Analyst

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/analyst

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/xxxxxx

ARTICLE TYPE

Potential use of Multivariate Curve Resolution for the analysis of Mass Spectrometry Images

Joaquim Jaumot*[a] and Romà Tauler[a]

Received (in XXX, XXX) Xth XXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX

DOI: 10.1039/b000000x

In this work the application of Multivariate Curve Resolution is proposed for the analysis of Mass Spectrometry Imaging (MSI) data. Recently, developments on ionization of samples have dramatically expanded the number of applications of MSI due to the possibility of collecting the mass spectrum for each pixel of a considered surface in a reasonable time. Using this method, both spatial distribution and spectral information of analyzed samples can be obtained. However, there are major drawbacks inherent to MSI related to the high complexity of the data obtained from real samples and to the extremely huge size of the data sets generated by this technique. Therefore, the potential of chemometrical tools in different steps of the analysis process is unquestionable, from data compression to data resolution of the different components present at each pixel of the image. In this work, this data analysis is carried out by means of the Multivariate Curve Resolution method. The benefits of the application of this method are shown for two examples consisting on a MS image of two platted microbes and on a MS image of a mouse lung section. Results show that Multivariate Curve Resolution allows obtaining distribution maps of different components and their identification from resolved pure high-resolution mass spectra.

1. Introduction

In recent years a lot of attention has been focused on the development of hyperspectral imaging techniques due to its ability of carrying out fast and relatively cheap analyses of multiple compounds spread over the surface of a sample[1, 2]. These imaging technologies have been applied to several research fields such as food processing and control[3], environmental and biomedical studies[4-6]. In these cases, a complete spectrum (usually from a vibrational spectroscopic technique such as NIR or Raman spectroscopies) is collected for each pixel location of the sample surface. However, for the success of the imaging technologies, application of data processing tools able to deal with big amount of data is necessary. Application of chemometrical tools is necessary at the different stages of the data analysis such as in compression (i.e. wavelets), in pretreatment (i.e. correcting baseline drifts) and in exploration[2, 7]. In this last step, several methods have been proposed to extract the maximum amount of information from the available spectral imaging data. Thus, Multivariate Image Analysis (MIA)[8] and Principal Component Analysis (PCA) have been applied[9] and, more recently, applications of Multivariate Curve Resolution (MCR-ALS) method have grown significantly[10]. MCR-ALS method allows the flexible application of constraints to obtain chemical (or biological) meaningful solutions which are easier to interpret especially when comparing with those obtained from other methods (i.e. PCA). Recent examples of application of MCR methods to hyperspectral imaging data in the literature dealing with monitoring of retina inflammation[11], monitoring of polymorphic transformations[12] or environmental remote sensing[13].

Mass Spectrometry Imaging (MSI), also known as Imaging Mass Spectrometry (IMS), technology is an extremely useful tool for the study of complex mixtures in real biological samples such as cells or tissues[14-16]. Its usefulness is due to its high chemical specificity in a way that allows analyzing simultaneously multiple molecular species in a very wide mass range, from small (i.e. metabolites) to large molecules (i.e. proteins). In addition to the qualitative information about the presence or absence of a particular molecule, MSI allows obtaining its spatial distribution in the analyzed sample surface[17]. Thus, MSI couples the spatial information provided by the spectral imaging techniques with the chemical specificity based on the mass accuracy of the high resolution mass spectrometry techniques (and possible MS/MS analysis) that allows the detected molecules unambiguous identification[18].

MSI experiments can be carried out by using one of several MS ionization techniques available that offer complementary capabilities[15]. Most commonly ionization techniques are secondary ion mass spectrometry (SIMS), laser desorption/ionization (LDI), desorption electrospray ionization (DESI) or matrix-assisted laser desorption/ionization (MALDI). For instance, SIMS imaging shows the higher spatial resolution imaging over a low mass range (until approximately 1000 Da), but requires high-vacuum conditions with pressures lower than 10^{-6} mbar. MALDI imaging allows working in a wider mass range (approximately 100000 Da) but at a lower spatial resolution and both high vacuum or atmospheric pressure. However, in some applications the sample preparation is rather complex.

Finally, DESI is used for almost untreated sample surfaces at atmospheric pressure for analysis of small molecules but at lower spatial resolution.

In this work, it is proposed the application of the Multivariate Curve Resolution to the exploration and analysis of MSI data sets. It is expected that the application of this resolution method will enhance the quantity and quality of information obtained from MSI data, in comparison with traditional methods despite the drawbacks in data processing and analysis that need to be overcome. Several articles are already available regarding the application of chemometric methods (mostly Principal Component Analysis and hierarchical clustering) to the analysis of MSI data (mainly TOF-SIMS generated data)[19-23], and only a couple of works are dealing with the application of MCR-ALS analysis to MSI data [24, 25].

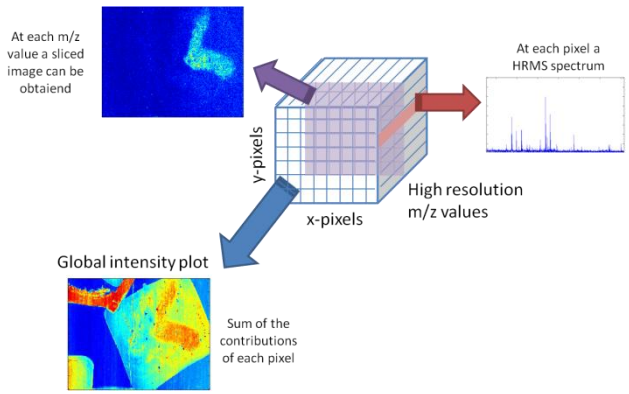
2. Materials and methods

2.1. Data sets under analysis

In this work, two data sets from the OpenMSI web-based platform[26] have been analyzed.

The first data set was generated by Louie and coworkers[27] and is an example of Mass Spectrometry Imaging using Nanostructure-initiator mass spectrometry (NIMS)[28] as desorption technique and an ABSciex 5800 TOF/TOF mass spectrometer as a detector. In that work, two microbes, *Shewanella oneidensis* (MR1) and *Pseudomonas stutzeri* (RCH2), were placed on agar film and the surface was scanned generating an image of 250 x 160 pixels (pixel size 100 microns) and 116152 m/z values. So, at least four different contributions could be expected: support material, LB agar media and the two considered microbes (MR1 and RCH2). More details about the fabrication of the NIMS wafer surface or the experimental details of the MS imaging could be found at the original work[27]. The second data set was generated by Marko-Varga and co-workers[23, 29] and is an example of the combination of MALDI desorption and detection by a Thermo LTQ Orbitrap XL. In this case, a section of a mouse lung was scanned and the obtained hyperspectral image has a size of 149 x 132 pixels (pixel size 50 microns) and 500000 m/z values. In this case, MSI is expected to allow observing different parts of the lung as the external membrane and blood vessels. Details about the sample preparation and measurement from the lung extraction to the experimental MS parameters could be found at the original work[23, 29].

