On the Importance of Mathematical Methods for Analysis of MALDI-Imaging Mass Spectrometry Data

Dennis Trede[1,2,*], Jan Hendrik Kobarg[2], Janina Oetjen[2,3], Herbert Thiele[4], Peter Maass[1,2] and Theodore Alexandrov[1,2,3,5]

[1]Steinbeis Innovation Center SCiLS (Scientific Computing in Life Sciences), Richard-Dehmel-Str. 69, 28211 Bremen, Germany, http://www.scils.de

[2]Zentrum für Technomathematik, Universität Bremen, Bibliothekstr. 1, 28359 Bremen, Germany, http://www.zetem.uni-bremen.de

[3]MALDI Imaging Lab, Universität Bremen, Leobener Str., 28359 Bremen, Germany, http://www.maldi.uni-bremen.de

[4] Bruker Daltonik GmbH, Fahrenheitstr. 4, 28359 Bremen, Germany, http://www.bdal.de

[5]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, 9500 Gilman Drive, MC 0657, La Jolla, Ca 92093-0657, USA, http://pharmacy.ucsd.edu

Summary

In the last decade, matrix-assisted laser desorption/ionization (MALDI) imaging mass spectrometry (IMS), also called as MALDI-imaging, has proven its potential in proteomics and was successfully applied to various types of biomedical problems, in particular to histopathological label-free analysis of tissue sections. In histopathology, MALDI-imaging is used as a general analytic tool revealing the functional proteomic structure of tissue sections, and as a discovery tool for detecting new biomarkers discriminating a region annotated by an experienced histologist, in particular, for cancer studies.

A typical MALDI-imaging data set contains 10^8 to 10^9 intensity values occupying more than 1 GB. Analysis and interpretation of such huge amount of data is a mathematically, statistically and computationally challenging problem. In this paper we overview some computational methods for analysis of MALDI-imaging data sets. We discuss the importance of data preprocessing, which typically includes normalization, baseline removal and peak picking, and hightlight the importance of image denoising when visualizing IMS data.

1 Introduction

In the last decade, matrix-assisted laser desorption/ionization (MALDI) imaging mass spectrometry [6, 21], also called as MALDI-imaging, has proven its potential in the spatially-resolved proteomic analysis of thin biological tissue sections. MALDI-imaging is used as a general analytic tool revealing the functional proteomic structure of tissue, and as a discovery

*To whom correspondence should be addressed. Email: trede@scils.de

tool for detecting new biomarkers discriminating a region annotated by an experienced histologist, in particular for cancer studies [11, 15]. Currently, the development of computational methods for MALDI-imaging is lagging behind the technological progress [1, 25]. The following problems in the field of MALDI-imaging need specially developed computational methods, see [1] for more details:

1. Preprocessing: baseline removal, spectra normalization, and noise reduction [2, 7, 14]
2. Data reduction using mass spectrometry peak picking [9] or scale-space transformations, e.g. the discrete wavelet transform [16]
3. Data representation using multivariate statistics, e.g. principal component analysis (PCA) and its variants [10, 12, 24]
4. Spatial segmentation of a MALDI-imaging data set based on spectra clustering [2, 8]
5. Supervised classification of spectra of a MALDI-imaging data set (or data sets) after training a classifier on manually annotated regions [3, 15] and detection of discriminative m/z-values
6. Postprocessing, e.g. image magnification and co-registration with a high-resolution microscopy image

In this paper, we demonstrate the effect of preprocessing and spatial segmentation algorithms for the evaluation of MALDI-imaging data sets, illustrate pre-processing for data analysis based on spatial segmentation, and highlight the importance of image denoising for visualization. The paper is based on our short workshop communication [23], however, in this extended form it presents new original results comparing segmentation of a real-life data set with and without preprocessing. The considered data set represents MALDI-imaging applied to a rat kidney section simulating a typical proteomics experiment (measured at the MALDI Imaging Lab, University of Bremen). The computations and visualization were done using the SCiLS Lab software (SCiLS, Bremen, Germany).

