UPTEC X 08 023
Examensarbete 20 p
Augusti 2008

Normalization methods for imaging mass

Tomas Bergvall

UPPSALA
UNIVERSITET

Bioinformatics Engineering Program

Uppsala University School of Engineering

UPTEC X 08 023	**Date of issue 2008-08-24**

Author

Tomas Bergvall

Title (English)

Normalization methods for imaging mass spectrometry

Title (Swedish)

Abstract

Spectra were gathered from healthy rat brains using the MALDI technique. A script library was created to visualize normalized spectra in BioMap and two models for between tissue section normalization were analyzed.

Keywords

Mass spectrometry, imaging, normalization, MALDI

Supervisors

Per Andrén
Uppsala Universitet

Scientific reviewer

Mats Gustafsson
Uppsala Universitet

Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **42**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 555217

Normalization methods for imaging mass spectrometry

Tomas Bergvall

Sammanfattning

MALDI-IMS är en teknik som kan visa uttryck och fördelning av hundratals proteiner eller peptider samtidigt, direkt på alla typer av vävnader. Ett stort problem är att det är svårt att kvantifiera och jämföra de funna nivåerna. Det beror på att det finns många källor till varians i dessa nivåer. Fokus i detta projekt var dels att skapa ett program som kan normalisera spektrum inom ett vävnadssnitt och dels att hitta metoder för att normalisera mellan vävnadssnitt.

Resultaten av detta projekt är för det första ett program som kan normalisera spektrum inom ett vävnadssnitt och visualisera dessa spektrum i en mjukvara kallad Biomap. För det andra visade en metod för att normalisera mellan vävnadssnitt lovande resultat att bygga vidare på.

Examensarbete 20p
Civilingenjörsprogrammet i Bioinformatik

Uppsala universitet Augusti 2008

Contents

1 Introduction **3**
- 1.1 MALDI IMS . . . 3
  - 1.1.1 Common experimental design . . . 5
- 1.2 Preprocessing . . . 5
  - 1.2.1 Baseline subtraction . . . 6
  - 1.2.2 Smoothing . . . 7
  - 1.2.3 Normalization . . . 8
    - 1.2.3.1 Normalization within a tissue section . . . 8
    - 1.2.3.2 Normalization between tissue sections . . . 9
- 1.3 Aim . . . 9

2 Materials and Methods **11**
- 2.1 Experimental design . . . 11
- 2.2 Animal treatment . . . 12
- 2.3 Sample preparation . . . 12
- 2.4 Matrix coating . . . 12
- 2.5 MALDI IMS . . . 13
- 2.6 Analysis of the tissue sections . . . 14
- 2.7 Preprocessing . . . 14
  - 2.7.1 Part 1: TIC-normalization and creating an image . . . 14
  - 2.7.2 Part 2: Normalization between sections . . . 15
    - 2.7.2.1 Data transformation . . . 15
    - 2.7.2.2 Normalization . . . 18

3 Results **23**
- 3.1 Part 1 . . . 23
  - 3.1.1 Make image file . . . 23
  - 3.1.2 Normalizing the image file . . . 25
  - 3.1.3 Experimental results . . . 26
- 3.2 Part2 . . . 26
  - 3.2.1 Right-left bias . . . 27
  - 3.2.2 Low-variance normalization . . . 29
  - 3.2.3 Loess normalization . . . 32
  - 3.2.4 Similarities between consecutive tissue sections . . . 33

4 Discussion **35**
- 4.1 Normalization within a tissue section . . . 35
- 4.2 Normalization between tissue sections . . . 35
- 4.3 Sample preparation . . . 36
- 4.4 Right-left bias . . . 37
- 4.5 Future work . . . 37

5 Acknowledgement **37**

A Readme for the script library **41**

1 Introduction

This master thesis project has been conducted at Department of Pharmaceutical Biosciences, Medical Mass Spectrometry (MMS) at Uppsala University under guidance of Prof. Per Andrén, Dr. Malin Andersson, grad. student Anna Nilsson and co supervised by Dr. Ingrid Lönnstedt. The project consisted of two parts, both regarding normalization of high throughput protein expression data to reduce systematic errors and noise. The aim of the first part was to implement a within section normalization method called total ion current (TIC) normalization [14]. The aim of the second part was to find new methods for normalization between different tissue sections. The experimental technique used was matrix assisted laser desorption ionization (MALDI) imaging mass spectrometry (IMS) to study brain tissue sections from animals models of Parkinson's disease [20, 17, 16].

1.1 MALDI IMS

Methods for finding proteins and peptides can be classified into two branches. There are labeling methods like immunohistochemistry [24] and magnetic resonance imaging (MRI)[13]. These methods are all based on a specific label for each protein. The label is usually an antibody with a fluorescent property. Since these antibodies are specific for each protein, only one protein can be found at a time. For a proteomics or peptidomics study this would require endless effort to acquire the correct antibodies and performing the experiments. These methods are on the other hand able to visualize the spatial distribution of proteins and peptides across tissue sections. There are other methods which are label-free like two-dimensional difference gel electrophoresis (DIGE) [9] and imaging mass spectrometry (IMS). These methods require no prior knowledge of the proteins of interest which means that it is easier to discover novel proteins and peptides. However, they usually have lower sensitivity, which means higher concentrations of the proteins for detection are required. The advantage of using IMS over the other techniques is the combination of specificity, where it is possible to analyze a couple of hundred proteins in one experiment, and the visualization of the spatial distribution of proteins in tissue sections.

MALDI IMS [10, 4] is a MS method where the result can be displayed as images showing the spatial distribution for any selected protein. For example, the spatial distribution for three different proteins with diverse distribution in the different regions of the rat brain can be displayed simultaneously (Figure 1).

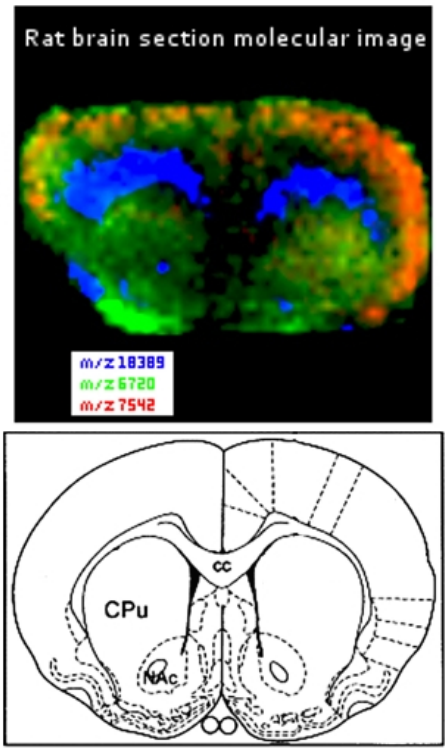

Figure 1: MALDI-IMS molecular image of a rat brain. Blue represents myelin in the white matter of corpus callosum (cc), green is Pep19 localized to striatum (CPu and NAc) and red most intense in the grey matter of cortex. The anatomical map of a rat brain is applied from Paxinos *et al. 1997.*

The experimental setup will be explained in Section 2.5 and the design in Section 4, but the basics of the method is as follows. Any tissue can be taken, cut into very thin slices (typically 12μm) and mounted to a MALDI target slide, a metal plate on which the tissue is attached. To extract and crystallize the proteins on the slide a matrix, a solution of cinnaminic acid, acetonitril and trifluoroacetic acid (TFA) in water, is applied by spraying or spotting. Spraying adds a thin layer of matrix over the whole tissue whereas spotting places spots of matrix on the tissue section in a rectangular grid of points i.e. a raster (Figure 6). Both of these techniques can be performed manually or automatically. All of these techniques cause variance in the data but automatical matrix application by an instrument is of advantage [1]), since the ability to reproduce the results is much better. The MALDI-TOF (time-of-flight) MS software, FlexImaging, can be used to add a raster over the tissue section and a laser beam ionizes and vaporizes proteins in each spot (usually $M+H^+$, the original mass plus a proton). The ionized proteins are put in an electric field and the mass of the protein is determined

according to how fast the charged proteins travel through field-free vacuum. The electric field causes all proteins with the same charge to have the same kinetic energy. According to the equation of kinetic energy (E_k) (Equation 1) the velocity (v) is inversely proportional to the mass (m).

$$E_k = \frac{1}{2}mv^2 \tag{1}$$

The detection is based on the time it takes for a protein to reach the detector (time-of-flight). This means that proteins can be separated according to their mass-to-charge (m/z) ratio since a protein with a larger m/z ratio will travel through the MS instrument with a lower velocity. A lot of proteins usually reside in many different charged states after ionization which will yield multiple peaks for these proteins. In this project this is disregarded from and each peak is viewed as a unique protein. If further analysis, for finding abundances of each protein, would be conducted this would have had to be accounted for. Each raster point is shot by the laser about 100 times and the detection of each shot is summed into a spectrum. Each spectrum has m/z ratio on the x-axis and intensity, the number of molecules with a specific m/z detected, on the y-axis (Figure 2). The gathered spectra consist of a baseline with small fluctuations, chemical noise (see Section 1.2.1), and peaks which corresponds to proteins. The chemical noise is a detection of a random event with no distinct m/z between two occurences and can therefore be detected with different m/z each time. The chemical noise can for example be a protein which has been degraded by the laser. A protein on the other hand will have the same m/z and be detected at the same m/z at all times thereby accumulating to a higher peak.

1.1.1 Common experimental design

A common experimental design in experimental Parkinson's disease is to induce dopamine denervation using a neurotoxin in one side of the brain and to use the other side as a control. The dopamine denervated hemisphere is analyzed for up- or down regulation of proteins and the other side is viewed as a negative control, i.e. the normal state of the brain.

1.2 Preprocessing

There are many sources causing systematical errors, i.e. errors affecting a whole experiment, therefore the gathered spectra need to be preprocessed in order to be comparable [14, 5]. The systematical errors come from variations in the sample preparation before the experiment, variation in laser intensity, differences in ionization capability for different proteins etc. The workflow is first to gather the data from the experiment, then preprocessing it to remove the systematical errors and finally a statistical analysis to discover biological results. The main steps of the preprocessing will be described below.

