BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 18 2013, pages 2335–2342
doi:10.1093/bioinformatics/btt388

Bioimage informatics

Advance Access publication July 19, 2013

Testing for presence of known and unknown molecules in imaging mass spectrometry

Theodore Alexandrov[1,2,3,4,5,*] and Andreas Bartels[1]

[1]Center for Industrial Mathematics, University of Bremen, 28359 Bremen, Germany, [2]MALDI Imaging Lab, Deparments of Biochemistry and Mathematics, University of Bremen, 28359 Bremen, Germany, [3]Steinbeis Innovation Center SCiLS Research, 28211 Bremen, Germany, [4]SCiLS GmbH, 28359 Bremen, Germany and [5]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA 92093, USA

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Imaging mass spectrometry has emerged in the past decade as a label-free, spatially resolved and multi-purpose bioanalytical technique for direct analysis of biological samples. However, solving two everyday data analysis problems still requires expert judgment: (i) the detection of unknown molecules and (ii) the testing for presence of known molecules.

Results: We developed a measure of spatial chaos of a molecular image corresponding to a mass-to-charge value, which is a proxy for the molecular presence, and developed methods solving considered problems. The statistical evaluation was performed on a dataset from a rat brain section with test sets of molecular images selected by an expert. The measure of spatial chaos has shown high agreement with expert judges. The method for detection of unknown molecules allowed us to find structured molecular images corresponding to spectral peaks of any low intensity. The test for presence applied to a list of endogenous peptides ranked them according to the proposed measure of their presence in the sample.

Availability: The source code and test sets of mass-to-charge images are available at http://www.math.uni-bremen.de/∼theodore.

Supplementary information: Supplementary materials are available at *Bioinformatics* online.

Contact: theodore@uni-bremen.de

Received on January 31, 2013; revised on June 14, 2013; accepted on July 2, 2013

1 INTRODUCTION

Imaging mass spectrometry (IMS) is a label-free technique for spatially resolved chemical analysis by acquiring mass spectra across the sample surface. A tremendous development of IMS and considerable increase of applications in biology and medicine have been observed over the past decade (Chaurand, 2012; Jungmann and Heeren, 2012; Watrous *et al.*, 2011). Various types of IMS were developed for diverse sample types ranging from biological and plant (Lee *et al.*, 2012) tissues to bio (Watrous and Dorrestein, 2011) and polymer thin films (Crecelius *et al.*, 2012). In this article, we consider the matrix-assisted laser desorption/ionization IMS (MALDI-IMS) (Caprioli *et al.*, 1997; Stoeckli *et al.*, 2001), which plays the leading role (Watrous *et al.*, 2011) in the rapidly developing field of IMS-based metabolomics, lipidomics and proteomics. MALDI-IMS is used for imaging of drugs and their metabolites (Ait-Belkacem *et al.*, 2012; Prideaux and Stoeckli, 2012), histopathological analysis of biopsy sections (Balluff *et al.*, 2012; Schwamborn and Caprioli, 2010) and discovery of new biomarkers (Schwamborn, 2012) and drugs (Castellino *et al.*, 2011; Kroiss *et al.*, 2010; Rubakhin *et al.*, 2005; Yang *et al.*, 2009).

Given a sample, usually a tissue section, MALDI-IMS acquires mass spectra at discrete spatial points across the sample surface, providing a so-called datacube or hyperspectral image, with a mass spectrum measured at each pixel (Caprioli *et al.*, 1997; Stoeckli *et al.*, 2001). A mass spectrum represents the relative abundances of ionizable molecules with various mass-to-charge ratios (m/z), ranging for MALDI-IMS from several hundred up to a few tens of thousands m/z. A channel of a MALDI datacube corresponding to an m/z-value is called an m/z- or molecular image and expresses the relative spatial abundances of a molecular ion with this m/z-value.

MALDI-IMS data are large with a typical dataset comprising 5000–50 000 spectra where each spectrum has a length of 1000–50 000 m/z-bins. In current practice, such datasets are analyzed either manually or using data-mining methods. Manual analysis requires an expert to go over all m/z-values, plot their m/z-images and examine them visually. If there is any prior knowledge about the sample [a histological annotation or pre-specified regions of interest (Schwamborn and Caprioli, 2010)], then the visual analysis aims to correlate the intensity patterns visible in m/z-images with the region of interest. Usually, one is interested in m/z-images showing high intensities in the region of interest and low intensities in the rest of the sample. However, often there is no prior knowledge given or it does not describe the spatial localization of molecules of interest completely. For example, detection of drug metabolites cannot be done by specifying the region, as these metabolites will not necessarily all localize in the drug target area (Khatib-Shahidi *et al.*, 2006). In the case of no prior knowledge on localization of molecules of interest, the visual analysis aims at finding m/z-images, which exhibit spatially structured intensity patterns (so-called spatially structured m/z-images) as opposed to those which exhibit just noise (so-called spatially unstructured m/z-images). We refer to this problem as to the **detection of unknown molecules** in MALDI-IMS data. In this article, we propose an automatic method that solves this

*To whom correspondence should be addressed.

problem without specifying a region of interest a priori and without visual examination.

Another unsolved problem of MALDI-IMS data analysis is the **test for presence of known molecules** when their localization is unknown. For example, given a list of possible drug metabolites or proteins, one needs to test for their presence in the tissue section. With only few molecules to be tested, one can visually examine *m*/*z*-images of these molecules and judge whether they are spatially structured. However, the visual analysis in not feasible for a large number of molecules and cannot be exploited in automatic pipelines. We propose an automatic method solving this problem.

