Tools and strategies for visualization of large image data sets in high-resolution imaging mass spectrometry

Ivo Klinkert, Liam A. McDonnell, Stefan L. Luxembourg,
A. F. Maarten Altelaar, and Erika R. Amstalden
FOM Institute for Atomic and Molecular Physics FOM-AMOLF, Kruislaan 407, 1098 SJ Amsterdam, The Netherlands

Sander R. Piersma
Oncoproteomics Laboratory, Free University Medical Centre, De Boelelaan 1117, 1081 HV Amsterdam, The Netherlands

Ron M. A. Heeren[a)]
FOM Institute for Atomic and Molecular Physics FOM-AMOLF, Kruislaan 407, 1098 SJ Amsterdam, The Netherlands

(Received 27 March 2007; accepted 16 April 2007; published online 31 May 2007)

Mass spectrometry based proteomics is one of the scientific domains in which experiments produce a large amount of data that need special environments to interpret the results. Without the use of suitable tools and strategies, the transformation of the large data sets into information is not easily achievable. Therefore, in the context of the virtual laboratory of enhanced science, software tools are developed to handle mass spectrometry data sets. Using different data processing strategies for visualization, it enables fast mass spectrometric imaging of large surfaces at high-spatial resolution and thus aids in the understanding of various diseases and disorders. This article describes how to optimize the handling and processing of the data sets, including the selection of the most optimal data formats and the use of parallel processing. It also describes the tools and solutions and their application in mass spectrometric imaging strategies, including new measurement principles, image enhancement, and image artifact suppression. © *2007 American Institute of Physics.*
[DOI: 10.1063/1.2737770]

I. INTRODUCTION

Enhanced science (e-science) is a rapidly developing discipline in information technology. e-science in this context means science in global collaboration enhanced by advanced computing and communication technologies[1] by sharing and reusing existing resources. These resources can vary from scientific analytical instruments, data storage resources, information resources such as databases or knowledge bases, or high-end computational infrastructure. The Dutch virtual laboratory for e-sciences VL-e[18] is an environment where these resources are brought together for a wide variety of application domains. One of the aims of the VL-e is to provide generic functionalities that support a wide class of specific e-science application environments.[1] As such, VL-e provides a computer environment that makes it possible to run specially developed generic software modules on a computer grid. In addition, the VL-e environment can be used to facilitate new interdisciplinary scientific collaborations that allow the joint usage of infrastructure, data, and information.

One particular area of research that is experiencing a data explosion is mass spectrometry based proteomics. With the continuing completion of genomes from different species, many biological and biomedical studies have focused on the analysis of the building blocks of life, the proteins. The determination of protein sequence combined with a variety of post-translational modifications, molecular conformations, physiological properties, interacting ligands, and different spatial organizations is one of the most challenging tasks for the bioanalytical sciences. The field is often referred to as "the science of proteomics."[2] This complex problem requires the use of many different analytical strategies providing different pieces of information. Mass spectrometry (MS) is currently the method of choice for the systematic analysis of a proteome.[2] Moreover, mass spectrometry based proteomics is now one of the key players in systems biology, i.e., the integrated approach of different technical disciplines to study the physiological processes in a cell or tissue.[3] Fast instrumental developments have made MS accessible to a broader research community and enabled automatic data acquisition and high-throughput analysis. In clinical research mass spectrometry has opened new ways of (early) detection of diagnostic molecules for specific diseases and disorders. Their identification is based on differential analysis of protein expression patterns in patient and control samples. A detailed proteomics study requires new data analysis strategies where workflow-based protocols and novel bioinformatic strategies are paramount. VL-e has already created a

[a)]Author to whom correspondence should be addressed; electronic mail: heeren@amolf.nl

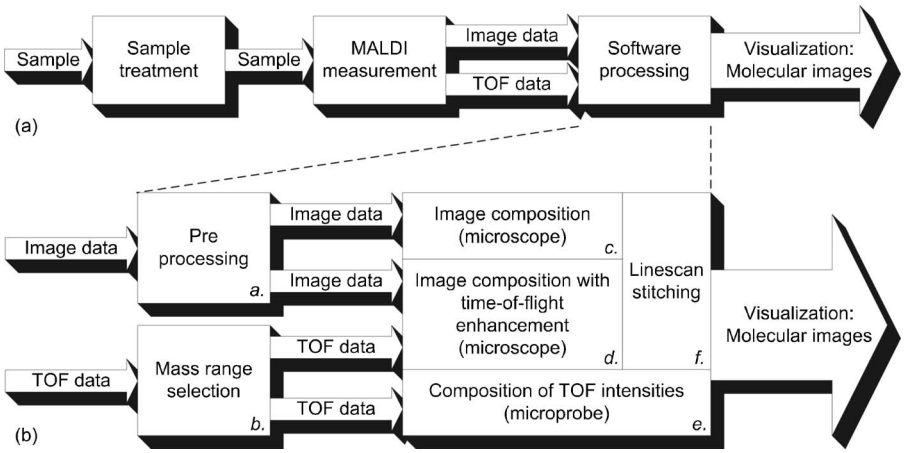


FIG. 1. (a) Workflow of experiment and (b) detailed workflow of the software processing.

problem solving environment in which parallel processing strategies in a distributed environment are used to process large scale proteomics data.[4]

Molecular imaging of biological surfaces with mass spectrometry provides both the chemical identity and the spatial distribution of each component on the surface and is revolutionizing the field of biological surface analysis. MS imaging, in particular, brings *chemical specificity* to this field. In other molecular imaging techniques, such as immunohistochemistry or immunoelectron microscopy, molecule-specific labeling is required. Imaging mass spectrometry can spatially map surface components without preselection (or labeling) using an intrinsic molecular property, the molecular mass. In proteome imaging it is increasingly important to be able to visualize protein expression directly on cellular surfaces or tissue sections. Recently, a high-resolution MS based rapid imaging technique was developed, the mass microscope.[5] This approach provides unmatched spatial detail combined with detailed molecular information directly from tissue. The sheer amount of data produced by this method requires new instruments and tools to extract and visualize relevant information.

This article presents a number of new tools and data processing strategies that have been developed to enable extraction and visualization of information from the data generated by the mass microscope. Based on their functionality these tools are comparable to software suites such as BIOMAP (Novartis, Basel, Switzerland). The novelty to our approach is found in how the tools deal with large data sets and their use in a problem solving environment created in VL-e.[1] The tools and strategies presented in this article are together called the spatial image composer (SIC) software package.

