Qualitative Comparison of Normalization Approaches in MALDI-MS

OSCAR HAGLUND

KTH Computer Science
and Communication

Master of Science Thesis
Stockholm, Sweden 2008

Qualitative Comparison of Normalization Approaches in MALDI-MS

O S C A R H A G L U N D

Master's Thesis in Biomedical Engineering (30 ECTS credits)
at the School of Computer Science and Engineering
Royal Institute of Technology year 2008
Supervisor at CSC was Jens Lagergren
Examiner was Stefan Arnborg

TRITA-CSC-E 2008:038
ISRN-KTH/CSC/E--08/038--SE
ISSN-1653-5715

Royal Institute of Technology
School of Computer Science and Communication

KTH CSC
SE-100 44 Stockholm, Sweden

URL: www.csc.kth.se

ABSTRACT
QUALITATIVE COMPARISON OF NORMALIZATION APPROACHES IN MALDI-MS

Matrix Assisted Laser Desorption Ionization Mass Spectrometry is a technique that measures the abundance and weight of particles by subjecting them to a magnetic field. There are very few comparisons of the different forms of normalization approaches that exist in MALDI-MS. By applying different normalization methods to a series of spectra with known peptide concentrations it is possible to calculate the errors of prediction for normalizations as well as for when no normalization is used. The result of this study is that normalization reduces the prediction error by up to 50%, and even more for individual peptides, depending on the technique used. However, the repeatability errors are still very large with coefficients of variation (relative standard deviation) of over 30% on repeated measurements of the same concentration.

SAMMANFATTNING
KVALITATIV JÄMFÖRELSE AV NORMALISERINGSMETODER I MALDI-MS

Matrix Assisted Laser Desorption Ionization Mass Spectrometry är en teknik som mäter mängden och massan hos partiklar genom att utsätta dem för ett magnetfält. Det finns väldigt få jämförelser mellan de olika normaliseringsmetoder som används innom MALDI-MS. Genom att applicera olika normaliseringar på en serie spektra med kända peptidkoncentrationer så är det möjligt att räkna ut prediktionsfelen för koncentrationsbestämmningarna för olika normaliseringar samt när ingen normalisering används. Resultatet av denna undersökning är att normalisering minskar prediktionsfelet med upp till 50%, och ännu mer för individuella peptider, beroende av vilken teknik som används. Repeterbarhetsfelen är dock fortfarande mycket stora med en variationskoefficient (relativ standardavvikelse) på över 30% för upprepade mätningar av samma koncentration.

PREFACE

This master's project was conducted at NADA, Royal Institute of Technology. Commissioner of the project, and also supervisor, was Dr Jenny Forshed, Karolinska Institute, and most of the work was being done at Karolinska Institute and Karolinska University Hospital. My supervisor at NADA was Professor Jens Lagergren.

I would like to thank Dr Jenny Forshed for all the help provided during the entire project, as well as Professor Jens Lagergren for his help with the thesis. Additionally I would like to thank Maria Pernemalm for conducting the laboratory parts of the experiment, as well as writing section 2.2-2.4 in the thesis. Finally I want to thank Scott and Lori McFarlane for proof reading my thesis.

Table of Contents

1 Background

This section contains brief descriptions of some of the basic concepts important for understanding this thesis, such as: biomarkers, mass spectrometry and normalization. The section also contains the aims and goals of this master's project.

1.1 Biomarkers and Cancer

Cancer, like other diseases, leaves traces in the body, and by finding some of these traces an early diagnosis may be possible. Having an early diagnosis is important since the survival rate of subjects decreases the longer the cancer remains untreated. The process of looking for specific diseases in subjects who are not yet diagnosed is called *screening*. However many screening procedures, like endoscopic examination of the colon, are invasive, uncomfortable and sometimes even hazardous [1]. It is therefore of importance to find other simpler and less invasive methods of screening. One such method could be using Mass Spectrometry (MS) to look for so called biomarkers in blood samples. Methods to do this are still in a development phase, but there are successful examples, such as A. Martin *et al*. [1] who used a computer program coupled with Mass Spectrometry to separate healthy and sick subjects with 95% sensitivity and 91% specificity using normal blood serum. In a more recent study published in Nature Medicine, Alzheimer patients were separated from non-Alzheimer subjects [2] and shows that this is a field to be reckoned with in the future.