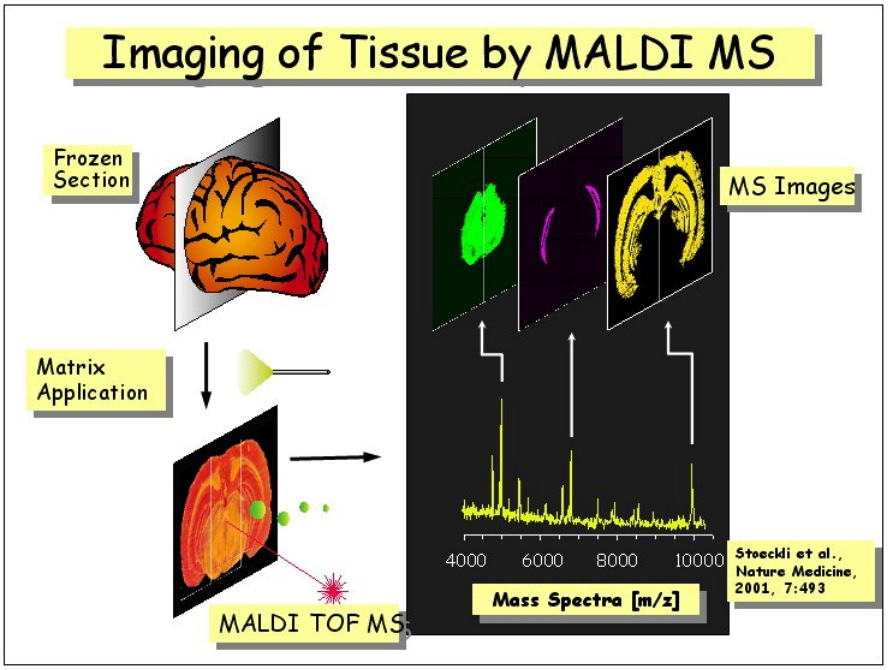

Figure 2: General overview of the MALDI-IMS method used. First the brain is sectioned and placed on a slide and then the matrix is applied. Next, the slide is placed in the MALDI IMS instrument and a mass spectrum is gathered in each raster position.

1.2.1 Baseline subtraction

At low m/z values of a typical spectrum there is usually a large amount of chemical noise [11] and this yields a much higher baseline, the red line in Figure 3, than at higher m/z values (Figure 3). The chemical noise comes from desorption/ionization of the proteins, the matrix and impurities in the sample. The exact nature of the chemical noise is not known but it can be fragments of proteins/matrix and has to be accounted for in order to compare different spectra.

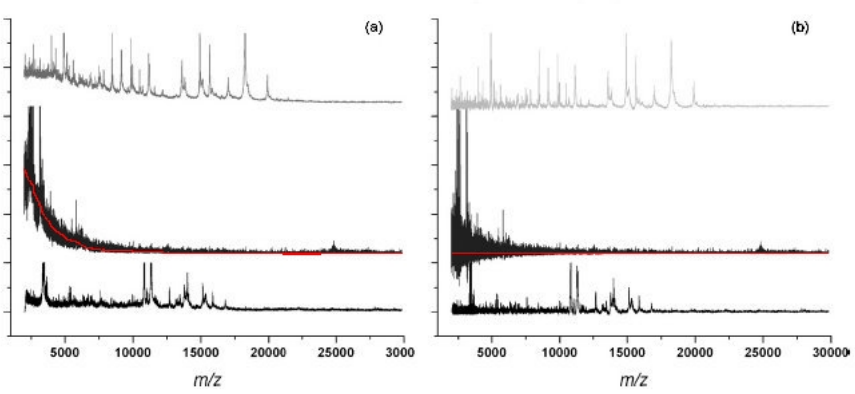

Figure 3: Baseline subtraction. Illustration used with permission from Elsevier Limited [14] showing how spectra look before (a) and after (b) baseline subtraction.

The effect of the chemical noise can be suppressed by estimating a baseline, the red line, which is usually a polynomial function or a moving average. The moving average is calculated by selecting a number of points closest to the first point and taking the average of their signal levels, then the same for the rest of the points. The averages are then used as the new baseline and all signal levels are measured from this curve.

1.2.2 Smoothing

Chemical noise also yields the intensities in each spectrum to be randomly distributed. Preferably the baseline would be a smooth curve not showing any small peaks, only large peaks for proteins. This is not the case for a normal spectrum where the baseline has fluctuations of the baseline which are still there after baseline subtraction. To be able to distinguish between peaks and background a method called signal-to-noise is used (ClinProTools). Signal-to-noise is a measure of how high a peak is compared to the baseline surrounding it. It is calculated by a moving average where the points around the peak are used to calculate the baseline and the points of the peak are used to calculate the peak height. A large peak height compared to the surrounding baseline represents a protein. Since the points of the baseline show fluctuations it will be more difficult to distinguish between peaks and the baseline. This problem can be taken care of by smoothing [19] the spectra. The smoothing algorithm (ClinProTools) will reduce the fluctuations by smoothing the spectra thereby increasing the signal-to-noise ratio for peaks.

1.2.3 Normalization

Normalization is often referred to as a mathematical function which is able to return data to its normal state, where there are no systematical errors and no noise. In this case normalization is needed to remove systematical errors derived from handling the samples and the MALDI IMS analysis. The goal is to remove all systematical errors and only keep the variations caused by biological variation. This is for almost all biological applications a utopia due to the complexity of the biological samples and a part of the systematical errors will therefore always remain in the data. The data used to evaluate the normalization methods is derived from the experiment described in materials and methods (Section 2.1).

1.2.3.1 Normalization within a tissue section

There is a need for normalization of each spectrum in an image because the information gathered in each laser shot can vary. Apart from biological variations there are also variations which can occur from either matrix crystallization or ionization [14]. These effects are due to the MALDI process and have to be accounted for in order to detect the biological variations. There are valid reasons to assume that each spectrum should show an equal amount of information. The information in each spectrum is measured in how much that is detected, i.e. the intensity in each point on the m/z axis denoted I_1,...,I_M where M is the number of points on the m/z axis. By summing all these measurements the TIC is acquired (Equation 2).

$$TIC = \sum_{i=1}^{M} I_i \tag{2}$$

Each spectrum is then scaled by its TIC and hence normalized [8]. Let us denote the number of spectra in each tissue section as S where each spectrum is containing a set of intensities I_1^S,...,I_M^S. Then I_1 will have S intensities corresponding to the same m/z ratio. The TIC normalization usually improves the %CV (coefficient of variance, Equation 3-5), which is a measure of reproducibility where 0% is best.

$$\overline{x}_1 = \frac{1}{S} \sum_{i=1}^{S} I_1^i \tag{3}$$

$$s_1 = \sqrt{\frac{1}{S-1} \sum_{i=1}^{S} (I_1^i - \overline{x}_1)^2} \tag{4}$$

$$\%CV = \frac{s_1}{\overline{x}_1} \times 100 \tag{5}$$

The %CV is usually reported as the mean of all M measures of %CV. Dekker et al however did not find TIC normalization to be an optimal procedure (reduced %CV from 42% to about 30%) but it has been used by others with good results [2, 14].

1.2.3.2 Normalization between tissue sections

An obstacle when performing semi-quantitative studies (i.e. to compare intensities between different tissue sections) is that there is usually a difference between tissue section images. This can be due to differences in the preparation steps like the matrix application procedure or variance in the thickness of the tissue slice. All of these affect the extraction and ionization capabilities of the proteins and will hence affect the spectra as well. Since these effects are present there is a need for normalization between images. Previous work in MALDI MS has usually focused on using internal calibration standard as the mean of quantification [15]. The internal calibration standard is typically the molecule of interest spotted on the same glass slide with decreasing concentration. The focus of this master thesis project was to find novel methods for normalization without having to add an external calibrant. However, this method would only give quantitative measurement of one substance, the internal calibrant, where it would be preferable to quantify every protein in an image. One normalization method presented in this report was to find proteins which act similar to housekeeping genes in microarray studies [18]. Housekeeping genes are thought to be genes involved in basic procedures in a cell and they should be active throughout the entire lifecycle of a cell. Hence they should have the same expression level in all cells. If there exists proteins with similar properties, equal distribution in all cells, then differences in signal intensities between tissue sections should only reflect the systematical errors and noise. Another normalization method (Section 2.7.2.2) was based on finding trends in the whole dataset simultaneously with a loess (locally weighted scatterplot smoothing) regression [21]. Loess normalization has often been used in microarray studies and is a method for finding trends in large datasets. It is performed by taking each point and calculating the new loess fitted point by looking at each point's neighboring points. These neighboring points are weighted in the regression by how close they are to the point of interest, with large weights for close points. The loess fitted point is the closest point on the regression curve to the point of interest.

1.3 Aim

There is a free software on the market called Analyze This! [6] which can create image files from the Bruker file format. The disadvantages of this software are that it distorts the mass scale if the number of data points has to be reduced and that the TIC normalization still has to be done. There is

also one software from Bruker (FlexImaging) which can export image files but not with TIC normalized data. The first aim of this project was to make a script library (a computer program) which can easily be used to make and normalize a MALDI image (Analyze 7.5 format) and view the image with appropriate software. Normalizing between tissue sections is a new area in MALDI imaging MS and there are no methods available which are proven to work for all datasets. The second aim was to evaluate different normalization methods for normalization between tissue sections and possibly identify a method which shows results worth further investigation.

2 Materials and Methods

First the experimental design used to acquire the data for evaluation of the normalization methods will be described. Then the normalization is divided in two parts, part 1 describes how to normalize spectra within a tissue section and part 2 will describe how to normalize between tissue sections.

2.1 Experimental design

The more sections the better but more sections would also implicate a longer period between the first and the last slide. There are variations arising from day-to-day variations of the experiment but these were not the focus of this project. A longer time period between the first and the last slide would introduce more variance. To eliminate as many of these variations as possible a pairwise 4 x 4 design was implemented, i.e. 16 sections altogether. On each glass slide there were four tissue sections and there would be four glass slides in total (Figure 4). The placement on each slide was important since the internal order of the sections was to be preserved. The consecutive sections could possibly be used as replicates to get reliable intensities. The numbering in Figure 4 represents the order in which the brain tissue sections were sliced but the name of the actual sections henceforth are according to which slide it reside on. For example on slide 1 the brain sections are named 1 through 4 and the second has 5 through 8.

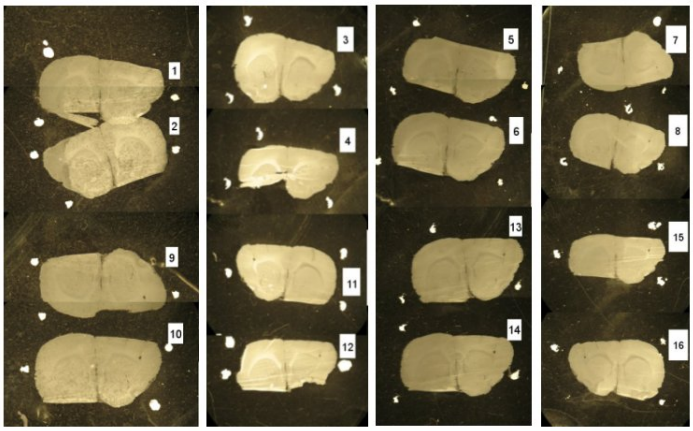

Figure 4: The experimental 4x4 design. Each brain tissue section has been numbered according to the cutting order in the tissue, i.e. to highlight those tissue sections which are consecutive. Henceforth each tissue section will be numbered according to which slide it reside on e.g. instead of $[1, 2, 9, 10]$ on slide 1 it will be $[1, 2, 3, 4]$ and slide 2 will have $[5, 6, 7, 8]$ and so forth.