Three major challenges had to be addressed during the development. First the large amount of data (tens of gigabytes), so parallel processing, data compression, and automatic configuration tools were incorporated. Second the implicit nonreproducibility of the unprocessed imaging results delivered by the MALDI microscope experiment renders automatic stitching impossible. The third challenge is the requirement to deal with complementary data modalities such as high-spatial resolution data and high mass resolution data. Although the processing strategies are generally applicable, we tested the software using imaging mass spectrometry data. The quest for chemical information in these experiments is, for example, comparable with hyperspectral imaging in remote sensing[12] and geographic information science.[13]

Many successful experiments were performed which resulted in publications[6–11] which would not have been possible without this way of developing the spatial image composer. The tools and strategies in this article constitute an approach to imaging mass spectrometry that is shown to allow the direct visualization of the distribution of endogenous neuropeptides directly from rat brain tissue sections.

II. EXPERIMENT

A. MALDI microscope experiment

The MALDI microscope experiment is a unique approach to imaging mass spectrometry.[5] The experiment workflow [Fig. 1(a)] consists of three stages: sample preparation, measurement, and processing of the measurement data. During MALDI sample preparation, analytes are dissolved in excess matrix, containing a chromophore that absorbs most of the laser radiation directed at the surface. This results in ionization of a larger part of the molecules without fragmentation.[14,15] Subsequently the measurement data are processed by the spatial image composer (SIC) software to visualize the results. This procedure is described in Sec. III.

The imaging MS experiments are performed on an extensively modified TRIFT-II ion-microscope/time-of-flight mass spectrometer equipped with a frequency-tripled Nd:YAG (yttrium aluminum garnet) laser (Wedge 355-100, Bright Solutions, Italy) and a phosphor screen/charge coupled device (CCD) camera combination, as described in detail by Luxembourg.[16] The ion optics of this mass microscope have been designed such that a magnified (ion-optical) image of the surface is mapped onto a two-dimensional detector with high-spatial resolution.[16]

A single partial surface image is recorded each time the laser is fired onto the surface of the tissue sample. By repetitive firing of the laser while continuously moving a tissue sample, a series of images is recorded. Despite the difference in scale and resolution, this measurement principle has many similarities with earth remote sensing, where the earth surface is scanned image by image. Imaging the tissue sample is

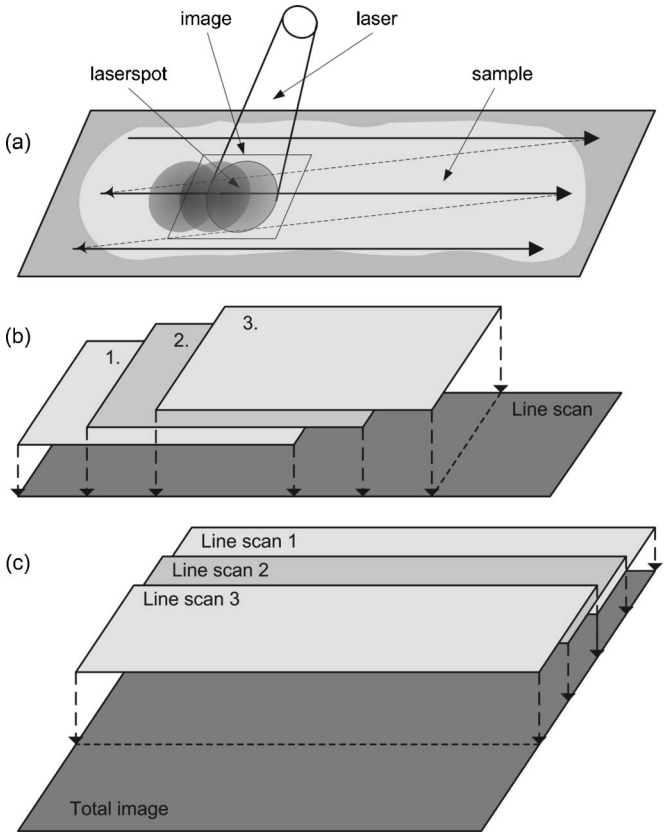


FIG. 2. Measurement principles: (a) measuring parallel line scans, (b) concatenating images within a line scan into an image strip, and (c) concatenating image strips together into the resulting total image. First creating the image strips enables the possibility of using parallel processing of the data.

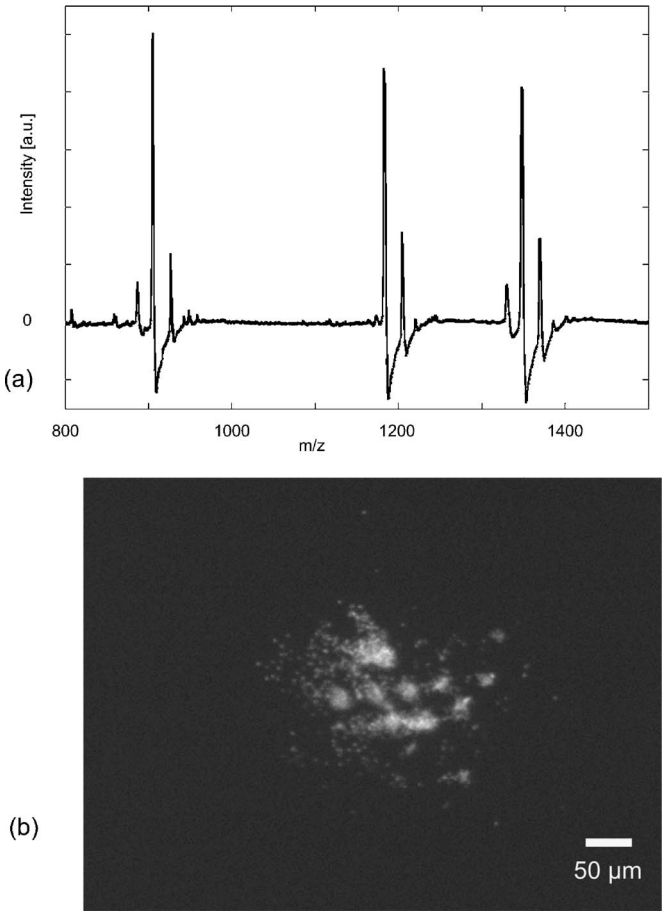


FIG. 3. Single shot measurement data. (a) A part of the spectrum of a single shot measurement, converted from time-of-flight to m/z values, which is shown in the horizontal axis. The vertical axis shows the number of ions as arbitrary units. The peaks show from left to right the peptides des-Arg bradykinin, angiotensin I (1–9), and substance P. (b) A single shot image taken with the CCD camera. The image has the size of 640×480 pixels, which is equivalent to 640×480 μm. The laser spot size used for this image is approximately 200 μm.

