research articles Journal of proteome research

Spatial Segmentation of Imaging Mass Spectrometry Data with Edge-Preserving Image Denoising and Clustering

Theodore Alexandrov,*,†,‡ Michael Becker,§ Sören-Oliver Deininger,§ Günther Ernst,|| Liane Wehder,|| Markus Grasmair,⊥ Ferdinand von Eggeling,|| Herbert Thiele,§ and Peter Maass†

Center for Industrial Mathematics (ZeTeM), University of Bremen, 28334 Bremen, Germany, Center for Computational Mass Spectrometry, University of California, San Diego, La Jolla, California 92093, United States, Bruker Daltonik GmbH, 28359 Bremen, Germany, Core Unit Chip Application, Institute of Human Genetics, University Hospital Jena, 07740 Jena, Germany, and Computational Science Center, University of Vienna, Vienna, Austria

Received July 15, 2010

In recent years, matrix-assisted laser desorption/ionization (MALDI)-imaging mass spectrometry has become a mature technology, allowing for reproducible high-resolution measurements to localize proteins and smaller molecules. However, despite this impressive technological advance, only a few papers have been published concerned with computational methods for MALDI-imaging data. We address this issue proposing a new procedure for spatial segmentation of MALDI-imaging data sets. This procedure clusters all spectra into different groups based on their similarity. This partition is represented by a segmentation map, which helps to understand the spatial structure of the sample. The core of our segmentation procedure is the edge-preserving denoising of images corresponding to specific masses that reduces pixel-to-pixel variability and improves the segmentation map significantly. Moreover, before applying denoising, we reduce the data set selecting peaks appearing in at least 1% of spectra. High dimensional discriminant clustering completes the procedure. We analyzed two data sets using the proposed pipeline. First, for a rat brain coronal section the calculated segmentation maps highlight the anatomical and functional structure of the brain. Second, a section of a neuroendocrine tumor invading the small intestine was interpreted where the tumor area was discriminated and functionally similar regions were indicated.

Keywords: Imaging mass spectrometry • bioinformatics • spatial segmentation • edge-preserving denoising • clustering • in situ proteomics • rat brain • neuroendocrine tumor

Introduction

For many years imaging of biological samples with mass spectrometry has been the Holy Grail of mass spectrometry research. Invention of such a technique would allow one studying spatial chemicat composition of any biological sample. Only in the late 90s of the previous century, development of matrix-assisted laser desorption/ionization (MALDI)−imaging mass spectrometry (IMS)[1,2] has opened new horizons for mass spectrometry in biology and medicine.[3] Since then, MALDI-imaging has become a mature technology, allowing for reproducible high-resolution measurements to localize proteins and smaller molecules for many purposes, in particular to detect and discover new biomarkers with a major focus in cancer research.[4−7] At the present time, a variety of MALDI-imaging instruments and preparation devices is manufactured and offered by major producers of mass spectrometers (Applied Biosystems, Bruker Daltonics, Shimadzu Biotech, and Waters). Along with attempts to apply SIMS to biological samples,[8] recently other IMS techniques have been developed and successfully applied in biology, including desorption electrospray ionization (DESI),[9] graphite-assisted laser desorption/ionization (GALDI),[10] laser ablation electrospray ionization (LAESI),[11] and nanostructure-initiator mass spectrometry (NIMS).[12] Surface enhanced laser desorption ionization (SELDI)−IMS was shown to be useful in histological analysis.[13] Despite the impressive technological advance of MALDI-imaging and other IMS techniques, at the present time only a few papers have been published concerned with computational methods for MALDI-imaging data. In this paper, we contribute to this area, considering the important issue of pixel-to-pixel variability in MALDI-imaging data and proposing a new method to reduce this variability. Upon the basis of this method, we present a new pipeline for spatial segmentation of a MALDI-imaging data set which compresses the full data set into one image, a segmentation map.

Development of new computational methods for MALDI-imaging is especially important since the state-of-the-art throughput of MALDI-imaging allows it to be used in clinical studies[14] with one of the main fields of interest in discovery

* To whom correspondence should be addressed. Dr. Theodore Alexandrov, Bibliothekstr. 1, 28359 Bremen, Germany. Phone: +49-421-218-63820. Fax: +49-421-218-98-63820. E-mail: theodore@math.uni-bremen.de.
† University of Bremen.
‡ University of California, San Diego.
§ Bruker Daltonik GmbH.
|| Institute of Human Genetics, University Hospital Jena.
⊥ University of Vienna.

10.1021/pr100734z

Published on Web 10/18/2010

and validation of biomarkers of human tumors.[4−7] This task requires measuring, processing, and understanding large numbers of patient samples to compare cohorts at several time points (see the review by McDonnell et al.).[15] At present, the most common way of examining MALDI-imaging data set is the manual inspection of a mean spectrum of the data set, selection of large peaks, and visual examination of molecular images corresponding to the selected m/z-values. This rather simple but straightforward approach allows for finding molecular masses specific to certain tissue states. However, it comes with several major drawbacks. First, the manual search of peaks is time-consuming and, therefore, is not feasible in a clinical study. Second, a molecular signal taking place in only a small portion of spectra can be under-represented in the mean spectrum and produce no visible peak. Third, all m/z-values are examined independently, although their combination may reveal more valuable information. Fourth, visual observation allows one to detect the most visible patterns of spatial localization, but fine details and differences between masses can hardly be detected. Thus, it is highly desirable to simplify complex MALDI-imaging data sets to allow for their interpretation in a reasonable time frame and, at the same time, retain the most important molecular features.

For this aim, the use of multivariate statistical methods is crucial. A widely used way of visualizing an IMS data set is to plot loadings derived with principal component analysis (PCA)[16] or improved variants of PCA[17] that provides images showing the main structure of the data set. However, the PCA loadings can hardly be interpreted from a mass spectrometric point of view, since certain mass spectra negatively contribute to the resulting image, that has no analytical meaning (see the paper by Deininger et al. for a discussion on disadvantages of PCA in the MALDI-imaging context).[18] The use of probabilistic latent semantic analysis (pLSA) remedies this problem[19] since pLSA loadings can be directly interpreted as contribution of masses, and this allows for interpreting spectra that are formed by a mixture of tissue types even if those are not spatially resolved.

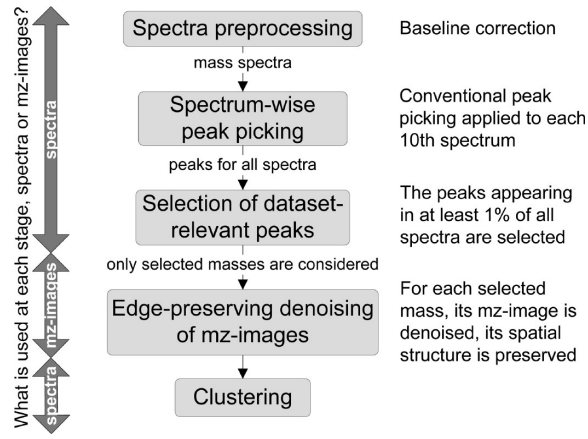
Later, partition of spatial points based on clustering of their mass spectra was proposed.[20] One can display the clustering results as a spatial segmentation map, coloring identically points grouped into one cluster. A segmentation map visualizes a MALDI-imaging data set with just one image and highlights regions of potential interest. Recently, hierarchical clustering has been introduced[18] and discussed[21] for the analysis of cancer data. The main advantage of hierarchical clustering in this context is interactive analysis when one can split a region of interest into subregions.

So far, all described statistical methods are purely based on the similarities of mass spectra alone and do not take their spatial relations into account. However, it is natural to expect that multivariate analysis of imaging data sets can be improved if spatial relations are considered.

In this paper, we propose a new approach to clustering MALDI-imaging spectra which provides segmentation maps of superior quality in terms of smoothness, lack of noise, level of detail, and correlation with morphological structures of the tissue. The core of this pipeline is based on the following natural assumption: for many neighboring spatial points of a morphologically defined area their spectra most likely represent similar molecular composition and, thus, should be similar.

Our procedure consists of the following steps (Scheme 1). First, the spectra are preprocessed with a baseline correction

Scheme 1. Spatial Segmentation Procedure for MALDI-Imaging Data



algorithm. No normalization is done.[22] Second, the peak picking is done selecting a list of data set-relevant peaks. Third, for each m/z-value from the selected peaks list, we consider an image of intensities of all spectra at this m/z-value and denoise it with locally adaptive edge-preserving image denoising algorithm, which is the most important step of the procedure. Finally, the reduced and processed spectra are clustered, and the clustering results are displayed as a spatial segmentation map in which spatial points whose spectra are grouped into one cluster are identically colored.

Here, we describe the procedure for the first time and apply it to two MALDI-imaging data sets. First, we analyze a rat brain coronal section and compare the resulting segmentation map to the anatomical structure of the brain. Brain tissue is a typical model system in MALDI-imaging because of its clear and well-studied anatomical structure, containing morphological features of different levels of detail. Using this data set as an example, we study the properties of MALDI-imaging mass spectra, and especially the pixel-to-pixel variation of spectra intensities. Second, we apply our procedure to a section of a neuroendocrine tumor (NET) invading the small intestine (ileum) proving the potential of our procedure for the analysis of highly complex tumor tissue samples.