In this article, we present a new approach to analysis of MALDI-IMS data, which resembles the visual analysis described earlier in the text but selects spatially structured *m*/*z*-images automatically. The core of this approach is the ranking of *m*/*z*-images by their level of spatial structure, similarly to how it can be done by the naked eye. The ranking is based on the original measure of spatial chaos, does not have parameters and can be used in a completely unsupervised manner. The spatial chaos is defined as a lack of spatial pattern in the pixels intensities; the proposed measure of spatial chaos is low for an image exhibiting spatially structured intensity pattern and is high for an image with spatially chaotic pixels intensities. The problem of estimating a measure of spatial chaos of an image is almost not addressed in the literature. Most publications on structure detection deal with finding structured patterns either of a given shape (Ballard, 1981) or based on edge detection (Mondal *et al.*, 2011), but not with estimating the level of structure. To the best of our knowledge, there is only one publication considering a similar problem (Chubb *et al.*, 1997), which presents a statistical method for testing the null hypothesis that an image is devoid of structure. Thus, defining a measure of chaos of an *m*/*z*-image is new, challenging and certainly requires an evaluation. We propose and apply a statistical evaluation approach by using a well-studied MALDI-IMS dataset from a rat brain section (Alexandrov and Kobarg, 2011; Alexandrov *et al.*, 2010).

Based on the ranking of *m*/*z*-images by their measure of spatial chaos, we propose solutions to the two described problems. First, we will show how to detect unknown molecules by calculating the measure of chaos of all *m*/*z*-images and selecting least chaotic *m*/*z*-images. This can be used, in particular, for highly sensitive peak picking by selecting a peak of any intensity if its *m*/*z*-image exhibits any spatial structure, whereas all established peak picking methods at least partially rely on intensity and suffer from low sensitivity. Second, we will show how to test for the presence of known molecules by calculating the measure of chaos for their *m*/*z*-images and taking least chaotic *m*/*z*-images. This can be used for searching for drug metabolites of unknown localization. We propose how the test for presence can improve current MALDI-IMS-based protein identification approaches.

2 METHODS

2.1 Samples, MALDI IMS, preprocessing

Sample preparation and MALDI-IMS data acquisition from a rat brain coronal section are described in detail in Alexandrov *et al.* (2010). Here, we follow our short description from Alexandrov and Kobarg (2011). The 10 μm thick cryosections were cut on a cryostat, transferred to a conductive indium-tin-oxide-coated glass slide (Bruker Daltonik GmbH, Bremen, Germany) and measured using a MALDI time-of-flight mass spectrometer Autoflex III (Bruker Daltonik) using the flexControl 3.0 and flexImaging 2.1 software (Bruker Daltonik). The lateral resolution was set to 80 μm. For mass calibration, the ClinProt standard mixture of peptides and proteins (Bruker Daltonik) covering a mass range of 2–20 kDa was used. The spectra pre-processing was done in the ClinProTools 2.2 software (Bruker Daltonik). The spectra were first baseline-corrected with the TopHat algorithm (minimal baseline width set to 10%) and then normalized using the total ion count normalization. No binning was performed. An *m*/*z*-image, even with no binning performed after data acquisition, represents a narrow range of *m*/*z*-values (an *m*/*z*-bin) due to a limited mass resolution. Then, spectra were saved into text files and loaded in Matlab R2012a (The Mathworks Inc., Natick, MA, USA), where the processing was performed using custom-made scripts. The dataset comprises 20 185 spectra acquired within the slice area (120 × 201 pixels), each spectrum of 3045 data bins in the *m*/*z* range 2500–10 000 Da. Before visualization, every *m*/*z*-image was processed with the automatic hot spot removal method using quantile thresholding (Watrous *et al.*, 2011) with the quantile value 0.99.

2.2 Test sets of *m*/*z*-images

For statistical evaluation of the proposed measure of chaos and the method for detection of unknown molecules, we selected test sets of unstructured and structured *m*/*z*-images of the rat brain MALDI-IMS dataset after visual examination of all 3045 *m*/*z*-images. Based on our expertise in MALDI-IMS and on our knowledge of this dataset, which we analyzed previously (Alexandrov and Kobarg, 2011; Alexandrov *et al.*, 2010), we selected 50 unstructured *m*/*z*-images; see an example in Figure 1. Then, we considered all images that looked structured and divided them into four classes based on the patterns they exhibit, namely, images with (i) compact curve-like sets high-intensity pixels, ‘curves’, (ii) large separated regions of high-intensity pixels, ‘regions’, (iii) large sets of high-intensity pixels with visible gradients of intensities around them, ‘gradients’ and (iv) small compact groups of high-intensity pixels, ‘islets’. Then, we selected 50 images of each class; see examples in Figure 2. The formulation of four classes of structured images as well as selection of 50 images for each class was performed before implementing the method to avoid possible bias in evaluation.

2.3 Measure of spatial chaos of an *m*/*z*-image

The steps of calculation of the measure of chaos are illustrated in Figure 1. At the first step, given an *m*/*z*-image $I \in \mathbb{R}^{k \times l}$, where $k \times l$ is the image size, we used a two-step edge detection filter for noisy images to detect intensity edges inside an image defined as sharp changes (or, more formally, discontinuities) of pixels intensities. First, we denoised the image with the bilateral edge-preserving denoising (Tomasi and Manduchi, 1998) producing $B = BF(I)$, where the width of the Gaussian bilateral filter window ω was 20, the spatial-domain standard deviation σ_d was three and the intensity-domain standard deviation σ_i was 0.3; for the influence of σ_i on the measure of chaos, see Supplementary Figure S4. Then, we calculated the values of the discrete gradient operator for all pixels of the denoised image $D = \text{grad}(B) \in \mathbb{R}^{k \times l \times 2}$. The high pixels intensities in the edge-detector image D correspond to sharp changes of pixels intensities of the original *m*/*z*-image. From D, we created a binary edge mask $M \in \{0,1\}^{k \times l}$ by using the quantile thresholding. We assigned one to Nq the most intense pixels of the edge detector image D and zero to other pixels, where N is the number of all pixels in the image and $0 < q < 1$ is the quantile parameter. The one-valued pixels of the binary edge mask M indicate the most sharp changes of the pixels intensities of the original *m*/*z*-image. At the