performed by moving the sample stage over a predefined two-dimensional (2D) region, where the sample stage can move in two directions perpendicular to each other. Data are recorded from multiple bands parallel to each other covering the region of interest [Fig. 2(a)]. During the measurement of a line scan, the sample stage is moved in one direction at a speed of typically 100 μm/s. Typically the laser spot size is 200 μm in diameter and the laser firing repetition rate is 10 Hz, giving an overlapping area between successive laser shots of 95%. This large overlapping area is necessary to accumulate a sufficient amount of data out of the single nonreproducible partial surface images. The displacement between two line scans is typically 100 μm and the overlap of the successive line scans is at most 50%. The basic purpose of the SIC software is to stitch these overlapping series of images together to visualize the full sample area to create a resulting mosaic image. Typically 1500 laser shots are used for a line scan, and it takes 50 of these line scans to cover a region of 16×8 mm^2. A complete measurement (without data processing) on the TRIFT takes a few hours, depending on the size of the sample and the speed and movements of the sample stage.

Data are recorded using a multichannel plate/phosphor screen (MCP-PS) detector. The MCP-PS is a position-sensitive detector, in which light flashes from the phosphor screen show at which position the ions arrive at the detector. A CCD camera (Imager 3 with corresponding DAVIS Version 6.2.3 software, La Vision, Goettingen, Germany) is used to record the distribution of the ions on the detector, which is the magnified image of the sample, recorded using a constant exposure time. Owing to the low time resolution of this detection assembly, only one image can be recorded with each laser shot. A total ion image is recorded when all ions are allowed to reach the detector, an example is shown in Fig. 3(b). By using fast ion optics to select a single m/z (or small m/z range), a microscope image of a specific species can be recorded. This principle is also known as blanking.[16]

In order to generate a mosaic image composition correctly, the accurate positions of all individual shots are essential. The positions of all shots are measured relative to one another by storing scan directions and distances. Constant movement of the sample together with the constant laser repetition rate allows the use of one single, constant value for the displacement between single images in a line scan. In the user interface of the SIC software, all scan direction and distance settings can be specified, after which the software takes the correct sequence of the data to create the resulting mosaic image.

The time-of-flight (TOF) of the ions is measured using a fast digitizer (DP240, Acqiris, Geneva, Switzerland) to ac-

quire the ion signals originating from the multichannel plate. These single shot mass spectrometric data are recorded with the corresponding software (ACQIRISLIVE Version 2.11) as the TOF information. After acquisition the TOF data are converted into calibrated mass values (m/z) later by the SIC software.

B. Data set characteristics

For each individual laser shot, two types of data are measured simultaneously, synchronized by a master trigger originating from the TRIFT.

- *Mass spectrometric (time-of-flight) data*. One TOF-MS spectrum for each laser shot [*TOF data* in Fig. 1(b)]. The TOF information is stored as binary values for each shot in a separate file. Values are stored as signed bytes and represent the ion intensity (ion current). The data are stored in the proprietary Acqiris file format. A graphic representation of these data is shown in Fig. 3(a). Typically the sample frequency of the time-of-flight data is 0.5 GHz and 100 000 samples per shot are stored. Using different settings influences accuracy, mass range, and file size. Higher sample rates (2 GHz as maximum) can be used to increase the acquisition accuracy. Acquiring more samples increases the file size linearly, so an optimum between accuracy, mass range, and file size is determined for each kind of experiment.
- *An ion-optical image*. A set of (total ion) images, one image for each laser shot [*image data* in Fig. 1(b)]. Images are stored by the CCD camera software in tagged image file format (TIFF) as a 640×480 (length $\times$ width) pixel image. The images are stored in grayscale, uncompressed (each image is 600 kbytes), and the image pixel intensities (brightness) are values between 0 and 255. An example is shown in Fig. 3(b).

Both types of data together comprise the data set. The time-of-flight data are called *microprobe*, because in the recording of these data the MCP detector effectively is a point detector. The image data are called *microscope*, because the recording is performed with the MCP-PS detector operating in 2D mode.

C. Software development

The software development process of the spatial image composer was not performed using the traditional waterfall software development model but a more interactive one. In the conventional waterfall model all requirements are formulated in advance, before the software is built and tested. In research often specifications are born out of new ideas during the experiments, so a step-by-step approach was chosen. This has the advantage of getting the first results earlier and gives the opportunity to implement new functionality based on the intermediate results. This method is only applicable when no strict time and budget constraints are applied to the software development. However, the ratio of effort to quality is still very favorable, because the functionality of the software meets the needs of investigators very closely, despite the evolving requirements during the software development process.

The software is written in the JAVA programming language, Version 1.4. The ECLIPSE JAVA programming environment is used for the SIC development. The software is implemented in modules (also called components) that allow sharing of different parts of the software. This workflow-based module architecture also allows the easy addition of new functionality by adding new modules. It also makes an architectural change to a web service architecture possible, an architecture which becomes more standard in the grid community.[17] The nowadays widely used extensible markup language (XML) is used for storing all configuration settings and metadata and also for intermodule communication. This increases the possibilities to extend and connect to other, including non-JAVA, software.

The software is now available both as a single software application on standard WINDOWS desktop computers and as modules for the VL-e problem solving environment. This VL-e environment helps software packages such as the spatial image composer by providing shared computer resources to speed up processing. By using JAVA as the programming language, it is reasonably easy to run the software on different platforms, for example, the use of "Swing" user interface components, in combination with JAVA's "layout managers," causes no extra user interface programming for cross platform execution of the software. By merely changing the startup scripts, the application runs on WINDOWS (2000/XP), LINUX (Red Hat), and UNIX (FreeBSD/Mac OS X). It was successfully tested by performing one part of the calculations on a WINDOWS machine, and the other part on a LINUX machine. Running the application on a standard computer requires the Java runtime environment (JRE), which is available as a free download.

Two characteristics of the data set render the use of the off-the-shelf image/photostitching software impossible. First, the large amount of data require parallel processing, data compression, and automatic configuration tools. These features are not available in the off-the-shelf software. Secondly, the implicit nonreproducibility of the unprocessed imaging results delivered by the MALDI microscope experiment. The absence of clear features in the large amount of unprocessed imaging results makes automatic stitching impossible. The imaging strategies of the new tools use position information from the instruments to stitch the images together.