Methods

Samples Preparation and Mass Spectrometry Measurements. Both for the rat brain and NET, cryosections of 10 μm thickness were cut on a cryostat (CM 1900 UV, Leica Microsystems GmbH, Weltzar, Germany) and transferred to a precooled, conductive indium-tin-oxide (ITO) coated glass slide (Bruker Daltonik GmbH, Bremen, Germany). The acquisition and evaluation were carried out using flexControl 3.0 and flexImaging 2.1 software (Bruker Daltonik GmbH).

Rat Brain. The sections were washed twice for 1 min in 70% ethanol, and once for 1 min in 96% ethanol and then dried in a vacuum desiccator. The matrix (Sinapinic acid at 10 mg/mL in 60% acetonitrile and 40% water with 0.2% trifluoroacetic acid) was applied using the ImagePrep device (Bruker Daltonik GmbH) following a standard protocol. Mass spectra were acquired on a MALDI-TOF instrument (Autoflex III; Bruker Daltonik GmbH) equipped with a 200 Hz smartbeam II laser. MALDI measurements were performed in linear positive mode at a mass range of 2.5 kDa to 25 kDa. The lateral resolution for the MALDI image was set to 80 μm. A total of 200 laser shots

were summed up per position. For data processing, we considered only the mass range from 2.5 kDa−10 kDa.

Neuroendocrine Tumor (NET). The sections were washed twice for 30 s in 70% ethanol, and once for 20 s in 96% ethanol, and then dried in a vacuum desiccator. The matrix was applied in the same way as for the rat brain sample. Mass spectra were acquired on a MALDI-TOF instrument (Autoflex III, Bruker Daltonik GmbH) equipped with a 200 Hz smartbeam II laser. MALDI measurements were performed in linear positive mode at a mass range of 1 kDa to 30 kDa with a lateral resolution of 50 μm and 300 laser shots per position. For data processing, we considered only the mass range from 3.2−18 kDa. After MALDI analysis, the matrix was washed off using 70% ethanol, and a conventional Haematoxylin and Eosin (H&E) staining was performed. The stained sections, coregistered with the MALDI-imaging results, were evaluated histologically by an experienced pathologist (GE) using a virtual slide scanner (MIRAX desk, Carl Zeiss MicroImaging GmbH, Munich, Germany).

Mass Spectrometry Data Preprocessing. The preprocessing of spectra was performed in ClinProTools 2.2 (Bruker Daltonik GmbH). Spectra were baseline corrected with the TopHat algorithm (minimal baseline width set to 10%, the default value in ClinProTools). No normalization or binning was done. Then spectra were saved into ASCII files and loaded in Matlab R2007b (The Mathworks Inc., Natick (MA), USA) where the rest of the processing was performed. The rat brain data set comprises 20 185 spectra acquired within area of the sample, each of 3045 data points covering the mass range 2.55−10 kDa; the NET data set comprises 27 360 spectra each of 5027 data points covering 3.2−18 kDa.

Peak Picking. In this step, we performed peak picking for the whole data set generating a list of data set-relevant peaks. The aim of this operation is to reduce the length of spectra selecting only informative peaks and discarding m/z-values which show no peaks in any spectra. First, we considered each 10th spectrum to speed up the procedure. For each of the considered spectra, we selected 10 peaks.

Naturally, for processing of a still huge number of spectra we need an efficient method, which disqualifies the use of computationally inefficient methods as continuous wavelet transformation or ridge lines. At the same time, peak picking should be robust to strong noise, preventing the use of too simple local maxima or signal-to-noise ratio methods, which produce too many false positives. We used our original peak picking method based on the orthogonal matching pursuit (OMP) algorithm,[23] which models each peak with a shape function. Note that this approach is also used in the popular mass-spectrometry processing software OpenMS[30] and MapQuant.[31]

In our approach, each spectrum is modeled as a sequence of Dirac delta peaks convolved with the Gaussian kernel (as in MapQuant[31]) plus noise. Assuming this model, the problem of peak picking is equivalent to the problem of deconvolution. For the deconvolution, we use OMP because it is simple, fast, allows for specification of the number of sought-for peaks, and is widely applied in signal processing. Denis et al. discussed advantages of OMP over other deconvolution algorithms.[23] To the best of our knowledge, this publication is the first one describing application of OMP-based peak picking to real-life mass spectrometry data.

The Gaussian kernel is selected as a reasonable approximation of the peak shape (Figure 2). In our experience, the OMP algorithm is robust to deviations in the shape and symmetry of peaks. As a simplification, we assume the width of a peak to be mass-independent and estimate it manually considering several large peaks. The parameter sigma of the Gaussian kernel is calculated with the two-sigma rule dividing the peak width by four.

After collecting the peaks lists for all considered spectra, we have a joint list of potential peaks. Among them, we select only those consensus peaks which appear in at least 1% of considered spectra. This reasonable assumption allows us to omit spurious peaks which take place in just a few spectra.

Edge-Preserving Denoising of *m/z*-Images. At this stage, we consider a MALDI-imaging data set as a datacube with 3-coordinates: x, y, and m/z (note that the data set is reduced in the number of m/z-values by the peak picking). Given the m/z-value, an image of intensities of all spectra at this m/z-value can be reconstructed, which we call the m/z-image.

The core of our procedure is denoising of m/z-images. So far, the existing procedures of clustering MALDI-imaging data are prone to noise[18−20] that complicates interpretation of their results and hides structural details. This is explained by the fact that MALDI-imaging data are contaminated with strong noise. A typical tissue sample represents a highly complex mixture of analytes with strong differences in abundance which in itself has strong effects on analyte ionization, leading to chemical noise. Na^+ and K^+ ions present in every tissue result in adduct formation aside from the commonly observed protonated analyte ions. In addition, a tissue section represents a far from perfect surface for matrix crystallization as compared to steel target plates commonly used in regular MALDI measurements. In addition to forming a relatively uneven surface from which ions are extracted, there are numerous effects such as uneven crystallization of the matrix or charge accumulation, which generally leads to reduced spectra quality and increased noise levels.

Recently, the smoothing of the resulting classification map was proposed,[24] which, although it brings some improvement, cannot reconstruct the details lost at the stage of data processing. More natural would be to denoise each m/z-image. However, the large variance of noise which, moreover, varies inside each individual m/z-image and between different m/z-images, makes denoising of m/z-images a challenging problem (see discussion). Moreover, when performing denoising, the aim is not to obscure the structure of an m/z-image by mixing up intensities of two neighboring morphological regions. This would smooth out the edges between regions and erode details, which is not acceptable when the tissue has complex structure with fine anatomical or histological details (e.g., tumor tissue).

Thus, standard image-denoising filters (median or convolution filter) are inappropriate for denoising of m/z-images (see discussion). We propose to exploit edge-preserving image denoising. One of the most popular methods for this purpose is the total variation (TV)-minimizing[25] Chambolle algorithm.[26] Informally speaking, TV is the sum of absolute differences between neighboring pixels. Noise increases TV significantly and TV-minimization algorithms, given an image, search for its approximation with small TV. The Chambolle algorithm, however, has the drawback that the level of smoothness of the output image can be adjusted only globally by manually choosing a parameter. We exploit a modification of the Chambolle algorithm proposed by Grasmair that adjusts the level of denoising to the local noise level and the local scale of the features to be resolved.[27] The Grasmair algorithm locally adapts the denoising parameter of the Chambolle algorithm

in an automatic way, increasing it in the areas with high noise level and decreasing it in the areas with low noise level, thus providing locally adaptive edge-preserving denoising. The main parameter of the Grasmair method is the level θ of smoothness of the resulting image (between 0.5 and 1; the higher, the smoother); for other parameters, we used their default values. Our own implementation was used.

Clustering. The peak picking reduces the full data set to intensities at considered m/z-values. Then the edge-preserving denoising is individually applied to each m/z-image replacing it with its denoised version.

The final step of our segmentation procedure is to cluster all reduced and processed spectra with a clustering algorithm. We do not attempt to estimate the number of clusters from the data (using, for example, the Akaike method), but rather specify it a priori. This is more reasonable in a general study because, first, in a state-of-the-art MALDI-imaging study one is interested in a small numbers of clusters (up to 10), so segmentation maps for all numbers of clusters can be computed quite fast, and second, visual observation of a segmentation map by a histologist provides an all-purpose way of evaluation and selection of the best number of clusters.

For the clustering we used the high dimensional discriminant clustering (HDDC) method[28] whose implementation is freely available through the MATLAB Central File Exchange repository. HDDC can be seen as a generalization of the linear discriminant analysis, where each cluster is modeled by a Gaussian distribution of its own covariance structure. The efficient calculation of the Gaussian parameters is based on the idea of modeling each cluster in its own subspace of reduced dimension (its so-called intrinstic dimension). The HDDC is developed for high-dimensional data (informally speaking, the clustering problem is referred to be high-dimensional if dimensionality of the data is larger than 10) where the curse of dimensionality disqualifies simple clustering methods.

Results

Rat Brain Data Set.

Peak Picking. The rat brain data set consists of 20 185 spectra, where for peak picking we considered only 2019 (10% of all) spectra. The joint list of potential peaks, which includes all peaks found in the considered spectra, contains 373 peaks and 110 of them were selected as consensus peaks taking place in at least 1% of spectra, that is, in 20 out of 2019 (Figure 1).