2.2 Animal treatment

Adult male rats were housed under a 12 hour light, 12 hour dark cycle with food and water *ad libitum*. The animal procedures were approved with local animal ethics committee and carried out in accordance with the European Communities Council Directive of November 24, 1986 (86/609/EEC). The animal was put to death with isoflurane and the brain was quickly extracted and frozen.

2.3 Sample preparation

The striatum is a part of the brain which is involved in Parkinson's disease and it was decided to focus on the striatal region of the rat brain. The rat brain was cut into 12 μm thin slices, using a cryostat at -17 C°, and mounted on indium-tin-oxide (ITO) coated microscope glass slides (75x25 mm, Bruker Part No. 237001). For each of the four glass slides four brain tissue sections were mounted, totaling an amount of 16 slices. Each tissue section was mounted in an ordered way such that on each slide there were two consecutive pairs of slices (Figure 4). This was performed since it would yield a possibility to analyze differences between consecutive tissue sections. These differences could be used to estimate systematical errors both within a slide and between slides since the biological variation between two consecutive slices is expected to be low. Each slide was stored in a freezer (-18 C°) until coated with matrix.

2.4 Matrix coating

The matrix was prepared by measuring 50 mg 3,5-dimethoxy-4-hydroxy-cinnaminic acid (Sigma Aldrich) into a tube and adding 3 ml acetonitril and 2 ml 0.5% trifluoroacetic acid (TFA) in water. The solution was then sonicated to make it homogenous. Acetonitrile is used to dissolve the hydroxycinnaminic acid. The addition of TFA gives a low pH of the solution. When the solution is applied to the tissue slice, all protein charged groups (amines, carboxy groups etc.) will be titrated to a common state. The procedure followed the ImagePrep, the instrument used for matrix application, User Manual Version 1.0 (Bruker Daltonik GmbH) with small alterations to suite our needs. To avoid any delocalization of the proteins, caused by thawing, the slides were stored in a vacuum desiccator for half an hour prior to coating. When the glass slides were dry each slide was washed twice in 70% ethanol (EtOH) for 30 seconds and then one time in 95% EtOH for 30 seconds. The slides were then dehydrated for about half an hour in the vacuum desiccator again. To be able to orient the software of the mass spectrometer each tissue section was fitted with three spots of tippex (teachpoints) and a picture was taken using a digital camera mounted on a microscope. Before the actual coating the ImagePrep was cleaned using methanol and wiped

dry with a tissue. The slide was placed inside the chamber and the coating was started. A typical coating application procedure can be seen in Figure 5. The rotations of the glass slide, between 1-2, 3-4 and 4-5, were performed because it was difficult to see whether the spray of the ImagePrep yielded an even distribution of the matrix.

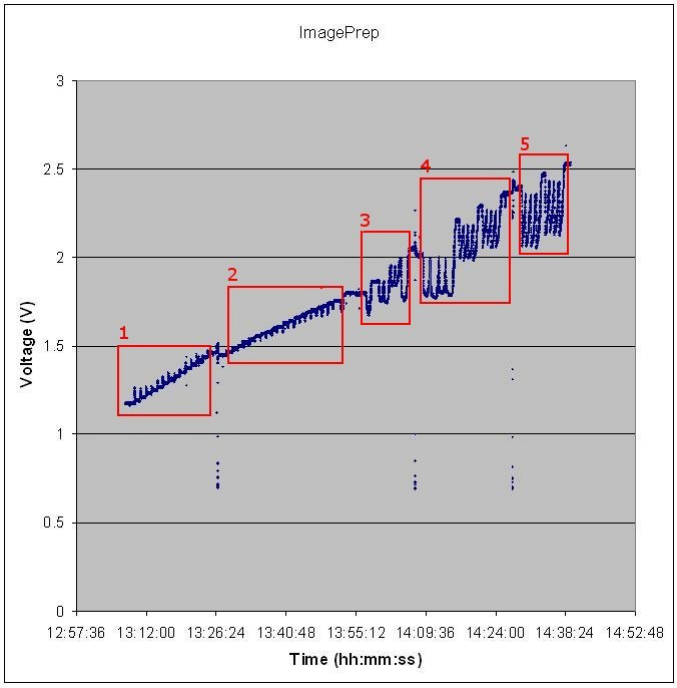

Figure 5: The matrix application procedure where Voltage represents the matrix thickness. The procedure consists of five phases where 1 and 2 are initialization with a rotation of the glass slide 180 degrees in between. Phase 3, 4 and 5 are similar with respect to that the matrix is not allowed to dry completely between each matrix application. In phase 3 the matrix solution is dried every second application and in 4 and 5 every fourth application. There is also a rotation of the glass slide between phase 3 and 4 and between phase 4 and 5.

2.5 MALDI IMS

The mass spectra were acquired using an Ultraflex II equipped with Smartbeam^{TM} technology (Bruker-Daltonics) in linear mode. The instrument was optimized for the best resolution with a standard mix of proteins, insulin, cytochrome C ($M+H^+$ and $M+2H^+$), ubiquitin and myoglobin($M+H^+$ and $M+2H^+$).The first two of the four glass slides were placed one by one in a carrier plate for the MALDI IMS instrument. The laser was guided by teach-points placed around each tissue section and the MS was set to accumulate

data for 200 laser shots in each raster point in a random pattern. An example of the placing of raster points and the area selected can be seen in Figure 6. The third and fourth glass slides were placed in the same carrier and were

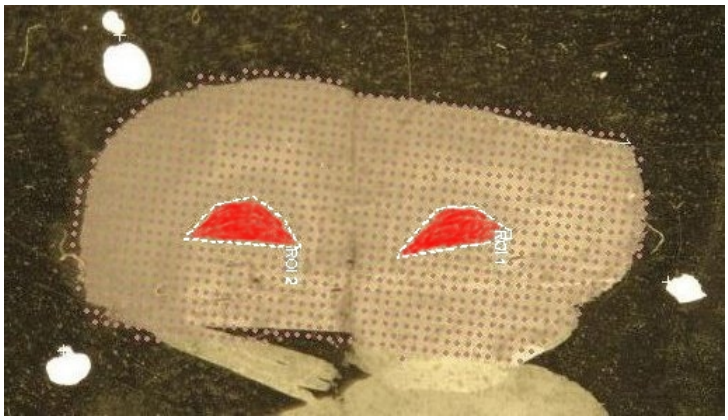

Figure 6: Mounted tissue section after matrix coating. The pink dots are the spots where each spectrum is collected and the white dashed, red colored areas are the striatum which are the selected areas to be analyzed further.

analyzed together over night with the same procedure as mentioned above.

2.6 Analysis of the tissue sections

A closer look at spectra from the tissue sections revealed a difference between the two sides of the brain. There were signs of hemoglobin in one hemisphere of the tissue which has ion suppression effects on other molecules [3]. An ion suppression molecule affects the crystallization and evaporation which decrease the signal levels of all molecules in its surrounding space. This is an extra source of variance in the tissue but it was decided to proceed with the brain data despite this feature.

2.7 Preprocessing

2.7.1 Part 1: TIC-normalization and creating an image

The output from the Bruker MALDI-TOF MS is stored in a special format which resembles a tree structure with every folder representing one spectrum. To be able to access the underlying data with ease each spectrum has to be exported with a program called FlexAnalysis. During the export each spectrum was processed by removing the baseline and smoothing the peaks. The baseline removal algorithm used was ConvexHullV3 and for smoothing the SavitskyGolay algorithm [19] implemented in the FlexAnalysis software. The exportation yielded files with each measurement point, i.e. the m/z, and the corresponding intensity.

The created images were stored in the Analyze 7.5 format (www.mayo-.edu/bir/PDF/ANALYZE75.pdf) and viewed in a software called BioMap (a-vailable at www.maldi-msi.org). Since this software is unable to normalize spectra on its own, normalization had to be performed before importing it into BioMap. A perl script was created which reads each spectrum, i.e. the exported data file from FlexAnalysis, and calculates the sum of all intensities, here denoted I_1,...,I_N. The TIC is calculated according to Equation 6. Each intensity I_i is divided with the TIC-value for normalization.

$$\forall I_i^{new} = \frac{I_i}{TIC} \times scalefactor, \text{ where i=1,2..N} \tag{6}$$

Each normalized intensity, I_i^{new}, will be much smaller than one and must be scaled to account for the fact that BioMap only allows integer values between -32767 and 32767 (signed short integers). The decision fell upon having a user defined scale factor to be able to have the same scale factor for all images. There are alternatives for a user defined scale factor where one is to have it fixed to something proportional to the number of m/z points. Another is a factor which would give the highest peak in all the spectra the highest possible value (32767). A useful rule of thumb is to use the number of data points times ten as the scaling factor since that will usually give a signal level range of about 0 to 25000 Da. The now normalized intensities were then written to the Analyze 7.5 format and the image could be viewed in BioMap.

2.7.2 Part 2: Normalization between sections

First a specific area in each tissue section was selected for further analysis. In the present study the striatum was chosen since it resides on both halves of the brain and it is a relatively large structure easily delineated. An example of the delineation can be seen in figure 6 where the area has been selected to have its endpoints close to the ends of the corpus callosum.

The areas of interest were selected to be as similar as possible between the tissue sections in respect to location and the quality of the tissue. Some of the areas were therefore split in two because of a mass shift detected during the experiments (Figure 7). A mass shift is an event causing all proteins to suddenly show a different m/z than before. A protein with a mass shift during the experiment will probably be treated as two separate proteins since the masses are so different. The red band reflects a mass shift of about 20 Da for a time period of about two minutes during the experiment.