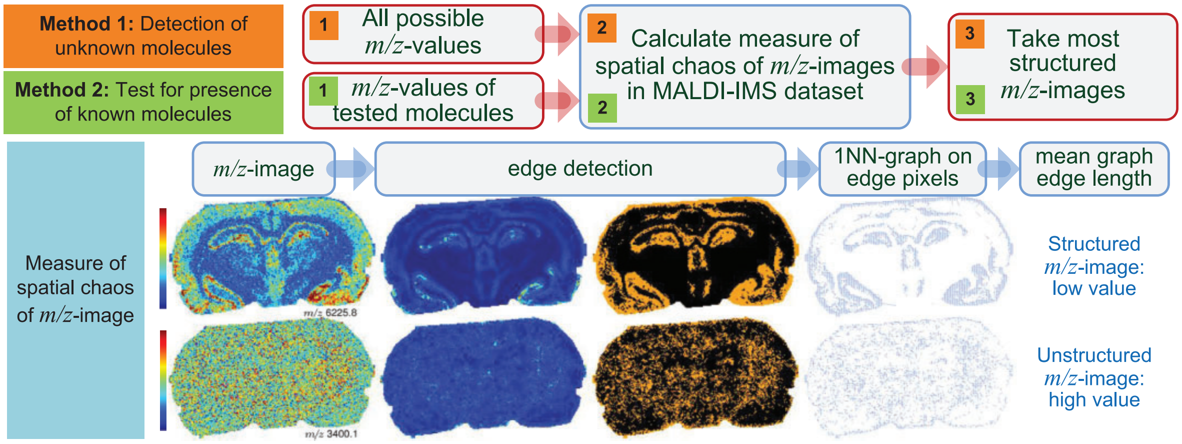
Method 1: Detection of unknown molecules
Method 2: Test for presence of known molecules
1 All possible m/z-values
1 m/z-values of tested molecules
2 Calculate measure of spatial chaos of m/z-images in MALDI-IMS dataset 2
3 Take most structured m/z-images 3
Measure of spatial chaos of m/z-image
m/z-image
edge detection
1NN-graph on edge pixels
mean graph edge length
m/z 6225.8
m/z 3400.1
Structured m/z-image: low value
Unstructured m/z-image: high value

Fig. 1. Detection of unknown molecules (Method 1), test for presence of known molecules (Method 2) as well as the steps of calculation of the measure of spatial chaos of an m/z-image. The intensities of each m/z-image are shown in the pseudo-colormap. The edge detection finds sharp changes of pixels intensities; edge pixels are highlighted. The one-nearest neighbor (1NN) graph connects detected edge pixels; each pixel is connected with its nearest neighbor

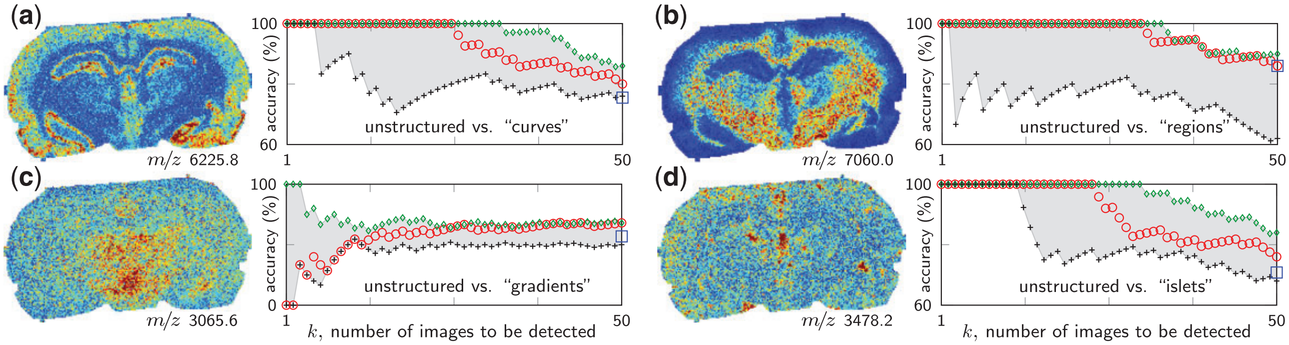

Fig. 2. Statistical evaluation of the method for detection of unknown molecules, separately for each test set of structured m/z-images. Each panel (**a–d**) shows (i) an exemplary m/z-image from the corresponding structured test set and (ii) the accuracies plotted against the parameter k, the number of m/z-images to be detected. For each k, we plotted the accuracy of the proposed method (red circle) and the baseline accuracies (cross and diamond). The baseline accuracies are calculated as the worst and the best accuracies among quantile values considered in edge detection; see Section 2. The square at k equal to 50 shows the double cross-validation accuracy from Table 1. For the test sets 'curves' (**a**), 'regions' (**b**) and 'islets' (**d**) our method detected the structured m/z-images with the perfect accuracy of 100% for k from 1 to 23. The test set 'gradients' (**c**) is challenging, especially when detecting a low number of m/z-images

second step, we built a one-nearest neighbor graph G on the one-valued pixels of M by linking each one-valued pixel of M with a graph edge to its closest neighboring one-valued pixel. To each graph edge, a weight was assigned calculated as the Euclidean distance between pixels connected by the edge. At the third step, we calculated the mean length $\bar{E}$ of the graph edges. This statistic depends on the earlier used quantile value q, i.e. $\bar{E} = \bar{E}(q)$. Finally, the measure of chaos was calculated as $\mathcal{C} = \min_{q \in \mathcal{Q}} \bar{E}(q)$ that is the minimum among the mean graph edge lengths over all considered quantile values $\mathcal{Q} = \{0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95\}$.

2.4 Method of detection of unknown molecules

Given a MALDI-IMS dataset, we calculated the value of the measure of spatial chaos for each m/z-image of this dataset and selected k m/z-images with the lowest values of the measure of chaos. The parameter k specifies the number of m/z-images to be detected.

2.5 Test for presence of known molecules

Given a MALDI-IMS dataset and a list of molecules to be tested for presence in this dataset, we computed their theoretical monoisotopic masses MH^+ corresponding to protonated ions, which are the dominant in MALDI. Then, we calculated the measure of chaos of m/z-images with m/z-values closest to the MH^+ masses. The tested molecules were sorted by the values of the measure of chaos of their m/z-images.