In Figure 2, we show two example spectra from the rat brain data set (a representative spectrum with spatial coordinates $x = 56$, $y = 105$, and a noisy spectrum, with $x = 170$, $y = 53$) together with the detected peaks 10 peaks per spectrum). One can see that (1) OMP successfully detects the major peaks, and (2) the Gaussian function provides reasonable approximation of the peak shape.

Figure 1 shows that most of the discarded peaks, namely, those appearing in less than 20 out of 2019 considered spectra, are located in the low-mass range (38% in 2.5−3 kDa, 73% in 2.5−4 kDa). Although they might correspond to some rare low-mass chemical compounds, we hypothesize that they are the noise peaks. MALDI-imaging spectra normally have baseline which is high in the low-mass region and then decreases taking small values for large masses. As discussed below, the noise variance is proportional to the peak intensity, which may lead

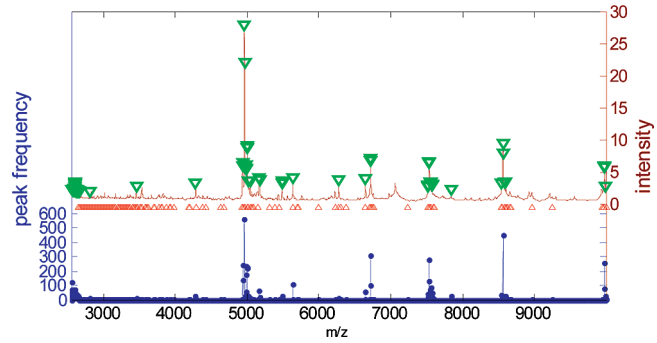


Figure 1. The data set-relevant peaks for the rat brain data set. Brown: the mean spectrum. Blue: the frequency of a peak (number of spectra in which the peak is detected); high values correspond to most observable peaks. Red triangles: potential peaks (found in at least one spectrum). Green triangles: selected consensus peaks peaks (found in at least 1% of spectra).

to high noise variance in the low-mass region and, correspondingly, to high random spikes in this region falsely detected as peaks.

Noise in MALDI-Imaging. The noise in MALDI-imaging spectra is strong (Figure 3), and this issue needs to be addressed. For a large peak, its intensity range can vary significantly from spectrum to spectrum (i.e., from one spatial point to another). The largest peak (at m/z 4963.5) takes values from 0.4 to 153. The peaks intensities histograms are unimodal and smooth which may indicate that peak intensities change randomly (affected by noise). The presence of strong noise is confirmed by visual observation of m/z-images corresponding to the selected peaks (Figure 3B). Note that the noise variance changes both within an image and between different images. Figure 3C illustrates this observation, showing the histograms of intensity values in four spatial areas for the m/z-image at 4963.5 (the largest peak in the mean spectrum), two areas of high intensity (A1, A4) and two areas of low intensity (A2, A3). The histograms demonstrate that in the highly intense areas, the variance of noise is higher. This effect is also observed in other m/z-images. Note that in the highly intense areas (A1, A4) the peak intensities range down almost to zero, thus making the variance of large peaks extremely high. Finally, Supplementary Figure 1, Supporting Information shows that the noise variance at a spatial point linearly depends (with correlation coefficient 0.96) on the mean intensity around this point that may point out the Poisson distribution of the noise. Thus, we conclude that (1) the noise is strong, (2) the noise variance changes within an m/z-image and between different m/z-images, (3) the noise variance is linearly proportional to the peak intensity.

Edge-Preserving Denoising. After selecting 110 peaks, we apply the edge-preserving denoising to m/z-images corresponding to these peaks. Examples of m/z-images and their denoised versions are shown in Figure 4. The Grasmair method efficiently removes the noise while not smoothing out edges.

Segmentation Map. The segmentation map after clustering with edge-preserving denoising is presented in Figure 5 together with an optical image of the analyzed rat brain section and a schematic of the anatomical structure. The major anatomical regions are well represented. When judging the quality of the representation, it is important to consider that only mass spectral information was used to recreate anatomical features in a completely automated way with no prior knowledge about the sample being utilized. Cortex (pale green,

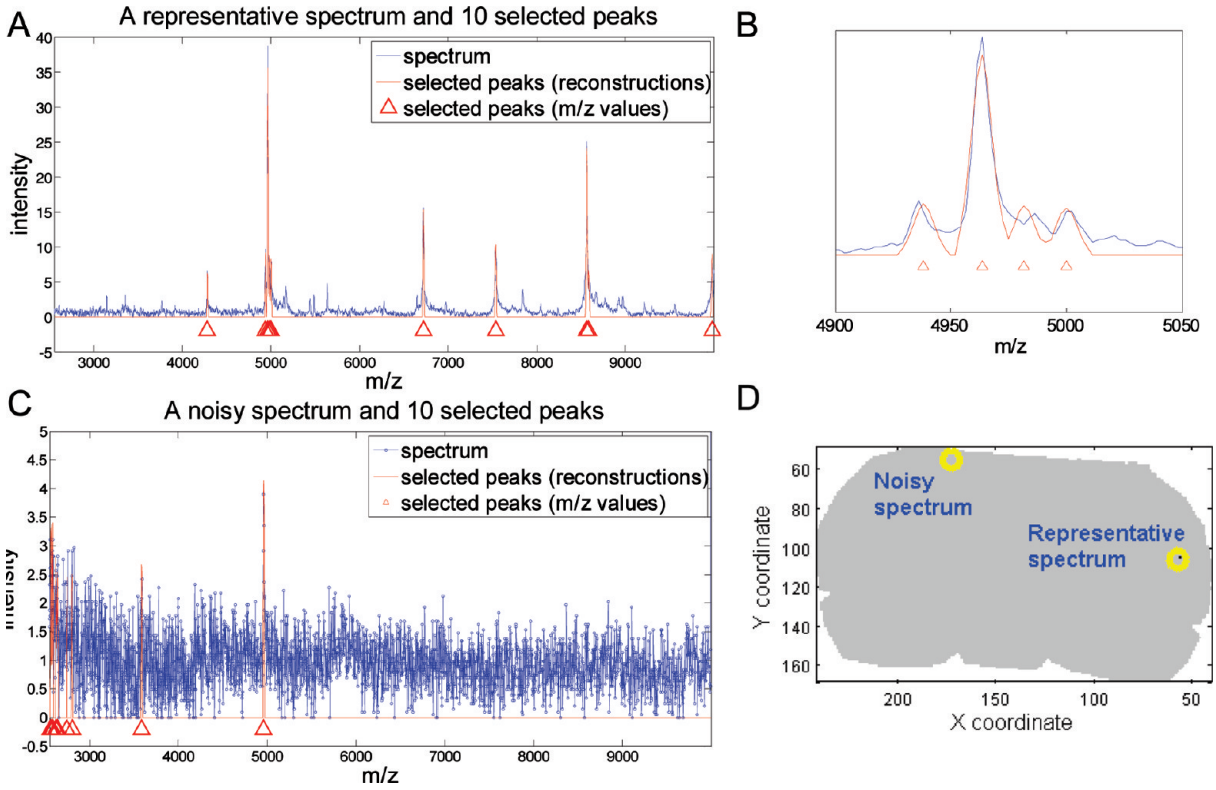


Figure 2. Two example spectra from the rat brain data set and selected peaks for them. (A) A representative spectrum, (B) its zoomed region, (C) noisy spectrum without prominent peaks, (D) spatial positions of the spectra. The reconstructions (red curves) are created summing up the Gaussian kernels found by the deconvolution.