2.7.2.1 Data transformation

The idea behind the between section normalization was to look only at the "real information" in each spectrum. Each tissue section was analyzed with

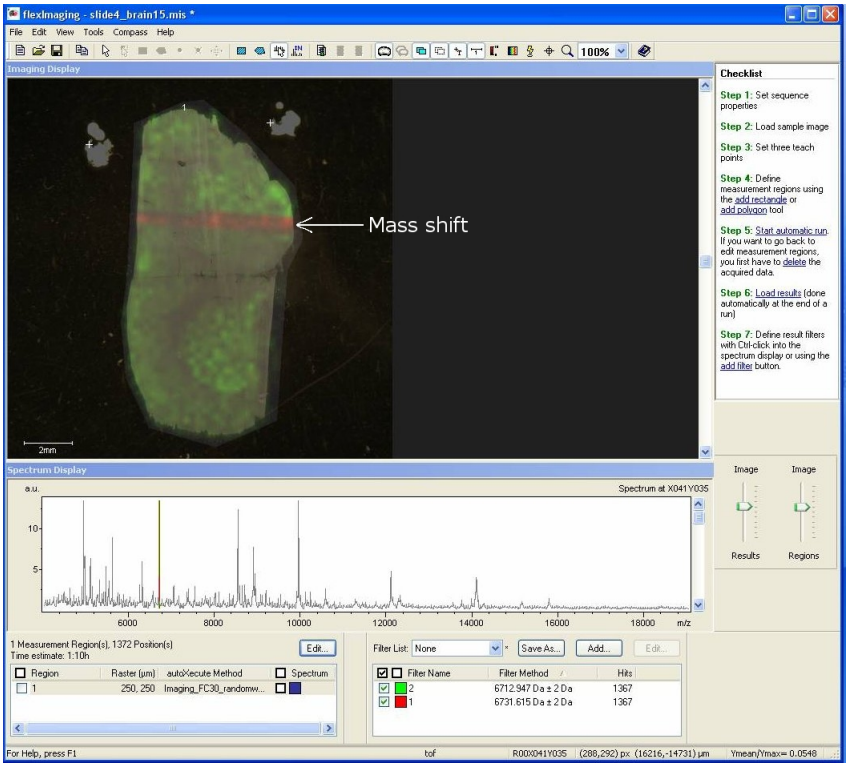

Figure 7: The mass shift shown for Pep19 (mass 6720 Da). The green area represents the peak on 6712 Da and the red 6731 Da. The red line corresponds to two rows of raster points which took about two minutes to complete during the experiment. The cause of the mass shift has not yet been discovered and to circumvent the problem the selected areas were chosen not to include these spectra.

ClinProTools (software from Bruker-Daltonics) which was used to calculate which peaks corresponds to proteins and which peaks that are noise. These calculations are based on a signal-to-noise ratio between a peak and the surrounding chemical noise. A list of these peaks from each tissue section was exported from ClinProTools. Each peak in the list represents a protein or peptide (see Section 1.1) and for each peak the corresponding maximal peak intensity (denoted PI) and peak area (denoted PA) were given. Both of these measures are related to how abundant the proteins were in the tissue. One tissue section usually consisted of about 60 spectra (S=60), the ones residing in the selected area, and each spectrum had about 100 peaks. The peaklists therefore contained 60 x 100 x 2 values in total. First, the correlation between PI, PA and variance was removed from the peaklists (Figure 8) for the peaks in the dataset by log transformation of all values [22]. This was performed to have the order of magnitude of PA variance the same independent of the mean PA. The other issue addressed with log-transformation is that there

are many peaks in the list having very low PA values and a few having large values. The log-transformation reduces these differences and makes it easier to fit statistical models to the data when a statistical analysis of up/down regulation of the proteins is performed (which is a step after normalization). Secondly, all peaks from the 16 different tissue sections were binned (Figure

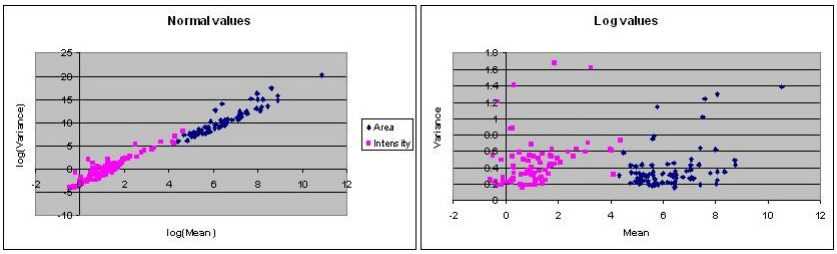

Figure 8: Log-transformation of the data. Each peak is represented by two points in the figure, one pink for PI and one blue for PA. In the left figure the PI and PA are the raw data, the x- and y-axis has just been log transformed after the calculation to get more overview. The right figure is built on log transformed data before calculating the mean and variance for each peak.

9) to produce one coherent dataset with the same set of peaks for all tissue sections. This was performed by selecting one of the peaklists as a template viewing these peaks as potential real proteins. Then, for each of the peaks in the rest of the lists, the peaks were placed in an array belonging to the template peak which had the smallest difference in mass. The peak shift tolerance was set to ± 20 Da, which means that no peak would be placed in the array if the difference to its closest peak was more than 20 Da. The rule is: if and only if the template peak gets a full array, i.e. 15 hits since there are 16 sections, it will be considered a real peak. These real peaks were then used as the dataset which all normalization methods were tested upon. The peak binning reduced the dataset down to 67 peaks from the original about 100 peaks.

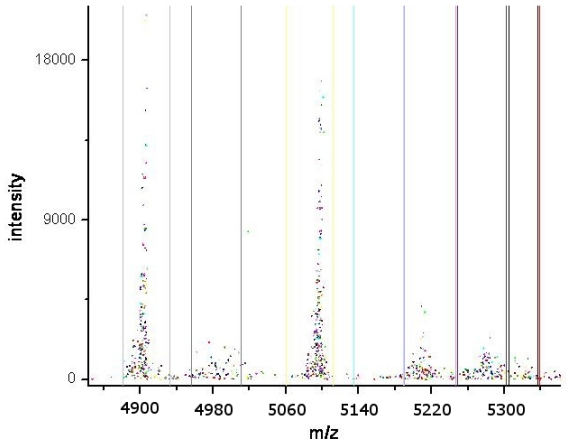

Figure 9: Binning of spectra. Each bin is 40 Da wide and each tissue sections contribute to one point/peak found by ClinProTools. This is only an illustration of how it might look and not an actual illustration of the data used in this project.

2.7.2.2 Normalization

Two methods for normalization between tissue sections on the resulting peak data were designed and evaluated. The data for both of them are the peak areas (PA) from ClinProTools, processed as described above (Section 2.7.2.1)

Low variance proteins

The first main idea for low-variance normalization was to find proteins which did not seem to vary in peak area over the selected area (the striatum). Each tissue section consists of S spectra and each spectrum consists of M peak areas, PA. First the mean for each PA within each tissue section was calculated (Equation 7), where S is the number of spectra within each tissue section. Then the variance for each PA according to Equation 8.

$$\overline{PA} = \frac{1}{S}\sum_{i=1}^{S} PA^i \tag{7}$$

$$Var(PA) = \frac{1}{S-1}\sum_{i=1}^{S}(PA^i - \overline{PA})^2 \tag{8}$$

This yielded one variance for each PA within each tissue section, i.e. 67 values per tissue section. Now there are 16 values, since there were 16 tissue sections, for each Var(PA). To be able to select the PA with lowest variance within the tissue sections the mean for each Var(PA) was calculated

(Equation 9), where ave() is the average.

$$ave(Var(PA)) = \frac{1}{16}\sum_{ts=1}^{16} Var(PA^{ts}) \tag{9}$$

The seven proteins with lowest mean variance (ave(Var(PA)), the mean variance) were selected for the low-variance normalization. After the selection of the seven proteins each tissue section has seven values (Var(PA)), one for each protein to base the low-variance normalization on. The goal of the next step in the low-variance normalization is to find one specific scale factor for each tissue section. The mean PA of the seven selected proteins is calculated in two steps, first the mean for each proteins PA is calculated (Equation 10) then the mean of these seven means is calculated called global mean (Equation 11).

$$\forall\overline{PA}_k = \frac{1}{16}\sum_{ts=1}^{16} PA_k^{ts}, \text{ where k=1,2..7} \tag{10}$$

$$\text{Global mean} = \frac{1}{7}\sum_{k=1}^{7} \overline{PA}_k \tag{11}$$

Then the mean PA for each tissue section was calculated called local mean (Equation 12).

$$\forall\text{Local mean}_{ts} = \frac{1}{7}\sum_{k=1}^{7} PA_k^{ts}, \text{ where ts=1,2..16} \tag{12}$$

The tissue specific scale factor was then the local mean divided with the global mean (Equation 13).

$$\text{tissue specific scale factor} = \frac{\text{local mean}}{\text{global mean}} \tag{13}$$

This factor was used to normalize each PA, i.e. the 67 peaks found in all tissue sections (Equation 14).

$$PA^{new} = \frac{PA}{\text{tissue specific scale factor}} \tag{14}$$

Loess location normalization

The second approach was to analyze the whole set of found peaks (the PA for each peak) and fit a loess curve [7] for each selected area (striatum, will be called tissue section henceforth). The loess regression was implemented in R and the function loessFit from package limma was used. The loess fit is a local regression method which takes a subset of points into consideration

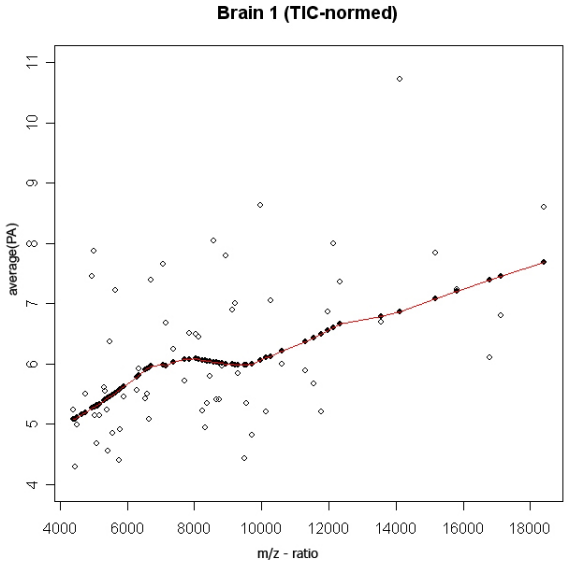

Figure 10: The loess regression curve where the circles represent the mean PA of each protein for tissue section 1 and the black dots are the actual loess fit LPA. The mean peak areas are calculated from spectra included in the area selection (the striatum).

when creating the regression model. It gives larger weight to points near the point to be calculated [7]. The idea is to, in each local neighborhood around the point of interest, fit a linear or quadratic polynomial according to a weighted least squares function (Figure 10). The loess normalization was based on the same idea as the low-variance normalization by calculating a tissue specific scale factor from a local and global mean of the loess regressions. The normalization procedure is described below where the idea is to make the loess regression for all tissue sections as similar as possible using the tissue specific scale factors. For each detected protein in a tissue section there is a m/z ratio x_i and for each protein a mean peak area, $\overline{PA}$, is calculated (Eq. 7). Suppose that Equation 15 holds for each protein, where f is function reflecting the relation between the m/z ratio and the peak area and e is a random error. No assumptions are made about the distribution of e. The function f is not in itself interesting, since no comparisons of proteins with different masses are made.