A biomarker is a substance usually indicating disease or increased risk of disease. If we know that a substance X is prevalent in three times the normal concentration in subjects with a certain form of Cancer we can use that knowledge to readily find out if a patient has this form of Cancer or not. Hence it is an important task both to find new biomarkers as well as finding techniques for detecting these markers in subjects.

1.2 Mass Spectrometry

One form of biomarkers is proteins. Mass Spectrometry is commonly used for identification of proteins, and it is even possible to measure their absolute or relative concentrations using this technique.

A Mass Spectrometer is an instrument that measures the mass, or more correctly, the mass per charge (m/z), of charged particles. The system is based on the fact that charged particles are deflected or accelerated at different speeds in a magnetic field depending on the charge and mass of the particle. A heavy particle accelerates slower than a light particle, assuming the same charge, and as such will take longer to traverse a certain distance. By measuring this time, it is easy to calculate the mass per charge ratio (m/z) of the particle. This technique is called Time of Flight (TOF) and is used in the system I will be examining: Matrix Assisted Laser Desorption/Ionization-Mass Spectrometry (MALDI-MS) [3].

The result of a run on a MALDI-MS machine is a series of m/z and intensity values. For every m/z value there is an intensity value that is proportional to the amount of hits on the detector, which in turn is proportional to the amount of particles with that mass in the sample. The unit for the intensity scale is variable between different mass spectrometers and the unit is arbitrary. Mass per charge is measured in daltons (Da) per elemental charge. Dalton and the atomic mass unit are equivalent. *Table 1* is an example of the raw data of a mass spectrum.

Table 1 Shows the 3 first pairs of values of a MALDI spectrum with a total of 124 489 pairs. By plotting all values with m/z on the x-axis and the intensity values on the y-axis we will end up with the spectrum in (Figure 1).

m/z	**Intensity**
699.908277	35.6863
699.923954	44.7059
699.939632	52.1569

The peaks visible in the MS-spectrum are usually not complete proteins but peptides which are the building blocks for proteins. By examining which peptides are present in a sample and their concentrations, it is often possible to calculate what or which proteins they come from. The spectrum in *Figure 1* shows that at 2900 m/z there is a substance with a very high concentration. Very simplified we could say that if the peptide at 2900 m/z only exists in high concentration levels in patients with cancer, it would be a biomarker, and that any high concentration levels at 2900 m/z would be an indication of possible cancer.

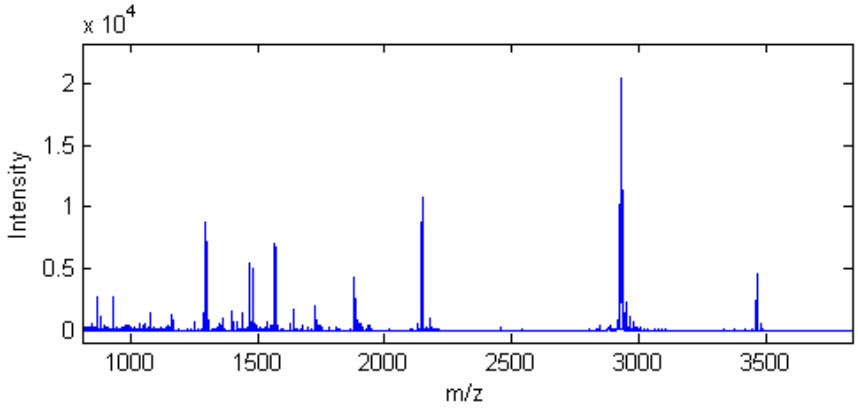


Figure 1 This is a typical MALDI mass spectrum used in the present study. It shows the intensity being plotted against the mass-per-charge ratio and consists of over 120 thousand pairs of values.

Figure 2 shows an enlarged view of the peaks in the 1400 m/z region of the same spectrum as *Figure 1*, and we can see that what seems like one peak in *Figure 1* in reality consists of several sub-peaks. The reason for this behavior is that all of the atoms in the peptides are naturally occurring isotopes with different masses. A peptide containing heavy isotopes will hence register slightly higher on the m/z scale than a peptide with fewer heavy isotopes. A short peptide (that has few atoms) will have a different isotope pattern than a

large peptide since some uncommon combinations of isotopes will have a lower probability to occur. The result is the slight displacement of the peaks which form a characteristic isotope pattern.