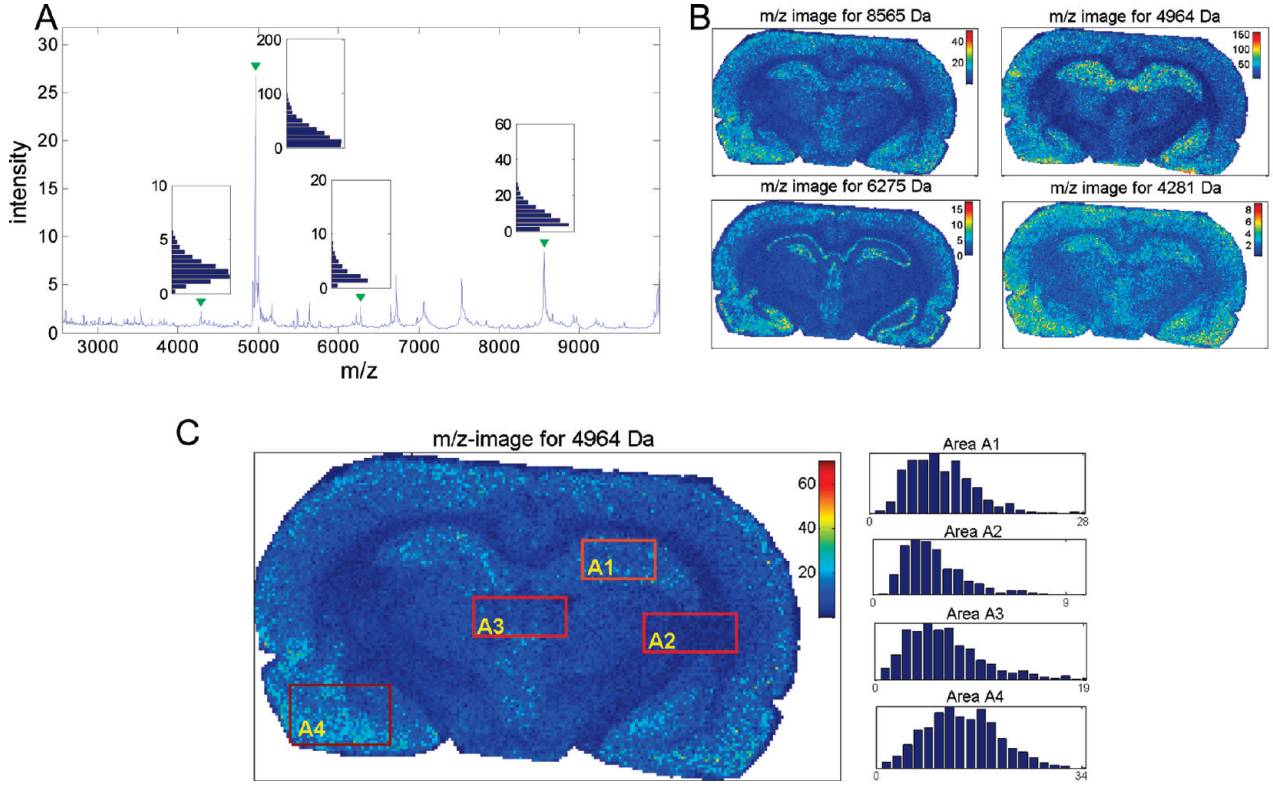


Figure 3. Noise properties for the rat brain data set. (A) The mean spectrum and histograms (rotated 90° clockwise for illustrative purposes) of intensities for peaks at *m*/*z* 4281.0, 4963.5, 6274.7, and 8563.8. (B) *m*/*z*-images for the same peaks. (C) *m*/*z*-image for 4963.5 and histograms of its values in four spatial areas. The areas A1, A4 (A2, A3) of high (low) intensity are selected manually.

dorsal), hippocampus (brown and light blue), thalamus (orange, in the central part), hypothalamus (dark blue, ventral), amygdala (red), and the paraventricular nuclei (light blue) are all well represented. It is of particular interest that both the dorsal and

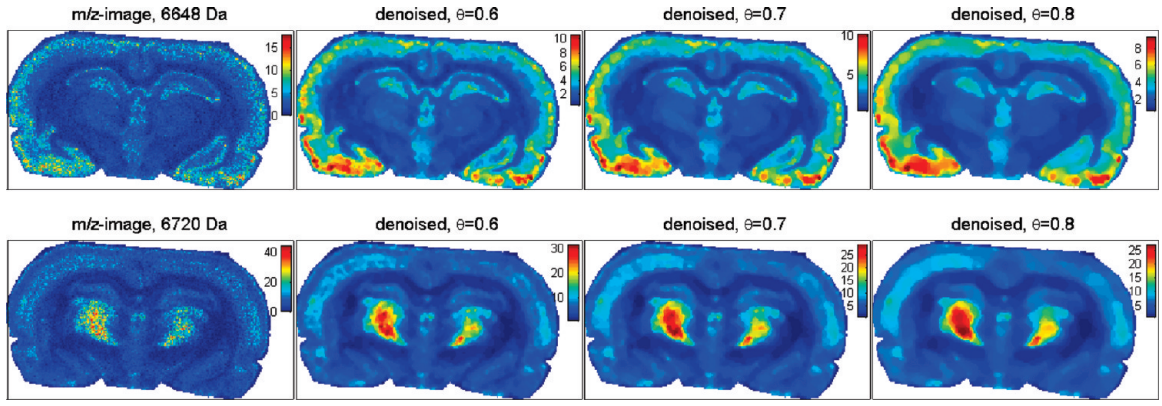


Figure 4. Two example *m*/*z*-images from the rat brain data set. For each *m*/*z*-image, results of weak ($\theta = 0.6$), moderate ($\theta = 0.7$) and strong ($\theta = 0.8$) edge-preserving image denoising are shown.

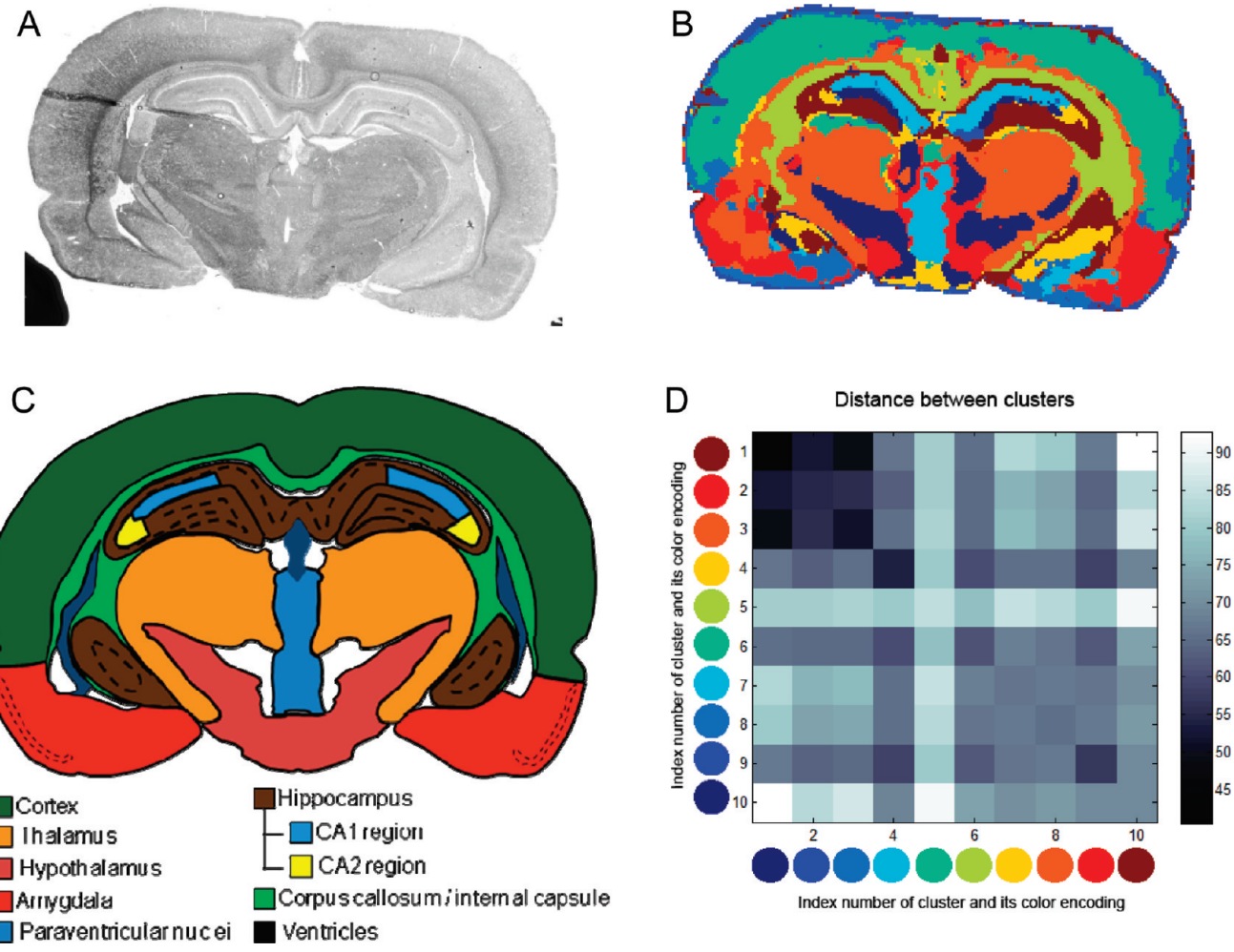


Figure 5. The rat brain data set. (A) Optical image. (B) Segmentation map (10 clusters, $\theta = 0.7$). (C) Schematic of the anatomical structure of the rat brain corresponding to coronal section ~4.16 mm from Bregma. (D) The matrix showing distances between clusters (dark color of an element of the matrix means that clusters corresponding to the row and column of this element are similar).

ventral parts of the hippocampus have been correctly assigned to the same clusters (CA2 region in brown, CA3 region in light blue) although they are not interconnected in the section shown. From the optical image, it is obvious that the ventral part of the hippocampus is larger and better represented on the right-hand side of the section, which explains the clearer representation in corresponding area on the segmentation map. The corpus callosum and the internal capsule are two prominent anatomical structures which are directly interconnected. They are not separated on the segmentation map (both shown in light green), which can be explained by their functional similarity. Both are part of the white matter and therefore contain numerous axonal fibres. It is not surprising that funtional similarities are represented in the similarity of profile spectra, which in turn results in spectra from both anatomical regions ending up in the same cluster.

Both the lateral and the third ventricles are well visible on the optical image but not picked up by the segmentation map.