MS imaging data can be arranged in a data cube in which the x- and y- axis correspond to the pixels building up the image and the z-axis corresponds to the mass spectrum obtained at each pixel. A graphical description of this data structure is shown in Scheme 1. For exploratory analysis purposes, it is possible to extract directly some information from the data cube[2]. First, it is possible to study the mass spectrum obtained at each pixel. However, this approach fails in the case of real (complex) samples in which more than one chemical compound is present at each pixel and, so, the amount of information that can be extracted is limited.



Scheme 1 Mass spectrometry imaging (MSI) data cube.

Second, it is possible to take a slice of the data cube at a considered m/z value. If prior knowledge about the scanned system is available, it can be useful to search for a specific compound (i.e., if a particular pollutant is known to be accumulated in a section of the tissue under analysis). If this prior knowledge is not available then the selection of m/z values that provide information is extremely difficult. Finally, if all the mass intensity values for a certain pixel are summed in a single value, a global intensity plot can be obtained. This total intensity graphical image displays the dominant spatial features of the image, but it does not allow obtaining detailed information about the chemical composition of the scanned surface.

2.2. Data pretreatment

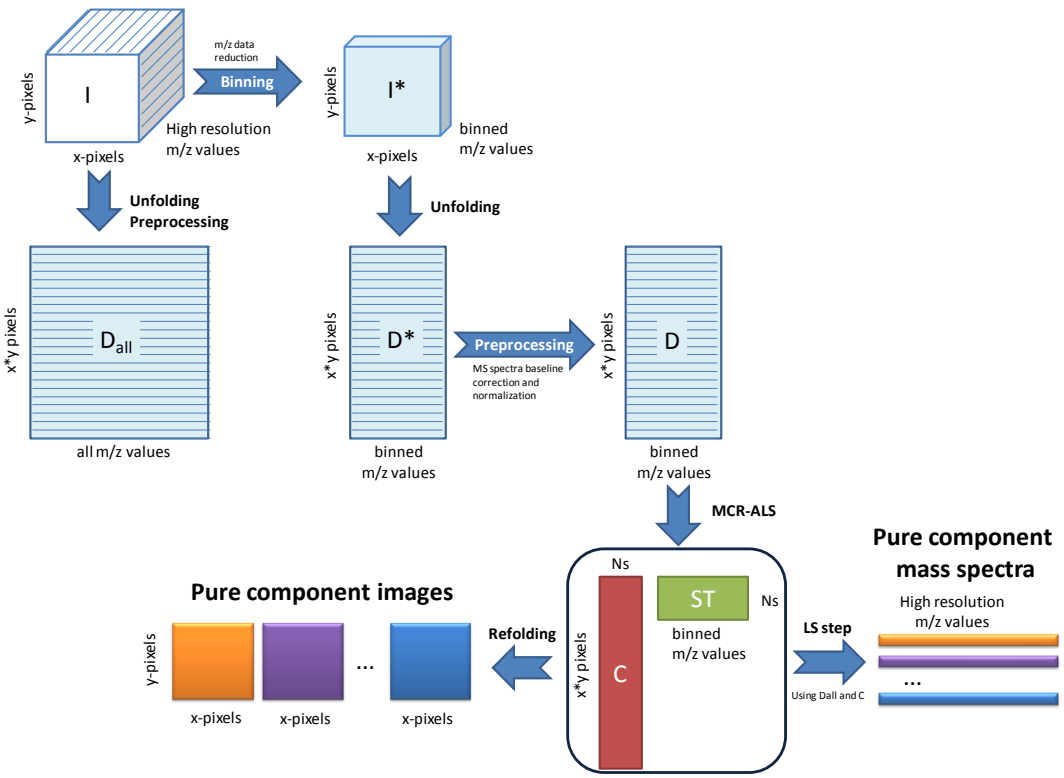
The analysis of raw MS imaging data is rather challenging due to the extremely large size of the mass spectra generated by high-resolution instruments[20]. In addition, data pretreatments are needed to enhance the obtained signal and facilitate its further analysis. A scheme of the steps needed in this data preparation is shown in Scheme 2.

Data pretreatment starts with a compression of the image prior to its analysis to reduce its size and obtain a data set which can be analyzed using a reasonable computational time. For instance, the number of elements of the data cube (**I**) is of more than 4500 millions (250 x 160 x 116152) in the case of the microbes data set and almost 10000 millions (149 x 132 x 500000) in the case of the lung data set. Despite the increasing power of current laboratory computers, chemometric analysis of these huge data sets would be extremely slow and impractical. It is clear that a data compression method is required. Since spatial information is low resolution compared to the acquired mass spectra, the binning of the mass spectrum at each pixel was preferred for both data examples. After binning, in the microbial interactions example the number of elements of the data cube (**I***) was reduced to 232 million (250 x 160 x 5800, approximately a 5% of the original size) while in the lung example was of only 1200 million (149 x 132 x 6000, approximately a 1% of the original size).

Prior to the application of other pretreatments, the data cube was unfolded into a two-dimensional data table (**D**) (see more details

in section 2.3.). In this step, the data cube **I*** of dimensions (number of x pixels, number of y pixels, number of binned m/z values) was unfolded to a data table of dimensions (number of x-pixels x number of y-pixels, number of binned m/z values). At each row of the data table, there is the mass spectrum corresponding to a single image pixel. For instance, the first row has the mass spectrum of the first pixel at the top left corner of the image and last row has the mass spectrum of the last pixel at the bottom right corner of the image. Likewise, each column of the refolded data table contains the signals for all pixels at the considered m/z values (see Scheme 2). In this way, microbial interaction data set has a final matrix size of 40000 x 5800 while mouse lung data set has a matrix size of 19668 x 6000.

