Automated, Feature-Based Image Alignment for High-Resolution Imaging Mass Spectrometry of Large Biological Samples

Alexander Broersen,[a] Robert van Liere,[a] A. F. Maarten Altelaar,[b] Ron M. A. Heeren,[b] and Liam A. McDonnell[b]

[a] Centrum voor Wiskunde en Informatica, Amsterdam, The Netherlands
[b] FOM Institute for Atomic and Molecular Physics, Amsterdam, The Netherlands

High-resolution imaging mass spectrometry of large biological samples is the goal of several research groups. In mosaic imaging, the most common method, the large sample is divided into a mosaic of small areas that are then analyzed with high resolution. Here we present an automated alignment routine that uses principal component analysis to reduce the uncorrelated noise in the imaging datasets, which previously obstructed automated image alignment. An additional signal quality metric ensures that only those regions with sufficient signal quality are considered. We demonstrate that this algorithm provides superior alignment performance than manual stitching and can be used to automatically align large imaging mass spectrometry datasets comprising many individual mosaic tiles.   (J Am Soc Mass Spectrom 2008, 19, 823–832) © 2008 American Society for Mass Spectrometry

Imaging mass spectrometry is a rapidly developing analytical tool because it provides the ability to map the profiles of specific biomolecules, in which the intrinsic mass of the molecule differentiates between any modified forms; to record the distributions of multiple analytes in parallel; and to perform these analyses without a label and with clinical samples [1]. This combination of specificity, parallel detection, and non-targeted analysis has led to great excitement for its potential as a discovery tool.

It is the goal of many research groups to be able to perform high-resolution analysis of large samples, thus combining the ability to examine distributions between organs/tumors and their surroundings as well as to investigate the subcellular/intracellular locations of the biomolecules. The central premise of this approach is that the subcellular locations will provide some of the information required to explain differences in the more global patterns.

High spatial resolution measurements are a well-established ability of secondary ion mass spectrometry (SIMS). Recent advances in both ionization efficiency (polyatomic primary ions) and sample preparation have significantly improved the sensitivity for detecting intact, medium-sized molecular ions (<1000 Da) from tissues and cells [1–6]. High-resolution images of small peptides, lipids, cholesterol, vitamins, and pharmaceuticals have all been reported, and through the use of large polyatomic primary ions three-dimensional (3D) molecular imaging results are beginning to appear [7–9].

The images are normally created by moving the ionization beam in a set pattern across the sample and performing mass analysis at each point of the raster (spatially correlated mass spectrometry). The raster pattern typically uses 8- or 10-bit encoding; as a result the maximum analysis field contains 256×256 (or 1024×1024) pixels. In a typical high spatial resolution SIMS measurement the pixel size is about 200 nm, meaning maximum analysis areas of approximately 50 or 200 μm, respectively. This is much smaller than many of the biological samples of interest; for example, a tissue section of an adult rat brain is approximately 2×1 cm. Reports [5, 6] on SIMS imaging mass spectrometry analysis of large samples first mention a low-resolution scan. This is accomplished by rastering the sample, followed by a high-resolution scan of selected areas. This approach of combining ionization beam and sample stage rastering has been developed further to provide high-resolution imaging of large areas: in mosaic mode imaging the sample is divided into a mosaic of small areas (termed tiles) with the sample stage raster; each tile is then analyzed with high spatial resolution using ion beam rastering and the results are combined to provide the final, complete (mosaic) dataset [2–4, 10].

Mosaic mode image (high-resolution imaging of large areas) has been limited by the user-intensive nature of data analysis. The accuracy of sample stages capable of moving through large areas (e.g., 5×5 cm) are often significantly less than the resolution of the SIMS imaging mass spectrometry experiment, for example, 5 and 0.2 μm, respectively. As a result, the

Address reprint requests to Dr. Liam McDonnell or Prof. Dr. Ron M. A. Heeren, FOM Institute for Atomic and Molecular Physics, Kruislaan 407, 1098 SJ Amsterdam, The Netherlands. E-mail: L.A.Mcdonnell@lumc.nl or heeren@amolf.nl

high-resolution images have to be aligned after data acquisition. In most examples this has been performed "by eye." This process is time consuming, user intensive, and subjective. At its best a manual alignment process would include the following steps:

(1) Import the data into a mathematical package capable of processing >2-Gb data files.
(2) Calculate total-ion-count (TIC) images of each tile.
(3) Create layered image to manipulate relative positions of TIC images.
(4) Vary relative positions until "optimum" is found.
(5) Use offsets determined in step 4 to combine tile data files into a single datacube.

However, it is more common to use the proprietary software to create the images from each tile's dataset, for each m/z range of interest. The TIC images are then aligned in a graphics package, the relative offsets applied to the m/z-specific images, and the results saved. In fact, no case studies are known to the authors in which multiple tiles were combined into one dataset and in which all raw data events could be preserved.

The work reported here describes a new automated image alignment algorithm for imaging mass spectrometry datasets. Usually the uncorrelated noise prevents the optimum offsets of two images from being determined. Here it is demonstrated how the reduction of uncorrelated noise provided by principal component analysis (PCA) can be exploited for automated alignment. PCA is an unsupervised multivariate technique that uses orthogonal linear transformations of the data onto a new coordinate system to project the greatest variance of the data onto the first component, the second greatest variance onto the second component, and so forth [11–13]. This processing extracts features with less uncorrelated noise from the datasets, thus allowing their automated alignment. A metric combines a measure of the equality of the overlapping regions as well as a measure of the signal quality. The automatic alignment routine is demonstrated on several high-resolution large area datasets.

Introduction of Automatic PCA-B Alignment

Figure 1 shows a schematic of how the automatic alignment of imaging mass spectrometry datasets is performed. The first step uses PCA to minimize the uncorrelated noise in the datasets and to reduce the size of the datasets. Results of the analysis are recorded in both the spatial and spectral domains, in which the spectral signatures are used to ensure that only images of similar chemical composition are compared. Typically the first 20 components are selected for each imaging dataset. More components can be extracted, but usually they lack sufficient spatial contrast to be useful. The second step of the procedure examines the

reconstructed mass spectra of the components to find components with similar chemical signatures. The final step is image registration: the relative x and y offsets of the spatial representation (score images) of the two chemically matched components are varied throughout all combinations of overlapping regions (or a user-selected region for the case in which the approximate offsets are known). For each relative offset the difference in the overlapping regions is calculated, followed by the removal of those areas with insufficient or low-quality signal. The minimum in this metric corresponds to the optimum alignment of the two datasets.