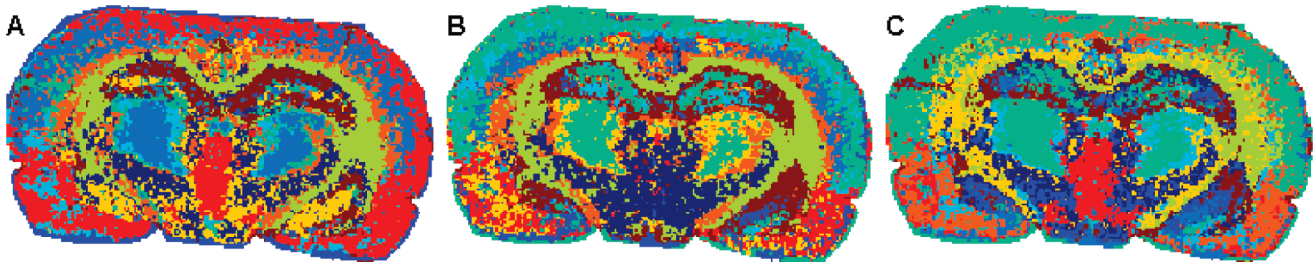


Figure 6. Segmentation maps for the rat brain data set. (A) No denoising. (B) Median filtering, 3 × 3 window. (C) Median filtering, 5 × 5 window.

We suppose that they were not smoothed out during denoising but characterized by a low signal intensities, as ventricles represent cavities in the brain. The clustering selects major features in the spectra (by intensity and number of spectra) and for this reason misses the ventricles.

Note that although the segmentation maps look coarser than the anatomical structure, they are able to detect thin features of one-two pixels width. So, the spatial resolution of the segmentation map is mainly restricted by the spatial resolution of the MALDI-imaging data set.

Importance of Edge-Preserving Denoising. Figure 6 shows segmentation maps produced without denoising of m/z-images as well as with simple median denoising. In the segmentation map calculated without denoising, most major anatomical regions can still be recognized, but their borders match the actual anatomical features not as well as in the segmentation map for denoised data (Figure 5B). More importantly, although using the same number of clusters (10), the assignment of the major anatomical features into independent clusters is not as good for the denoised data. For example, both thalamic and cortical areas (blue in Figure 6A) as well as hippocampal and hypothalamic areas (dark yellow in Figure 6A) have been assigned to the same cluster. Cortex and amygdala are not clearly defined but mixed up in two clusters (red and light blue in Figure 6A).

Use of simple and well-known median filtering algorithm instead of edge-preserving denoising leads to inferior results. With a 3 × 3 size (Figure 6B), the paraventricular nuclei (light blue in Figure 5B) are not visualized and the general preservation of edges is much worse. With a 5 × 5 size (Figure 6C) although the amygdala (orange in Figure 6C) and the paraventricular nuclei (red in Figure 6C) are somewhat visible, the hippocampal area is almost completely disintegrated and mixed up with other regions. Another asymmetric cluster (light green on the right in Figure 6C) could not be matched with an existing anatomical feature; as a result, simple filtering methods do not appear as useful as edge-preserving filtering.

Co-Localized Masses. Finally, after spatial segmentation of a data set, one might be interested in finding masses the most colocalized with a specific segment. In particular, these masses can be used to identify proteins (or peptides) using tandem mass spectrometry that can be done either from the extract of a full tissue sample (for abundant proteins), from microdissected cells (low-abundant proteins), or using tandem MS imaging.[29] Figure 7 shows the most colocalized masses for six clusters of the segmentation map from Figure 5B. The colocalization is measured by the correlation with the spatial mask specified by the cluster.

The Role of Parameters.

Peak Picking. The peak picking does data reduction and significantly speeds up further analysis. At the same time, note that large peaks usually express spatially structural information. Thus, peak picking simplifies the problem of clustering removing masses mostly representing noise. The three main parameters used in addition to the peak width are (1) portion of spectra considered for peak picking (selection of each 10th spectrum is recommended), (2) the number of peaks selected for an individual spectrum (10 is recommended), and (3) the percentage of spectra where a peak is to be found to be selected in the final consensus peak list (1% is recommended). Figure 8 shows segmentation maps for different values of the second and third parameters. One can see that the results are robust to changes of these parameters. The numbers of selected peaks (Figure 8, Supplementary Figure 2, Supporting Information) show that these two parameters are coupled in a way that an increase of the first parameter can be compensated by higher values of the second one. However, an increase of each of them slows down the procedure (by requiring more iterations of OMP and/or by selecting more peaks at the end). In our experience, the combination of 10 and 1% works well for many MALDI-imaging data sets (results not shown). Supplementary Figure 2, Supporting Information shows results when each 5th and 20th spectrum is considered for peak picking and reveals that this parameter does not affect the number of selected peaks.

Denoising and the Number of Clusters. Let us consider the segmentation maps for the rat brain data set produced with three levels of denoising (weak, $\theta = 0.6$; moderate, $\theta = 0.7$; and strong, $\theta = 0.8$) and three numbers of clusters (6, 8, and 10), Figure 9. As expected, a decrease in the number of clusters merges together some features separated before. At the same time, a similar effect is observed when denoising gets stronger since some neighboring details are oversmoothed. As a result, the level of denoising should not be increased too much in order to get smooth-looking images, especially in case of structures with fine details, such as tumor sections.

Human Neuroendocrine Tumor Data Set. Notwithstanding that the brain data set is complex, a brain section shows a clear anatomical structure that can be compared with a text-book. In contrast, tumor sections do not show a standard structure; that is, every tumor section is different and requires considerable expertise to be evaluated. In the context of clinical research, clustering methods are of particular interest to facilitate the interpretation of tumor data sets. Therefore, we have applied our method to the analysis of a human neuroendocrine tumor section.

The H&E stained tissue section was annotated by an experienced pathologist (GE), indicating different functional areas of the tissue (Figure 10A). The segmentation map represents histological structures in detail (Figure 10B,C). The tumor area and all main structural components of the small intestine wall could be allocated entirely. Functional processes/structures are highlighted in Figure 10A. The segmentation map

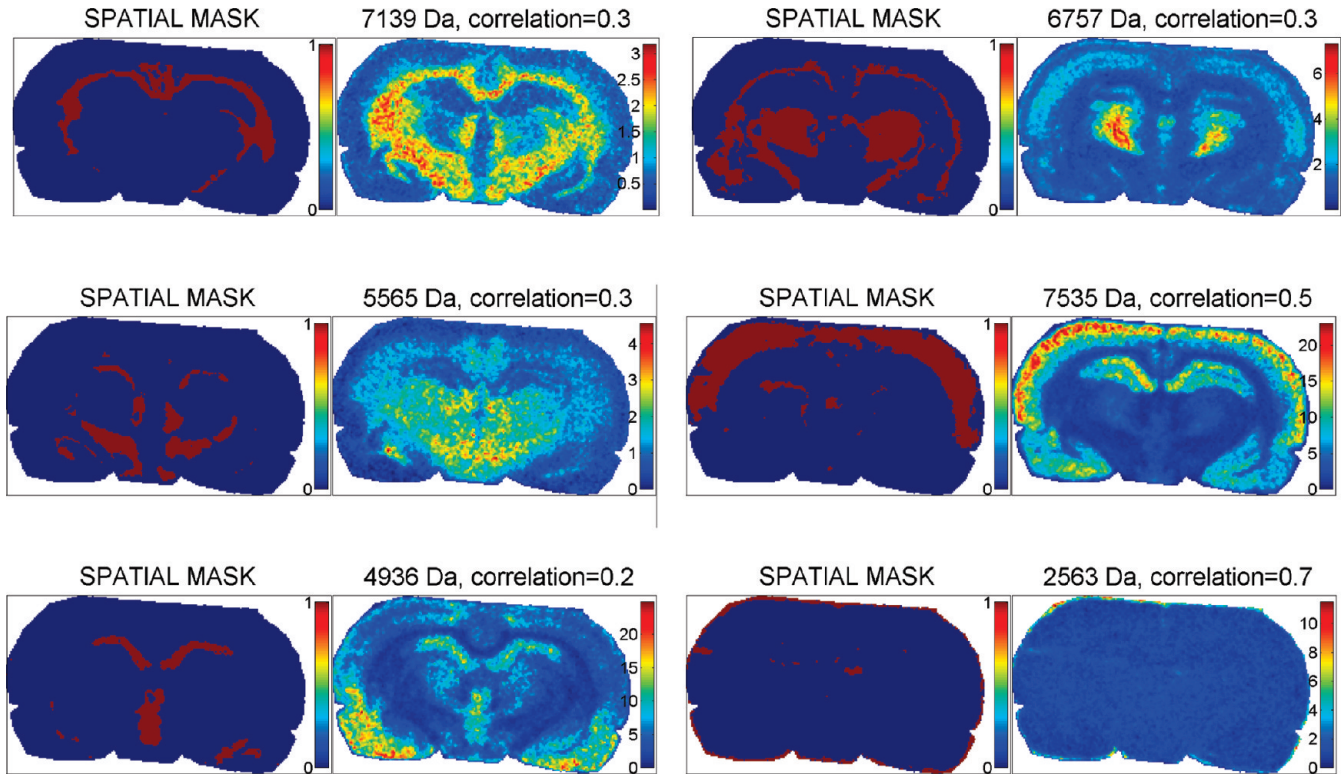


Figure 7. The most colocalized masses for six clusters of the segmentation map (Figure 5B) for the rat brain data set. Two-colored image shows the spatial mask (cluster), and the next image shows *m*/*z*-image of the colocalized mass.