Finally, the image data was further preprocessed to enhance its signal to noise ratio and features. Two pretreatments have been used. First, a baseline correction of all mass spectra was applied using the asymmetric least-squares algorithm (AsLS)[30]. Second, mass spectra of all pixels were normalized to have equal length[19, 20]. The application of this normalization step allowed studying the data independently of the absolute intensity of each pixel. This normalization has the advantage of minimizing the variation between pixels due to ionization differences between components and to spikes effects (pixels at certain m/z values which present extremely high intensity values). It is useful to identify pixels with similar composition in the MS image, although it should be used with caution because this normalization can also cause the loss of relevant quantitative information in some situations[31].



Scheme 2 Flowchart of the different data pretreatment, data analysis and postprocessing steps used in this work.

2.3. Bilinear model and Multivariate Curve Resolution analysis

In the case of hyperspectral imaging data using vibrational techniques (Raman, IR, NIR, …), the bilinear model defined by Lambert-Beer's law for spectroscopic measurements has been proposed. Absorption measured at different wavelengths is additive (linear) and is defined by the product of a term related to concentration and another to the spectral properties as can be seen in Equation 1.

$$d_{ij} = \sum_{n=1}^{Ns} c_{in} s_{nj} + e_{ij} \qquad \text{Equation 1}$$

Where each individual d_{ij} value represents the signal measured for the ith sample at the jth channel, c_{in} represents the concentration values of n species for the sample i and s_{nj} represents the spectral properties of n species at channel j. When the bilinear model is fulfilled, each d_{ij} value is the total sum of product of concentration and spectral values for the Ns considered components (constituents or species). Finally, e_{ij} value corresponds to the error contribution to the measurement that does not follow the bilinear model.

The previous bilinear model is extended in this study for the analysis of the data table (**D**) obtained from unfolding the MS image. In this case, each of the considered components will be characterized by a contribution or concentration value at each image pixel and by its pure mass spectrum. In matrix form Equation 1 can be written as:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \qquad \text{Equation 2}$$

where **D** is the data matrix containing the unfolded MS image to be analyzed of size (number x-pixels x number y-pixels) rows by (number of binned m/z values) columns. **D** data matrix is decomposed into the product of matrix **C** of dimensions number (x-pixels x number y-pixels, *Ns*) and by matrix $\mathbf{S^T}$ of dimensions (*Ns*, number of binned m/z values). *Ns* is the number of components related to the main sources of data variance, i.e., it is assumed that most of the information relevant to the image can be explained by a few number of *Ns* components. Matrix **E** of dimensions (number x-pixels x number y-pixels, number of binned m/z values) contains the variance not explained by the bilinear model, i.e., by the resolved profiles in **C** and $\mathbf{S^T}$ matrices.

There are many chemometric tools able to decompose **D** matrix according to Equation 2. For instance, one popular method commonly used to explore spectral images is Principal Component Analysis (PCA) in which the data matrix **D** is decomposed into a few number of principal components giving orthogonal scores (information related to image pixels) and orthonormal loadings (information related to m/z spectral values), respectively[9]. However, other resolution methods have been developed due to the possibility to impose more natural constraints to the components in order to obtain more easily interpretable solutions. Among these methods, Multivariate Curve Resolution by Alternating Least Squares (MCR-ALS) has been already successfully applied to the analysis of other type of spectral images[12, 32]. MCR-ALS solves iteratively Equation 2 by an Alternating Least Squares optimization algorithm under constraints and calculates matrices **C** and $\mathbf{S^T}$ that optimally fit the data matrix **D**[33, 34].

The number of components is usually preliminary estimated by means of PCA or by the Singular Value Decomposition (SVD) algorithm[35]. However, in the case of MS imaging this estimation is not straightforward and, in many cases, the complete data analysis has to be repeated using different number of components in order to select a simple model that allows obtaining interpretable information as well as fitting appropriately the data (without overfitting) ALS optimization starts by using initial guesses of either **C** or $\mathbf{S^T}$. In the case of analysis of MS images, an initial estimation of $\mathbf{S^T}$ matrix obtained by means of the estimation of the purest pixels[36] is preferred. MCR-ALS uses an iterative alternating least squares (ALS) constrained optimization solved in two separate linear least squares steps, one to estimate the **C** (concentration profiles) matrix and another one to estimate the $\mathbf{S^T}$ (spectra profiles) matrix of the multiple components contributing to the MSI signal. At each of these iterative steps, the least squared problem is solved under non-negativity and spectral normalization constraints[37, 38] (other constraints are also possible as implemented in the current version of the method[38]). When constraints are appropriately defined, the ALS optimization converges fast to a minimum with optimal data fitting. The advantages of MCR-ALS over other MCR methods proposed in the literature are that implementation of constraints and extension to very complex data sets is rather easy.

ALS optimization concludes when in two consecutive iterative cycles, relative differences in standard deviations of the residuals between experimental and ALS calculated data values are less than a previously selected value (usually 0.1%). Quality of the fitting of the data can be measured by the amount of explained variance calculated according to the following expression:

$$R^2 = 100\,\frac{\sum_{ij} d_{ij}^2 - \sum_{ij} e_{ij}^2}{\sum_{ij} d_{ij}^2} \qquad \text{Equation 3}$$

where d_{ij} designs an element of the input data matrix **D** and e_{ij} is the related residual obtained from the difference between the input element and the MCR-ALS reproduced matrix.

2.4. MCR-ALS solutions postprocessing

The interpretability of the information obtained from MCR-ALS profiles present in resolved **C** and $\mathbf{S^T}$ profiles can be enhanced by appropriate postprocessing. Each column of matrix **C** giving the relative contributions of a particular component in all the image pixels can be refolded appropriately into a two-dimensional distribution map image of this component on the entire scanned surface[2]. On the other hand $\mathbf{S^T}$ rows contain the mass spectra of the resolved components at the resolution used for MCR-ALS analysis. It is interesting to mention that several possible chemical compounds can be observed (i.e. several peaks at multiple m/z values) in the signal corresponding to the same component resolved spectra. However, from these resolved spectra, it is not possible to unambiguously identify the image constituents yet, because binned m/z values (necessary due to computer limitations) have not enough resolution to be used for exact mass compound characterization. In order to recover the high-resolution present in raw measurements and allow exact mass component identifications, a single non-negative least-squares step was used, where matrix $\mathbf{D_{all}}$ (size of x-pixels x y-pixels by all m/z values) with all m/z data is projected on MCR-ALS resolved **C** profiles. This single least-squares step allows estimating full resolution mass spectra $\mathbf{S^T_{all}}$ for all MCR-ALS resolved components. Using them, the identification of the image constituents is then possible by comparison with MS spectra compiled in public libraries such as MassBank[39] or Lipid Maps[40, 41].