2 Methods

2.1 Preprocessing

Once a sample has been prepared for MALDI analysis, mass spectra are acquired at discrete spatial points, providing a so-called data cube or hyperspectral image, with a mass spectrum measured at each pixel, see Figure1. A mass spectrum represents the relative abundances of ionizable molecules with various mass-to-charge (m/z) values, ranging for MALDI-TOF-IMS from several hundreds up to a few tens of thousands of m/z-values.

A MALDI-imaging data set can be considered as a collection of spectra that have been measured independently, hence normalization of spectra is an important task of image preprocessing. The most popular method is the so-called total ion count method (TIC), which normalizes every spectrum separately by dividing each spectrum intensity by the sum of all its intensities. In [7], more advanced ways than spectrum-wise normalization are discussed, namely,

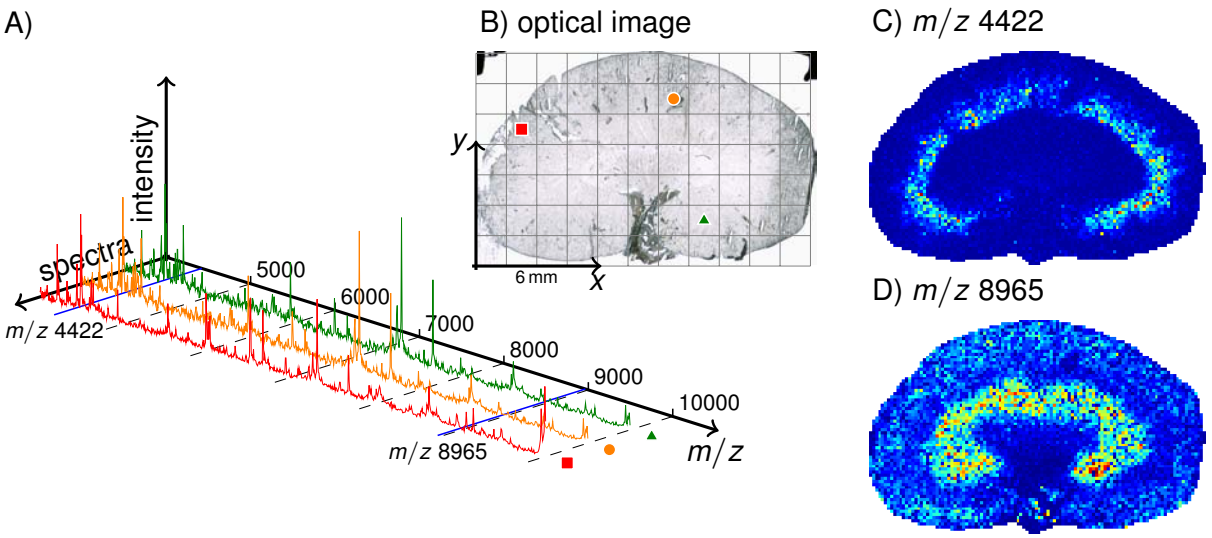

Figure 1: An MALDI-imaging data set is a data cube. Spectra (A) are measured at spatial points of a sample (B) with spatial coordinates (x,y). Given a mass (m/z value), one obtains an intensity image; examples for the channels m/z 4422 and m/z 8965 are shown in (C, D).

normalization based on the spectra noise level and normalization based on the median of signal intensities.

A typical mass spectrum consists of three fundamental components: peaks, a baseline and noise (see [22] for more details). A peak can be approximated by the Gaussian function. Note that the peak width increases with increasing m/z-values. The baseline is inherent to sample preparation and the MALDI measurement and can vary throughout the image [18]. The baseline is a slowly varying component of a spectrum, normally with high values in the region of low m/z-values and vanishing for high m/z-values. The noise level depends on the measurement device and the measuring process. Since the MALDI technique relies upon counting ions, the noise is assumed to be Poisson distributed and our study confirms this hypothesis [2]. Poisson noise is multiplicative, since the mean number of counts is equal to their variance. The peaks comprise the essential information about abundant ionizable molecular compounds present in the sample. The aim of MALDI-imaging data preprocessing is to clean the spectra from baseline and noise first, then to select peaks encoding relevant information.