Experimental

Measurement

All experiments were performed using a TRIFT-II time-of-flight mass spectrometer (Physical Electronics, Eden Prairie, MN, USA) and 15 keV primary ions (DC current 400 pA, pulse length 20 ns, spot diameter 500 nm). The primary ion dose, 3.5×10^{11} ions cm^{-2}, was well within the static SIMS limit [14] (statistically a fresh area is sampled with each primary ion impact), thus ensuring that the spectra are representative of the sample surface.

All imaging experiments used a rastered, focused ion beam. For each ion detected the ion's time-of-flight and origin are recorded. The ion images display how the intensities relate to position for a specified m/z range. The 8-bit primary ion deflectors used to raster the primary ion beam limit the number of pixels in any single imaging experiment to 256×256. To perform high-resolution imaging of large areas, the sample stage is used to manually define a mosaic and the ionization beam to record a high-resolution imaging mass spectrometry dataset for each of the mosaic's tiles. The algorithm reported herein has been developed to automatically align these tiles.

Sample

Freshwater snails (*Lymnaea stagnalis*) were decapitated and the cerebral ganglion and nervous tissue were dissected, directly embedded in 10% gelatin at 30 °C, and frozen at −80 °C. Tissue sections (10 μm thick) were cut on a Leica CM 3000 cryostat (Leica Microsystems, Nussloch, Germany) at −20 °C. Sections were placed on conductive, transparent indium tin oxide (ITO)-coated glass slides (Delta Technologies, Stillwater, MN, USA), dried in a stream of warm air for 1 min, frozen on dry ice, and stored at −80 °C. Conservation of morphology was checked by optical microscopy during the sectioning, drying, and storage process.

Data Analysis

The raw data, recorded using the maximum time and spatial resolution from the TRIFT, were read using MatLab (Version 7.3.0, R2006b, MathWorks, Natick,
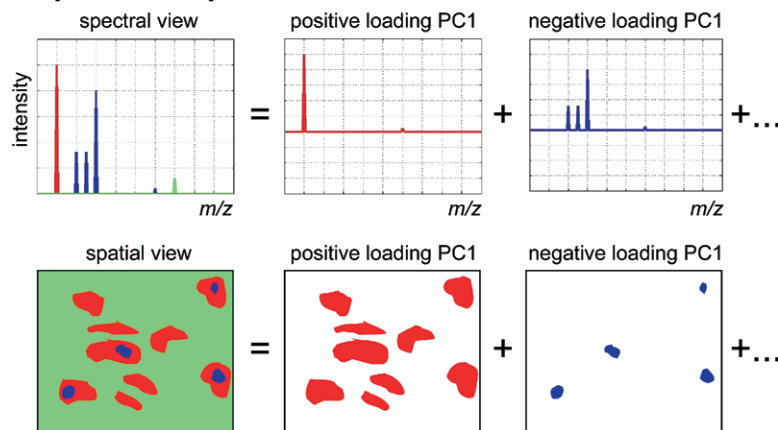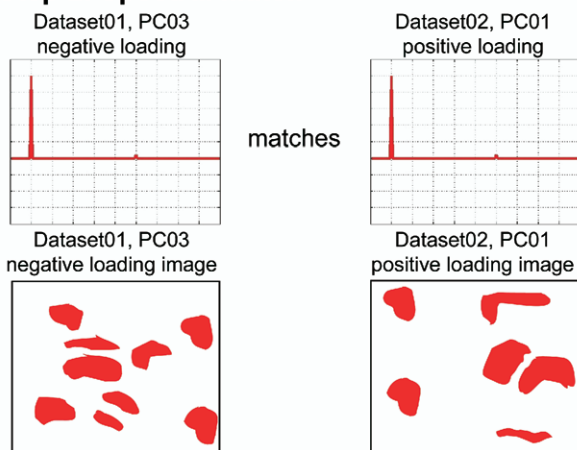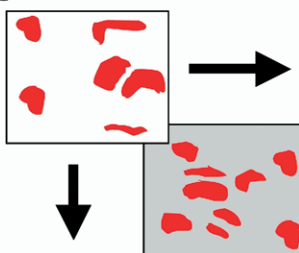
Step 1. Decomposition

Step 2. Spectral Match

Step 3. Registration

Figure 1.   Schematic of automatic image alignment algorithm.

MA). The raw data are event based, containing the spatial origin and time-of-flight of each detected ion. This dataset is converted into an $x \times y \times m/z$ datacube in which each voxel contains the number of counts for that spectral and spatial combination. To reduce memory requirements the data were saved as a sparse matrix. This omits the zeros without loss of information, because SIMS imaging data are quite sparse (large regions without counts). The datasets were binned in the time-of-flight domain (140-ns bins) because the performance of PCA has been shown to improve with increasing signal intensity, to ensure PCA examines chemical variation rather than variation within the

time-of-flight peaks [15] and to reduce memory requirements.

PCA is performed by unfolding the $x \times y \times m/z$ datacube in such a way that a 2D m/z by $x \times y$ matrix **X** is constructed. The components are then extracted in an orthonormal matrix P (see eq 1) using a standard eigenvector decomposition:

$$\mathbf{Y} = \mathbf{P} \cdot \mathbf{X}^{T} \tag{1}$$

This results in a matrix **Y** with spectral score vectors, which can be interpreted as spectral components. These spectral components have negative and positive peaks

(corresponding to the positive and negative parts of the maximized variance). Both groups can be considered as separate chemical signatures. The transposed datacube is multiplied with the spectral component matrix to obtain the spatial distributions of these spectral signatures:

$$\mathbf{I} = \mathbf{X}^T \cdot \mathbf{Y} \tag{2}$$

Each row in the resulting matrix $\mathbf{I}$ contains an unfolded image component that contains the spatial contributions of each spectral profile in $\mathbf{Y}$. To ensure that only images of similar chemical composition are used for image registration, the spectral signatures of the images are compared. The positive and negative spectral signatures of each component from dataset01 are compared with all positive and negative spectral signatures from dataset02 (note for this application the sign of the spectral score is unimportant). The two components with the smallest difference in their spectra are found using the mean-squares (MS) error metric, expressed as

$$MS(A, B) = \frac{1}{N} \sum_i^N (A_i - B_i)^2 \tag{3}$$

where A is a spectral signature from the first data tile, B is a spectral signature from the second data tile, N is the number of channels, and i is the channel. The image-alignment step is performed by calculating the similarity between the spectrally matched component images, for all possible relative positions. The applied metric was based on the same mean square metric used to match the spectra (eq 3), but where A is a region in the fixed image, B is a region in the moving image, N is the number of pixels in these regions, and i is the pixel position. Changes in the intensity distribution between both images will result in a poor match value and poor matches result in large values of the metric. This metric is used on all overlapping regions of the two image components. The results are displayed as a 2D or 3D plot, with the values of the metric for all combinations of overlapping regions.