III. DATA PROCESSING

Figure 1(a) schematically depicts the data processing workflow [Fig. 1(b)] as a detail of the overall workflow. The different parts of the data processing workflow are explained in the following sections.

The SIC software uses the following data fusion strategies with the two mutually synchronized image and time-of-flight data set.

- *Image composition*. Stitching the set of single partial surface images to create a high-resolution microscope

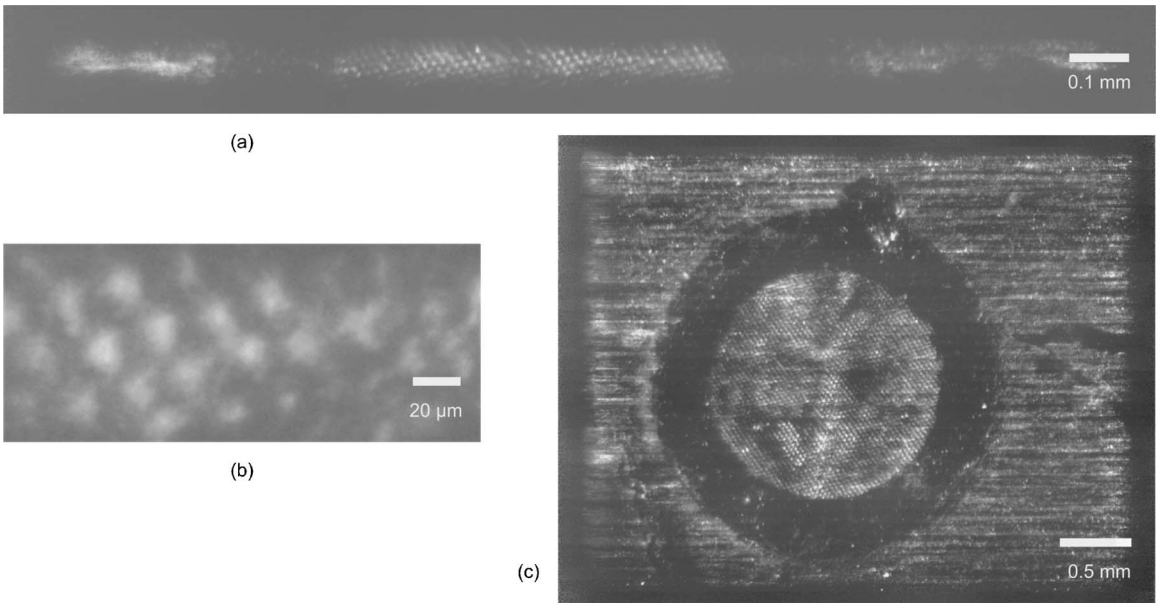


FIG. 4. Three segments of the same data set resulting in high-spatial resolution image of a sample which consists of PEG-1500 doped matrix crystals overlain with a transmission electron microscopy (TEM) grid. (a) An image strip generated out of one line scan which is one image strip (left to right) out of the resulting mosaic image. (b) A detailed look at the line scan image generated out of the image line scan data. The TEM grid shown in this image (as the dark grid) has a spacing of 25 μm. (c) The resulting mosaic image where the dark circle marks the outside of the TEM grid.

mosaic resulting image of the scanned surface [parts c and f in Fig. 1(b)].

- *Image composition enhanced with time-of-flight data*. Stitching the set of images and using the ion intensity for each laser shot for a selected mass range for image enhancement. In this way mass specific information of the time-of-flight data is used to perform an image transformation to suppress nonrelevant data.
- *Time-of-flight mass slice composition*. Using intensities in a small selected mass windows as pixels to create a microprobe image [part e in Fig. 1(b)]. This is performed in two ways, as is will be shown later. In this way all mass specific information of the experiment is visualized at lower resolution.

A. Image composition

The resulting mosaic image, which visualizes the entire measured area, is created by combining overlapping stigmatic ion images. If no blanking is used during the measurements, the images contain intensities from all masses, and a "total ion image" will be created. When blanking of ions is employed in the measurement, the images contain only intensities for a specific mass window that was selected before the measurement.

The SIC software processes the image data set in a line scan oriented way, similar to the manner the measurement is carried out. At first all images from each single line scan are combined, resulting in an image strip for each line scan measurement [c and d in Fig. 1(b)], which is performed using the following steps for each image strip.

(1) Create an image containing no data, having the width and height of the final image strip. The dimensions of this empty image strip are dependent on the displacement between the individual images and the scan direction. The dimension of the individual images (640 $\times$480 pixels) determines the fixed height or the width of this empty image strip.

- Line scans measured horizontally have a fixed height of 480 pixels and a width of $640+(n-1)\times$single displacement pixels, where n is the number of images in a line scan.
- Line scans measured vertically have a fixed width of 640 pixels and a height of $480+(n-1)\times$single displacement pixels, where n is the number of images in a line scan.

(2) In the empty image strip the image data of all line scan images are included. Combining images is implemented by taking the average pixel value from all contributing images which are relevant for a pixel in the resulting image. A contributing image is relevant for a pixel in the resulting image when that contributing image overlaps at the point of that pixel,

$$C(x) = \sum_k w_k(x) I_k(x) \Big/ \sum_k w_k(x), \tag{1}$$

where $I_k(x)$ are the images and $w_k(x)$ is 1 where an image overlaps at that point and 0 if not. Figure 2(b) represents a graphical illustration of this principle. All images have the same weight for the average value. The intermediate results of the image strips are stored as 32 bit float values (IEEE 754 floating point) to guarantee that no loss of precision is introduced.

(3) The resulting image strip data are also stored as an image itself. The intermediate results are stored to prevent loss of data. An image strip is shown in Fig. 4(a). Figure 4(b) shows a part of this image strip in more detail.

These images themselves are not used for further image composition. After generation of several resulting image strips, they are combined into a complete mosaic image [f in Fig. 1(b)], which is performed using the following steps.

(4) Create an image containing no data, having dimensions of the complete mosaic image. The dimensions are determined by calculating a square in which the overlapping image strips exactly fit, dependent on the x and y displacements between the image strips.

(5) The data of all intermediate image strips are combined in the empty mosaic image. Combination is performed using the same averaging algorithm (1) used for the generation of the image strips. Figure 2(c) shows a graphical representation of this principle. In Sec. IV, several improvements to this basic algorithm are described that reduce the appearance of seam artifacts in the mosaic image.