Baseline correction is a standard method of mass spectra preprocessing. One method to perform baseline correction is the top-hat operator from mathematical morphology [19] which is defined as the difference between the original spectrum and its morphological opening. Other baseline substraction methods are summarized in e.g. [14].

The presence of noise in MALDI-imaging data can be easily seen by visual inspection of m/z-images corresponding to selected m/z-values. Since the noise in MALDI data is strong, image denoising can significantly improve the visualization. An important issue to consider when selecting the image denoising methods is the change of noise variance both within an image and between different images. In [2], we showed that the noise variance at a spatial point linearly depends on the mean intensity around this point. This possibly indicates the Poisson distribution of the noise. In order to reduce this pixel-to-pixel variability, in [2] a method for edge-preserving image denoising has been introduced that adjusts the level of denoising to the local noise level and to the local scale of the features to be resolved.

2.2 Data Compression

MALDI-imaging data typically consists of thousands of different channels (10^3 to 10^4). To process such a huge amount of data one can constrain the channels to the most relevant ones without losing significant information. In MALDI-imaging, this can be achieved by peak picking where m/z-values for specific peaks are selected. For processing huge MALDI-imaging data sets efficient peak picking methods are crucial. At the same time, peak picking should be robust to strong noise, preventing the use of straightforward local maxima or signal-to-noise ratio methods, which can produce false positives.

In [9], a peak picking method based on the orthogonal matching pursuit (OMP) was proposed and in [2] this method was applied to MALDI-imaging mass spectrometry data. The main idea of the method is to model each spectrum as a sum of Gaussian-shaped functions. For each single spectrum, the peak-picking algorithm from [2] selects certain peaks of the Gaussian shape. This assigns to each m/z-value a number of spectra in which this m/z-value was selected as a peak. Finally, the most frequent peaks which occur in more than 1% of considered spectra are selected.

However, because the Gaussian shape is just an approximation of a real peak shape, and probably because of small mass shifts (the mass recalibration for each spectrum is not used in imaging MS), often several m/z-values close to the center peak m/z-value are selected. This reduces the frequency of the peak m/z-value. Moreover, for a peak, this approach selects not one but several m/z-values that can influence subsequent processing steps. This effect seems to be stronger for large peaks, which e.g. leads to their increased impact on clustering.

In order to prevent this redundant selection of several m/z-values per peak, in [5] the selected m/z-values have been aligned by moving them uphill the data set mean spectrum so that they are in the local maxima of the mean spectrum, see Figure 2. This simple improvement allows us to increase the sensitivity of the peak picking without a drop in specificity.

Alternative methods for reducing the amount of data for a later feature selection and classification are scale space methods as e.g. the discrete wavelet transform [16]. The idea of the wavelet transform is to use a wavelet for which its scaling function closely matches the peak pattern of spectra, as e.g. the bi-orthogonal *bior3.7* wavelet in [3].

2.3 Spatial Segmentation

Presently, data mining of MALDI-imaging data sets is a time-consuming endeavor as it is mostly done manually and a MALDI-imaging data set consists of thousands of m/z-channels. Manual data mining of such data requires the user to click through each image and look for distributions that may correlate to the morphology of the sample analyzed. Unsupervised processing methods do not require a user to provide data annotation and can be used as a first step of data mining providing data overview and extracting prominent features.

Such an unsupervised method is spectral clustering resulting in spatial segmentation of a data set [8]. The outcome of clustering can be displayed as a spatial segmentation map (an integer-valued image, usually shown using pseudo-color), coloring identically points grouped into one cluster.