The overlapping regions are small on the edges and corners of an image because the number of contributing positions is low. This causes the mean-squares error metric to be low because there is insufficient signal for a fair comparison. To alleviate this problem the entropy of the overlapping regions was calculated, which provides a measure of the randomness of the intensities using its histogram. The entropy H of each region in one image component was calculated using

$$H(A) = - \sum_A p_A \cdot \log p_A \tag{4}$$

where A is the region in an image component and p_A represents the histogram values of A. The entropies of all overlapping regions of both component images are

calculated and, by calculating their product, a combined entropy landscape is obtained. An additional fit on the histogram is then made to remove unwanted areas with low entropy values (insufficient signal). This is done by fitting all entropy values of all compared regions to a normal distribution. The values to the left of the standard deviation of the mean in the histogram are removed. The remaining entropy landscape is then applied as a weight for the MS landscape. The zero regions in entropy spaces indicate that a minimum found in the corresponding regions of the MS landscape will probably not contain enough information or too much "randomness" of signal to be considered as a solution. The final, weighted search space is created by dividing the MS landscape by the fitted entropy landscape. This results in a MS-entropy search space in which the minimum corresponds to the most probable offset between the two images. Multiple components can be used to verify the accuracy minimum to make the registration more robust.

Results

The automatic alignment procedure was first tested with a sample containing an array of spots of the MALDI matrix 2,5-dihydroxybenzoic acid on a silicon wafer [16]. This "standard" sample was chosen as the high contrast between the matrix analyte droplets and the sample plate provided an easy target for testing the analysis routines. Following the successful implementation of the routines for aligning these high-contrast samples, the routines were applied to the more challenging application of aligning high-resolution imaging mass spectrometry results of tissue sections.

Figure 2a and b show two total-ion-count (TIC) images from neighboring regions of a tissue section of the cerebral ganglion of the fresh water snail *Lymnaea stagnalis*. The high resolution of the imaging experiment clearly distinguishes the cellular detail. The complexity of the cellular details is advantageous: the intricate structures will cause the optimum alignment of the two datasets to be sharply defined. Figure 2c shows approximately the relative position of the two datasets. Clearly, one can manually organize the relative positions of the mosaic datasets, although as explained in the introduction this is both highly subjective and impractical.

Attempts at automatically aligning complete image datasets by calculating the difference metric for all spectral planes in the datasets proved unsuccessful because the significant amount of uncorrelated noise prevented the correct offsets from being determined. PCA is used here to minimize uncorrelated noise and thus aid the alignment procedure. Once PCA is performed the spectral representations of the components were compared using a least-square metric (eq 3), to find those components with similar chemical signatures, thus ensuring that their images describe the distribution of a similar chemical medley. Figure 3
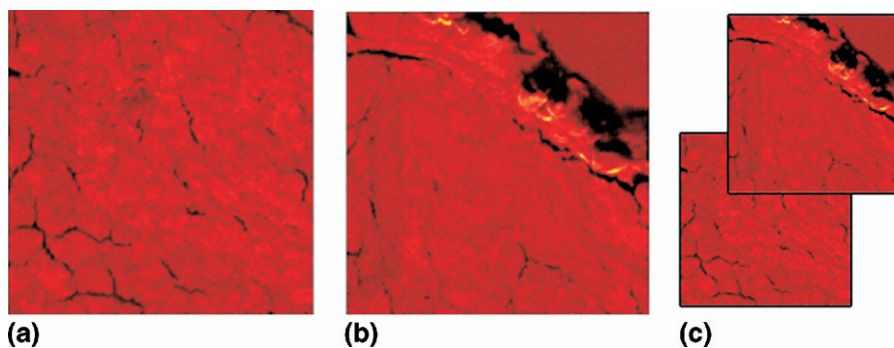
Figure 2. Parts (**a**) and (**b**) show total ion count (TIC) images of two overlapping regions of a tissue section of the cerebral ganglion of the freshwater snail *Lymnaea stagnalis*. Their approximate alignment is illustrated in (**c**).