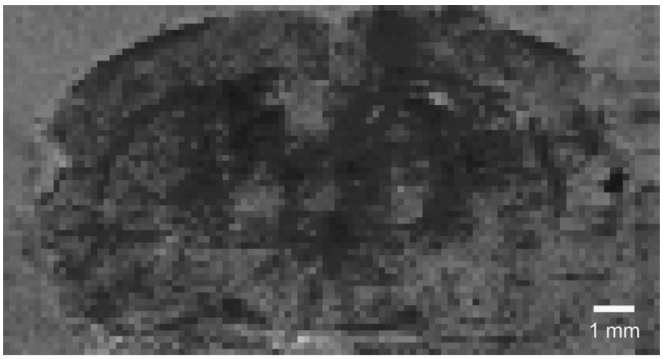
B. Image composition enhanced with time-of-flight data

The difference between the *image data* and the *time-of-flight data* is the fact that either spatial resolution is available with limited mass resolution (image data) or mass resolution is available with limited spatial resolution (time-of-flight data). To enhance the mass information for a mass range of interest while maintaining the high-spatial resolution of the mosaic image, the set of single partial images is combined as described in Sec. III A. At the same time, the ion intensity for each laser shot for a certain mass range is used to enhance the mosaic image by suppressing pixels from nonrelevant images. Which mass ranges are relevant for producing resulting images by the software is determined by the specific research problem under investigation. Care must be taken when interpreting these results because high intensities in "nonrelevant" masses can still dominate the image at certain locations.

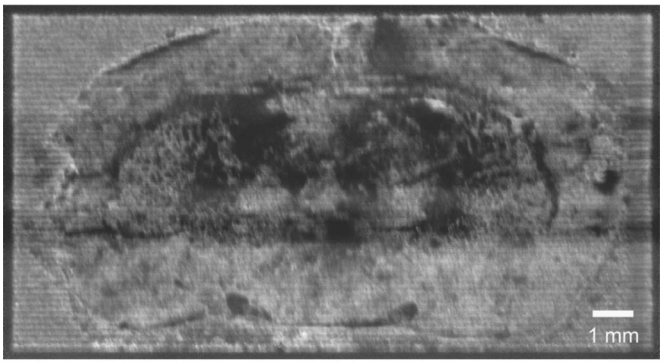
This image enhancement strategy uses an intensity data file that is generated out of the time-of-flight data [part b in Fig. 1(b)]. This data file contains the summed (binned) ion intensity values for the specified mass range of interest, one value for each laser shot, resulting in a relative distribution of ion intensities across the sample. These data are stored in a binary file containing the summed values as 32 bit float values stored sequentially in the file. These time-of-flight data are used as a relative intensity multiplication factor on the corresponding image data, where each pixel of the single 640×480 image data is multiplied with the corresponding mass window intensity before stitching the image into the image strip. For algorithm (1) this means that $w_k(x)$ turns into the relative multiplication factor.

C. Time-of-flight mass slice composition

This third data fusion option uses the time-of-flight (microprobe) data exclusively. This data type is characterized by a lower spatial resolution but a high mass resolution. This high mass resolution is employed to visualize different component (m/z) distributions. The method to create these mass slice images uses the intensity file which is generated for the "image composition enhanced with time-of-flight data."



(a)



(b)

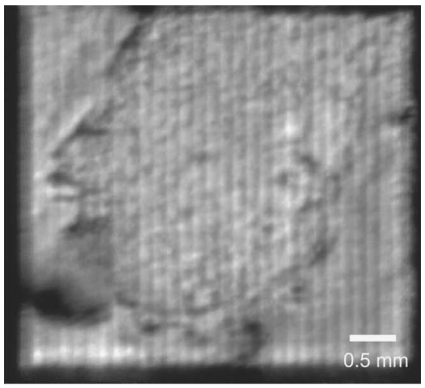
FIG. 5. Imaging results. (a) The sample of (b) in microprobe resolution, composed out of the Acqiris time-of-flight data summed from 500 to 3100 m/z, spatially binned using bilinear resampling. (b) An example of a resulting microscope total ion image of the spatial image composer software, showing a tissue section of a rat brain, covered with a MALDI matrix of α-cyano-4-hydroxycinnamic acid. The image consists of 148 Mpixels ($16\,740 \times 8890$ pixels), which is 82 Mbytes when compressed to the PNG file format. The image map algorithm is used to remove seam artifacts from the image. This result has been published previously in Ref. 9 but without the seam artifacts in the image removed by the distance map algorithm.

The first method uses the intensity file to directly create an image. The intensity file contains the intensity values for a selected mass range for all shots, which are used as pixel values for the mass resolved image. The image does not have the correct width to height ratio because the movements between the successive shots in the width and height direction are not identical. This is currently corrected using commercial image processing software [Paint Shop Pro, Jasc Software, see Fig. 5(a)] by resampling or binning the data (in the direction of the overlapping images only) to obtain the original width to height ratio.

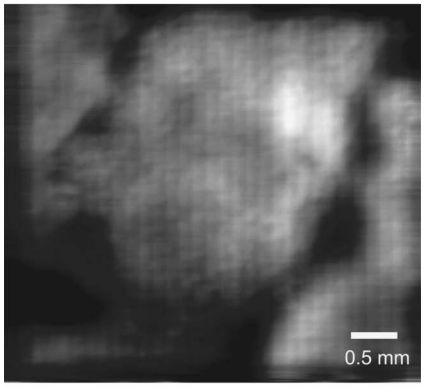
The second method employs the values in the intensity file to create virtual CCD images. These images are 640×480 pixels wide, but all image pixel intensity values for one virtual image get the same value: the value from the intensity file for this image. To create the mass slice image [see Fig. 6(c)], further processing is the same as the "image composition," differing only in the fact that the virtual images are used instead of real images.