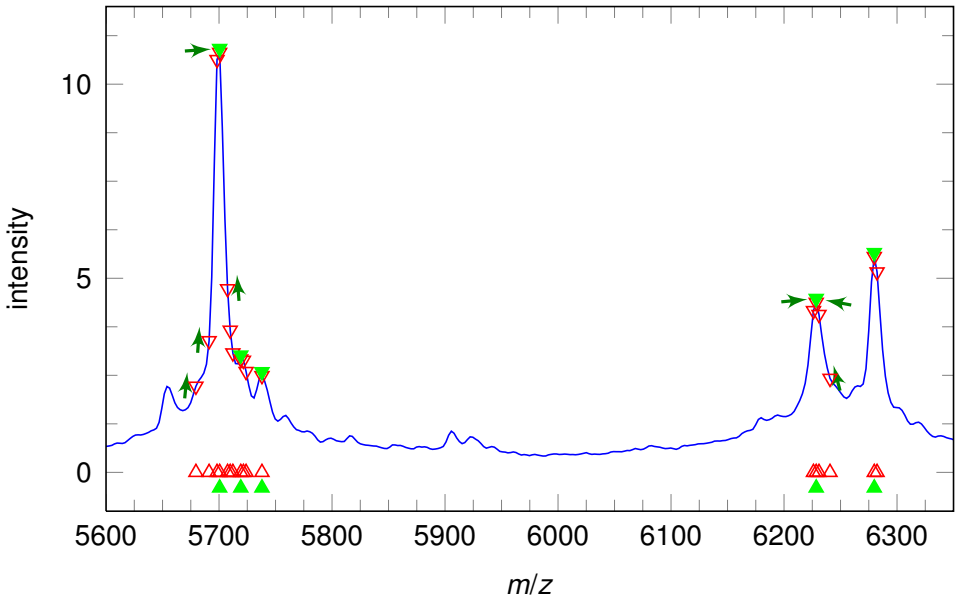

Figure 2: The method from [5] of alignment of data set frequent masses. The data set mean spectrum is shown in blue. Red triangles indicate peaks (and their masses) found after peak picking. Green arrows illustrate the process of alignment. Green triangles show aligned peaks and their masses. Reprinted from [5], Copyright 2011, with permission from Elsevier.

The main drawback of using straightforward clustering of mass spectra is that it is negatively affected by the pixel-to-pixel variability. Taking into account the spatial relations between spectra improves the segmentation maps considerably by suppressing the noise and pixel-to-pixel variability. In [2], a pipeline for segmentation of MALDI-imaging data sets is proposed that takes into account spatial information. After spectra normalization and baseline correction, the first step of this pipeline is a reduction of the data set by selecting peaks appearing in at least 1% of spectra. The second step and the core of this segmentation procedure is an edge-preserving denoising of m/z images for each m/z-value from the selected peaks list. 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. The edge-preserving image denoising operation improves the segmentation map significantly.

Most of the advanced clustering methods are computationally intensive. Use of simpler methods reduces the computation time but deteriorates the qualitity of the segmentation maps due to strong noise in data. In [4], an efficient segmentation approach for data has been proposed, that projects the data to fewer dimensions and at the same time considers a spectrum together with its spatial neighbors.

2.4 Postprocessing

An important issue for MALDI imaging mass spectrometry technique is its relatively low spatial or lateral resolution (i.e. a large size of a pixel) as compared with microscopy. The state of the art resolution is around 20 micron for MALDI-imaging [13] versus 0.25 micron for optical microscopy. So, when comparing an MALDI-imaging data set or its segmentation map with a microscopy image, a significant difference in spatial resolution complicates the visual inter-

pretation. In [5], a computational approach was proposed to improve the spatial resolution of a segmentation map of an imaging mass spectrometry data set.

Other imaging problems occur when extending the 2D MALDI-imaging technique to three spatial dimensions with consecutive sections of tissue. Here one has to align a stack of hyperspectral images to each other. Methods for image registration of grey-scale images are available [17], but—to the best of our knowledge—not yet specially adapted to 3D hyperspectral MALDI data.

From a technical perspective, visualizing 3D information is highly complex. From a medical perspective however, it still does not provide enough information for diagnosis. To draw conclusions from the data, it must first be correlated with 3D anatomical information (such as data obtained via magnetic resonance imaging). However, superimposing these two data sets originating from entirely different imaging modalities is complicated by the issue of image co-registration [20] and standard pipelines are not established, yet.