shows examples of spectrally matched components of the two datacubes in both the spatial and *m/z* domains. The best spectrally matched components, used in the subsequent analysis, are positive component 6 from dataset01 and positive component 7 from dataset02. All principal components have positive and negative scores, which correspond to the positive and negative aspects of the variance. Recall PCA maximizes variance and it is this systematic variance—and the removal of uncorrelated noise—that is advantageous for image
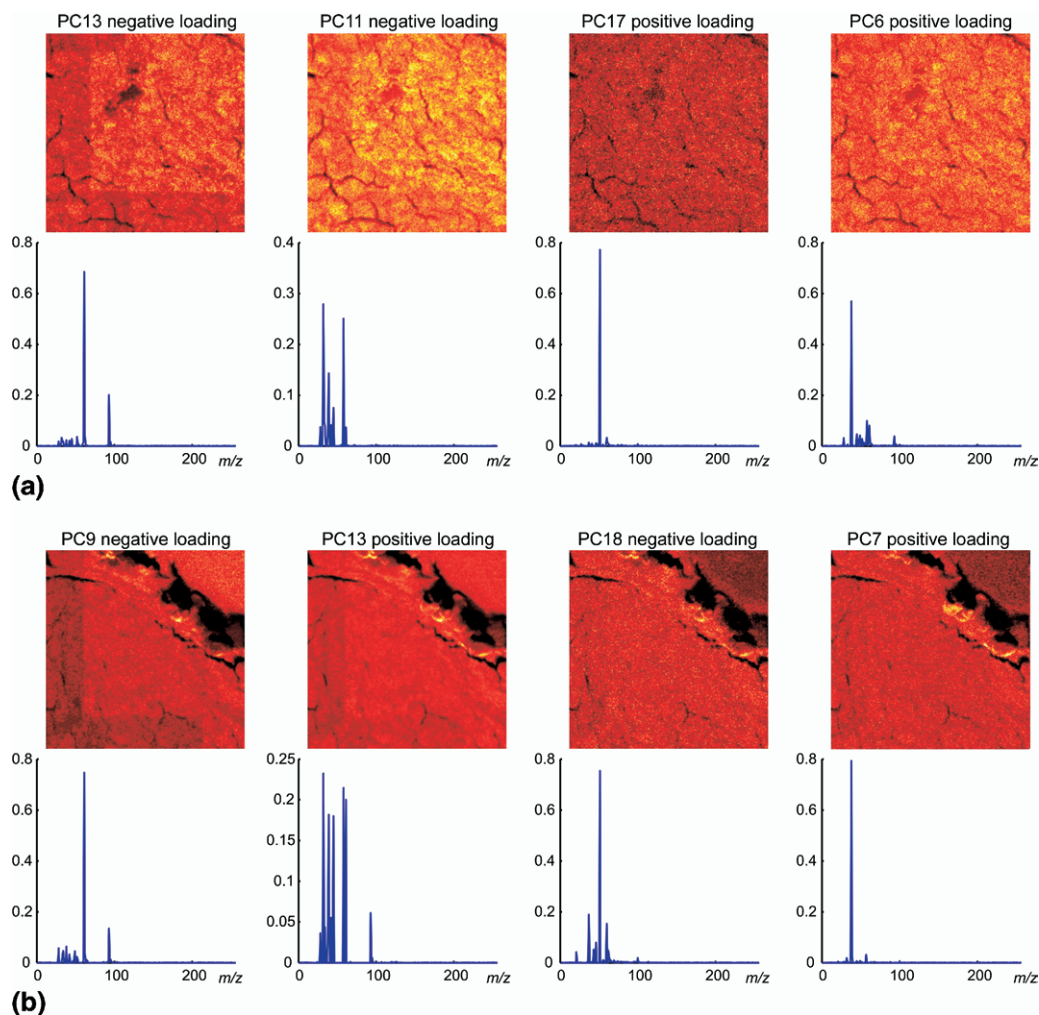

Figure 3. Spatial and spectral representations of chemically matched principal components from the datasets shown in Figure 2. Note the areas of lower intensity visible in some of the images are due to the overlapping areas where the tissue section has been repeatedly analyzed.
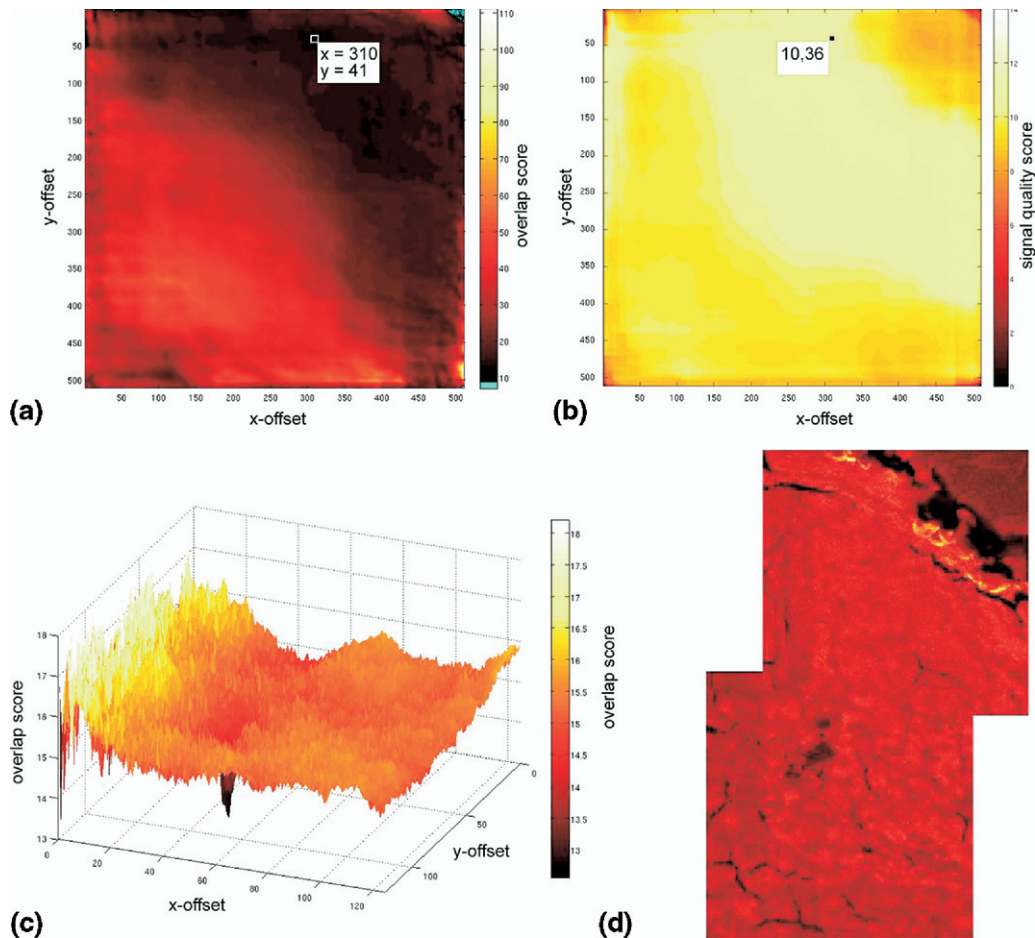
Figure 4. Results of the automatic image metric applied to the components with the best spectral match (positive component 6 of dataset01 and positive component 7 of dataset02). (**a**) The metric landscape, (**b**) the entropy (signal quality) landscape, (**c**) a close up of the minimum in the metric landscape, and (**d**) the mosaic TIC image.

alignment: whether this corresponds to the positive or negative scores is unimportant for this application.