3 RESULTS

3.1 The measure of spatial chaos of an m/z-image

Figure 1 illustrates the steps of calculation of the proposed measure of spatial chaos applied to two m/z-images from a MALDI-IMS dataset of a rat brain coronal section: a structured m/z-image 6225.8 and an unstructured m/z-image 3400.1. Briefly, we detected the sharp changes of pixels intensities (also known as intensity edges) [for more on edge detection see e.g. Canny (1986)], then created a one-nearest-neighbor (1NN) graph for the detected pixels by connecting each edge pixel with its nearest neighbor. The measure of spatial chaos is defined as the mean length of the graph edges. A structured image has clear intensity edges with edges pixels compactly located, which leads to a low measure of spatial chaos. The edge detection algorithm applied to an unstructured image finds many separated intensity edge pixels due to the pixel-to-pixel variation producing high-intensity pixels randomly distributed across the image, which leads to a high measure of spatial chaos.

3.2 Statistical evaluation of the measure of spatial chaos by classifying test sets of *m*/*z*-images

We statistically evaluated the proposed measure of spatial chaos on a well-studied MALDI-IMS dataset from a coronal rat brain section presented in Alexandrov and Kobarg (2011); Alexandrov *et al*. (2010). First, we selected by the naked eye five test sets each of 50 *m*/*z*-images: (i) a set of unstructured images and four sets of structured images exhibiting (ii) compact curve-like series of high-intensity pixels, 'curves', (iii) large distinct regions of high-intensity pixels, 'regions', (iv) large regions of high-intensity pixels with gradually decreasing intensities around them, 'gradients' and (v) small compact groups of high-intensity pixels, 'islets'; see exemplary images in Figure 1 and 2. These types of structured images were defined by us after visual examination of all *m*/*z*-images. Then, we calculated for each *m*/*z*-image its value of the measure of spatial chaos. For each class of structured images, we calculated the double cross-validation accuracy when classifying this class versus the class of unstructured images using Support Vector Machines as described in Alexandrov *et al*. (2009) but without using the statistical test for difference and discrete wavelet transformation; the double cross-validation was used to ensure unbiased evaluation.

The classification results (Table 1) confirm that based only on the measure of spatial chaos, we were able to automatically reproduce the expert judges (structured–unstructured). The best discrimination (accuracy 84%) was achieved for the structured class 'regions' containing *m*/*z*-images with large separated regions of high-intensity pixels. The worst discrimination (accuracy 55%, specificity 48%) was achieved for the structured class 'gradients' with images with large sets of high-intensity pixels with visible gradients of intensities around them. This can be explained by the fact that the edge detection works well for the 'curves' images and cannot detect clear edges for the 'gradients' images. We hypothesize that using advanced methods for edge detection such as active contours or so-called snakes can help solving this problem.

Table 1. Statistical evaluation of the measure of spatial chaos using test sets of *m*/*z*-images of the MALDI-IMS rat brain dataset

Classification of test sets	Accuracy[a]	Sensitivity	Specificity
Unstructured versus 'curves'	75	83	66
Unstructured versus 'regions'	84	85	83
Unstructured versus 'gradients'	55	48	61
Unstructured versus 'islets'	71	75	66

[a]Accuracies, sensitivities and specificities (in %) for classification of unstructured versus structured m/z-images. The test sets (each of 50 m/z-images) were selected by an expert.

3.3 Statistical evaluation of the method for detection of unknown molecules

For the detection of unknown molecules, we calculated the measure of chaos for all *m*/*z*-images and selected *m*/*z*-images with the lowest values of the measure; see Figure 1 for an overview. To statistically evaluate the proposed method, we used the earlier described test sets of *m*/*z*-images of the rat brain MALDI-IMS dataset. One particular aim of the evaluation was as follows. At the first step of calculating the measure of chaos of an *m*/*z*-image, we detect image intensity edges by (i) applying an edge detection algorithm and (ii) converting the edge detector values to a binary mask (edge/no edge) using the quantile thresholding, see Section 2. We proposed that among various considered quantile values, the quantile value producing the minimal mean graph edge is to be selected, see Section 2. It is crucial to evaluate this method of selection of the quantile value.

We proposed the following evaluation scheme. For each class of structured *m*/*z*-images ('curves', 'regions', 'gradients' and 'islets'), we mixed its *m*/*z*-images together with the *m*/*z*-images of the unstructured class and then applied our method to detect *k* *m*/*z*-images, where *k* varied from one to 50 (the number of structured *m*/*z*-images). This procedure simulates detection of *k* unknown molecules in a MALDI-IMS dataset. Given *k* *m*/*z*-images detected, we calculated the accuracy defined as the percentage of those detected *m*/*z*-images, which are from the structured class. For example, if among *k* detected *m*/*z*-images all are from the structured class, the accuracy is 100%. For each *k*, we calculated the best and the worst accuracies over all quantile values. These two accuracy values serve as baseline values. Figure 2 shows for each *k* (the number of *m*/*z*-images to be detected) the accuracy of our method as well as the baseline accuracy values. First, our method performs comparably well with the best possible quantile value (the upper baseline). Our method is feasible in practice, whereas selecting the best possible quantile value is not. To select the best possible quantile, one has to calculate the accuracy that requires knowing which *m*/*z*-images are structured. Second, for the test sets 'curves' (Fig. 2a), 'regions' (Fig. 2b) and 'islets' (Fig. 2d), our method provided the perfect accuracy of 100% for *k* from 1 to 23. That is, if one uses our method to detect 5, 10 or 20 (up to 23) structured images in the considered test set of 100 *m*/*z*-images, images from the structured class will be correctly found. As expected, detecting images from the test set 'gradients' (Fig. 2c) is challenging, especially when detecting a low number of *m*/*z*-images.