3 Results

The rat kidney sample has been preparated and MALDI measured at the MALDI Imaging Lab, University of Bremen. Mass spectra were acquired on a MALDI-TOF instrument (Autoflex IV LRF; Bruker Daltonik GmbH) in linear positive mode. MALDI measurements were performed at a mass range of 2 kDa to 20 kDa. The lateral resolution for the MALDI image was set to 150 µm. The rat kidney data set comprises 6,304 spectra. The following processing steps were performed with SCiLS Lab software.

Normalization and Baseline Subtracting. Normalization of spectra has been done with respect to the total ion count (TIC). Spectra were baseline corrected by subtracting a smooth lower envelope curve which consist of wide Gaussians. In figure 3, the TIC normalization and the baseline correction is visualized for the rat kidney data set. By means of these preprocessing methods the m/z images allow for better description of anatomical structures.

Data Compression by means of Peak Picking. For peak picking the method from [2] has been used. From the joint list of potential peaks, which includes all detected peaks, 344 peaks were selected as consensus peaks as they occurred in at least 1% of spectra. After applying the alignment to the mean spectrum as described in from [5], 63 important peaks remained.

Spatial Denoising. After peak picking, for subsequent data processing we only use the reduced MALDI-imaging hyperspectral datacube (the data set is reduced in the number of m/z-values by the peak picking). We processed this data with edge-preserving denoising of m/z-images corresponding to these peaks. Examples of m/z-images and their denoised versions are shown in figure 3. The method efficiently removes the noise while not smoothing out edges.

Spatial Segmentation. The segmentation map after clustering with edge-preserving denoising is presented on the right-hand side in figure 4 coregistered with an optical image of the analyzed rat kidney section. The major anatomical regions are well represented. On the contrary, the segmentation map produced without prior-to-clustering image denoising does not recover the anatomical structure (left-hand side). 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.