Figure 4a shows the landscape of the fitting metric, determined by calculating the entropy-weighted mean square of the images for all relative positions in which the two images overlap. The lowest point in the surface, corresponding to the optimally aligned position, was found to be $x = 310$ and $y = 41$. Figure 4b shows the entropy landscape for the two overlapping images (a measure of the signal quality for all alignment positions); as can be seen, the optimum position determined using the entropy-weighted fitting metric lies in a region of high signal. These alignment parameters were used as the x and y offsets to combine the *complete*, raw datacubes and thus obtain a high-resolution mosaic imaging datacube of a large area. Figure 4c shows the region containing the lowest point as a 3D contour map and Figure 4d shows a TIC image of the final, mosaic dataset. For clarity all images based on mosaic datasets will be referred to as mosaic images. Clearly, the alignment procedure is able to find the optimum alignment and that this corresponds to the minimum in the overlap metric. Note that the minimum corresponding

to the best fit is quite narrow but sufficiently broad to distinguish it from noisy, local minima. This reflects the fact that a position close to the optimally aligned position will still possess regions with good overlap. A very narrow or single-valued low score of the metric is indicative of noise and can be used as an additional criterion by which the best alignment of the two images can be determined.

The two datasets used to produce Figures 2 to 4 were part of a larger, six-tile dataset. The automation procedure was applied further to align such multi-component images. This was performed by using the above-cited procedure to *align and combine* two neighboring tiles into a single larger datacube. The alignment procedure was then repeated using the larger "working" datacube and a third tile. This process was repeated until the "working" datacube contained the *aligned and combined* datacubes of all six tiles. Note that to ensure that the final combined dataset contains the mass spectra corresponding to the fresh sample surface (the overlapping areas from neighboring tiles are analyzed multiple times) the combined dataset is reconstructed in reverse chronological order and the entries in the over-
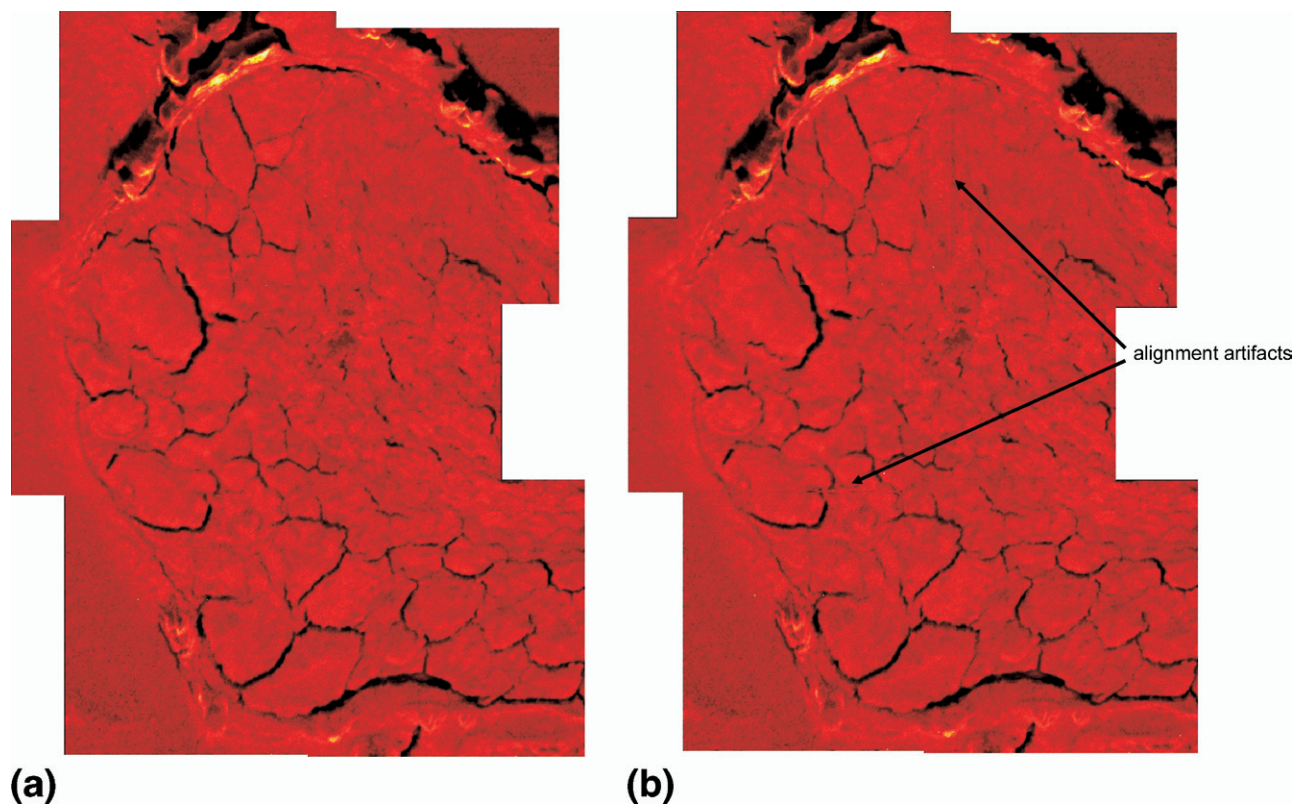
(a)

(b)

alignment artifacts

Figure 5. Comparison of automatically (**a**) and manually aligned (**b**) TIC mosaic images of 6-tile mosaic analysis of a *Lymnaea stagnalis* tissue section. To ensure the results of the automatic alignment routine did not bias the manual alignment, A. Broersen calculated the automatic alignment and L. McDonnell performed the manual alignment blind and independently.

lapping regions are overwritten by the older datasets (recorded first). Consequently, all regions of the final dataset have been analyzed with the same primary ion dose.

This method of aligning multiple tiles has several advantages:

(1) By using the sequentially aligned and combined data from the first two tiles, all subsequent alignment steps include additional overlap information.
(2) The tiles can be ranked according to contrast, allowing those of higher contrast to be aligned first, thus maximizing the overlapping regions for those tiles of lower contrast.
(3) It is rapid because the process is executed in parallel using a "divide and conquer" approach.