$$PA_i = f(x_i) + e, \text{ where i=1,2..67} \tag{15}$$

From Fig. 11 it can be seen that there are systematic differences in the "strength" of the peak areas, i.e. the regression curves are lower for some tissue sections and higher for some. Therefore a regression model was chosen

for each tissue section (Equation 16).

$$PA_{i,ts} = f_{ts}(x_i) + e, \text{ where i=1,2..67 and ts=1,2..16} \tag{16}$$

Normalization of PA for each tissue section was based on a tissue specific scaling factor k_{ts} (Equation 17).

$$f(x_i) = f_{ts}(x_i)/k_{ts}, \text{ where i=1,2..67 and ts=1,2..16} \tag{17}$$

The esimatate of k_{ts} was based on loess regressions. Let us denote the loess fitted points with LPA (loess fitted peak area). Since each tissue section had a set of 67 PA there will be 67 LPA for each tissue section which constitutes the loess curve. The loess regression of these 67 peak areas was analyzed for all tissue sections (Figure 11). The m/z range up to 10000 Dalton (the red box) were selected to be used for the further calculations. The reason was that in the higher m/z range the loess curves were not showing any clear correlations with each other whereas the curves follow each other in the m/z range up to 10000 Da [12]. That left 50 LPA points for the loess normalization, which still is 75% of the data since there are more peaks in the low m/z range. A local average for each tissue section was then calculated (Equation 18).

$$\forall \text{Local average}_{ts} = \frac{1}{50}\sum_{i=1}^{50} LPA_i^{ts}, \text{ where ts=1,2..16} \tag{18}$$

The global mean was then calculated according to Equation 19.

$$\text{Global average} = \frac{1}{16}\sum_{ts=1}^{16} \text{Local average}_{ts} \tag{19}$$

The tissue specific scale factor, k_j, was then estimated with the ratio between the local and global average (Equation 20).

$$k_{ts} = \text{Local average}_{ts}/\text{Global average} \tag{20}$$

And finally the normalization of each peak was done according to eq. 17

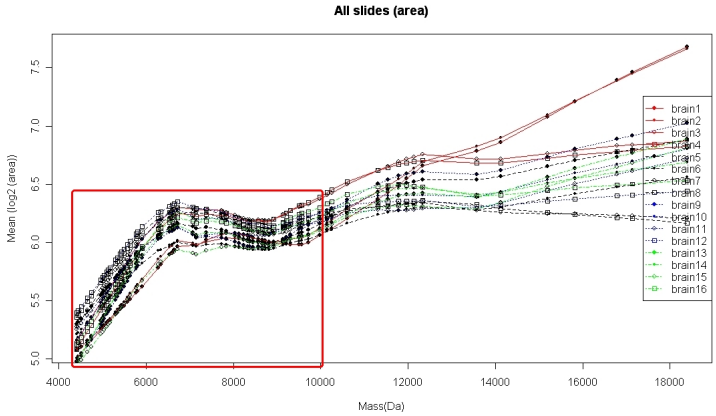

Figure 11: Selected range for loess regression. The red box shows which LPA points that has been used for the loess normalization.

3 Results

In order to use TIC normalization a script library was built where the different steps are described in part 1 below. The normalization between images is shown in part 2 with the loess normalization as a promising result.

3.1 Part 1

The first part of the project was to create a script library which can easily be used to create and normalize a MALDI image. This was done by creating two perl scripts where the first handled the creation of the correct file structure (Analyze 7.5 format) and the second handled the TIC normalization. The readme file for these scripts can be found in Appendix A.

3.1.1 Make image file

The in-files to the make-image script are the files from FlexAnalysis as described in Section 2.7 (Figure 12). There are three files created where the

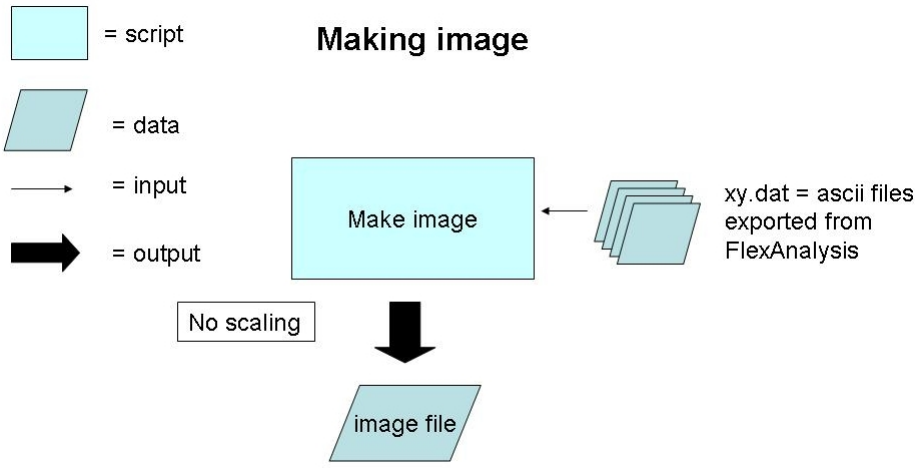

Figure 12: The workflow of the make-image script created. The results is an image file (containing the data) with a corresponding file for the m/z range and a guide file for BioMap.

main file, which stores all spectra, has the ending .img. The two other have the endings .t2m and .hdr where the first defines the x-axis of each spectrum, i.e. the mass scale. The second is a file to guide BioMap to how the image file is structured. Each text-file, from FlexAnalysis, consist of every data point in the corresponding spectrum which is the only information needed to create the image file. There are some choices to be entered by the user during the initialization phase of the script where most of them are about

where to store and what to call the image file. The only critical question asked by the script is for the user to type the number of data points in each spectrum (each spectrum has the same number of data points). This has to be accurate for the script to work. There are no limitations in the number of data points the script can handle but software like BioMap can not handle more than 32767 data points. If there are more data points than this number the whole image will be scaled in BioMap to something smaller than this number. Each spectrum has a name corresponding to the raster as set by the MALDI-IMS. This is used to print the spectra at the correct order where each data point is printed in a signed short integer form (16 bit binary form) to the image file. Since BioMap considers every image to be a rectangle the image has to be padded with null spectra to make the image rectangular. This is done by finding the edges of the picture and adding null spectra, of equal number of data points, with only zeros as intensities, around the real image (Figure 13).

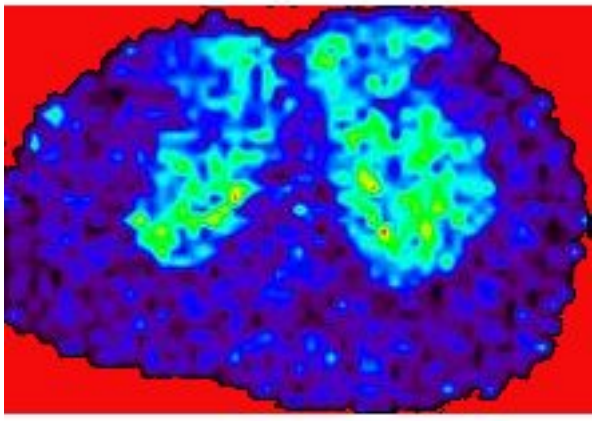

Figure 13: Padding of an image with null spectra. The red area shows were the image has been padded with null spectra.

3.1.2 Normalizing the image file

The in-data to the image-normalization script is the image file and the two files created above (.hdr and .t2m) (Figure 14).

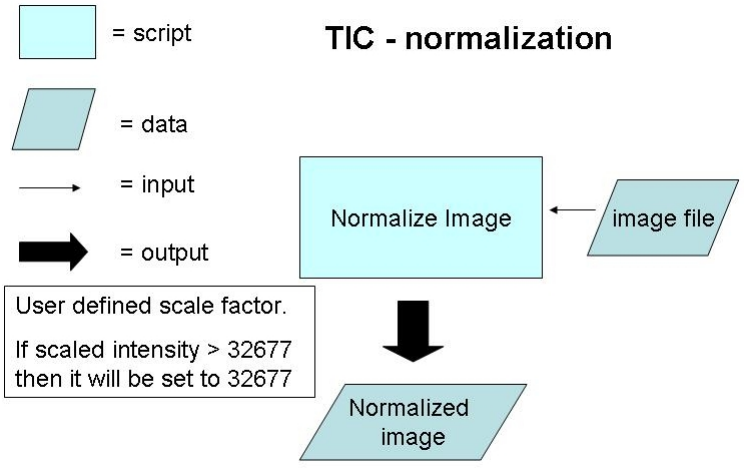

Figure 14: The workflow of the normalize-image script created. The results is an image file (containing the normalized data) with a corresponding file for the m/z range and a guide file for BioMap.

These files are just copied and named after the normalized image. To normalize the image each spectrum is read twice, the first time to calculate the

TIC which is the sum of all data points and the second to divide every data point with the TIC. Each data point is also scaled by a user defined value to ensure that the normalized values are in the range defined by BioMap (0-32767). If a scalefactor is not used the range is often too small and the resolution in BioMap will suffer since BioMap only handles integer values.

3.1.3 Experimental results

The result of TIC normalization is an image where the different regions of the tissue are better defined and smoother than before. Figure 15 shows the result before and after normalization for one of the proteins. The TIC normalized figure is smoother than the left figure which is good since large differences is not expected between two adjacent spectra. The script library

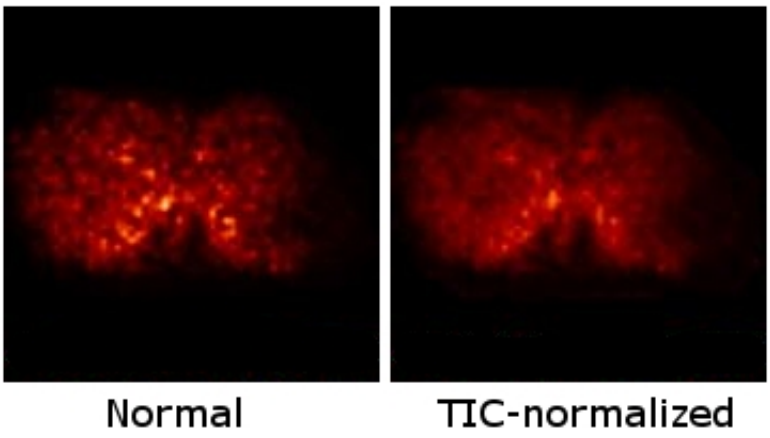

Figure 15: Spatial distribution after TIC normalization. The overall spatial distribution is smooth across the entire section after TIC normalization whereas the figure without normalization show signs of pixelation, i.e. large differences between adjacent pixels.

can be used for all experiments performed on the UltraFlex MALDI-MS and has been used in projects studying Alzheimer's disease and drug distribution in lung tissue with good results (unpublished data, not shown).