2.5. Software and hardware

All calculations were performed using MATLAB® software running on a HP Z620 Workstation equipped with two Intel® Xeon® E5-2620 processors and 32Gb Ram using Windows 7. PCA analysis has been carried out using the Eigenvector PLS Toolbox for the MATLAB® environment. Multivariate Curve Resolution routines are freely available at the webpage www.mcrals.info.

3. Results and Discussion

Two examples of application of the MCR-ALS method to the analysis of MSI data are given below. These two examples show the advantages of the application of MCR-ALS to this type of data considering different ionization and detection techniques

such as NIMS ionization and TOF/TOF detection in the case of the microbial interaction example, and MALDI ionization and Orbitrap detection in the cause of the mouse lung example.

3.1. Microbial interaction data set

As stated above, preprocessing and analysis of this raw data were impractical with the available computers because of its huge size. The analysis of this data set required its preliminary compression. The size of the data cube after binning was 250x160x5800 (binning in m/z ranges of 0.25 units). This data cube was then unfolded to a a two-way data matrix of size 40000x5800). Data background and baseline was corrected using the AsLS method (see methods section) and raw spectral data were normalized to equal length.

First step in the Multivariate Curve Resolution procedure was the preliminary estimation of the data complexity and number of components by SVD data analysis). From the plot of the singular values, in Figure 1a the decision about how many components should be selected for an accurate data description is not straightforward. It is preferred therefore to perform the MCR-ALS analysis for a different number of components and, then, decide about the model that extracts maximum interpretable and reliable information with the smallest number of components, trying at the same time not to overfit the data. In this case, MCR-ALS models with components ranging from 6 to 18 were tested and evaluated.

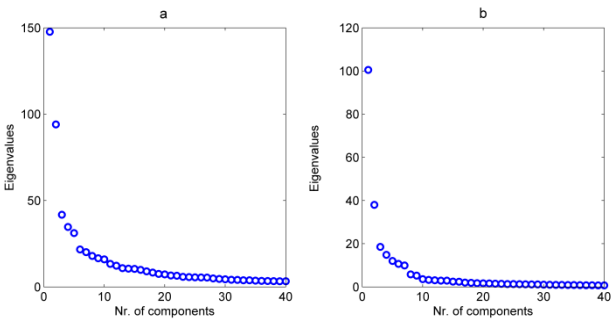


Fig. 1 SVD analysis results of a) microbes and b) lung data sets.

For each one of these models initial estimations were obtained from purest variables in the data. ALS optimization was then performed using non-negativity constraints for pixels (rows) and mass spectra (columns) and normalizing the mass spectra to equal length. Finally, the MCR-ALS model using 15 components (explained variance of 92.1%) was selected as appropriate. Resolved distribution maps (obtained from the refolding of each column of the resolved **C** matrix) and their related binned mass spectra ($\mathbf{S}^{\mathbf{T}}$ matrix) for this particular model with 15 components are shown in Figure 2. The comparison of these results with those given in Figure S1 (results obtained when initial spectral estimates were used directly to calculate the **C** matrix by a single least-squares step), clearly shows that the application of the ALS iterative optimization procedure under constraints gave better results (i.e. distribution maps) easier to interpret.

To allow a better explanation of the whole image, some of the resolved components have been interpreted together. For instance, MCR-ALS components 1, 3, 4, 8 and 12 in Figure 2 are all related to the image background (scanning support). However, they differ on their composition and contribution at each pixel, as it can be seen in their individually resolved MS spectra. Thus, a visual inspection of the resolved distribution maps did allow grouping all resolved MCR-ALS components into five groups.

First, MCR-ALS component number 11 is assigned to microbe RCH2. Its distribution map is clearly distinguished from the background agar plate and its related mass spectrum allows identifying the principal m/z values related to this particular microbe: 621.97, 575.30, 738.41 and 257.01. However, microbe MR1 image contribution could not be totally resolved by a single component and MCR-ALS components 10 and 13 were also needed. It can be seen that MCR-ALS component 13 resolves the two ends of the microbe image (see component 13 in Figure 2) and that its contribution in the central part of the image is resolved by MCR-ALS component number 10. The study of the corresponding MCR-ALS spectra for these two components confirmed some differences in their mass spectra. Whereas intensity signals at some m/z values were common such as 523.41, 686.28 or 254.03, other m/z values were present only at the ends of the image (598.14, 305.17 and 694.22 for component 13) or in the central region of the microbe image (242.36, 270.17 and 854.11 for component 10). MCR-ALS resolved component number 15 shows a highly located and strong signal between the two microbes. Its corresponding resolved mass spectrum shows an important peak located at m/z 550.47 and other minor peaks at m/z of 522.41, 598.14 and 746.11 which can be assigned to the interaction between both microbes [27]. Another important group of MCR-ALS resolved components (2, 5, 6, 7 and 9) can be considered to be linked to the agar surface on which the microbes lay. Their distribution maps can be assigned to the regions closer to the perimeter of one of the two microbes, indicating their different metabolic processes [27]. This can be confirmed from their resolved MS spectra, in which, despite of the fact that most of the detected peaks are common for the different components; there are only some peaks that appear in a particular component. For instance, MCR-ALS resolved component number 5 is closer to the RCH2 microbe and shows clearly the peak at a m/z value of 623.21, while MCR-ALS component number 9, is closer to the MR1 microbe and presents an intense peak at a m/z value of 746.36, which was not present in the other resolved agar group of components. In the fifth group of MCR-ALS resolved components (1, 3, 4, 8 and 12), three clear contributions (1, 3 and 4) can be associated to the signal contribution from the support material used for the NIMS ionization (background). Distribution maps show different background contributions to the signal. In this case, their resolved mass spectra resulted to be rather similar (mass spectra peaks at m/z values lower than 500) which means that the composition of the background surface was rather similar. However, in the distribution maps two additional regions could be identified at the bottom and top left corners of the image. In the former case (component 12), the region provided a defined mass spectrum while in the latter (component 8), no meaningfully signal was recovered after the treatment. Only MCR-ALS component number 10 is still difficult to be assigned to one of the previously described group of components, because its distribution map was not well defined and its resolved mass

spectrum showed a very large number of peaks at high m/z values.