Figure 5 compares the mosaic TIC image of the automatically aligned six-tile datacube (left) with that performed manually (right). The manually aligned mosaic image was created by importing the datacubes of the six tiles into Matlab, calculating the TIC images, and aligning them by eye. This ensured that both the automatically and manually aligned mosaic images used the same color panel and intensity scaling. As can be seen the automatically and manually aligned results are nearly identical. The small differences arise from those

areas in which a subjective compromise has been made in the manual alignment, which necessarily involves a subjective ranking of the importance of the features and often leads to alignment artifacts (indicated in figure). The mathematical basis of the alignment in the entropy-weighted mean-squares procedure ranks features according to the quality of the signal and so is less prone to user bias.

We have now used the automatic alignment routine on a number of samples, including tissue sections, droplet arrays, and polymer samples. Provided there are features for the alignment metric to rank fitness, the data tiles could be automatically aligned. Figure 6 shows a larger example—this time an aligned 12-tile mosaic image showing a large-area, high-resolution imaging mass spectrometry analysis of a tissue section of the nervous tissue, specifically the commissure (Cm) and dorsal body (Db), of the fresh water snail *Lymnaea stagnalis*. Figure 6a shows the final, aligned mosaic TIC image. PCA was then applied to the single dataset of this mosaic image. As stated in the introduction, PCA transforms the data onto a new coordinate system to maximize variance. We have used this feature in the automatic alignment procedure to remove the uncorrelated data that would otherwise obstruct the alignment metric. However, it is more common to use the maximized variance identified by PCA to identify, unaided,
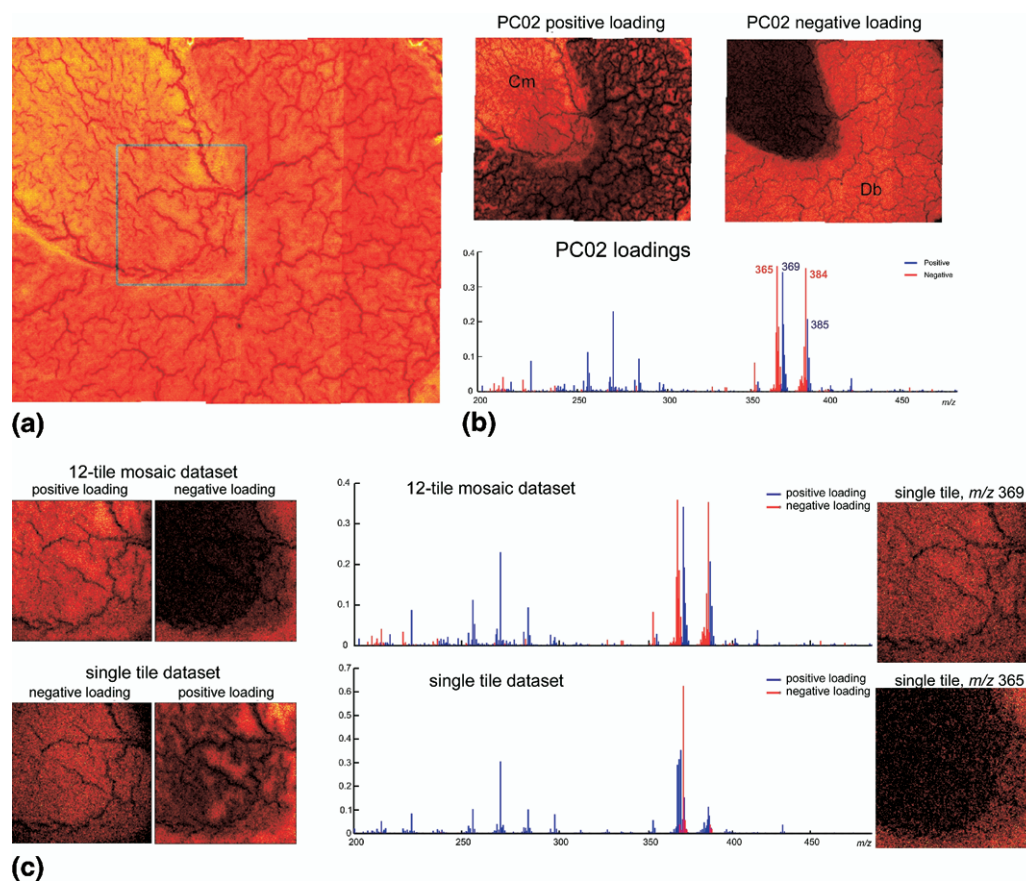
Figure 6. (**a**) Mosaic TIC image of an automatically aligned 12-tile analysis of a section of the nervous tissue of *Lymnaea stagnalis*. PCA of the molecular region (*m/z* 200–1000) of the mosaic dataset. (**b**) The localization of *m/z* 369 and 385 in the commissure (Cm) and *m/z* 365 and 384 in the dorsal body (Dm). (**c**) Compares PCA of the mosaic dataset with that of a single tile and with the images due to *m/z* 369 and 365. Note that the sign of the loadings correlated with the different parts of the tissue section are different for PCA of the mosaic dataset and the single tile.

the regions that generate the most variance in the dataset and are thus the most different. The components can be extracted in a relatively short amount of time with the fast implementation of PCA on a dataset in a sparse format. The application of PCA on the combined dataset takes approximately only 0.001% of the total amount of time needed to create the 12-tile dataset. Most (88%) of this total amount of time is spent on simple preprocessing and data-handling steps instead of calculating the proposed metrics.

Figure 6b shows the mosaic images and spectra corresponding to the positive and negative loadings of the second principal component obtained after performing PCA on the molecular region (*m/z* 200–1000) of the mosaic dataset. The molecular ion region of the mass spectra is more diagnostic than the low-mass region, which contains many fragments whose origins cannot be specified; limiting PCA to the molecular ion region will thus favor specific chemical differences. The mosaic images included in Figure 6b clearly show that PCA has distinguished two regions within the large mosaic dataset that are responsible for a significant fraction of the total variance in the dataset (variance = 5.75% of total). The spectra, shown

below the two mosaic images, clearly indicate that the differences are due to the cluster of molecular ion peaks at *m/z* 369 and 385 (positive loading) and *m/z* 365 and 384 (negative loading). A closer examination of the distributions of these four ions confirms that the distributions of *m/z* 365 and 384 are strongly correlated together but anti-correlated to *m/z* 369 and 385 (data not shown). The *m/z* 369 and 385 peaks are the $[M - OH]^+$ and $[M - H]^+$ ions of cholesterol, and are localized within the commissure of the snail's nervous tissue (in agreement with Altelaar et al. [10]). The peaks at *m/z* 365 and 384 are located in the dorsal body of the tissue section; however, the chemical identity of the peaks has not been determined. We believe this is in part due to the wide diversity in marine natural products and the lack of tandem mass spectrometry capabilities in SIMS instrumentation (see annual reviews by Faulkner et al. in *Natural Product Reports*).