3.2 Part2

Part 1 regarded normalization within a tissue section with methods tested extensively with good results by others [2, 14]. There are however no methods available for normalization between tissue section. This section is divided in three parts where the first is regarding a bias occurring in almost all the tissue sections analyzed. The bias is that one hemisphere of the tissue sections show an elevated expression level for many proteins. This is not a

big problem for this experiment but for a normal experiment design (Section 1.1.1) this could cause problems. The second and third parts are the two normalization methods derived from discussions with a statistical expert (Dr. Ingrid Lönnstedt).

3.2.1 Right-left bias

The first thing to investigate was if there were any dissimilarities in the actual tissue. When the brain tissue was sliced there were hemoglobin present which could act as an ion suppression molecule [23]. These molecules can suppress other molecules ability to ionize which has a detrimental effect on those signal levels. The spatial distribution of hemoglobin and Pep19 of each tissue section was analyzed (Figure 16) and there were visible differences between the left and right side of the respective tissue section. The possible

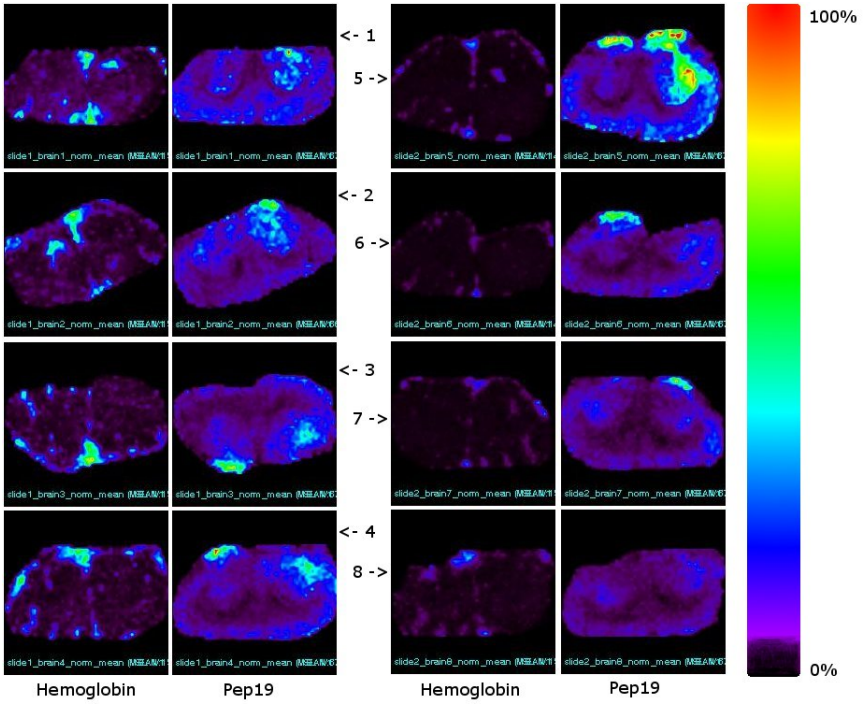

Figure 16: The spatial distributions of hemoglobin and Pep19 in brain 1 through 8 as placed on each slide. This means that 2, 5 and 4, 7 are consecutive sections as well. Please note that some of the sections should be transposed either horizontally or vertically to match its neighboring sections, e.g. tissue section 2 should be flipped horizontally to be directly comparable with tissue section 1.

right-left bias, caused by hemoglobin or the matrix application, was analyzed by looking at the right and left side of the brain and checking if the peak areas differed significantly between the two sides. The null hypothesis was

Table 1: Sections with a p-value smaller than 0.05 have significantly different peak areas between the right and left side on a 5% significance level. Those sections were considered to be in need of normalization.

Tissue section	T-test (p-value)	Normalize
1	0.025	Yes
2	0.382	No
3	0.00072	Yes
4	0.00037	Yes
5	0.015	Yes
6	0.00017	Yes
7	0.238	No
8	0.4584	No
9	0.0012	Yes
10	0.0038	Yes
11	0.00032	Yes
12	0.0031	Yes
13	0.284	No
14	0.074	Yes
15	0.000	Yes
16	0.027	Yes

that there should not be any difference between the two halves. A t-test was performed on the differences between the mean of the peak areas for each half and the results are displayed in Table 1. The t-test can be explained as follows: suppose there are N number of proteins detected. Then each half of the tissue section has N peak areas called $x_1, ..., x_N$ for the right half and $y_1, ..., y_N$ for the left half. The difference is then calculated as $d_1 = x_1 - y_1, d_2 = x_2 - y_2$ etc. and the null hypothesis is that these differences $d_1, ..., d_N$ are to be equal to zero. These results correspond with Figure 16 above where brain tissue section 7 and 8 seem more evenly distributed over the two sides. The difference seen in the intensity of Pep19 in tissue section 2 is harder to explain since the t-test showed no significant right-left bias.

The sections in need of right-left normalization (p-value lower than 0.05) were normalized by calculating, for all detected proteins, the local mean for each half. A hemisphere specific scale factor used to divide all peak areas (PA) for each half was calculated as the local mean (Equation 7) divided by the average of the two local means (Equation 13). This was done to preserve the signal intensities for each tissue section. The result of the normalization can be seen in Figure 17. The change is small but the points have shifted in general towards the line x = y.
The result is easier to see in Table 2 where the average and the sum of squares of the pairwise differences are shown.

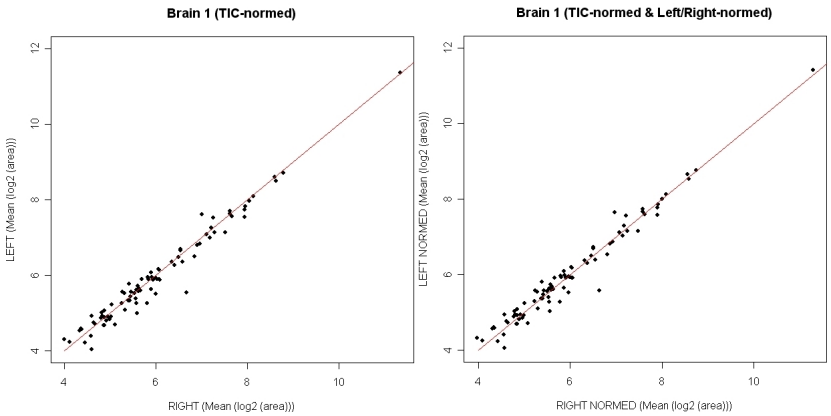

Figure 17: Each vector, x and y, consists of the mean peak area for every protein in the right side and left side of the tissue respectively. The normalization causes the points to be centered around the line x = y.

Table 2: The average is calculated from the difference between each peak area, i.e. the right peak area minus the left peak area. The average difference is much smaller after normalization than before. The sum of the squared distances from the line x = y has also decreased with normalization.

	Average pairwise difference	Sum of squares (distances from x = y)
Before normalization	0.053	5.32
After normalization	$2.38 * 10^{-16}$	5.07

3.2.2 Low-variance normalization

Finding proteins with low variance, of their peak area, across a tissue section is based on the idea of housekeeping genes, i.e. proteins which have similar expression levels throughout an entire tissue section. A protein with low variance presumably has the same expression level over the whole tissue section.

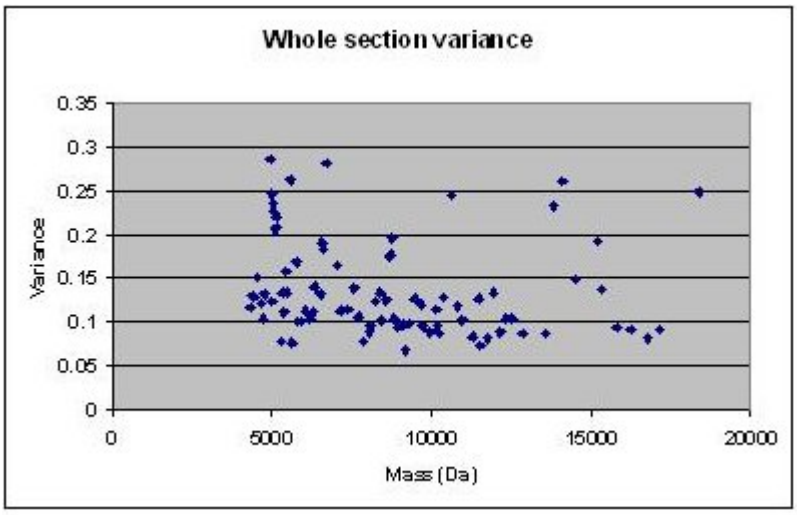

Figure 18: The blue dots represent all detected proteins for one of the tissue sections. The variances are calculated from the peak areas(PA) from each spectrum in the striatum (Equation 8).

The distribution of variance for all detected proteins can be seen in Figure 18 and the seven proteins with lowest average variance over all tissue sections were chosen for normalization. The same figure as above can be seen in Figure 19 but for all tissue sections instead of just one and for only the seven selected proteins.

Table 3: The variance before and after normalization where it would have been better if the variance would have been lower after normalization.