3.4 Detection of unknown molecules in the rat brain

After statistical evaluation, we applied the proposed method for detection of unknown molecules to the rat brain MALDI-IMS dataset. Figure 3a shows the sorted values of the measure of chaos for all *m*/*z*-images of the dataset; ~500 images have comparatively low values (a steep part of the blue curve). For illustration, we selected 10 equidistant values of the measure of chaos

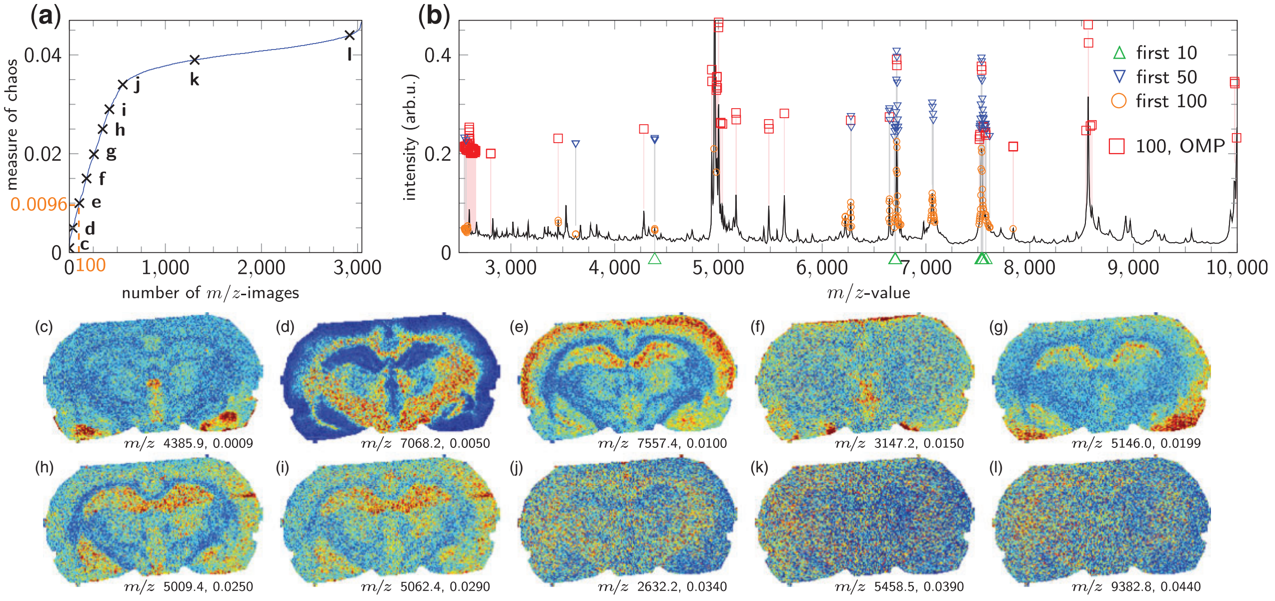

Fig. 3. Detection of unknown molecules in the MALDI-IMS dataset from a rat brain coronal section. (**a**) Sorted values of the measure of spatial chaos for all m/z-images. Crosses indicate ten equidistant values of the measure of chaos; their m/z-images are shown in panels (**c–l**). (**b**) The dataset mean spectrum with the most 10 (triangle), 50 (upside-down triangle) and 100 (circle) m/z-images having lowest values of the measure of chaos. For comparison, 100 peaks selected by a conventional spectrum-wise peak picking algorithm are shown with rectangles. (**c–l**) m/z-images corresponding to the equidistant values of the measure of spatial chaos pointed out in the panel (**a**)

(black crosses in Fig. 3a, annotated with c–l) and plotted the corresponding m/z-images in Fig. 3c–l. The first seven m/z-images (Fig. 3c–i) with low values of the measure of chaos look structured, whereas the last three m/z-images (Fig. 3k–l) appear to be unstructured. Among all 3045 m/z-images, 556 images have lower values of the measure of chaos than that in Figure 3j. Figure 3b shows the dataset-mean spectrum with the 10, 50 and 100 m/z-images having lowest values of the measure of chaos. All detected m/z-values correspond to peaks. Multiple m/z-images may correspond to different m/z-bins of the same peak (Fig. 3b) and should be aligned as proposed in Alexandrov and Kobarg (2011).

We compared the 100 detected m/z-images with 100 peaks found by a spectrum-wise peak picking algorithm designed for MALDI-IMS (Alexandrov *et al.*, 2010), which finds a number of peaks for each spectrum and then selects most frequently found peaks. Although the peak picking found most major peaks, it overlooked a large peak at m/z 7100 as well as two low-intensity peaks at m/z 3625.5 and 4385.9; the latter peak is discussed in Figure 5. This illustrates a general disadvantage of spectrum-wise peak picking methods, which suffer from low sensitivity, see Section 4. Our method for detection of unknown molecules can complement spectrum-wise peak picking increasing its sensitivity, as it does not depend on peak intensity, but only on the measure of spatial chaos of the corresponding m/z-image.

The processing time for the full dataset was 16 min (0.31 s for one m/z-image of size 120 × 201 pixels) on a laptop with i5 2.5 GHz CPU, whereas the visual analysis, described in Section 1, aiming to select 20–50 structured m/z-images normally takes >1 h and is still not as extensive as the automatic analysis.

To ascertain that the method of detection of unknown molecules is not overfit to the rat brain dataset, we applied it to another MALDI-IMS dataset from a rat kidney section acquired at our laboratory; see Supplementary Figure S1. The resulted ranking of m/z-images of the rat kidney dataset was relevant as evaluated by an expert.

3.5 Test for presence of known molecules in the rat brain dataset

We applied the proposed method to test for presence of peptides in the rat brain MALDI-IMS dataset. A list of 50 occurring in rat peptides in the m/z-range 2500–10 000 was selected using the Uniprot database (Consortium, 2012) (version 82, released 11.07.2012). For the list of peptides with their names, precursor accession numbers and names, and masses, see Supplementary Table S1. For each peptide, its theoretical monoisotopic mass MH^+ was stored into the list of m/z-values. Afterwards, for each m/z-value from the list, we calculated the measure of spatial chaos of the corresponding m/z-image.