PCA performed on the mosaic dataset allows correlations within the entire dataset to be identified, an impossibility with separated data tiles. Also, by including more spectra in the analysis there should be better separation between different components. Therefore,

we investigated how PCA of the aligned and combined 12-tile mosaic dataset compares with PCA of a single tile, selected to include a fraction of the two anti-correlated distributions. The tile selected for this comparison is indicated in Figure 6a. Figure 6c shows the results of this comparison, including the component images, spectra, and molecular images of *m/z* 369 and 365. It can be seen that PCA of the entire 12-tile mosaic dataset distinguishes the anti-correlated distributions, reflected by the distributions of the ions of *m/z* 369 and 365, more clearly than PCA of the single-tile dataset.

Discussion

The generation of mosaic images of large datasets necessitates stitching together a collection of separate imaging experiments. The automatic alignment routine reported here addresses the problem of the greater uncertainty in the sample stage position (than the spatial resolution of the SIMS measurement). There are other issues associated with mosaic imaging that the analyst should guard against to ensure the validity of mosaic measurements:

(1) *Sample height variation* across a large sample would result in a systematic variation of compound's measured *m/z* across the sample and lead to ionization artifacts. Any such variation (including a localized protrusion) can be readily identified by calculating the variation of an ion's mass with position (a so-called height map) [4, 15]. This height map can be used to correct the mass measurements but cannot remove any ionization artifacts. No such variation was found with the samples reported here.

(2) *Chemical damage of biologic samples*: The dose of the primary ion beam is a critical factor; if it is too high the chemical integrity of the sample is compromised and the SIMS spectra are no longer representative of the sample. A mosaic image requires overlap regions that will receive twofold the normal primary ion dose. If this dose is too high it could cause the overlapping regions to display different spectral signatures and potentially skew the alignment procedure. The ion dose delivered to the overlapping areas was sufficiently low, 7×10^{11} ions cm^{-2}, that the only effects observed were a slightly lower signal intensity in the overlapping areas (for the second analysis). These lower intensity areas were sufficiently different to be distinguished by PCA. The final aligned and combined dataset included the first analysis of all overlapping areas, thus ensuring that the dataset includes the SIMS results corresponding to an equal ion dose for the entire sample.

The automatic alignment routine reported here is suited for highly multidimensional datasets, which are sparse in any single channel and possess a significant degree of uncorrelated noise. This procedure was developed because it was found that existing image alignment routines, such as those used in photography, were unsuitable for the imaging mass spectrometry datasets. The reasons for this include:

(1) There are no sharply defined edges/corners that can be used for "landmarking."

(2) Each datacube can contain thousands of distinct images; photography normally compares a few single images.

(3) Individual spectral images contain less correlation between groups of neighboring pixels due to counting statistics.

(4) No weighting on areas with low signal.

(5) Smaller overlapping regions between the datasets.

PCA is used to reduce the datasets of each tile into a much smaller number of images that contain the spatial variation but not the uncorrelated noise, and the entropy calculation is used to ensure that only those regions with sufficient signal are used to estimate the quality of the alignment. Clearly, the results of the PCA step and the determination of the quality of the signal are essential elements in the success of the automated alignment algorithm. The results obtained with PCA can be highly affected by preprocessing the data, which includes denoising, selection of peaks, and even the choice of scales. For example, in SIMS the signal intensities decrease rapidly with increasing *m/z*; the use of a logarithmic intensity scale can be used to give more weight to the higher mass, but lower intensity, molecular ions. For the automated alignment routine preprocessing was limited to binning. Previous work on SIMS data has demonstrated that binning is "the most effective technique to improve PCA performance" [17], in agreement with our preliminary experimental work (results not shown). Many variants of PCA or multivariate analyses have been reported for SIMS imaging mass spectrometry [11, 13, 18–20], even applied to 3D distributions [21], and many other classification systems have been used in protein mass spectrometry [22]. The auto-alignment procedure benefits from PCA in noise reduction and the availability of more than one-component images for a more robust alignment and not classification. Consequently, we used fast PCA methods rather than more computationally intensive variants for these large datasets (in 32-bit binary sparse format the 12-tile dataset exceeds 1.2 Gb) [23, 24]. Several standard metrics on image texture properties were investigated, including contrast, correlation, homogeneity, energy, and entropy. It was found that the approach of local entropy, used in all the results reported here, was the most effective at removing regions with a high randomness in intensity values.

The results shown earlier clearly show that the entropy-weighted, mean-squares landscape of chemically matched component images can be used to automatically align high-resolution imaging mass spectrom-

etry datasets. This algorithm can be adapted for all datasets of similar nature in imaging mass spectrometry, particularly the mass microscope being developed as part of our high-resolution imaging mass spectrometry research efforts [25, 26].

Conclusion

The reduction of uncorrelated noise provided by principal component analysis allows high-resolution imaging mass spectrometry datasets to be automatically aligned and combined for high-resolution analysis of large areas. The three steps of PCA decomposition, spectral matching, and signal quality assurance are necessary because of the high dimensionality and sparsity of the SIMS imaging mass spectrometry data and indicate future methods of how to work with such data.

