_3 217-189 = CO 202-174 = CO 217-161 = 2CO 161-146 = CH_3	Xanthotoxin	[23]

		PCA 4, ICA 2, NMF 3, PARAFAC 2			146-118 = CO		
6	13.1 217	LRMS: PCA 9, ICA 3, NMF 6, PARAFAC 2, FS HRMS: PCA 8, ICA 3, NMF 3, PARAFAC 1	217.0495 ($C_{12}H_8O_4$, 0.3)	234.0760 $[M+NH_4]^+$	217-202 = CH_3 202-174 = CO 217-161 = 2CO 161-146 = CH_3 146-118 = CO	Bergapten	[23]
7	13.1 247	LRMS: PCA 10, ICA 3, NMF 3, PARAFAC 6, FS HRMS: PARAFAC 2	247.0605 ($C_{13}H_{10}O_5$, -1.4)	269.0429 $[M+Na]^+$ 264.0872 $[M+NH_4]^+$	247-232 = CH_3 232-217 = CH_3 217-189 = CO 189-161 = CO	Isopimpinellin	[23]
8	13.3 261	LRMS: PCA 5, ICA 3, NMF 3, PARAFAC 6, FS HRMS: PARAFAC 1	261.1123 ($C_{15}H_{16}O_4$, -0.5)	283.0944 $[M+Na]^+$	261-217 = CO_2 261-189 = C_4H_7O 189-161 = CO	Isomeranzin	[25]
9	13.4 203, 269	LRMS: PCA 5, ICA 3, NMF 6, PARAFAC 2, FS HRMS: PCA 1, ICA 4, NMF 1, PARAFAC 2	287.0916 ($C_{16}H_{14}O_5$, -0.7)	304.1182 $[M+NH_4]^+$	287-203 = C_5H_9O 203-175 = CO 203-159 = CO_2 175-147 = CO	Pabulenol	[24]
10	14.1 287	LRMS: PCA 4, ICA 4, NMF 5, PARAFAC 6, FS	287.0918 ($C_{16}H_{14}O_{5,}$ -1.3)	309.0739 $[M+Na]^+$	287-203 = C_5H_9O 203-159 = CO_2 203-147 = 2CO 159-131 = CO	Oxypeucedanin	[24]

		HRMS: PCA 2, ICA 1, NMF 2, PARAFAC 8					
11	14.6 323	LRMS: PCA 5, ICA 1, NMF 4, PARAFAC 6, FS HRMS: PCA 6, ICA 2, NMF 7, PARAFAC 1	323.0679 ($C_{16}H_{15}ClO_5$, -0.6)	345.0501 $[M+Na]^+$	323-287 = HCl 287-203 = C_5H_9O 203-159 = CO_2 203-147 = 2CO	Saxalin	[21]
12	15.8 203	LRMS: PCA 4, ICA 4, NMF 5, PARAFAC 4, FS HRMS: PARAFAC 7	271.0961 ($C_{16}H_{14}O_4$, 1.5)	288.1224 $[M+NH_4]^+$	271-215 = 2CO 271-203 = C_5H_8 203-175 = CO 203-159 = CO_2 175-147 = CO 175-131 = CO_2	Imperatorin	[24]
13	15.9 317	LRMS: PCA 2, ICA 2, NMF 2, PARAFAC 4, FS HRMS: PARAFAC 5	317.1384 ($C_{18}H_{20}O_5$, -1.1)	339.1208 $[M+Na]^+$	317-247 = C_4H_6O 247-229 = H_2O	Cnidiadin (linear isomer)	[26]
14	16.3 329	LRMS: PCA 10, ICA 1, NMF 6, PARAFAC 8, FS HRMS: PCA 3, ICA 4, NMF 5, PARAFAC 5	- 329.1387 ($C_{19}H_{20}O_5$, -1.2)	351.1208 $[M+Na]^+$	329-247 = C_5H_6O 247-229 = H_2O	Decursin	[32]
15	16.3 329	LRMS: PCA 10, ICA 2, NMF 2, PARAFAC 8, FS	329.1387 ($C_{19}H_{20}O_5$, -1.2)	351.1207 $[M+Na]^+$	329-229 = $C_5H_8O_2$ 229-187 = C_3H_6 187-159 = CO	Jatamansin	[34]

					159-131 = CO		
16	16.5 189 245	LRMS: PCA 5, ICA 3, NMF 3, PARAFAC 8, FS HRMS: PCA 8, ICA 3, NMF 3, PARAFAC 5	245.1177 ($C_{15}H_{16}O_3$, -2.0)	262.1350 $[M+NH_4]^+$ 267.0990 $[M+Na]^+$	245-189 = C_4H_8 189-159 = CH_2O 159-131 = CO	Osthol	[23]
17	16.6 329	LRMS: PCA 2, ICA 2, NMF 2, PARAFAC 8, FS	329.1387 ($C_{19}H_{20}O_5$, -1.2)	351.1211 $[M+Na]^+$	329-247 = C_5H_6O 247-229 = H_2O	Decursinol Angelate	[32]
18	16.7 329	LRMS: PCA 2, ICA 2, NMF 2, PARAFAC 3, FS	329.1387 ($C_{19}H_{20}O_5$, -1.1)	351.1209 $[M+Na]^+$	329-247 = C_5H_6O 247-229 = H_2O	Prantschimgin	[33]
19	16.7 329	LRMS: PCA 2, ICA 2, NMF 2, PARAFAC 3, FS	329.1385 ($C_{19}H_{20}O_5$, -0.5)	351.1206 $[M+Na]^+$	329-229 = $C_5H_8O_2$ 229-187 = C_3H_6 187-159 = CO 159-131 = CO	Libanorin	[34]
20	16.8 329	LRMS: PCA 10, ICA 2, NMF 2, PARAFAC 3, FS HRMS: PCA 3, ICA 4, NMF 5, PARAFAC 5	329,1387 ($C_{19}H_{20}O_5$, -1.1)	351.1208 $[M+Na]^+$	329-247 = C_5H_6O 247-229 = H_2O	Sprengelianin (Deltoin)	[33]
21	17.0 329	LRMS: PCA 2, ICA 2, NMF 2, PARAFAC 3, FS	329.1386 ($C_{19}H_{20}O_5$, -0.8)	351.1204 $[M+Na]^+$	329-229 = $C_5H_8O_2$ 229-187 = C_3H_6 187-159 = CO 159-131 = CO	Columbianadin	[34]

22	16.8 229 203	LRMS: PCA 10, ICA 4, NMF 5, PARAFAC 8, FS HRMS: PARAFAC 7	271.0961 ($C_{16}H_{14}O_4$, 1.6)	293.0782 $[M+Na]^+$	271-215 = 2CO 271-203 = C_5H_8 203-175 = CO 203-159 = CO_2 175-147 = CO 175-131 = CO_2	Iso-imperatorin	[24]
23	17.2 245	LRMS: PCA 4, ICA 4, NMF 5 PARAFAC 4, FS HRMS: PARAFAC 5	245.1172 ($C_{15}H_{16}O_3$, -0.1)	267.0996 $[M+Na]^+$	245-215 = C_2H_6 245-187 = C_4H_{10} 245-131 = $C_6H_{10}O_2$	Suberosin	[21]