Exact mass estimations can be used to reduce the number of candidate compounds assigned to a particular peak at a certain mass to charge value. After MCR-ALS analysis using lower resolution, it is possible to recover the exact mass spectrum in the original m/z scale by a single non-negative least squares step, relating the raw spectra matrix, $\mathbf{D}_{all}$, and the concentration matrix, **C** resolved by MCR-ALS. In this way, the original information available in the raw mass spectra is conserved. For instance, as example, the exact mass of the major peaks resolved for the components related to the couple of microbes were 523.3732 for the MR1 microbe and 575.3735 for the RCH2 microbe. These recovered mass to charge values were searched in MassBank and Lipid Maps online databases considering M(neutral), $[M+H]^+$, $[M+Na]^+$ and $[M+K]^+$ ions and allowing a mass error lower than 50 ppm. In the case of mass to charge value 523.3732, six candidate compounds were obtained (see Table S1 in Supplementary material for the detailed list of candidates). Neutral, $[M+H]^+$ and $[M+K]^+$ adducts have been obtained but it is difficult to decide the most suitable candidate. With respect to mass to charge value 575.3735, nine candidate compounds were obtained (see Table S2 for more details). In this case, eight of the candidates (adduct with postassium) are different isomeric carotenes: lycopene, α-carotene, β-carotene, … Unambiguos determination of the compound would require more information such as fragmentation patterns obtained in a MS/MS experiment. Assignation of any other resolved mass peak of any component could be done in the same way.

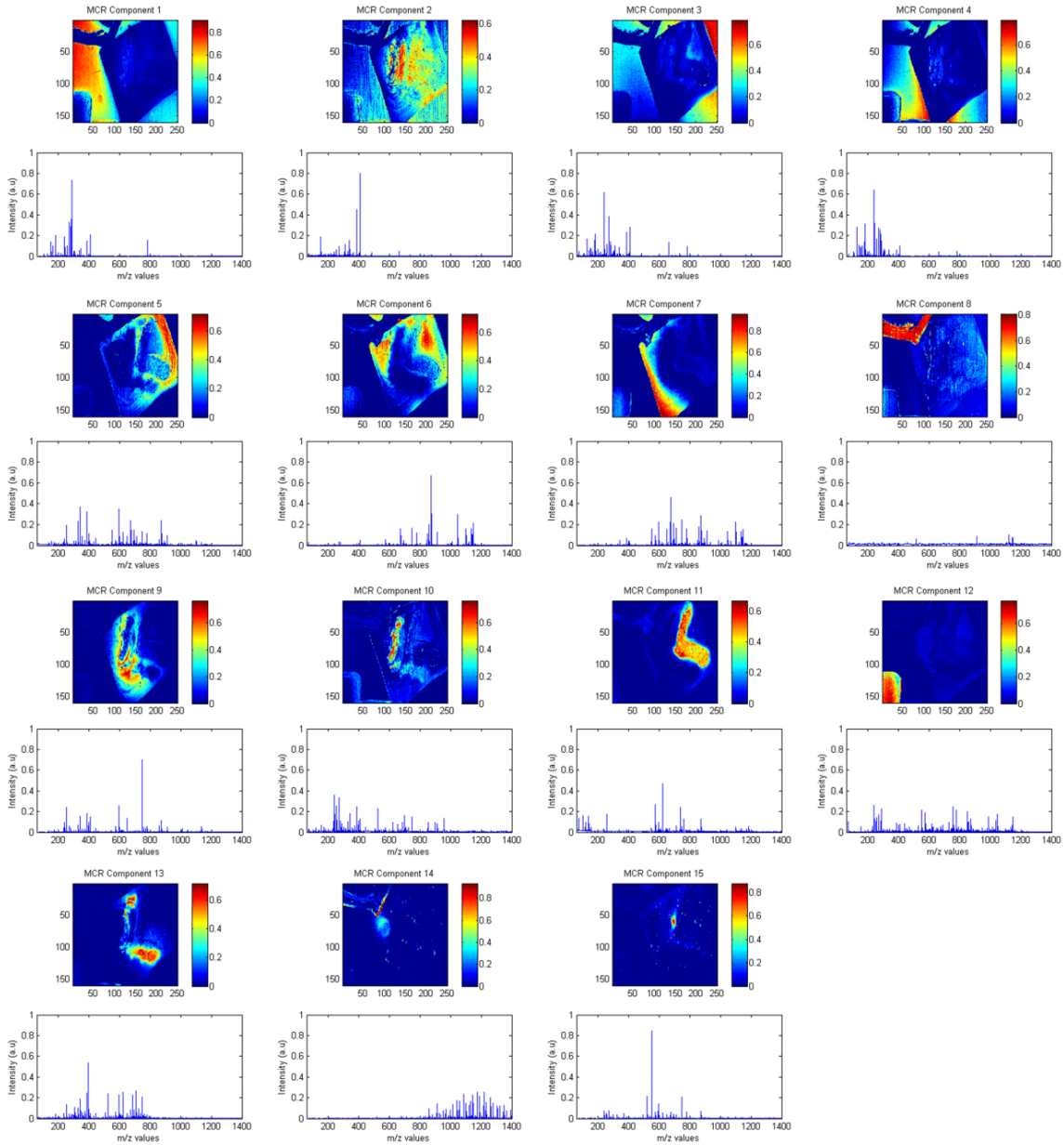


Fig. 2 MCR-ALS results of microbes data set. Distribution maps after refolding and MS spectra of all resolved components

Finally, as stated above, the total explained variance by the MCR-ALS model using 15 components was 92.1% (PCA explained variance of 92.5%). When considering a smaller number of components the total explained variance was also high (i.e. using 6 component the variance explained by the model was already 87.2%) but the interpretation of the resolved distribution maps and mass spectra was then more difficult since some of the previously described components appeared mixed in a single component. The amount of variance explained by each of the resolved components of the MCR-ALS model is shown in Table 1.

Table 1 Individual and total explained variances by MCR-ALS resolved components for the microbes and lung data sets. MCR-ALS components are sorted by the amount of individual explained variance in decreasing order.