Acknowledgments

References

1. McDonnell, L. A.; Heeren, R. M. A. Imaging Mass Spectrometry. *Mass Spectrom. Rev.* **2007**, *26*, 606–643.
2. Altelaar, A. F. M.; Klinkert, I.; Jalink, K.; deLange, R. P. J.; Adan, R. A. H.; Heeren, R. M. A.; Piersma, S. R. Gold-enhanced Biomolecular Surface Imaging of Cells and Tissue by SIMS and MALDI Mass Spectrometry. *Anal. Chem.* **2006**, *78*, 734–742.
3. McDonnell, L. A.; Heeren, R. M. A.; de Lange, R. P. J.; Fletcher, I. W. Higher Sensitivity Secondary Ion Mass Spectrometry of Biological Molecules for High Resolution, Chemically Specific Imaging. *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 1195–1202.
4. McDonnell, L. A.; Piersma, S. R.; Altelaar, A. F. M.; Mize, T. H.; Luxembourg, S. L.; Verhaert, P. D. E. M.; van Minnen, J.; Heeren, R. M. A. Subcellular Imaging Mass Spectrometry of Brain Tissue. *J. Mass Spectrom.* **2005**, *40*, 160–168.
5. Touboul, D.; Halgand, F.; Brunelle, A.; Kersting, R.; Tallarek, E.; Hagenhoff, B.; Laprévote, O. Tissue Molecular Ion Imaging by Gold Cluster Ion Bombardment. *Anal. Chem.* **2004**, *76*, 1550–1559.
6. Touboul, D.; Kollmer, F.; Niehuis, E.; Brunelle, A.; Laprévote, O. Improvement of Biological Time-of-Flight Secondary Ion Mass Spectrometry Imaging with Bismuth Cluster Ion Source. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 1608–1618.
7. Breitenstein, D.; Rommel, C. E.; Möllers, R.; Wegener, J.; Hagenhoff, B. The Chemical Composition of Animal Cells and Their Intracellular Compartments Reconstructed from 3D Mass Spectrometry. *Angew. Chem. Int. Ed.* **2007**, *46*, 5332–5335.
8. Debois, D.; Brunelle, A.; Laprévote, O. Attempts for Molecular Depth Profiling Directly on a Rat Brain Tissue Section Using Fullerene and Bismuth Cluster Ion Beams. *Int. J. Mass Spectrom.* **2007**, *260*, 115–120.
9. Fletcher, J. S.; Lockyer, N. P.; Vaidyanathan, S.; Vickerman, J. C. TOF-SIMS 3D Biomolecular Imaging of *Xenopus laevis* Oocytes Using Buckminsterfullerene (C_{60}) Primary Ions. *Anal. Chem.* **2007**, *79*, 2199–2206.
10. Altelaar, A. F. M.; Minnen, J. V.; Jiménez, C. R.; Heeren, R. M. A.; Piersma, S. R. Direct Molecular Imaging of *Lymnaea stagnalis* Nervous Tissue at Subcellular Spatial Resolution by Mass Spectrometry. *Anal. Chem.* **2005**, *77*, 735–741.
11. Keenan, M. R.; Kotula, P. G. Accounting for Poisson Noise in the Multivariate Analysis of ToF-SIMS Spectrum Images. *Surf. Interface Anal.* **2004**, *36*, 203–212.
12. Meglen, R. M. Examining Large Databases: A Chemometric Approach Using Principal Component Analysis. *Mar. Chem.* **1992**, *39*, 217–237.
13. Willse, A.; Tyler, B. Poisson and Multinomial Mixture Models for Multivariate SIMS Image Segmentation. *Anal. Chem.* **2002**, *74*, 6314–6122.
14. Vickerman, J. C.; Briggs, D., Eds. *ToF-SIMS: Surface Analysis by Mass Spectrometry*; IM Publications: Chichester, UK, 2001.
15. McDonnell, L. A.; Luxembourg, S. L.; Mize, T. H.; Koster, S.; Eijkel, G. B.; Verpoorte, E.; de Rooij, N. F.; Heeren, R. M. A. Using Matrix Peaks To Map Topography: Increased Mass Resolution and Enhanced Sensitivity in Chemical Imaging. *Anal. Chem.* **2003**, *75*, 4373–4381.
16. Broersen A.; van Liere, R. Feature Based Registration of Hyperspectral Data-Cubes. *Proceedings of IASTED International Conference on Visualization, Imaging, & Image Processing*, Palma de Mallorca, 2006; pp 543–548.
17. Wickes, B. T.; Kim, Y.; Castner, D. G. Denoising and Multivariate Analysis of Time-of-Flight SIMS Images. *Surf. Interface Anal.* **2003**, *35*, 640–648.
18. Klerk, L. A.; Broersen, A.; Fletcher, I. W.; Liere, R. V.; 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*, 222–236.
19. Tyler, B. J. Multivariate Statistical Image Processing for Molecular Specific Imaging in Organic and Biosystems. *Appl. Surf. Sci.* **2006**, *252*, 6875–6882.
20. Wagner, M. S.; Graham, D. J.; Ratner, B. D.; Castner, D. G. Maximizing Information Obtained from Secondary Ion Mass Spectra of Organic Thin Films Using Multivariate Analysis. *Surf. Sci.* **2004**, *570*, 78–97.
21. Smentkowski, V. S.; Ostrowski, S. G.; Braunstein, E.; Keenan, M. R.; Ohlhausen, J. A. T.; Kotula, P. G. Multivariate Statistical Analysis of Three-Spatial-Dimension TOF-SIMS Raw Data Sets. *Anal. Chem.* **2007**, *79*, 7719–7726.
22. Hilario, M.; Kalousis, A.; Pellegrini, C.; Müller, M. Processing and Classification of Protein Mass Spectra. *Mass Spectrom. Rev.* **2006**, *25*, 409–449.
23. Lehoucq, R. B.; Sorensen, D. C.; Yang, C. ARPACK Users's Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods; SIAM Publications: Philadelphia, 1998.
24. Sorensen, D. C. Implicit Application of Polynomial Filters in a k-Step Arnoldi Method. *SIAM J. Matrix Anal. Appl.* **1992**, *13*, 357–385.
25. Altelaar, A. F. M.; Luxembourg, S. L.; McDonnell, L. A.; Piersma, S. R.; Heeren, R. M. A. Imaging Mass Spectrometry at Cellular Length Scales. *Nat. Protocols* **2007**, *2*, 1185–1196.
26. Luxembourg, S. L.; Mize, T. H.; McDonnell, L. A.; Heeren, R. M. A. High-Spatial Resolution Mass Spectrometric Imaging of Peptide and Protein Distributions on a Surface. *Anal. Chem.* **2004**, *76*, 5339–5344.