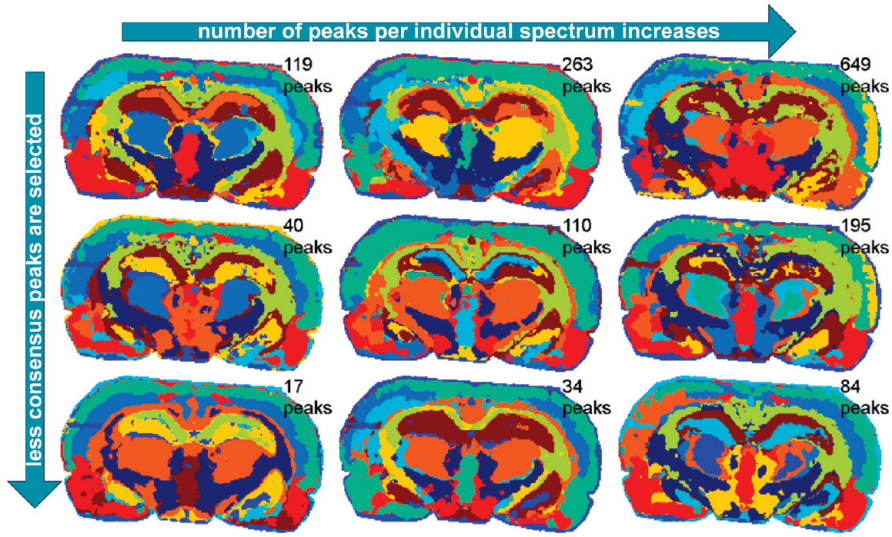


Figure 8. Segmentation maps for the rat brain data set for different parameters of peak picking: number of peaks selected for an individual spectrum (first column: 5, second column: 10, third column: 20 peaks) and different percentage thresholds for the consensus peaks (first row: 0.1%, second row: 1%, third row: 5%). The moderate level of denoising ($\theta = 0.7$) is used; each 10th spectrum is considered for peak picking; see also Supplementary Figure 2, Supporting Information.

displays the same segments (orange, dark orange) for surface (S) and mucosal stroma (St) of the small intestine. This may result from resorption and transport of the same low-molecular protein substance (for instance food component) from the intestinal surface to the mucosal stroma and lymphatics and reflect a physiological function of the small intestine. The tumor area is segmented into two main clusters (dark blue and red, Figure 10B). For this differentiation no clear morphologic or functional correlation was found in the optical image. Higher optical magnification (not shown) of the tumor area shows the heterogeneous composition of this tissue consisting of at least three components: (i) small neuroepithelial tumor cell nests, (ii) tumor stroma and pre-existent structures of the intestinal wall especially smooth muscular tissue, and (iii) connective tissue. The segmentation map with weak denoising (Figure 10C) also shows heterogeneous composition, although we do not have evidence that it is tumor specific.

Finally, we found for the NET data set four masses the most colocalized with dark blue and red regions (corresponding to tumor based on histological analysis) of the segmentation map shown in Figure 10B, which are 3791.1, 5920.8, 7550.0, and 13976.9 Da (Figure 11). Interestingly, while the first three

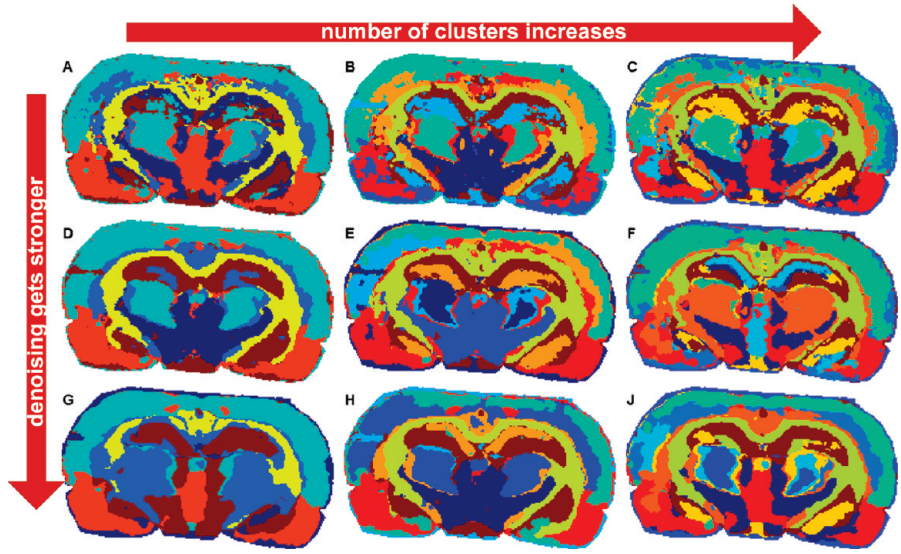


Figure 9. Segmentation maps for the rat brain data set for different numbers of clusters (first column: 6, second column: 8, third column: 10 clusters) and for different levels of edge-preserving denoising (first row: weak, second row: moderate, third row: strong denoising).

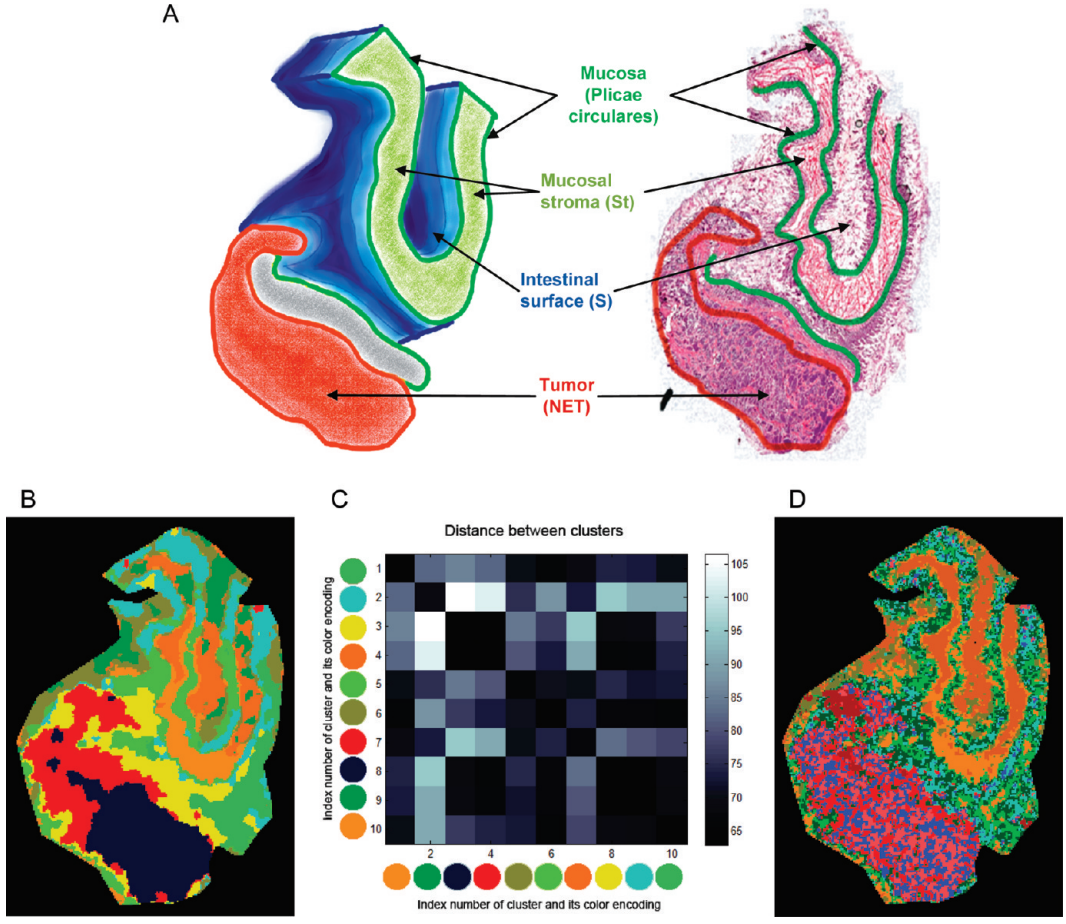


Figure 10. The human neuroendocrine tumor data set. (A) 3D-structure of the tissue used for MALDI-imaging measurement and optical image of the H&E stained section with main functional structures. (B) Segmentation map, strong denoising, 10 clusters. (C) The matrix showing distances between clusters for panel B. (D) Segmentation map, weak denoising, 10 clusters.

masses have low intensity in the nontumor area, the last mass shows high intensity also in the most part of the data set except for the mucosal stroma, that highlights that the corresponding molecular compound is common for all (neuro-)epithelial cells.