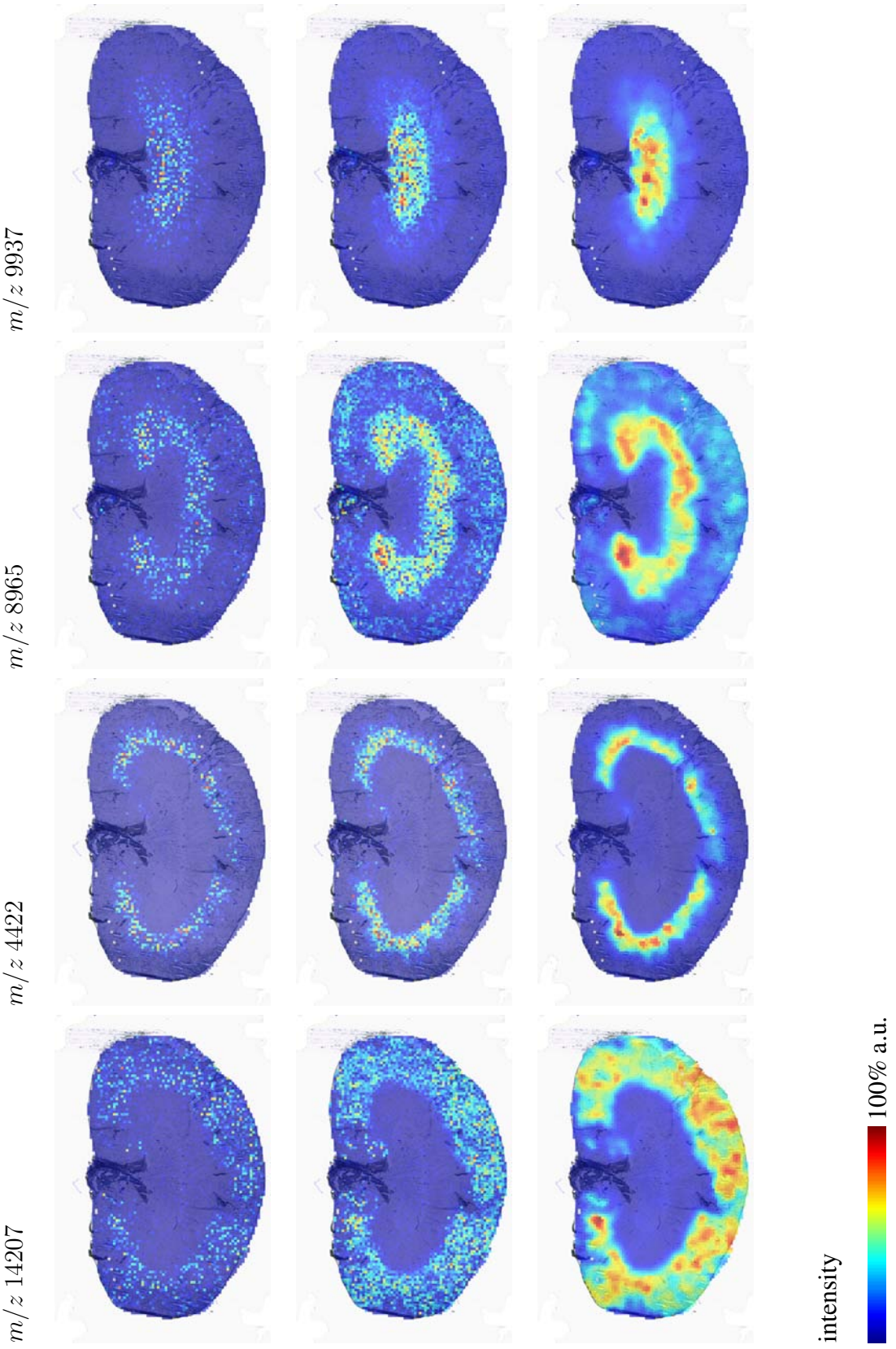

Figure 3: Images corresponding to one m/z-value. Left column: measured raw data, middle column: after TIC normalization and baseline correction, right column: after edge-preserving denoising. a.u. = arbitrary units.

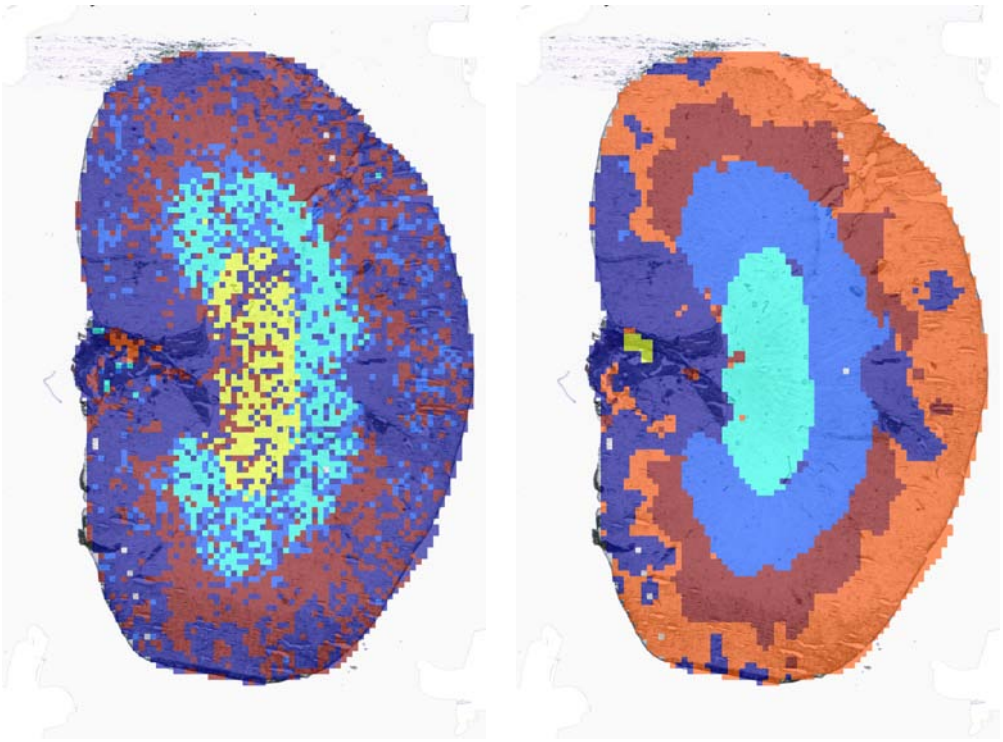

Figure 4: Comparison of clustering with and without previous smoothing. Left: Simple clustering does not recover the anatomical structure of the rat kidney section due to high-dimensional data, low spatial resolution (mixture of tissues) and multiplicative-like Poisson noise. Right: Locally-adaptive edge-preserving denoising of m/z images and following clustering preserves spatial structure, i.e. edges and small details are not eroded.