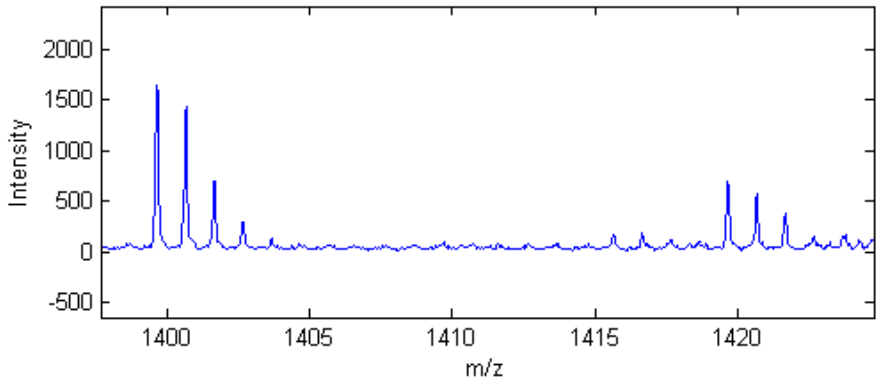


Figure 2 Here the same spectrum as in Figure 1 is shown but zoomed in around the 1400 m/z area

An important aspect to remember here is that it is the sum of all the isotope intensities, the area under the peaks, which is proportional to the real concentration of the substance and not just the height of the highest isotope peak. However this measurement is often used because it is considered proportional to the area and easier to measure.

Another thing of notice in the spectrum is the background noise that among other things is caused by electric interference in the machine.

In *Figure 3* it is possible to see a region of the spectrum from *Figure 1* where the isotope patterns are not as clear as previously. This part also shows a situation where the resolution of the MALDI-MS is not able to clearly distinguish between peaks, as we can see that the resolution is not good enough to allow the intensity to drop to 0 between the peaks in the middle. In an ideal case with infinite resolution all peaks would consist of a single "spike" like value but as shown in *Figure 3* this is not the case even with high performance machines.

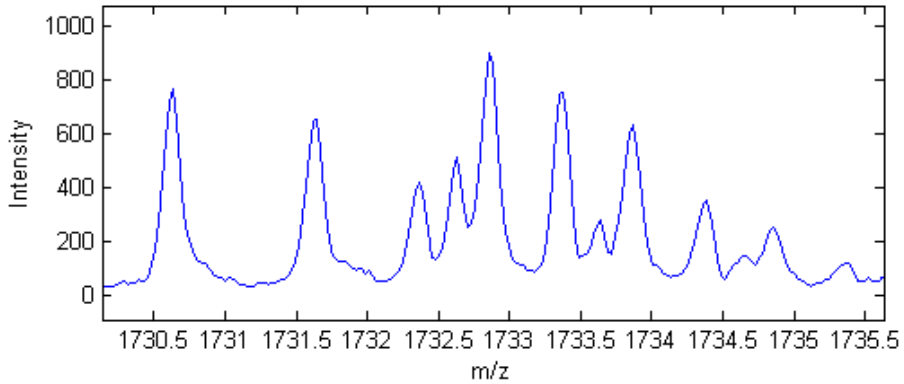
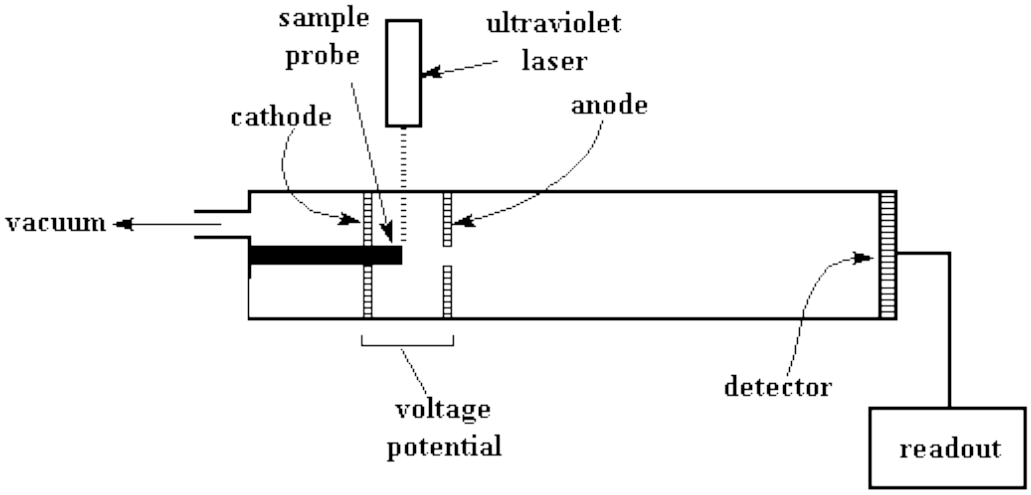