Discussion

Peak Picking. Given a set of spectra, the most popular approach of peak picking, also used in the ClinProTools

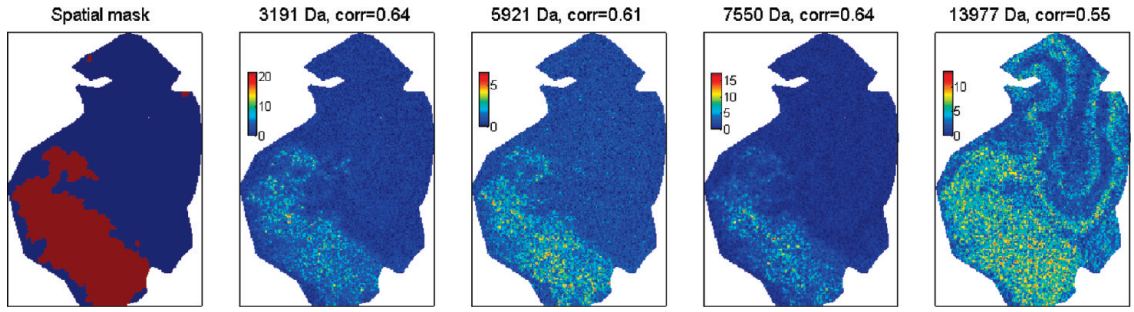


Figure 11. The spatial mask (corresponding to the dark blue and red segments in Figure 10B) and four *m*/*z*-images mostly colocalized with this mask. The colocalization is measured by the correlation coefficient (shown in the image title).

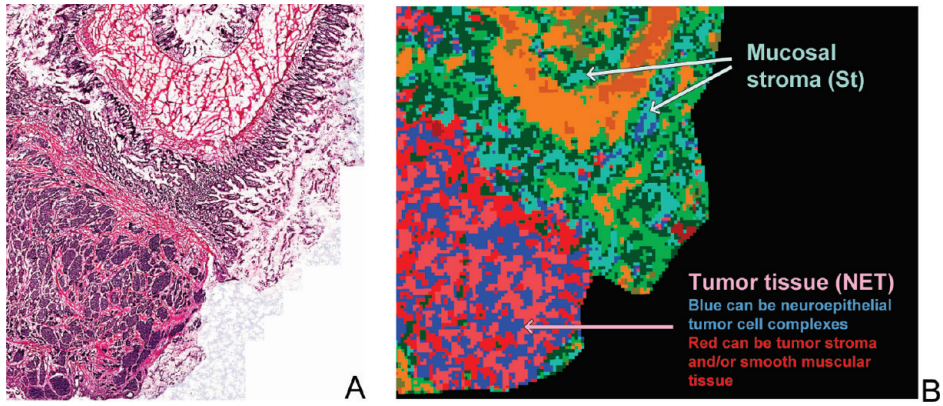


Figure 12. A part of the human neuroendocrine tumor data set. (A) Optical image of the H&E stained section. (B) Segmentation map with weak denoising. We hypothesize that the porous (blue-red) segmentation in the tumor area is due to heterogeneous nature of the tumor.

software, is to select peaks based on the data set-mean spectrum. However, as discussed in the introduction, the mean spectrum of a MALDI-imaging data set can show no high peak for a mass localized only in a small spatial area. Thus, heterogeneity of a MALDI-imaging data set poses a new challenge and requires new approaches to peak picking. Our approach allows for selection of peaks which are observed only in a small portion (at least in 1%) of all spectra. To the best of our knowledge, this approach has not been described yet.

Edge-Preserving Denoising. At the present time, there is only one study[32] published where denoising of *m*/*z*-images (a moving average filtering of 3 × 3 pixels) was used to reduce the pixel-to-pixel variation. In another study,[24] denoising was applied posthoc in order to improve the classification map. We suppose that the concept of denoising *m*/*z*-images has not yet been exploited to its full extent due to mathematical complexity of this problem.

Clustering Methods. The problem of clustering has a long history, and at the present time there are many methods and approaches of clustering. We have selected HDDC since it is developed for high dimensional data. In our experience, HDDC leads to better results if compared to simple methods such as k-means (small features are resolved, k-means sometimes splits the large anatomical parts of the rat brain into several segments, strong smoothing seems to affect k-means results leading to additional anatomically not reasonable layers along edges; results not shown), although HDDC is significantly slower.

Currently, hierarchical clustering (HC) is used in MALDI-imaging,[18,21] in particular, because it is available in the flexImaging software (Bruker Daltonik GmbH). The main feature of HC in this context is an interactive analysis of the HC-dendrogram and manual splitting of the sample into regions based on this analysis. In contrast, our approach automatically divides the data set into the given number of clusters.

However, when using hierarchical clustering, one should keep in mind the following. First, it requires more memory for storing the full distance matrix although there are memory-optimized methods like BIRCH.[33] Second, at each step (increasing the number of clusters) one cluster is split into two parts. Not the same in HDDC or k-means, where for each number of clusters an optimal partition of the full data set is searched for. From one side, hierarchical partitioning can be better interpreted (at each step an already established cluster is split into two subclusters). However, it leads to not optimal partitioning for the fixed number of clusters, in contrast to HDDC or k-means.

We do not perform comparison of clustering methods because the focus of this paper is on improving clustering results with the use of spatial information through edge-preserving denoising of *m*/*z*-images. Note that after edge-preserving denoising is applied, any clustering method instead of HDDC (e.g., hierarchical clustering) can be exploited.

Importance for Cancer Studies. As shown, the computed segmentation maps are able to reveal the morphological composition of analyzed tissue (Figure 10). Moreover, a segmentation map can highlight functional similarity of morphological structures (like the similarity of intestinal surface and mucosal stroma shown in the segmentation map, as discussed in the results for NET) that can lead to understanding of functional processes in tissue.

When compared our MALDI-imaging segmentation maps with standard histological tools in cancer studies such as H&E and immunohistochemistry, where tissue is stained with

antibodies with respect to a specific protein, our map (1) takes into account the full range of proteins insides tissues, (2) is not a targeted but a data-driven approach that finds regions of similar molecular composition, (3) pictures the tissue with several colors. Thus, our segmentation map represents a proteomic functional topographic map on the basis of tissue morphology that cannot be reached by any other method.

Naturally, interpretation of a segmentation map showing the complex proteomic diversity in one image and allocation of the derived segments to single structures requires histological expertise and depends on the spatial resolution. The state-of-the-art spatial resolution of MALDI-imaging (20 μm) is much less than that of microscopy used in histological studies. We believe that with improvement of its spatial resolution (to 10 μm or lower) this technology will become a histological tool along with H&E and immunohistochemistry.

Let us consider the segmentation map produced with weak denoising (Figure 12), which is clearly less homogeneous, especially in the tumor area. It needs to be evaluated whether this reflects the functional heterogeneity of the tissue or is caused by noise. However, we have found evidence (unpublished results) that inflammatory infiltrates or the enrichment of serum components in the stroma of head and neck cancer can lead to localized changes in protein concentrations and compositions that can be represented by such heterogeneous segmentation.

Although interpretation of segmentation maps is a challenging task, these maps provide a unique way to depict the complex functional proteomic heterogeneity of a tissue in one image. Therefore, integral aspects of tissue functions could be explored under diverse conditions such as tumor proliferation, invasion, and drug metabolisation.

Relation to Supervised Methods. Note that in this study we are interested in unsupervised processing of MALDI-imaging data. Supervised processing, when several regions of interest or microdissected cells are intercompared, is better developed in the context of MALDI-imaging.[4−7] We believe that our segmentation approach also can be useful in a supervised framework.

First, the produced segmentation regions can be taken as regions of interest with subsequent intercomparison of spectra from these regions. This makes sense when, even after histological analysis, the regions of interest cannot be determined precisely enough. In biomarker discovery studies, this plays an especially important role, due to the heterogeneous structure of tumor tissue, insufficient spatial resolution of MALDI-imaging, and the recently discovered molecular exchange between tumor and the surrounding tissue.[34]

Second, our segmentation map provides a way to establish discriminative information that can be found in the spectra also answering the question at which level of detail the regions of interest should be selected. As demonstrated, in the neuroendocrinal tumor data set the outline of the tumor can be found as well as fine substructures inside the tumor itself. If these features can be found in an unsupervised manner, they are prominent enough to be investigated with a supervised approach.

Application to other IMS Modalities. Besides MALDI-IMS, we have applied our segmentation pipeline to other IMS data, among others to DESI-, LDI-IMS, and SIMS (results not shown). LDI and SIMS data sets are quite similar to MALDI, and the pipeline shows good segmentation results. SIMS has a higher dynamic range, but our peak picking is able to detect small peaks if they have the proper shape. DESI technology is particular; because of using the spray for desorption and no matrix, there is much less pixel-to-pixel variation and original m/z-images look quite smooth. The edge-preserving denoising improves the segmentation maps but not as considerably as for MALDI, LDI, and SIMS.

Acknowledgment. The authors thank Charles Bouveyron (Laboratoire SAMM, Université Paris 1 Panthéon-Sorbonne, France) for his help with the HDDC clustering algorithm, Dennis Trede (ZeTeM, University of Bremen) for his implementation of the OMP algorithm, Marc Gerhard (formerly Bruker Daltonik GmbH) for his assistance with ClinProTools software, and Merten Hommann and Daniel Kämmerer (both Zentralklinikum Bad Berka, Germany) for providing the neuroendocrine tumor samples.