4 Discussion

Data from imaging mass spectrometry can be represented as a hyperspectral image with thousands of channels. Since manual data mining of MALDI-imaging data sets is very time-consuming, the development of automated computational methods is necessary. Mathematics offers a variety of methods from image processing, statistics and machine learning that can be used for simplifying and automating the analysis of imaging mass spectrometry data. Other areas of science where hyperspectral images incur use similar methods for related problems. Here, an interdisciplinary exchange of experiences can inspire each other and avoid gratuitous parallel developments.

MALDI-imaging data is characterized by independently measured spectra and the presence of strong and multiplicative noise. Hence preprocessing procedures of MALDI data as e.g. normalization, baseline correction and denoising are fundamental first steps in data processing pipelines for MALDI-imaging data. For achieving spatial segmentation using spectra clustering, preprocessing of MALDI data has a crucial influence.

Acknowledgements

DT gratefully acknowledges the financial support of the Bremen Economic Development (WFB, project “3D MALDI-Imaging basic experiment”, grant FUE0485B). JHK and DT gratefully

acknowledge the financial support of the European Union Seventh Framework Programme (project "UNLocX", grant 255931). JO gratefully acknowledges the financial support of the Federal Ministry of Education and Research (BMBF, project "MALDI-AMK", grant 01IB-10004C).

References

[1] T. Alexandrov. MALDI imaging mass spectrometry: statistical data analysis and current computational challenges. *BMC Bioinformatics*, to appear.

[2] T. Alexandrov, M. Becker, S.-O. Deininger, G. Ernst, L. Wehder, M. Grasmair, F. von Eggeling, H. Thiele, and P. Maass. Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering. *Journal of Proteome Research*, 9(12):6535–6546, 2010.

[3] T. Alexandrov, J. Decker, B. Mertens, A.M. Deelder, R.A.E.M. Tollenaar, P. Maass, and H. Thiele. Biomarker discovery in MALDI-TOF serum protein profiles using discrete wavelet transformation. *Bioinformatics*, 25(5):643–649, 2009.

[4] T. Alexandrov and J.H. Kobarg. Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering. *Bioinformatics*, 27(13):i230–i238, 2011.

[5] T. Alexandrov, S. Meding, D. Trede, J.H. Kobarg, B. Balluff, A. Walch, H. Thiele, and P. Maass. Super-resolution segmentation of imaging mass spectrometry data: solving the issue of low lateral resolution. *Journal of Proteomics*, 75(1):237–245, 2011.

[6] R.M. Caprioli, T.B. Farmer, and J. Gile. Molecular imaging of biological samples : Localization of peptides and proteins using MALDI-TOF MS. *Analytical Chemistry*, 69(23):4751–4760, 1997.

[7] S.-O. Deininger, D.S. Cornett, R. Paape, M. Becker, C. Pineau, S. Rauser, A. Walch, and E. Wolski. Normalization in MALDI-TOF imaging datasets of proteins: practical considerations. *Analytical and Bioanalytical Chemistry*, 401(1):167–181, 2011.

[8] S.-O. Deininger, M.P. Ebert, A. Fütterer, M. Gerhard, and C. Röcken. MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers. *Journal of Proteome Research*, 7(12):5230–5236, December 2008.

[9] L. Denis, D.A. Lorenz, and D. Trede. Greedy solution of ill-posed problems: Error bounds and exact inversion. *Inverse Problems*, 25(11):115017 (24pp), 2009.

[10] M. Hanselmann, M. Kirchner, B.Y. Renard, E.R. Amstalden, K. Glunde, R.M.A. Heeren, and F.A. Hamprecht. Concise representation of mass spectrometry images by probabilistic latent semantic analysis. *Analytical Chemistry*, 80(24):9649–9658, 2008.