D. Scaling

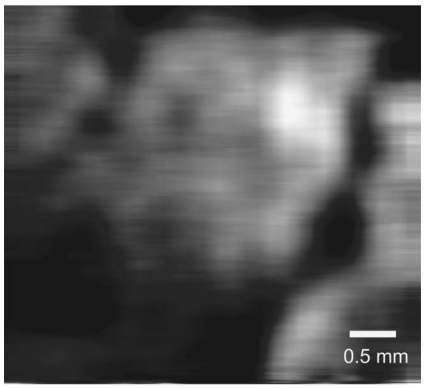
A conversion from micrometers to pixels ("spatial scaling") is performed before combining all images. All dis-



(a)



(b)



(c)

FIG. 6. Results of different composition techniques showing a human cervical tissue section previously treated with trypsin and incubated overnight at 37 °C in a humid environment and covered with α-cyano-4-hydroxycinnamic acid matrix and 5 μm of gold. All three images are 4690×4270 pixels in size corresponding to 4.7×4.3 mm^2 (1 pixel equals 1 μm). (a) Resulting mosaic total ion image from "image composition." (b) Result of image composition enhanced by a transformation with time-of-flight ion intensity data from 1338.0 to 1339.0 m/z. (c) Result of time-of-flight mass slice composition from the same 1338.0 to 1339.0 m/z mass range. Both images [(a) and (b)] show the same high-spatial resolution; however, in (b) the image assists in the localization of a possible identification of a human serum albumin. Image (c) shows the same localization of (b), solely using the microprobe data. Image (a) fully shows the result in the microscope mode resolution ($\geqslant$4 μm), while image (c) in the microprobe mode resolution ($\geqslant$200 μm).

placements in the experiment are given in micrometers, but all image dimensions and displacements in the software are specified by means of pixels (the software internally handles the images as rows of pixels). The conversion from micrometers to pixels is performed by a reference measurement on a known-size sample structure containing, for example, a grid, and determining the micrometer to pixel conversion factor. This conversion factor also takes the magnification of the ion optics and the CCD camera lens into account.

No scaling of the intensity (brightness) values ("intensity scaling") of the image strips [except for the intermediate image strip images as shown in Fig. 4(a)] is used during the processing of data into the resulting image. After processing the intensity values are scaled up or down to a range from 0 to 255 using linear scaling, a linear contrast stretch. In this way the resulting image always has the maximum contrast.

IV. RESULTS AND DISCUSSION

A. Visualization results

The spatial image composer software generates high-resolution microscope images. Figure 4(b) shows a detailed part of the SIC processed image of a known-size grid (grid spacing 25 μm). The resolution for single shot images of this grid was determined to be 4 μm in Refs. 5 and 16. A single image strip was already shown in Fig. 4(a), and the resulting mosaic image is shown in Fig. 4(c).

Figure 5(b) shows a resulting high-resolution image consisting of 16 740$\times$8890 pixels and clearly indicates the high number of pixels (148 Mpixels). The scanned area is 16 $\times$8 mm^2 (1 mm$\sim$1 μm), and details at the cellular level (10–20 μm) are visible.

Figures 6(a)–6(c) demonstrate the results of the different composition techniques of a medium-size high-resolution resulting mosaic image (4690×4270 pixels). The three different composition techniques (mosaic, enhancement with time-of-flight data, and time-of-flight mass slice composition) all have their own specific purpose. The image composition, as shown in Fig. 6(a), visualizes in high-spatial resolution and allows subcellular details to be distinguished. This composition is based on the total ion images, indicating that when only small mass ranges must be visualized, blanking of the ions in the measurement is necessary.[16] Figure 6(b) shows the result of the image composition enhanced with time-of-flight data. In this image the 1338–1339 m/z range is chosen to visualize the localization of a possible marker of a human serum albumin. The enhancement transforms Fig. 6(a) into Fig. 6(b) where the high-spatial resolution is still present, while suppressing images outside the 1338–1339 m/z range. Figure 6(c) shows the result of the mass slice composition containing solely data of the 1338–1339 m/z range. Figures 6(b) and 6(c) both show the 1338–1339 m/z localization, but Fig. 6(b) clearly shows more of the sample morphology than Fig. 6(c).

As described in Ref. 5, the high resolution of the mosaic microscope images [Figs. 5(b) and 6(a)] cannot be achieved with typical MALDI microprobe experiments, without the loss of sensitivity and speed that is associated with the use of a highly focused laser to achieve such a high resolution.

To show the magnitude of difference in resolution between microscope and microprobe mode imaging of the software, Fig. 5(a) shows the microprobe image version of Fig. 5(b) (microscope). This microprobe image is created by making pixels of each shot, by summing the ion intensities of

the time-of-flight data in the m/z range between 500 and 3100 m/z for each pixel. To create the correct original height to width ratio, spatial binning is performed in the line scan direction using bilinear resampling (weighted distance averaging) having 150 μm square pixels, equal to the distance between two line scans.

B. Improved speed of microscope mode imaging

An advantage of microscope mode imaging over the more often used microprobe approach is the improved speed of analysis. Because of the varying mass resolution of the images, total ion, or a blanked mass range, a full comparison between microscope and microprobe mode imaging is not possible, but the following calculation indicates the magnitude of difference between the two modes. When a sample of 1×1 mm^2 is scanned using microprobe mode imaging, 250 000 number of pixels are necessary to obtain a resolution of 4 μm (which is explained in Ref. 5). For using microscope mode imaging with the spatial image composer, the number of shots in a line scan times the number of line scans is needed. Using a 200 μm laser spot size and a 10 μm step between successive images, 1000 μm needs $(1000+200)/10=120$ images to be recorded. Using a 100 μm movement for successive line scans, 1000 μm needs $(1000+200)/100=12$ line scans. This makes the total number of shots $120\times12=1440$ which is a factor 170 less than for microprobe imaging.

C. Imaging strategies

The ion-optical images recorded by the CCD camera differ to a large extent from the typical images used in image stitching examples. First, the amount of signal varies between successive images, as a result of sample consumption and surface variation. In addition, only part of the image contains signal, which also varies from image to image. This is the reason why it is favorable to always process the entire image instead of cropping parts of the image away. These large variations between the acquired images make the automatic feature recognition for the images not possible. To overcome this difficulty an averaging stitching algorithm is used, which is described in Sec. III.

Improvement of the signal-to-noise ratio in the resulting mosaic image is achieved by the use of oversampling. The stitching algorithm averages the values from the successive images which have an overlap of more than 95% with typical settings (stage moving speed and laser repetition rate). The improvement can be seen by the disappearance between the noise in a single shot image [Fig. 3(b)] and the stitched image strip image [Fig. 4(a)].

The method of combining the images assumes that

- no spatial distortion during image acquisition occurs and that
- the laser fire frequency and the sample stage movement are stable to allow for effective resampling.

If these conditions are not met, blurring will fade away details of the sample in the resulting mosaic image. Figure 4(b) demonstrates that these two assumptions are legitimate.