Supporting Information Available: Supplementary Figure 1. Relation between peak intensity and noise variance; Supplementary Figure 2. Segmentation maps for the rat brain dataset for different parameters of peak picking. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Caprioli, R. M.; Farmer, T. B.; Gile, J. Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. *Anal. Chem.* **1997**, *69*, 4751–4760.
(2) Stoeckli, M.; Chaurand, P.; Hallahan, D. E.; Caprioli, R. M. Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues. *Nat. Med.* **2001**, *7*, 493–496.
(3) Heeren, R. M. A.; Smith, D. F.; Stauber, J.; Kukrer-Kaletas, B.; MacAleese, L. Imaging mass spectrometry: hype or hope. *J. Am. Soc. Mass Spectrom.* **2009**, *20* (6), 1006–1014.
(4) Yanagisawa, K.; Shyr, Y.; Xu, B. J.; Massion, P. P.; Larsen, P. H.; White, B. C.; Roberts, J. R.; Edgerton, M.; Gonzalez, A.; Nadaf, S.; Moore, J. H.; Caprioli, R. M.; Carbone, D. P. Proteomic patterns of tumour subsets in non-small-cell lung cancer. *Lancet* **2003**, *362* (9382), 433–439.
(5) Lemaire, R.; Menguellet, S. A.; Stauber, J.; Marchaudon, V.; Lucot, J.-P.; Collinet, P.; Farine, M.-O.; Vinatier, D.; Day, R.; Ducoroy, P.; Salzet, M.; Fournier, I. Specific MALDI imaging and profiling for biomarker hunting and validation: fragment of the 11S proteasome activator complex, reg alpha fragment, is a new potential ovary cancer biomarker. *J. Proteome Res.* **2007**, *6* (11), 4127–4134.
(6) Cazares, L. H.; Troyer, D.; Mendrinos, S.; Lance, R. A.; Nyalwidhe, J. O.; Beydoun, H. A.; Clements, M. A.; Drake, R. R.; Semmes, O. J. Imaging mass spectrometry of a specific fragment of mitogen-activated protein kinase/extracellular signal-regulated kinase kinase kinase 2 discriminates cancer from uninvolved prostate tissue. *Clin. Cancer Res.* **2009**, *15*, 5541–5551.
(7) Rauser, S.; Marquardt, C.; Balluff, B.; Deininger, S.-O.; Albers, C.; Belau, E.; Hartmer, R.; Suckau, D.; Specht, K.; Ebert, M. P.; Schmitt, M.; Aubele, M.; Höfler, H.; Walch, A. Classification of HER2 receptor status in breast cancer tissues by MALDI imaging mass spectrometry. *J. Proteome Res.* **2010**, *9* (4), 1854–1863.
(8) Heeren, R. M. A.; McDonnell, L. A.; Amstalden, E.; Luxembourg, S. L.; Altelaar, A. F. M.; Piersma, S. R. Why don't biologists use SIMS? A critical evaluation of imaging MS. *Appl. Surf. Sci.* **2006**, *252* (19), 6827–6835.
(9) Wiseman, J. M.; Puolitaival, S. M.; Takats, Z.; Cooks, R. G.; Caprioli, R. M. Mass spectrometric profiling of intact biological tissue by using desorption electrospray ionization. *Angew. Chem.* **2005**, *44* (43), 7094–7097.
(10) Cha, S.; Yeung, E. S. Colloidal graphite-assisted laser desorption/ionization mass spectrometry and MSn of small molecules. 1. Imaging of cerebrosides directly from rat brain tissue. *Anal. Chem.* **2007**, *79* (6), 2373–2385.
(11) Nemes, P.; Barton, A. A.; Li, Y.; Vertes, A. Ambient molecular imaging and depth profiling of live tissue by infrared laser ablation electrospray ionization mass spectrometry. *Anal. Chem.* **2008**, *80* (12), 4575–4582.
(12) Yanes, O.; Northen, T. R.; Uritboonthai, W.; Estrada, M. N.; Manchester, M.; Siuzdak, G. Nanostructure initiator mass spec-

trometry for biological tissue imaging and biofluid analysis. *Anal. Chem.* **2009**, *81* (8), 2969–2975.
(13) Ernst, G.; Melle, C.; Schimmel, B.; Bleul, A.; von Eggeling, F. proteohistography−direct analysis of tissue with high sensitivity and high spatial resolution using ProteinChip technology. *J. Histochem. Cytochem.* **2006**, *54* (1), 13–17.
(14) Franck, J.; Arafah, K.; Elayed, M.; Bonnel, D.; Vergara, D.; Jacquet, A.; Vinatier, D.; Wisztorski, M.; Day, R.; Fournier, I.; Salzet, M. MALDI imaging mass spectrometry: State of the art technology in clinical proteomics. *Mol. Cell. Proteomics* **2009**, *8*, 2023–2033.
(15) McDonnell, L. A.; Corthals, G. L.; Willems, S. M.; van Remoortere, A.; van Zeijl, R. J. M.; Deelder, A. M. Peptide and protein imaging mass spectrometry in cancer research. *J. Proteomics* **2010**, *73* (10), 1921–1944.
(16) Yao, I.; Sugiura, Y.; Matsumoto, M.; Setou, M. In situ proteomics with imaging mass spectrometry and principal component analysis in the Scrapper-knockout mouse brain. *Proteomics* **2008**, *8* (18), 3692–3701.
(17) Klerk, L. A.; Broersen, A.; Fletcher, I. W.; van Liere, R.; Heeren, R. M. A. Extended data analysis strategies for high resolution imaging MS: New methods to deal with extremely large image hyperspectral datasets. *Int. J. Mass Spectrom.* **2007**, *260* (2−3), 222–236.
(18) Deininger, S.-O.; Ebert, M. P.; Fütterer, A.; Gerhard, M.; Röcken, C. MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers. *J. Proteome Res.* **2008**, *7* (12), 5230–5236.
(19) Hanselmann, M.; Kirchner, M.; Renard, B. Y.; Amstalden, E. R.; Glunde, K.; Heeren, R. M. A.; Hamprecht, F. A. Concise representation of mass spectrometry images by probabilistic latent semantic analysis. *Anal. Chem.* **2008**, *80* (24), 9649–9658.
(20) McCombie, G.; Staab, D.; Stoeckli, M.; Knochenmuss, R. Spatial and spectral correlations in MALDI mass spectrometry images by clustering and multivariate analysis. *Anal. Chem.* **2005**, *77* (19), 6118–6124.
(21) Walch, A.; Rauser, S.; Deininger, S.-O.; Höfler, H. MALDI imaging mass spectrometry for direct tissue analysis: a new frontier for molecular histology. *Histochem. Cell Biol.* **2008**, *130*, 421–34.
(22) Norris, J. L.; Cornett, D. S.; Mobley, J. A.; Andersson, M.; Seeley, E. H.; Chaurand, P.; Caprioli, R. M. Processing MALDI mass spectra to improve mass spectral direct tissue analysis. *Int. J. Mass Spectrom.* **2007**, *260* (2−3), 212–221.
(23) Denis, L.; Lorenz, D. A.; Trede, D. Greedy solution of ill-posed problems: error bounds and exact inversion. *Inverse Probl.* **2009**, *25* (11), 115017.
(24) Hanselmann, M.; Köthe, U.; Kirchner, M.; Renard, B. Y.; Amstalden, E. R.; Glunde, K.; Heeren, R. M. A.; Hamprecht, F. A. Toward digital staining using imaging mass spectrometry and random forests. *J. Proteome Res.* **2009**, *8* (7), 3558–3567.
(25) Rudin, L. I.; Osher, S.; Fatemi, E. Nonlinear total variation based noise removal algorithms. *Physica D* **1992**, *60* (1−4), 259–268.
(26) Chambolle, A. An algorithm for total variation minimization and applications. *J. Math. Imaging Vis.* **2004**, *20* (1−2), 89–97.
(27) Grasmair, M. Locally adaptive total variation regularization. *LNCS 5567 (Scale Space and Variational Methods in Computer Vision)* **2009**, 331–342.
(28) Bouveyron, C.; Girard, S.; Schmid, C. High-dimensional data clustering. *Comput. Stat. Data Anal.* **2007**, *52* (1), 502–519.
(29) Groseclose, M. R.; Andersson, M.; Hardesty, W. M.; Caprioli, R. M. Identification of proteins directly from tissue: in situ tryptic digestions coupled with imaging mass spectrometry. *J. Mass Spectrom.* **2007**, *42* (2), 254–262.
(30) Lange, E.; Gropl, C.; Reinert, K.; Kohlbacher, O.; Hildebrandt, A. High-accuracy peak picking of proteomics data using wavelet techniques. *Pac. Symp. Biocomput.* **2006**, *11*, 243–254.
(31) Leptos, K. C.; Sarracino, D. A.; Jaffe, J. D.; Krastins, B.; Church, G. M. MapQuant: Open-source software for large-scale protein quantification. *Proteomics* **2006**, *6*, 1770–1782.
(32) McDonnell, L. A.; van Remoortere, A.; van Zeijl, R. J. M.; Deelder, A. M. Mass spectrometry image correlation: quantifying colocalization. *J. Proteome Res.* **2008**, *7* (8), 3619–3627.
(33) Zhang, T.; Ramakrishnan, R.; Livny, M. BIRCH: an efficient data clustering method for very large databases. *SIGMOD Rec.* **1996**, *25* (2), 103–114.
(34) Oppenheimer, S. R.; Mi, D.; Sanders, M. E.; Caprioli, R. M. Molecular analysis of tumor margins by MALDI mass spectrometry in renal carcinoma. *J. Proteome Res.* **2010**, *9* (5), 2182–2190.

PR100734Z