[11] R.M.A. Heeren, D.F. Smith, J. Stauber, B. Kukrer-Kaletas, and L. MacAleese. Imaging mass spectrometry: hype or hope. *Journal of the American Society of Mass Spectrometry*, 20(6):1006–1014, 2009.

[12] J. H. Kobarg and T. Alexandrov. Algorithms to incorporate spatial information into clustering of hyper-spectral data. 2012. Submitted for publication.

[13] M. Lagarrigue, M. Becker, R. Lavigne, S.-O. Deininger, A. Walch, F. Aubry, D. Suckau, and C. Pineau. Revisiting rat spermatogenesis with MALDI imaging at 20 µm resolution. *Molecular and Cellular Proteomics*, page mcp.M110.005991, 2010.

[14] J.L. Norris, D.S. Cornett, J.A. Mobley, M. Andersson, E.H. Seeley, P. Chaurand, and R.M. Caprioli. Processing MALDI mass spectra to improve mass spectral direct tissue analysis. *International Journal of Mass Spectrometry*, 260(2–3):212–221, 2007.

[15] S. Rauser, C. Marquardt, B. Balluff, S.-O. Deininger, C. Albers, E. Belau, R. Hartmer, D. Suckau, K. Specht, M.P. Ebert, M. Schmitt, M. Aubele, H. Höfler, and A. Walch. Classification of HER2 receptor status in breast cancer tissues by MALDI imaging mass spectrometry. *Journal of Proteome Research*, 9(4):1854–1863, 2010.

[16] F.-M. Schleif, M. Lindemann, M. Diaz, P. Maass, J. Decker, T. Elssner, M. Kuhn, and H. Thiele. Support vector classification of proteomic profile spectra based on feature extraction with the bi-orthogonal discrete wavelet transform. *Computing and Visualization in Science*, 12(4):189–199, 2009.

[17] O. Schmitt, J. Modersitzki, S. Heldmann, S. Wirtz, and B. Fischer. Image registration of sectioned brains. *International Journal of Computer Vision*, 73(1):5–39, 2007.

[18] S.A. Schwartz, M.L. Reyzer, and R.M. Caprioli. Direct tissue analysis using matrix-assisted laser desorption/ionization mass spectrometry: practical aspects of sample preparation. *Journal of Mass Spectrometry*, 38(7):699–708, 2003.

[19] J. Serra. *Image Analysis and Mathematical Morphology*. Academic Press, Inc., 1983.

[20] T.K. Sinha, S. Khatib-Shahidi, T.E. Yankeelov, K. Mapara, M. Ehtesham, D.S. Cornett, B.M. Dawant, R.M. Caprioli, and J.C. Gore. Integrating spatially resolved three-dimensional MALDI IMS with in vivo magnetic resonance imaging. *Nature Methods*, 5(1):57–59, 2008.

[21] M. Stoeckli, P. Chaurand, D.E. Hallahan, and R.M. Caprioli. Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues. *Nature Medicine*, 7:493–496, 2001.

[22] C.S. Sun and M.K. Markey. Recent advances in computational analysis of mass spectrometry for proteomic profiling. *Journal of Mass spectrometry*, 46(5):443–456, 2011.

[23] D. Trede, J.H. Kobarg, K. Steinhorst, and T. Alexandrov. Mathematical methods for imaging mass spectrometry. In *Proceedings of the 14th Joint International IMEKO TC1+TC7+TC13 Symposium*, 2011.

[24] R. Van de Plas, B. De Moor, and E. Waelkens. Imaging mass spectrometry based exploration of biochemical tissue composition using peak intensity weighted PCA. In *Proceedings of the Third IEEE/NIH BISTI Life Science Systems and Applications Workshop 2007*, pages 209–212, 2007.

[25] J.D. Watrous, T. Alexandrov, and P.C. Dorrestein. The evolving field of imaging mass spectrometry and its impact on future biological research. *Journal of Mass Spectrometry*, 46(2):209–222, 2011.