	Microbes data set	Lung data set
	Explained variance (%)	Explained variance (%)
1	22.5	58.6
2	14.0	26.0
3	13.9	15.4
4	12.9	9.7
5	11.1	8.4
6	10.7	7.8
7	9.8	4.5
8	9.3	4.5
9	8.2	3.6
10	5.9	3.6
11	5.5	-
12	4.5	-
13	4.4	-
14	1.3	-
15	0.9	-
Sum of Individual component variances	134.9	142.1
Total variance (all components)	92.1	98.9
PCA explained variance	92.5	99.1

It can be seen that the components that explain a higher amount of variance are mostly related to the background and agar regions. For instance, the sum of the data variances explained by the background assigned components was 63%, while the data variance explained by the microbes regions was only 15% of the total. On the other hand, when variances explained by each individual component were added, the obtained value was significantly higher than the total variance explained by the MCR-ALS model, and over than 100% (134.9% vs 92.1%). This is because MCR-ALS resolved components are not orthogonal and they do overlap. This is an important difference with PCA, in which the orthogonality of the components forces explained variances not to overlap between components (their covariance is 0). However, MCR-ALS component profiles are closer to real ones which may have multiple overlapping contributions and MS signals.

3.2. Mouse lung data set

The second example is different and it was tested in order to check the reliability of MCR-ALS results. The analyzed data set corresponds to the scanning of a lung tissue sample using MALDI as desorption and ionization technique and an Orbitrap MS detector.

The first step in the analysis of raw MSI data is to perform its compression. In this example, the size of the data cube after binning m/z values is 149x132x6000 (binning in m/z ranges of 0.275 units). After this transformation, the data cube was unfolded to a matrix of size 19668x6000, and data baseline correction was performed using AsLS, as well as spectra normalization. An initial estimation of the total number of components is obtained by using the SVD algorithm. In this case, the number of selected components was significantly smaller than in the case of the microbial interaction (see Figure 1b). However, as it was mentioned above, it was not possible to decide about the total number of components and, therefore several MCR-ALS analysis were carried out in order to decide what number of components was finally considered according to their possible interpretation. In this example, the considered number of components ranged between 6 and 12.

The procedure of analysis was the same as before, with initial estimations obtained from the purest variables in the data set and with ALS optimization under non-negativity constraints for both, pixel intensities and mass spectra, and normalization of the later (to equal vector length). MCR-ALS model with 10 components (explained variance of 98.9%) was finally selected and distribution maps and binned mass spectra ($\mathbf{S}^T$ matrix) of the resolved components are shown in Figure 3.

In this second example, the interpretation of the MCR-ALS resolved components resulted not to be as straightforward as in previous case. Three different lung regions were distinguished based on differences in resolved mass spectra and distribution maps. In Figure 3, resolved MCR-ALS components 4, 8 and 9 are mostly present at the external membrane of the lung (pleura). These MCR-ALS components showed two groups of signals at m/z values around 650 and 875 which can be assigned to lipids. Resolved MCR-ALS components 1 and 2 are related to the parenchyma region and explain most of the data variance. Resolved mass spectra showed peaks in the m/z range from 400 to 800. A third group of MCR-ALS components 3, 5, 6 and 10 can be related to the blood vessels present in the lung as it can be seen from their distribution maps. Whereas MCR-ALS components 3 and 6 can be attributed to thinner blood vessels, MCR-ALS components 5 and 10 are probably more related to thicker (main) blood vessels. Their mass spectra showed a profile with several peaks at the region between 400 and 800 m/z which corresponds mostly to different families of biological compounds (proteins, lipids …) [23, 42]. Other components (for instance, such as MCR-ALS component number 8) can also have a small contribution in the same blood vessels region.

Finally, it is also interesting to remark the behaviour of MCR-ALS resolved component number 7, whose distribution map spreads around the whole lung region with some pixels at higher

concentrations and, also, with an intense spot outside the lung. Its MCR-ALS resolved mass spectrum presents a well defined MS peak at m/z value of 392.02 (LS obtained exact mass 392.0991). After literature searching, this peak can be assigned to tiotropium which results to be a bronchodilator in previous studies by Végvári and Nilsson [43, 44].

Total explained variance by this MCR-ALS model was similar to that explained by the PCA model with the same number of components (99.1% vs 98.9%). The sum of the variances explained by each component individually grows up to 142.1%, showing that the amount of variance overlapping between different MCR-ALS components is considerable. It is worth to mention the significant amount of variance explained individually by the MCR-ALS component number 3 (almost 60%). As it can be seen in its resolved distribution map, this component is present at a relatively high intensity in almost the whole lung tissue.