Figure 3 is a more zoomed-in part of the spectrum that Figure 2 shows. Here the form of all the individual peaks is more clearly visible

1.3 MALDI-MS

There are many different kinds of mass spectrometers with different functionality and operations. The biggest differences between the different types are the procedures used to ionize the sample, which is how the substances in the sample become charged and how they are released from the surface. Matrix Assisted Laser Desorption/Ionization Mass Spectrometry (MALDI-MS) is a technique that, like the name implies, is based on imbedding the sample in a matrix and then "firing" a laser on this matrix. The laser will cause both the sample and the matrix to be released (desorbed) from the surface. Ionization is achieved in the preprocessing step by placing the sample in a solution. Figure 4 shows the basic inner workings of a MALDI-TOF-MS machine.



A simplified diagram of a MADLI apparatus
(After Creel, H., *Trends in Polym. Sci.*, 1993, *1(11)*, 336-342.)

Figure 4 This figure represents a MALDI-TOF-MS setup. The Laser releases sample and matrix from the sample probe and is then accelerated by the voltage potential. The time between firing and the subsequent hits on the detector can be used to calculate the weight of the particle.

The matrix that the sample is embedded into is also ionized, and the signal is picked up by the detector and causes a background signal that is not related to the sample. The matrix signal which can be seen in *Figure 5* is located in the low-mass region of the spectrum. This signal causes noise and it is normally removed using something called baseline correction.

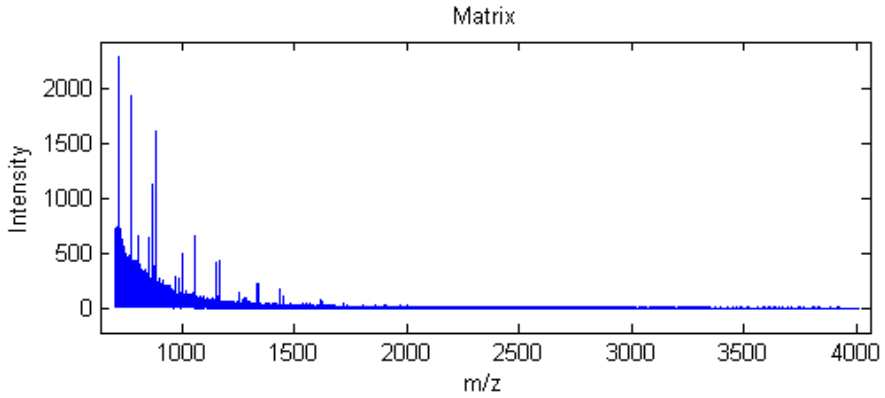


Figure 5 MALDI spectra where no substances apart from the matrix are present meaning that this signal is only coming from the matrix. If a sample was to be added this signal would be superimposed on top of the sample signal causing noise.

One of the biggest advantages of using MALDI over other MS techniques is that the ionization is mild and does not break down the sample to the same extent as other methods. Another important characteristic is that it almost exclusively creates singly charged ions, which mean that the m/z value becomes a "very real" measurement of the actual mass of the particles. By combining MALDI with a Time of Flight detector (MALDI-TOF-MS), it is also possible to examine relatively large molecules (5000 Da using a reflectron and 250 000 Da without one). Finally the technique is sensitive and is capable of detecting molecules at amol (10^{-18}) concentrations.