Dataset	Mean of variances before normalization	Mean of variances after normalization
all	0.133	0.142

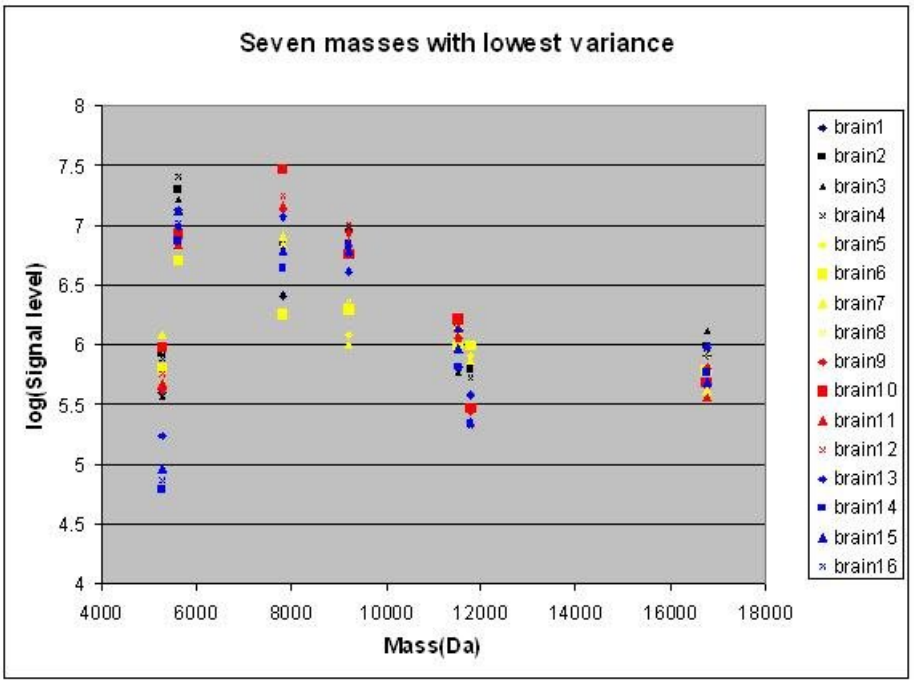

Figure 19: Each slide has a specific color, e.g. slide 1 has blue, and each tissue section has a different symbol. These are the seven masses the lowest-variance normalization has been built on.

After choosing these proteins, scale normalization was performed using the ratio between the local mean and the global as the normalization factor (Equation 13-14). The results of the lowest-variance normalization can be seen in Table 3.

The results show an increase in variance while a decrease in variance would have been preferable. This could be because, as shown in Figure 20, the local means (the blue dots and line) are not showing enough variation over the different tissue sections. This is however only a speculation and there might be other issues reasons that still are unknown.

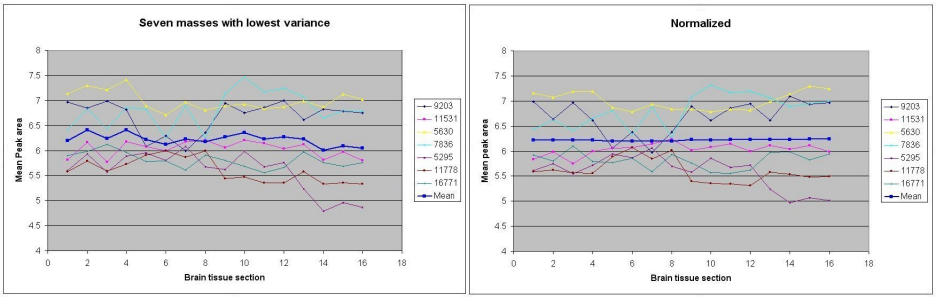

Figure 20: The seven selected peaks before and after low-variance normalization. The x-axes show the index of each brain tissue section and each tissue section has the mean peak area ($\overline{PA}$) of the seven peaks corresponding to the lowest variance masses on the y-axis. The blue line is the mean of the seven masses for each brain tissue section. The left figure represent the situation before low-variance normalization and the right after low-variance normalization.

3.2.3 Loess normalization

The loess normalization was based every detected peak (67 in total) and the peaks were investigates for trends in the dataset. In the range up to 10000 Da the loess lines seemed parallel. Therefore it was decided to normalize each section as described in Section 2.7.2.2 (Equations 18 and 19. The results of the normalization can be seen in Figure 21 and in Table 4. The loess lines are closer together in the low m/z range (0-10000 Da) which corresponds to the peaks used for normalization. The variance is lower in this range as well but higher in the higher range.

Table 4: The variance before and after loess normalization. Although the overall result is increased variance after normalization for the majority of the proteins it has decreased. Remember that the range between 0-10000 Da holds 75% of the proteins.

Dataset	Mean of variances before normalization	Mean of variances after normalization
All	0.133	0.134
0-10000	0.129	0.123
10000-20000	0.145	0.168

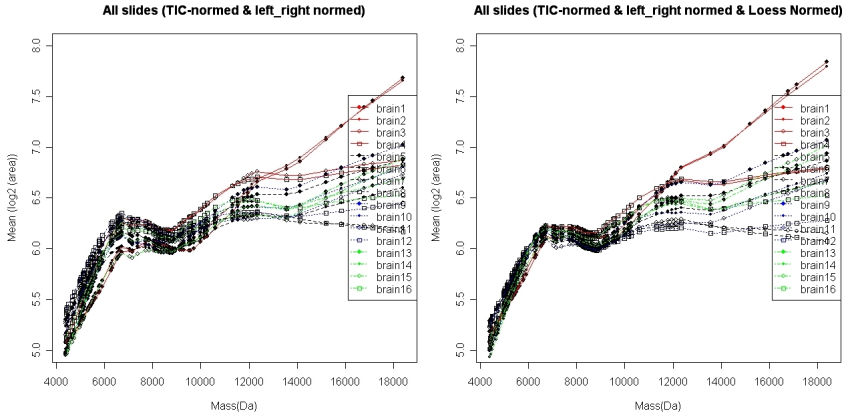

Figure 21: The loess fit before and after loess normalization. The lines are closer together in the m/z range used for normalization but further apart in the high m/z range. This indicates that the normalization has been successful.

3.2.4 Similarities between consecutive tissue sections

An interesting discovery is that the consecutive sections on the same slide usually group together (Figure 22). This means that consecutive sections, placed on the same slide, could probably be viewed as replicates which would make a statistical analysis of up/down regulation of proteins in animals induced with Parkinson's more reliable. This finding still needs further investigation, for example, if it holds for different types of tissue or if it only occurs in brain tissue.

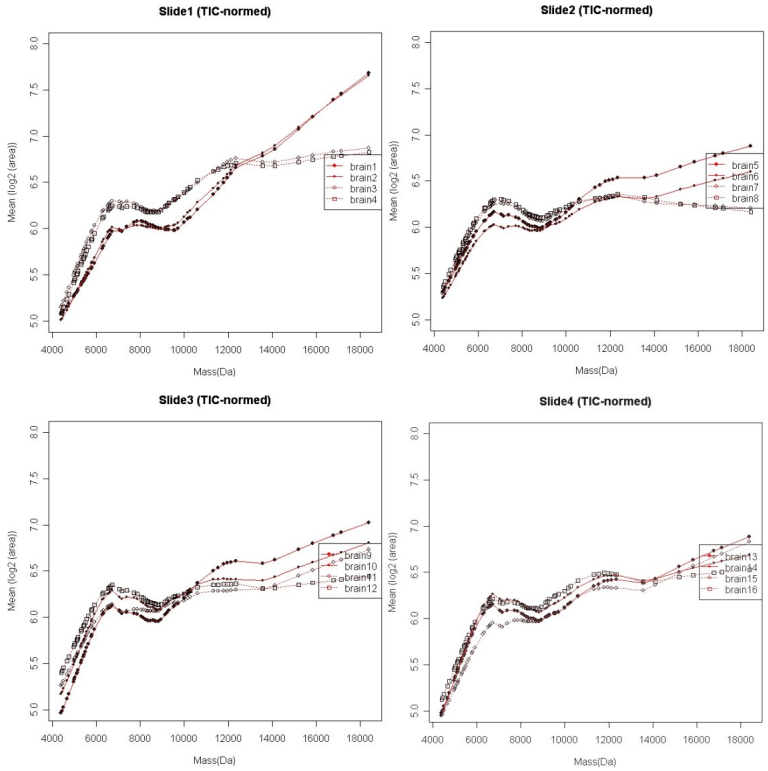

Figure 22: Similarities between consecutive tissue sections. Each curve corresponds to a tissue section and the points represent the loess fit of the mean peak areas. If examined carefully it can be seen that the consecutive sections, e.g. brain tissue section1 and 2, are more correlated than non consecutive sections.

4 Discussion

4.1 Normalization within a tissue section

The first part of the project regarding TIC normalization was successfully implemented. A working library of perl scripts able to make images from the files outputted by FlexAnalysis was created. There are however improvements left to implement, e.g. a filter for noise spectra (see Section 1.1 for more information on noise spectra). The TIC will be lower in a noise spectrum than in a normal which means that the ratio between the scalefactor and TIC, i.e. the ratio used for normalization, will be much larger than for a normal spectrum. This suggests that the noise spectrum will get larger signal levels than the surrounding normal spectra. Noise spectra can arise from spots where the tissue has cracked or if the tissue has detached from the surface.

There is a free software on the market called Analyze This! [6] which can create image files from the Bruker file format. The disadvantages of this software are that it distorts the mass scale if the number of data points has to be reduced and that the TIC normalization still has to be done. The script library created can handle both the creation of the image and the normalization but it can not handle data reduction, i.e. reducing the x-axis to less than 32767 points. The data reduction can be handled by BioMap but it is preferable to use datasets with less than 32767 data points to remove the need for reduction.

4.2 Normalization between tissue sections

The normalization on the seven masses with lowest variances did not work at all. The variances after normalization actually increased instead of decreased. The cause might be that the mean variances for the seven proteins for each tissue section are not varying as much as it should to use this normalization method. There are however two features in Figure 20 which are worth further investigation. The first is that the mean peak areas of sections 14-16 are lower than the other sections which are not seen as clearly after normalization. The second is that the variance of the seven peaks in sections 5-7 is lower than in the other sections. These two effects should be considered when normalizing the spectra and might improve these results if analyzed further. The loess normalization yields a slight improvement in the variance for the range which has been the base for the normalization. This normalization did not yield an overall improvement of the variance since the higher masses show a larger increase of the variances than the improvement in the lower m/z range (Table 4). These results are not as poor as it might appear since there is an improvement of the variance for the masses below 10000 Da which constitutes 75% of the detected proteins. There are several other ways of normalizing the loess lines which might improve the variance.

One example would be to subtract instead of divide the mean of the distance from the local loess curve to the global loess curve, where the global curve is the average of all the local loess curves. The possible advantage of using subtraction instead of division is that with subtraction the whole dataset will have the exact same shape as before normalization but just shifted a little on the y-axis. Another possible improvement would be to use all the data instead of just using the mean intensity of each mass for each tissue section when carrying out the loess regression. This could improve the loess regression since it will be more representative of the dataset the more data it is built upon.

4.3 Sample preparation

A source of variance I would have liked to investigate further is the possible bias caused by the ImagePrep, i.e. the instrument handling the matrix application. It was suspected that there might be a bias in the procedure in which the matrix gets distributed over the glass slides. Figure 23 is showing a possible distribution of the matrix during an experiment. This is not the actual distribution but an exaggeration of what could happen during an experiment.