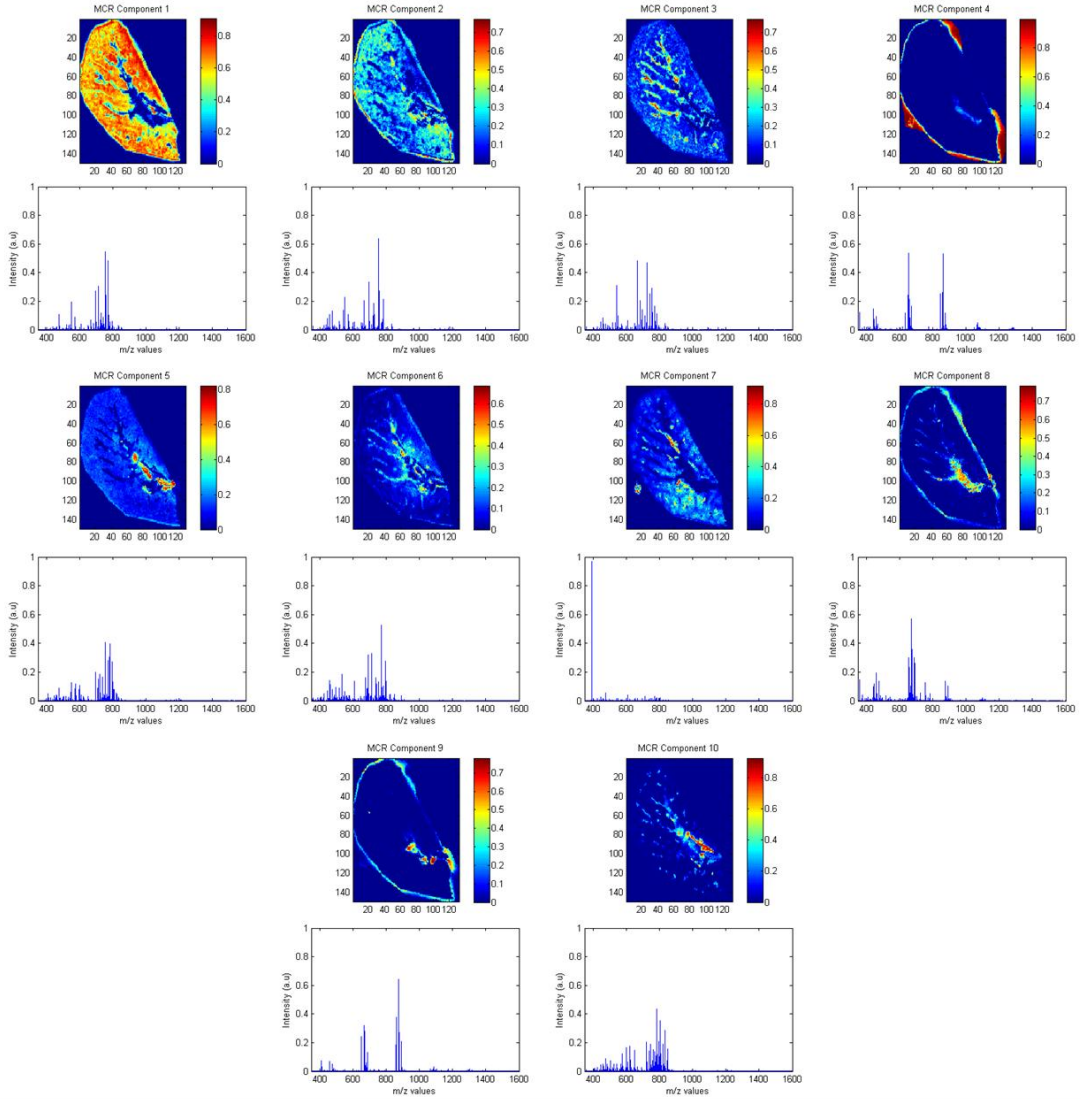


Fig. 3 MCR-ALS results of lung data set. Distribution maps after refolding and MS spectra of all resolved components.

Conclusions

The combination of Mass Spectrometry Imaging (MSI) and advanced data analysis tools such as Multivariate Curve Resolution has allowed the extraction of valuable information from the two highly complex massive data sets investigated in this work. Such a combination is proposed to be used in –omics studies in which the spatial information about the location of a target molecule is required as well as for the detection and/or confirmation of possible biomarkers. Multivariate Curve Resolution application to MSI data is rather simple and results are easy to interpret, providing the distribution maps and mass spectra of the individual resolved components present in the analyzed samples, from which qualitative and semi quantitative relative information about their distribution over different sections of the scanned surface can be further recovered. More work is needed however, to overcome some of the drawbacks still present in this type of analysis such as the decision about the total number of components to be finally included in the analysis, and the computational limitations due to the huge size of Mass Spectrometry Imaging data sets at full resolution.

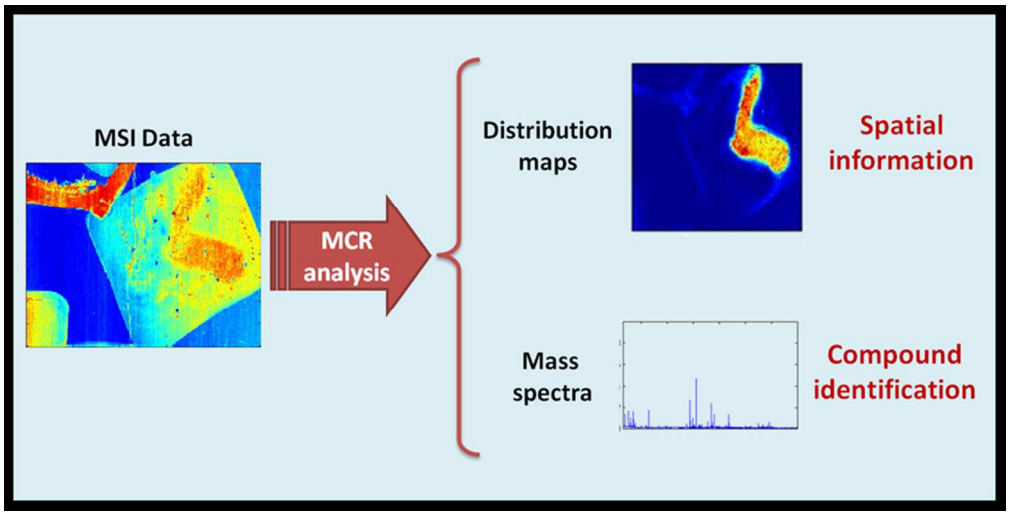
Notes and references

[a] *Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, Barcelona 08034, Spain. Tel: (34)-934006140-1643; E-mail: joaquim.jaumot@idaea.csic.es*

Acknowledgments. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 32073. Also, recognition from the Catalan government (grant 2009SGR45) is acknowledged. JJ acknowledges a CSIC JAE-Doc contract cofounded by the FSE.

† Electronic Supplementary Information (ESI) available: Figure S1, Table S1 and Table S2. See DOI: 10.1039/b000000x/