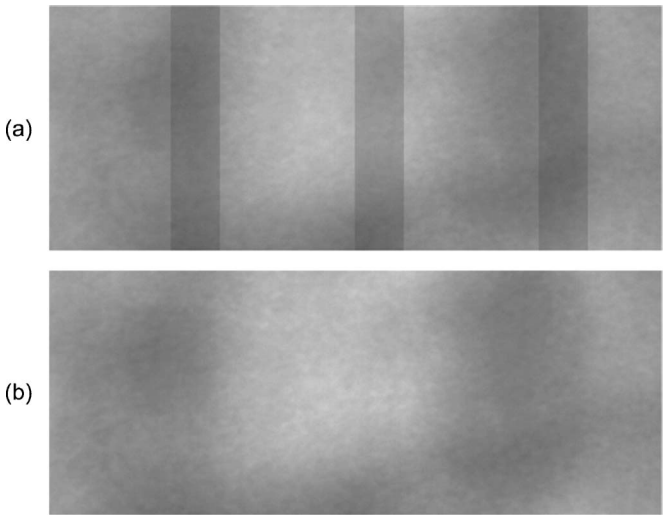


FIG. 7. Solving seam artifacts. (a) Detailed part of a total ion image of Fig. 6 without using the distance map. (b) Detailed part of a total ion image using the distance map, where the seam artifact bands disappeared.

Resulting mosaic images are scaled up linearly to the maximum contrast in the image (linear contrast stretch to grayscale values 0–255). Interpreting intensities in the resulting line scans and resulting mosaic images may only be done relatively; absolute intensity values are scaled arbitrary units. Although mutual comparison for absolute intensity values between resulting mosaic images is not applicable, showing that the maximum contrast optimizes locating spatial features in the image.

1. Seam artifacts

Seam artifacts were visible after stitching the line scans together in many of the resulting images. These artifacts were visible as bands in the line scan direction [Fig. 7(a), a detailed part of Fig. 6(a)] and were caused while averaging the values of the line scans. The cause of this problem was the combination of two facts. First, line scans overlap, but in such a way that the number of overlapping images is not constant, which is shown in Fig. 8(a). Second, at the edges of the line scans no signal is present, which is shown in Fig. 4(a). So, at position A in Fig. 8(a) a new image is added to the averaging calculation, but the image does not contribute any signal, because at the edges no signal is available. This results in the darker bands [the lower average values in Fig. 7(a)].

Theoretically, the best solution to the problem, cropping the nonrelevant parts of the images, is not allowed because of the nonsteady boundaries of the signal spots in the acquired images. Solving the cause of the problem, the movement between the line scans must be chosen in such a way that no variation in the number of overlapping line scans occurs. This occurs when the movement is chosen in such a way that the line scan width is an integral number times the shift, for example, a 640 pixel wide line scan can have a shift of 160 pixels (factor of 4 which results in 75% overlap between adjacent line scans images).

The seam artifacts that occur when the movement was not chosen correctly can be minimized by using weighted

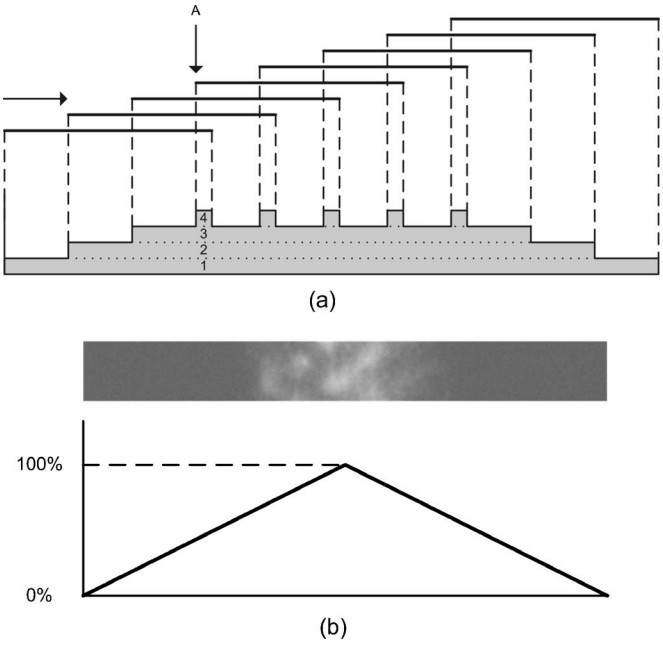


FIG. 8. (a) Seam artifacts, simplified view showing the principle. The left arrow indicates the scan direction. For clarity the images are not shown overlain to each other but with some white space in between. Below, the gray bars indicate the number of images from which the total ion image is created, which is changing between 3 and 4 in the middle part. (b) Weighting factor using a distance map. Above a cross section of one image strip is shown, and below the percentage of the distance map related to the cross section.

averaging with a distance map algorithm (also called feathering). Looking perpendicular to a line scan, pixels near the center of the line scan are more heavily weighted than pixels near the edges, where no signal is present. Figure 8(b) shows this principle, where the weight of a pixel is a linear function to the distance to the center. An improved distance map can be programed in such a way that the distance map truly reflects the signal variation over the line scan, which will vary with each measurement. The improvements of using the distance map are shown in Fig. 7(b), where the figure clearly shows the vanished bands by using the distance map algorithm.

2. Background subtraction

Improved results in enhancing the images with the mass time-of-flight data are achieved by performing background subtraction on the original images before combining them. This results in an ion intensity that corresponds to a pixel intensity of zero instead of the noise floor value. A preprocessing tool, which will be described later, determines the value of the subtraction to be the smallest pixel value intensity of all images.

V. PERFORMANCE AND OPTIMIZATION

A. Image formats

The portable network graphics (PNG) file format was used for storing the generated image data from the SIC software. The PNG file format results in small file sizes, uses lossless compression, and is patent-free and well documented[19] by the World Wide Web Consortium (W3C). The compression technique (a LZ77 derivative) is a standard part of the JAVA programming language, so the implementation can be effectively realized. The PNG file format also supports the storage of text strings associated with the image that is used to store metadata about the image.

Performance measurements on the software show that the file transfer time largely determines the total processing time. To keep the transfer time, and hence the processing time, low, the source TIFF image files were converted into PNG image files. Depending on the size of the images and the compression ratio, it typically reduces the file size by a factor of 4 and so the transfer time of these files. It takes some extra time to convert the images from one format to the other, but a preprocessing tool was developed to automate the conversion. Often the images are used many times in the image processing, so the conversion saves time in the long run. The extra time to decompress the PNG image files during stitching the images is negligible compared to the extra transfer time of the noncompressed TIFF images.

B. Parallel processing

The spatial image composer software has been developed to handle the large amount (tens of gigabytes) of MS image data. During design of the software the use of parallel (distributed) processing was taken into account. Parallel processing can be very time effective when processing large data sets and is known to reduce the overall processing times. The combination of images in the software is performed per line scan to be prepared to make use of parallel processing. Packages such as NIMROD, which is also supported in the VL-e environment, can now be used to implement the parallel processing of the data set in a distributed computing environment. Using NIMROD to run the spatial image composer software saves the software developer from the burden of monitoring the parallel processing. Without a distributed computer environment like the VL-e, a local computer cluster can be used to perform parallel execution of the software.