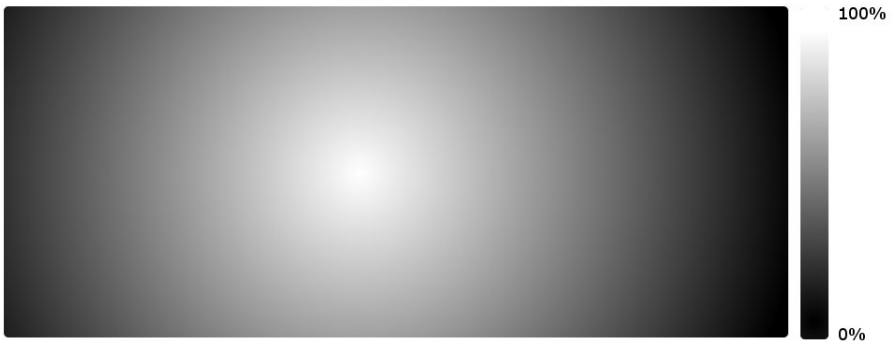

Figure 23: Possible thickness of the matrix layer when spray coating is used. Please note that the figure is created using computer software and not an actual image of a glass slide.

Unfortunately, there was not time enough to analyze if this was the case and if it would have an effect on the signal levels. The effect of this could be that there would be a signal intensity gradient in each tissue section which would make it even more difficult to draw any conclusions from this dataset.

4.4 Right-left bias

The right-left bias seen in the result can be an important issue if it occurs in real experiments (Section 1.1.1). If the two brain hemispheres are biased from start this inherent feature might corrupt the statistical analysis. It would be impossible to detect this effect since the normal experimental design (Section 1.1.1) would induce changes in one side of the brain. It would also be impossible to separate between the biological regulation caused by the dopamine denervation or by other variations in the sample. Although there is evidence for bias in the matrix application in the results, it is difficult to draw any conclusions at the moment.

4.5 Future work

The experimental design in the present study included biological tissue. However, it would have been preferable to have an experimental design with more control over the variables. It would have been really interesting to examine the normalization methods against an example with a known answer. A suggestion would be to analyze each step in the experiment closer. To analyze the matrix application, a striatum could be dissected and dissolved in some solution and spotted to a glass slide to ensure that each spot would have the same content.

Another approach with the present dataset could be to try to find the systematical errors based on the consecutive sections. The only difference between these sections should be systematical errors with only a small contribution from biological variance. This would probably yield two different levels of normalization where there would first be normalization within a slide and then normalization between slides. To analyze this further a recommendation would be to set up a new experiment with a brain not suffering from ion suppression effects caused by hemoglobin.

5 Acknowledgement

I would like to thank my supervisor Prof. Per Andrén for the opportunity to get a view in to this exciting field. My co-supervisor Ingrid Lönnstedt for her many thought on how to proceed with the mathematical parts of the project. I would also like to thank Malin Andersson, Anna Nilsson and Maria Fälth (also at the Department of Pharmaceutical Biosciences, Medical Mass Spectrometry) for all their knowledge in this field and help with writing the report.

References

[1] H. R. Aerni, D. S. Cornett, and R. M. Caprioli. Automated acoustic matrix deposition for maldi sample preparation. *Anal Chem*, 78(3):827–34, 2006.

[2] M. Andersson, M. R. Groseclose, A. Y. Deutch, and R. M. Caprioli. Imaging mass spectrometry of proteins and peptides: 3d volume reconstruction. *Nat Methods*, 5(1):101–8, 2008.

[3] T. M. Annesley. Ion suppression in mass spectrometry. *Clin Chem*, 49(7):1041–4, 2003.

[4] R. M. Caprioli, T. B. Farmer, and J. Gile. Molecular imaging of biological samples: localization of peptides and proteins using maldi-tof ms. *Anal Chem*, 69(23):4751–60, 1997.

[5] P. Chaurand, K. E. Schriver, and R. M. Caprioli. Instrument design and characterization for high resolution maldi-ms imaging of tissue sections. *J Mass Spectrom*, 42(4):476–89, 2007.

[6] S. Clerens, R. Ceuppens, and L. Arckens. Createtarget and analyze this!: new software assisting imaging mass spectrometry on bruker reflex iv and ultraflex ii instruments. *Rapid Commun Mass Spectrom*, 20(20):3061–6, 2006.

[7] W.S. Cleveland and S.J. Devlin. Locally weighted regression: An approach to regression analysis by local fitting. *J AM Stat Assoc*, 83(403):596–610, 1988.

[8] L. J. Dekker, J. C. Dalebout, I. Siccama, G. Jenster, P. A. Sillevis Smitt, and T. M. Luider. A new method to analyze matrix-assisted laser desorption/ionization time-of-flight peptide profiling mass spectra. *Rapid Commun Mass Spectrom*, 19(7):865–70, 2005.

[9] P. Dowling, L. O'Driscoll, P. Meleady, M. Henry, S. Roy, J. Ballot, M. Moriarty, J. Crown, and M. Clynes. 2-d difference gel electrophoresis of the lung squamous cell carcinoma versus normal sera demonstrates consistent alterations in the levels of ten specific proteins. *Electrophoresis*, 28(23):4302–10, 2007.

[10] M. Karas and F. Hillenkamp. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem*, 60(20):2299–301, 1988.

[11] A. N. Krutchinsky and B. T. Chait. On the nature of the chemical noise in maldi mass spectra. *J Am Soc Mass Spectrom*, 13(2):129–34, 2002.

[12] C. Laurent, D. F. Levinson, S. A. Schwartz, P. B. Harrington, S. P. Markey, R. M. Caprioli, and P. Levitt. Direct profiling of the cerebellum by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry: A methodological study in postnatal and adult mouse. *J Neurosci Res*, 81(5):613–21, 2005.

[13] C. H. Liu, Z. You, J. Ren, Y. R. Kim, K. Eikermann-Haerter, and P. K. Liu. Noninvasive delivery of gene targeting probes to live brains for transcription mri. *Faseb J*, 2007.

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

[15] R. Pang, P. Johnson, C. Chan, E. Kong, A. Chan, J. Sung, and T. Poon. Technical evaluation of maldi-tof mass spectrometry for quantitative proteomic profiling. *Clinical Proteomics Journal*, 1:259–270, 2004.

[16] J. Pierson, J. L. Norris, H. R. Aerni, P. Svenningsson, R. M. Caprioli, and P. E. Andren. Molecular profiling of experimental parkinson's disease: direct analysis of peptides and proteins on brain tissue sections by maldi mass spectrometry. *J Proteome Res*, 3(2):289–95, 2004.

[17] J. Pierson, P. Svenningsson, R. M. Caprioli, and P. E. Andren. Increased levels of ubiquitin in the 6-ohda-lesioned striatum of rats. *J Proteome Res*, 4(2):223–6, 2005.

[18] M. Rodriguez-Lanetty, W. S. Phillips, S. Dove, O. Hoegh-Guldberg, and V. M. Weis. Analytical approach for selecting normalizing genes from a cdna microarray platform to be used in q-rt-pcr assays: A cnidarian case study. *J Biochem Biophys Methods*, 2007.

[19] A. Savitsky and M. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Anal Chem*, 36(8):1627–1639, 1964.

[20] K. Skold, M. Svensson, A. Nilsson, X. Zhang, K. Nydahl, R. M. Caprioli, P. Svenningsson, and P. E. Andren. Decreased striatal levels of pep-19 following mptp lesion in the mouse. *J Proteome Res*, 5(2):262–9, 2006.

[21] G. K. Smyth and T. Speed. Normalization of cdna microarray data. *Methods*, 31(4):265–73, 2003.

[22] G. K. Smyth, Y.-H. Yang, and T. P. Speed. Statistical issues in cdna microarray data analysis. *Methods in Molecular Biology*, 224:111–136, 2003.

[23] C. I. Wu, C. C. Tsai, C. C. Lu, P. C. Wu, D. C. Wu, S. Y. Lin, and J. Shiea. Diagnosis of occult blood in human feces using matrix-assisted laser desorption ionization/time-of-flight mass spectrometry. *Clin Chim Acta*, 384(1-2):86–92, 2007.

[24] H. C. Zheng, H. Takahashi, X. H. Li, T. Hara, S. Masuda, Y. F. Guan, and Y. Takano. Downregulated parafibromin expression is a promising marker for pathogenesis, invasion, metastasis and prognosis of gastric carcinomas. *Virchows Arch*, 2007.

A Readme for the script library

Files included

Make Image

071002_MakeImg_MultipleFolders.pl
: Creates an image based on gathered spectra.

071106_MakeImg_MultipleFolders_SetLargest.pl
: Same as above but the largest value of the whole image will be set to 32677.

071105_MakeImg_OneFolder_SetLargest.pl
: Same as above but with the .dat files placed in a separate folder.

071217_MSMS_MakeImg_MultipleFolders_SetLargest.pl
: To be used when an MSMS experiment has been conducted, since the folder structure changes a little bit between different kinds of experiments.

Normalize image

071002_NormalizeImg_NoConstraints.pl
: Creates an image with the normalized intensities

071029_NormalizeImg_Constraint.pl
: Same as above but the intensities can not be scaled to more than 32677.

Requirements

Perl: ActivePerl-5.8.8.822
Can be found on "http://www.activestate.com/Products/activeperl/" Click "Get ActivePerl", then click "download" and "continue" (you don't have to sign up). Download the MSI package for windows (x86) or your appropriate OS. Make sure you install it under "C:\usr" to ensure compability with UNIX.

FlexAnalysis:
Can be bought from Bruker Daltonik.

User manual

When preprocessing the data in FlexAnalysis the following bit of code have to be added to the Method script. Place it just before the "end sub" command already in the method script. Have to be done since the file xy.dat which will be created is used in the perl scripts.

```
Dim strOutputFile$
Dim strSpectrumName$
strSpectrumName = Split(Spectra(1).Name)(0)
strOutputFile = Path + "\" + strSpectrumName + "\xy.dat"
ResultSpectra(1).Export(strOutputFile, 2, 0)
```

When all the spectra have "scan\0_R00X015Y004\1\1SLin\xy.dat" (important that the path looks similar to this, otherwise it will not work) you can run the perl scripts by typing: "perl 071002_MakeImg_MultipleFolders.pl" in the command prompt.
Type "perl 071002_NormalizeImg_NoConstraints.pl" when you have an existing image that is to be normalized and follow the instructions on screen.

Note
Run perl scripts in windows prompt (opened by clicking on the start button at the lower left corner of your screen) and then click run and type 'cmd' to start windows command prompt. Then you only have to follow the instructions printed to screen to build your image.

Performance
Expected running time
FlexAnalysis
    Batchprocess: 1000 spectras in less than 1 hour
Perl
    071002_MakeImg_MultipleFolders.pl
        1000 spectra in 4-6 minutes
    071002_NormalizeImg_NoConstraints.pl
        3MB per second