Molecules' competing over ionization energy is a potential problem in quantification by mass spectrometry. There is only a finite amount of energy in a laser beam, and all the molecules in the sample need to compete over this resource. Some molecules are easier to ionize than others, and this can create a bias towards certain substances. The result might be that because one molecule is better at absorbing energy, it gives a higher intensity value in the spectrum relative to another molecule with the same concentration.

1.4 Normalization

Intensity values in a spectrum may vary greatly between different measurements on the same substance and/or sample. Some of the reasons for this are that the samples may have been deposited differently on the surface or perhaps most importantly because of discrimination of some substances depending on the sample matrix as stated above. Hence it is not reasonable to directly compare two spectra with one another with regards to concentration.

To compare the two spectra in *Figure 6,* we need to perform a normalization (sometimes called standardization). There are many techniques for normalization available, and they all strive to make spectra comparable to each other. For this example it would mean that ideally the peaks in sample 8 and sample 10 shown in *Figure 6* would be of the same intensity.

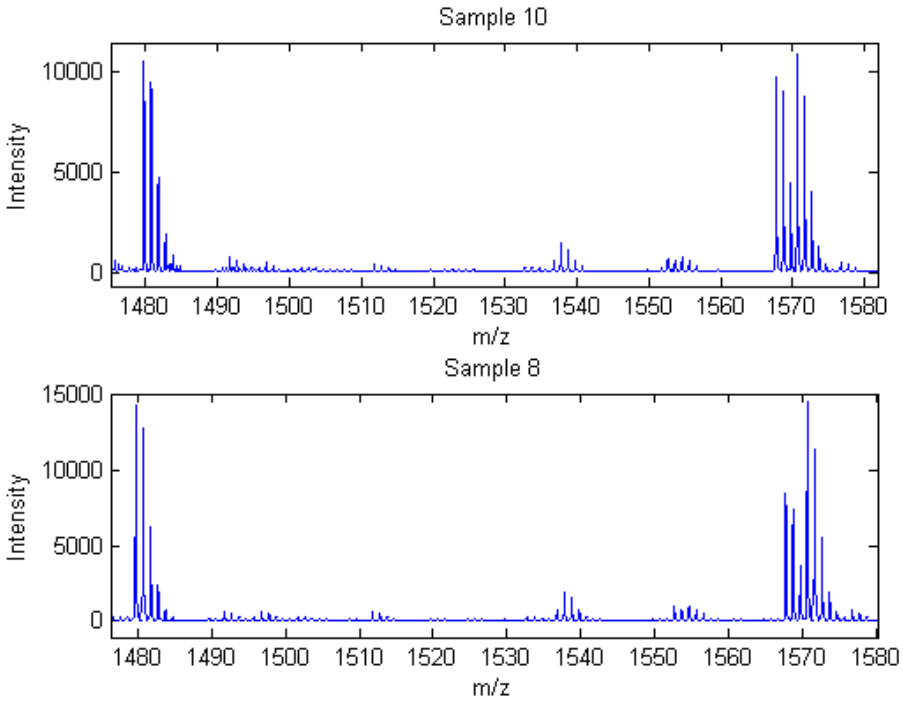


Figure 6: Two peaks in two spectra of the same peptide mix but with very different intensity values are shown. Sample 10 has 50% higher intensity values than sample 8 despite having the same concentration.