1. H. F. Grahn and P. Geladi, *Techniques and Applications of Hyperspectral Image Analysis*, 2007.
2. R. Salzer and H. W. Siesler, *Infrared and Raman Spectroscopic Imaging*, 2009.
3. D. Liu, X. A. Zeng and D. W. Sun, *Applied Spectroscopy Reviews*, 2013, **48**, 609-628.
4. A. A. Gowen, C. P. O'Donnell, P. J. Cullen and S. E. J. Bell, *European Journal of Pharmaceutics and Biopharmaceutics*, 2008, **69**, 10-22.
5. A. R. Kherlopian, T. Song, Q. Duan, M. A. Neimark, M. J. Po, J. K. Gohagan and A. F. Laine, *BMC Systems Biology*, 2008, **2**.
6. F. D. van der Meer, H. M. van der Werff, F. J. van Ruitenbeek, C. A. Hecker, W. H. Bakker, M. F. Noomen, M. van der Meijde, E. J. M. Carranza, J. B. de Smeth and T. Woldai, *International Journal of Applied Earth Observation and Geoinformation*, 2012, **14**, 112-128.
7. J. M. Amigo, *Analytical and Bioanalytical Chemistry*, 2010, **398**, 93-109.
8. K. Esbensen and P. Geladi, *Chemometrics and Intelligent Laboratory Systems*, 1989, **7**, 67-86.
9. P. Geladi, H. Isaksson, L. Lindqvist, S. Wold and K. Esbensen, *Chemometrics and Intelligent Laboratory Systems*, 1989, **5**, 209-220.
10. A. De Juan, R. Tauler, R. Dyson, C. Marcolli, M. Rault and M. Maeder, *TrAC - Trends in Analytical Chemistry*, 2004, **23**, 70-79.
11. M. Marro, A. Taubes, A. Abernathy, S. Balint, B. Moreno, B. Sanchez-Dalmau, E. H. Martínez-Lapiscina, I. Amat-Roldan, D. Petrov and P. Villoslada, *Journal of Biophotonics*, 2013.
12. S. Piqueras, L. Duponchel, R. Tauler and A. de Juan, *Analytica chimica acta*, 2014, **819**, 15-25.
13. X. Zhang and R. Tauler, *Analytica chimica acta*, 2013, **762**, 25-38.
14. M. Setou, ed., *Imaging Mass Spectrometry. Protocols for Mass Microscopy*, Springer Japan, Tokyo, Japan, 2010.
15. S. S. Rubakhin and J. V. Sweedler, eds., *Mass Spectrometry Imaging. Principles and protocols*, Humana Press, New York, US, 2010.
16. K. B. Louie, B. P. Bowen, S. McAlhany, Y. Huang, J. C. Price, J. H. Mao, M. Hellerstein and T. R. Northen, *Scientific reports*, 2013, **3**, 1656.
17. K. Chughtai and R. M. Heeren, *Chemical reviews*, 2010, **110**, 3237-3277.
18. L. A. McDonnell and R. M. A. Heeren, *Mass Spectrometry Reviews*, 2007, **26**, 606-643.
19. J. M. Fonville, C. Carter, O. Cloarec, J. K. Nicholson, J. C. Lindon, J. Bunch and E. Holmes, *Analytical chemistry*, 2011, **84**, 1310-1319.
20. E. A. Jones, S. O. Deininger, P. C. W. Hogendoorn, A. M. Deelder and L. A. McDonnell, *Journal of proteomics*, 2012, **75**, 4962-4989.
21. V. Pirro, L. S. Eberlin, P. Oliveri and R. G. Cooks, *The Analyst*, 2012, **137**, 2374-2380.
22. A. M. Race, R. T. Steven, A. D. Palmer, I. B. Styles and J. Bunch, *Analytical chemistry*, 2013, **85**, 3071-3078.
23. F. Suits, T. E. Fehniger, A. Vegvari, G. Marko-Varga and P. Horvatovich, *Analytical chemistry*, 2013, **85**, 4398-4404.
24. W. Rao, D. J. Scurr, J. Burston, M. R. Alexander and D. A. Barrett, *The Analyst*, 2012, **137**, 3946-3953.
25. W. Rao, A. D. Celiz, D. J. Scurr, M. R. Alexander and D. A. Barrett, *Journal of the American Society for Mass Spectrometry*, 2013, **24**, 1927-1936.
26. O. Rubel, A. Greiner, S. Cholia, K. Louie, E. W. Bethel, T. R. Northen and B. P. Bowen, *Analytical chemistry*, 2013, **85**, 10354-10361.
27. K. B. Louie, B. P. Bowen, X. Cheng, J. E. Berleman, R. Chakraborty, A. Deutschbauer, A. Arkin and T. R. Northen, *Analytical chemistry*, 2013, **85**, 10856-10862.
28. T. R. Northen, O. Yanes, M. T. Northen, D. Marrinucci, W. Uritboonthai, J. Apon, S. L. Golledge, A. Nordstrom and G. Siuzdak, *Nature*, 2007, **449**, 1033-1036.
29. G. Marko-Varga, T. E. Fehniger, M. Rezeli, B. Dome, T. Laurell and A. Vegvari, *Journal of proteomics*, 2011, **74**, 982-992.
30. P. H. Eilers, *Analytical chemistry*, 2004, **76**, 404-411.
31. J. Burger and P. Geladi, *Journal of Near Infrared Spectroscopy*, 2007, **15**, 29-37.
32. A. A. Gowen, F. Marini, C. Esquerre, C. O'Donnell, G. Downey and J. Burger, *Analytica chimica acta*, 2011, **705**, 272-282.
33. C. Ruckebusch and L. Blanchet, *Analytica chimica acta*, 2013, **765**, 28-36.
34. R. Tauler, *Chemometrics and Intelligent Laboratory Systems*, 1995, **30**, 133-146.
35. G. H. Golub and C. Reinsch, *Numerische Mathematik*, 1970, **14**, 403-420.
36. W. Windig and J. Guilment, *Analytical chemistry*, 1991, **63**, 1425-1432.
37. A. De Juan, Y. Vander Heyden, R. Tauler and D. L. Massart, *Analytica chimica acta*, 1997, **346**, 307-318.
38. J. Jaumot, R. Gargallo, A. De Juan and R. Tauler, *Chemometrics and Intelligent Laboratory Systems*, 2005, **76**, 101-110.
39. H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito and T. Nishioka, *Journal of mass spectrometry : JMS*, 2010, **45**, 703-714.
40. E. Fahy, M. Sud, D. Cotter and S. Subramaniam, *Nucleic acids research*, 2007, **35**, W606-612.
41. M. Sud, E. Fahy, D. Cotter, A. Brown, E. A. Dennis, C. K. Glass, A. H. Merrill, Jr., R. C. Murphy, C. R. Raetz, D. W. Russell and S. Subramaniam, *Nucleic acids research*, 2007, **35**, D527-532.
42. T. E. Fehniger, F. Suits, Á. Végvári, P. Horvatovich, M. Foster and G. Marko-Varga, *PROTEOMICS*, 2014, **14**, 862-871.
43. Á. Végvári, T. E. Fehniger, M. Rezeli, T. Laurell, B. Döme, B. Jansson, C. Welinder and G. Marko-Varga, *Journal of Proteome Research*, 2013, **12**, 5626-5633.
44. A. Nilsson, T. E. Fehniger, L. Gustavsson, M. Andersson, K. Kenne, G. Marko-Varga and P. E. Andren, *PloS one*, 2010, **5**, e11411.



Application of MCR-ALS to Mass Spectrometry Imaging data provides spatial distribution and MS spectra of pure species allowing compound identification.
79x39mm (300 x 300 DPI)