Prototype tests have proved the principle of the time effectiveness of this parallel (distributed) processing. A data set consisting of 12 line scans was processed on, respectively, one, two, and three computers, where on the two computers 6 line scans, and on the three computers 4 line scans per computer were processed in parallel. The computers were connected to the same network switch and all accessed the same data set on a Silicon Graphics 8 Tbyte data storage system. Parallel processing on two computers reduced the processing time to 50%, and on three computers to 34% of the time needed to process the data on one computer. These results not only show the effectiveness of parallel processing but also show the possibility to use grid computing technology to speed up processing.

C. File access time

The main challenge during the software development was to handle the large amount of data and the related long processing times required to generate the results. Because the calculation itself is relatively simple, the processing time de-

pends almost completely on the transfer time of the data files (I/O intensive application). This was demonstrated by the fact that the CPU usage increased every time the access time to the files decreased. Better transfer times over the network, or even storing the data on a local hard drive, decreased the total processing time substantially so that the CPU usage of 100% became the limitation in speed. Fast data access over the network was also achieved by using a dedicated 1 Gbit/s CXFS connection to an SGI CXFS Filesystem, which is available to some computers in addition to the shared 1 Gbit/s standard LAN.

Because the speed of the software is largely dependent on the file access time, it will be useful to test the use of a data format such as netCDF or HDF for storing the separate image data. In the period of testing and using the software, it is observed that the total access time of small files is largely dependent on the time to open an image file and less on reading the data itself. By putting all images into one large file, the time to open all of the small images separately is gained.

D. Software application

Because of the iterative development process of the software, supportive functionalities were implemented during the programming, testing, and use of the software. Below is an overview of the supportive functionalities which optimize the imaging process.

- An analyze function configures the software by automatically grouping all images and line scans of an experiment. For a typical data set the software saves a user from analyzing the data set structure of 75 000 images. This not only saves time; not every file browser can handle this number of files.
- The automatically generated data set configuration is completed by a user and stored in a central database. The user completes the configuration by adding both scan directions, the movements between successive images and line scans, the name and locations of the resulting mosaic images, and a descriptive text on the data set. The next time the data set is used, these metadata are available by selecting the configuration file.
- Progress indicators are available to the user while processing a data set. Progress indicators are essential for the effective use of the software. This is because most users only accept programs processing the data when an indication is given about the progress of the data processing and when the software seems to be "alive" during the processing.
- The preprocessing part of the software takes care of the preprocessing of the image data. The preprocessing tool can perform the following auxiliary functions described earlier: determination of background subtraction value, background subtraction of the image data, data compression by converting the image format, and copying data while preprocessing. The determination of the background subtraction value is determined first, while the background subtraction itself, the data compression, and the copying of data can be performed at the same time.

These supportive functionalities create, together with the different processing strategies, a problem solving environment which enables the visualization for data intensive high-spatial resolution imaging mass spectrometry experiments.

ACKNOWLEDGMENTS

This work was carried out in the context of the Virtual Laboratory for e-science project.[18] This project is supported by a BSIK grant from the Dutch Ministry of Education, Culture and Science (OC&W) and is part of the ICT innovation program of the Ministry of Economic Affairs (EZ). This work is also part of the research program of the "Stichting voor Fundamenteel Onderzoek der Materie (FOM)," which is financially supported by the "Nederlandse organisatie voor Wetenschappelijk Onderzoek (NWO)." The authors would like to acknowledge P. G. Kistemaker for his careful reading of the manuscript.

[1] H. Rauwerda, M. Roos, B. O. Hertzberger, and T. M. Breit, Drug Discovery Today **11**, 228 (2006).
[2] R. Aebersold and M. Mann, Nature (London) **422**, 198 (2003).
[3] T. Ideker *et al.*, Science **292**, 929 (2001).
[4] Y. E. M. van der Burgt *et al.*, J. Am. Soc. Mass Spectrom. **18**, 152 (2007).
[5] S. L. Luxembourg, T. H. Mize, L. A. McDonnell, and R. M. A. Heeren, Anal. Chem. **76**, 5339 (2004).
[6] A. F. M. Altelaar, I. Klinkert, K. Jalink, R. P. J. de Lange, R. A. H. Adan, R. M. A. Heeren, and S. R. Piersma, Anal. Chem. **78**, 734 (2006).
[7] A. F. Maarten Altelaar, I. M. Taban, L. A. McDonnell, R. P. J. de Lange, R. A. H. Adan, W. J. Mooi, R. M. A. Heeren, and S. R. Piersma, Int. J. Mass. Spectrom. **260**, 203 (2007).
[8] R. M. A. Heeren, L. A. McDonnell, E. R. Amstalden, S. L. Luxembourg, A. F. M. Altelaar, and S. R. Piersma, Appl. Surf. Sci. **252**, 6827 (2006).
[9] S. L. Luxembourg, A. F. M. Altelaar, L. A. McDonnell, and R. M. A. Heeren, Ned. Tijdschr. Natuurkd. **72**, 188 (2006).
[10] S. L. Luxembourg, A. R. Vaezaddeh, E. R. Amstalden, C. G. Zimmermann-Ivol, D. F. Hochstrasser, and R. M. A. Heeren, Rapid Commun. Mass Spectrom. **20**, 3435 (2006).
[11] I. M. Taban, A. F. M. Altelaar, J. Fuchser, YE. M. van der Burgt, L. A. McDonnell, G. Baykut, and R. M. A. Heeren, J. Am. Soc. Mass Spectrom. **18**, 145 (2007).
[12] T. Lillesand, R. Kiefer, and J. Chipman, *Remote Sensing and Image Interpretation* (Wiley, New York, 2003).
[13] P. Longley, M. Goodchild, D. Maguire, and D. Rhind, *Geographic Information Systems and Science* (Wiley, New York, 2005).
[14] K. Dreisewerd, Chem. Rev. (Washington, D.C.) **103**, 395 (2003).
[15] M. Karas and R. Krüger, Chem. Rev. (Washington, D.C.) **103**, 427 (2003).
[16] S. Luxembourg, PhD thesis, Utrecht University, Utrecht, 2005 (www.amolf.nl).
[17] T. Hey and A. E. Trefethen, Science **308**, 817 (2005).
[18] www.vl-e.nl
[19] www.w3.org/TR/PNG