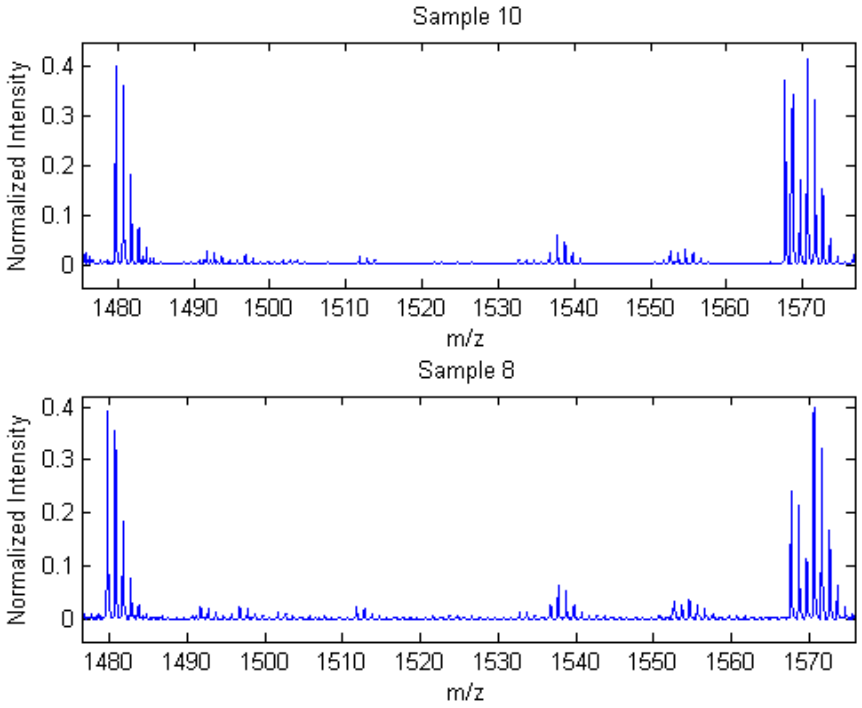


Figure 7: This is the same spectra as in Figure 6, but this time they are normalized with a specific normalization method called LOWESS.

The normalized intensities shown in *Figure 7* are much more similar in intensity than in the un-normalized case *Figure 6*. Although some differences still remain, it is reasonable to draw the correct conclusion that the two samples are of the same concentration.

Normalization methods can be split into two major subcategories, *global* and *local* normalization. Global normalization means that all features in the spectrum are used simultaneously to compute a single normalization factor between two spectra [4]. A global normalization of *Figure 6* could, for example, be to multiply sample 8 by 1.5, where 1.5 could be the ratio of the summed intensities between the two spectra. Contrary to the global variant, local normalization means that subsets of spectral features are used to normalize different parts of the spectrum. The Total Ion Current Normalization (TIC) is one of the more common types of global normalizations in MALDI-MS. This normalization makes the total amount of detected ions (the sum of all intensity values) equal in all spectra. An example of a local normalization is the more complicated Locally Weighted Scatterplot Smoothing (LOWESS) normalization. However there are methods where there is some ambiguity in this definition. The standard deviation of noise normalization, for example, does not operate on the entire spectrum but rather a subset of the spectrum that contains only noise, but it still uses a single normalization factor applied to the total spectra like the traditional global normalization. For this thesis, normalizations such as these will still be counted as global as long as they normalize the entire spectrum by a single factor.

These normalized spectra can now be used for many different purposes such as identification of peptides and proteins. In this work we use the spectra to calculate predicted concentrations of certain peptides by examining the intensity in certain regions of the spectra that correspond to a particular peptide. As shown earlier in this chapter (*Figure 6* and *Figure 7*), intensity values in different spectra will be comparable after normalization and hence concentrations can be compared by examining the normalized intensities.

1.5 Overview of the thesis project

Different methods for normalization obviously give different results, and it is hard to know which technique really is the best. There is a lack of systematic examinations of different normalization methods in the literature, and due to this lack of study it is hard to say which methods are the best [4]. Hence the goal of my thesis is to try to bring some clarity into this matter.

One of the major drawbacks of existing studies of normalization is that they do not actually measure the correctness of the normalization but rather the coefficient of variance (CV) [5]. A low CV value is certainly a good thing, but it does not really tell us anything about how accurately the normalization reflects the true concentrations of the sample. Because of this, it was important for us to actually measure how good the normalizations were in predicting the correct value, and we designed our experiments with this in mind.

By creating a mix of known protein or peptide concentrations and then analyzing this with the mass spectrometer we received a series of spectra where we knew the exact true concentration of each peak. These spectra were then subjected to different normalization techniques each yielding a series of new normalized spectra. To compare these different normalized spectra with each other we used cross validation on the concentration curves created by the peptide intensities. From these concentration curves it was possible to calculate the predicted concentration of the different peptides and compare this value to the known measured concentration. The smaller the overall difference between measured and predicted concentration for the normalized spectra compared to the non-normalized spectra the better the normalization.

2 Sample preparation

This section of the thesis details the steps leading up to the normalization. First, peptides need to be selected and concentrations decided upon. After that the sample mixtures are created and analyzed by the MALDI-MS. Finally the spectral data from the MALDI-MS is pre-processed where things like peptide peak regions are calculated.
Sections 2.2-2.4 are written by Maria Pernemalm.