The sorted values of the measure of chaos as well as 10 least chaotic m/z-images are shown in Figure 4. All peptides ranked by their values of the measure of spatial chaos are listed in Supplementary Table S1, and their m/z-images are shown in Supplementary Figure. S2. Visually, the ranking of m/z-images looks in agreement with a possible expert ranking based on a level of structure.

This information does not assume or provide peptide identification. The peptides detected as ‘present’ need to be validated either by *in situ* MS/MS or by an extraction coupled with LC-MS/MS (Gustafsson *et al.*, 2012; Rauser *et al.*, 2010;

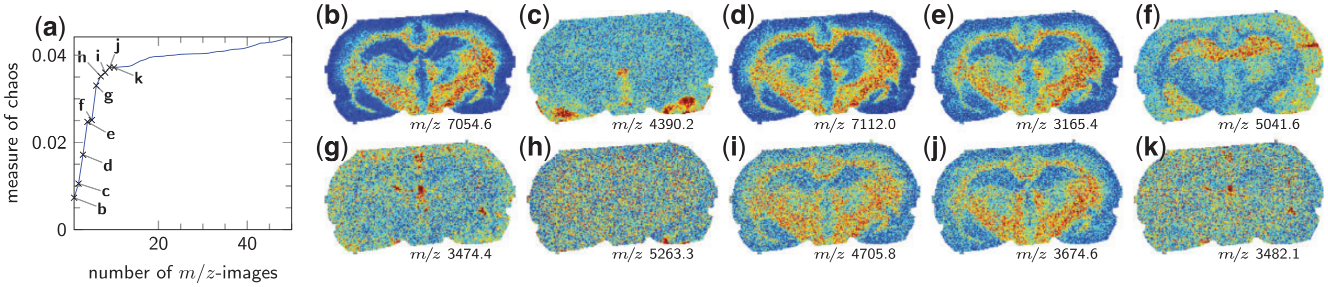

Fig. 4. Test for presence of 50 peptides in the rat brain MALDI-IMS dataset. (**a**) The sorted values of the measure of spatial chaos of *m*/*z*-images of the peptides. (**b–k**) *m*/*z*-images corresponding to 10 peptides with lowest values of the measure of chaos. For the peptide names, masses and values of the measure of chaos, see Supplementary Table S1; for all *m*/*z*-images, see Supplementary Figure S2

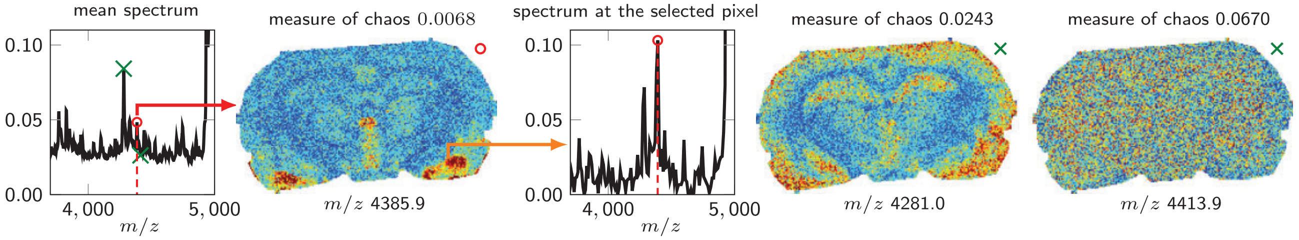

Fig. 5. Illustration of the need for highly sensitive peak picking in IMS. A peak at *m*/*z* 4385.9, which was detected among 100 *m*/*z*-values by our method (Fig. 3) has a low intensity in the mean spectrum. However, the spectrum acquired at a high-intensity pixel of this *m*/*z*-image (shown with an arrow) exhibits a peak of significantly higher intensity at the same *m*/*z*-value. Circle indicates the *m*/*z*-value detected among 100 *m*/*z*-values by our method; crosses indicate *m*/*z*-values not detected but shown for illustration

Schober *et al*., 2011; Stauber *et al*., 2010). We propose to use the test for presence to improve and make automatic current MALDI-IMS-based protein identification approaches, see Section 4.

4 DISCUSSION

Our motivation to exploit a measure of spatial chaos to detect unknown molecules and to test for presence of known molecules is as follows. Selecting structured *m*/*z*-images after visual examination of *m*/*z*-images is the well-accepted approach of manual analysis MALDI-IMS data and is a part of everyday work of an imaging mass spectrometrist. Assuming any spatial differentiation of the sample, one is normally interested in molecules whose spatial distribution across the sample is not homogeneous. Normally, only molecules corresponding to baseline signals (e.g. matrix) have homogeneous distribution. In MALDI-IMS, the baseline signals can have non-zero intensities. Owing to the strong pixel-to-pixel variation, which is inherent to MALDI-IMS, their *m*/*z*-images exhibit noise and look unstructured.

4.1 The need for highly sensitive peak picking

Earlier in the text, we concluded that our method for detection of unknown molecules complements the conventional spectrum-wise peak picking for IMS data by allowing a highly sensitive peak picking; see Figure 3. This is an important achievement because the two existing approaches to peak picking in IMS suffer from low sensitivity.

In the average-spectrum approach, a peak picking algorithm is applied to the mean, medium or quantile spectrum of the dataset. As illustrated in Watrous *et al*. (2011), this approach is fast but not sensitive, as it does not favor high peaks presented in a small part of a sample only. For example, if a peak is present only in 1% of spectra (for an image of 100 × 100 pixels, this is a large area of 10 × 10 pixels), then its contribution to the mean spectrum is reduced by 100 times as compared with a low peak presented in all spectra (e.g. a matrix peak). Figure 5 illustrates this effect. A peak at *m*/*z* 4385.9 detected among 100 *m*/*z*-values by our method (Fig. 3) has a low intensity in the mean spectrum. However, the spectrum acquired at a high-intensity pixel of the *m*/*z*-image 4385.9 (orange arrow in Fig. 5) exhibits a peak of a muchhigher intensity at the same *m*/*z*-value.

In the spectrum-wise approach, a peak picking algorithm is applied to each spectrum. Then, the obtained spectrum-wise peak lists are combined into a consensus peak list, e.g. by taking those peaks that are found in at least 1% of spectra (Alexandrov *et al*., 2010). This procedure also has limitations in terms of sensitivity and can overlook a peak presented in a smaller than 1% part of spectra. Additionally, an imperfection of the peak picking algorithm might lead to overlooking low-intensity peaks. Another issue inherent to the spectrum-wise peak picking is the inverse proportion between efficiency and sensitivity. To decrease runtime, one can process only a part of spectra (e.g. each 10th spectrum), but this reduces the sensitivity. The sensitivity can be increased by selecting more peaks per spectrum, but this increases the runtime for most peak picking algorithms. This issue becomes even more troublesome as spatial resolution increases resulting in a quadratic increase in the number of spectra per MALDI-IMS dataset.

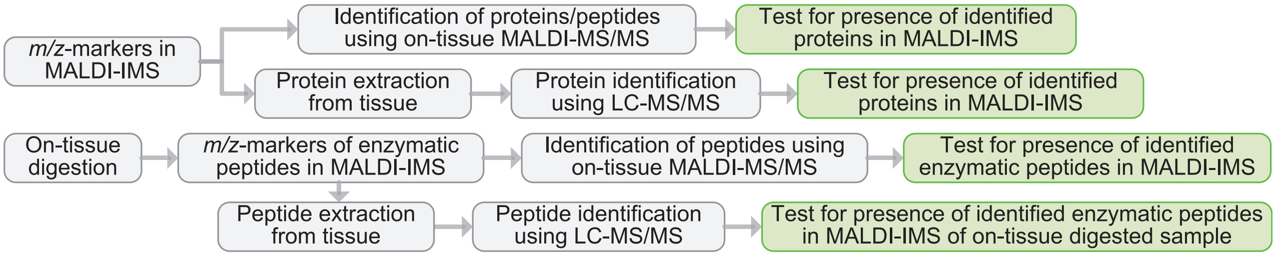
m/z-markers in MALDI-IMS
Identification of proteins/peptides using on-tissue MALDI-MS/MS
Test for presence of identified proteins in MALDI-IMS
Protein extraction from tissue
Protein identification using LC-MS/MS
Test for presence of identified proteins in MALDI-IMS
On-tissue digestion
m/z-markers of enzymatic peptides in MALDI-IMS
Identification of peptides using on-tissue MALDI-MS/MS
Test for presence of identified enzymatic peptides in MALDI-IMS
Peptide extraction from tissue
Peptide identification using LC-MS/MS
Test for presence of identified enzymatic peptides in MALDI-IMS of on-tissue digested sample

Fig. 6. Our vision how the proposed test for presence can be incorporated into current MALDI-IMS-based protein identification approaches

4.2 Detection of drug metabolites and protein identification

Several applications can be enabled by the proposed test for presence of known molecules. First application is the IMS-based detection of drug metabolites, which can be predicted *in silico*, e.g. using DrugBank (Wishart *et al*., 2006). Given a list of m/z-values of candidate drug metabolites, one can test for their presence in the sample and reduce the number of candidate metabolites or to select the most prominent of them. Second application is the IMS-based protein identification, which is an crucial challenge in MALDI-IMS-based proteomics (Gustafsson *et al*., 2012; Rauser *et al*., 2010; Schober *et al*., 2011; Stauber *et al*., 2010). Figure 6 shows existing approaches to MALDI-IMS-based protein identification. All of them involve testing for presence of identified proteins in MALDI-IMS data, i.e. currently done visually, but can be done automatically using our test for presence. When using on-tissue digestion approaches, one can *in silico* predict protein fragments and to test both for their presence and for their co-localization. Just testing for co-localization may be not enough because of the strong pixel-to-pixel variation that, together with the need to control the number of false positives, requires a test for presence. Applications of the test for presence will greatly benefit from using the high mass accuracy MALDI-IMS systems such as FT-ICR (Seeley and Caprioli, 2008) and Orbitrap (Schober *et al*., 2011).

4.3 Importance of high mass resolution

One can argue that the test for presence is feasible only for high mass resolution IMS. In case of insufficient resolution, a molecule of interest is not separated from other molecules of similar m/z, and they are represented with one m/z-image. This can lead to false positives, i.e. molecules can be detected as ‘present’, although they are not. However, testing for absence does not suffer from this problem. If an m/z-image has a high value of the measure of spatial chaos, then there is no molecule within the corresponding m/z-bin, which has a spatially structured pattern and contributes any significantly to the m/z-image. Thus, the hypothesis of presence of all molecules within the tested m/z-bin can be rejected.

4.4 Other hyperspectral imaging data

The presented methods can be applied to other types of hyperspectral imaging with strong pixel-to-pixel variation, such as confocal Raman microspectroscopic (Puppels *et al*., 1990), infrared spectroscopic (Fernandez *et al*., 2005), secondary ion mass spectrometry (Colliver *et al*., 1997; Watrous and Dorrestein, 2011) or nanostructure-initiator mass spectrometry (Woo *et al*., 2008) imaging.

ACKNOWLEDGEMENTS

The authors thank Chen Sagiv (SagivTech Ltd.) for the discussions and advice at the initial stage of this study, Michael Becker (Bruker Daltonik) for providing the rat brain dataset, Dennis Trede (SCiLS GmbH) for consulting on edge detection, Janina Oetjen (University of Bremen) for acquiring data presented in supporting information, as well as Chris Rath (Novartis Institute for Biomedical Research), Jeramie Watrous (University of California San Diego), Janina Oetjen, and Ilya Chernyavsky (University of Bremen) for their valuable input on the final version of the manuscript.

Funding: European Union Seventh Framework Programme FP7 under grant agreement 305259.

Conflict of Interest: Theodore Alexandrov is the scientific director of SCiLS GmbH which develops and markets software for imaging mass spectrometry.

REFERENCES

Ait-Belkacem,R. *et al.* (2012) Mass spectrometry imaging is moving toward drug protein co-localization. *Trends Biotechnol*., **30**, 466–474.

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

Alexandrov,T. *et al.* (2009) Biomarker discovery in MALDI-TOF serum protein profiles using discrete wavelet transformation. *Bioinformatics*, **25**, 643–649.

Alexandrov,T. *et al.* (2010) Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering. *J. Proteome Res*., **9**, 6535–6546.

Ballard,D.H (1981) Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recogn*., **13**, 111–122.

Balluff,B. *et al.* (2012) Direct molecular tissue analysis by MALDI imaging mass spectrometry in the field of gastrointestinal disease. *Gastroenterology*, **143**, 544–549.

Canny,J (1986) A computational approach to edge detection. *IEEE Trans. Pat. Anal. Mach. Int*., **PAMI-8**, 679–698.

Caprioli,R.M. *et al.* (1997) Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. *Anal. Chem*., **69**, 4751–4760.

Castellino,S. *et al.* (2011) MALDI imaging mass spectrometry: bridging biology and chemistry in drug development. *Bioanalysis*, **3**, 2427–2441.

Chaurand,P (2012) Imaging mass spectrometry of thin tissue sections: a decade of collective efforts. *J. Proteomics*, **75**, 4883–4892.

Chubb,C. *et al.* (1997) Structure detection: a statistically certified unsupervised learning procedure. *Vision Res*., **37**, 3343–3365.

Colliver,T.L. *et al.* (1997) Atomic and molecular imaging at the single-cell level with ToF-SIMS. *Anal. Chem.*, **69**, 2225–2231.

Consortium,T.U (2012) Reorganizing the protein space at the universal protein resource (uniprot). *Nucleic Acids Res.*, **40**, D71–D75.

Crecelius,A.C. *et al.* (2012) Application of MALDI-MSI for photolithographic structuring. *Anal. Chem.*, **84**, 6921–6925.

Fernandez,D.C. *et al.* (2005) Infrared spectroscopic imaging for histopathologic recognition. *Nat. Biotechnol.*, **23**, 469–474.

Gustafsson,J.O. *et al.* (2012) Internal calibrants allow high accuracy peptide matching between MALDI imaging MS and LC-MS/MS. *J. Proteomics*, **75**, 5093–5105.

Jungmann,J.H. and Heeren,R.M (2012) Emerging technologies in mass spectrometry imaging. *J. Proteomics*, **75**, 5077–5092.

Khatib-Shahidi,S. *et al.* (2006) Direct molecular analysis of whole-body animal tissue sections by imaging MALDI mass spectrometry. *Anal. Chem.*, **78**, 6448–6456.

Kroiss,J. *et al.* (2010) Symbiotic Streptomycetes provide antibiotic combination prophylaxis for wasp offspring. *Nat. Chem. Biol.*, **6**, 261–263.

Lee,Y. *et al.* (2012) Use of mass spectrometry for imaging metabolites in plants. *Plant J.*, **70**, 81–95.

Mondal,T. *et al.* (2011) Automatic craniofacial structure detection on cephalometric images. *IEEE Trans. Image Proc.*, **20**, 2606–2614.

Prideaux,B. and Stoeckli,M (2012) Mass spectrometry imaging for drug distribution studies. *J. Proteomics*, **75**, 4999–5013.

Puppels,G.J. *et al.* (1990) Studying single living cells and chromosomes by confocal Raman microspectroscopy. *Nature*, **347**, 301–303.

Rauser,S. *et al.* (2010) Classification of HER2 receptor status in breast cancer tissues by MALDI imaging mass spectrometry. *J. Proteome Res.*, **9**, 1854–1863.

Rubakhin,S.S. *et al.* (2005) Imaging mass spectrometry: fundamentals and applications to drug discovery. *Drug Discov. Today*, **10**, 823–837.

Schober,Y. *et al.* (2011) Protein identification by accurate mass matrix-assisted laser desorption/ionization imaging of tryptic peptides. *Rapid Commun. Mass Spectrom.*, **25**, 2475–2483.

Schwamborn,K (2012) Imaging mass spectrometry in biomarker discovery and validation. *J. Proteomics*, **75**, 4990–4998.

Schwamborn,K. and Caprioli,R (2010) Molecular imaging by mass spectrometry–looking beyond classical histology. *Nat. Rev. Cancer*, **10**, 639–646.

Seeley,E.H. and Caprioli,R.M (2008) Molecular imaging of proteins in tissues by mass spectrometry. *Proc. Natl Acad. Sci. USA*, **105**, 18126–18131.

Stauber,J. *et al.* (2010) On-tissue protein identification and imaging by MALDI-ion mobility mass spectrometry. *J. Am. Soc. Mass Spectrom.*, **21**, 338–347.

Stoeckli,M. *et al.* (2001) Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues. *Nat. Med.*, **7**, 493–496.

Tomasi,C. and Manduchi,R. (1998) Bilateral filtering for gray and color images. In: *Proceedings of the1998 IEEE International Conference on Computer Vision, Bombay, India*. pp. 839–846.

Watrous,J.D. and Dorrestein,P.C. (2011) Imaging mass spectrometry in microbiology. *Nat. Rev. Microbiol.*, **9**, 683–694.

Watrous,J.D. *et al.* (2011) The evolving field of imaging mass spectrometry and its impact on future biological research. *J. Mass Spectrom.*, **46**, 209–222.

Wishart,D.S. *et al.* (2006) DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Res.*, **34**, D668–D672.

Woo,H.K. *et al.* (2008) Nanostructure-initiator mass spectrometry: a protocol for preparing and applying NIMS surfaces for high-sensitivity mass analysis. *Nat. Protoc.*, **3**, 1341–1349.

Yang,Y.L. *et al.* (2009) Translating metabolic exchange with imaging mass spectrometry. *Nat. Chem. Biol.*, **5**, 885–887.