2.1 Peptide samples

Eleven samples were prepared, each containing five different peptides. In addition to these peptides, a Bovine Serum Albumin (BSA) protein digest was added to function as a background to simulate a sample matrix in a biological sample.

Sample 0-6 would have amounts (also referred to as concentrations as all substances were diluted into the same amount of water) of the different peptides ranging from 0-1 in relative amounts, while sample 7-10 would have all peptides at 0.5 amounts. Because we want to create calibration curves, none of the peptides were allowed to have the same concentration in two samples in sample 0-6. The concentrations 0, 1/6, 1/3, ½, 2/3, 5/6 and 1 were randomly distributed between samples with regards to the constraint (Table 2). Samples 7-10 form a repeatability study to examine the spread between samples of equal concentration (Table 2). The repeatability study included sample preparation, technical and instrumental variance.

Table 2. This table shows the relative amounts of the different peptides in the different samples. In addition to the figures mentioned here, each sample also contains digested BSA.

	Peptide amount				
Sample	A	B	C	D	E
0	1	1/3	1/2	0	5/6
1	0	1/2	2/3	1/2	1
2	1/3	1	0	2/3	1/3
3	1/2	0	1	5/6	1/6
4	1/6	5/6	5/6	1	0
5	2/3	1/6	1/6	1/3	1/2
6	5/6	2/3	1/3	1/6	2/3
7	1/2	1/2	1/2	1/2	1/2
8	1/2	1/2	1/2	1/2	1/2
9	1/2	1/2	1/2	1/2	1/2
10	1/2	1/2	1/2	1/2	1/2

Table 3 shows the peptides that we used in the study. These peptides are commonly used as standards [6] and have good signal properties. In *Table 2* we can see the distribution of concentrations in the different samples that were used in the study.

Table 3. This table shows the peptides studied and their masses.

Peptide	Name	m/z
A	Angiotensin	1296,5
B	Glufibrino peptide B	1570,6
C	Dynorphin A	2147,5
D	ACTH	2933,5
E	Beta-Endorphin	3465

Once the sample mixtures were created, each sample was analyzed on the MALDI-TOF-MS, generating one spectrum per sample. Spectra from only digested BSA and from the matrix without any sample were also run, to get a picture of the sample matrix and the background of the MALDI-MS runs respectively.

2.2 Sample matrix background

1.5 mg of Bovine Serum Albumin (BSA) was dissolved in 50 mM Ammonium Bicarbonate, providing an optimal environment for trypsin, the digestive enzyme used in this reaction. DL-Dithiothreitol was added to the sample to an end concentration of 5 mM followed by an incubation at 60 ºC for 30 minutes, in order to reduce cystein bridges, making sure that the protein will only be present in its linear form. The sample was then let to cool to room temperature, and Iodoactetamide was added to end concentration 15 mM, thereby blocking the cysteins from re-forming any cystein bridges. The sample was then incubated in the dark for 30 minutes. 7 µg of trypsin was added, and the sample was digested over night, cleaving the protein into peptides.

2.3 Preparation of samples

Lyophilized A) Angiotensin, B) [Glu^1]-Fibrinopeptide B, C) Dynorphin A, D) Adrenocorticotropic hormone (ACTH) and E) β-Endorphin were each dissolved in MilliQ water to 10 pmol/µl. BSA concentration was calculated to 263.16 pmol/µl. Samples were prepared separately as described in *Table 4* below.

Table 4: Description of sample composition. Amount constituent given in µl.

	Peptide (µl)						
Sample	**A**	**B**	**C**	**D**	**E**	**BSA digest**	**MilliQ**
0	10	3.33	5	-	8.33	0.38	72.96
1	-	5	6.67	5	10	0.38	72.95
2	3.33	10	-	6.67	3.33	0.38	76.29
3	5	-	10	8.33	1.67	0.38	74.62
4	1.67	8.33	8.33	10	-	0.38	71.29
5	6.67	1.67	1.67	3.33	5	0.38	81.28
6	8.33	6.67	3.33	1.67	6.67	0.38	72.95
7	5	5	5	5	5	0.38	74.62
8	5	5	5	5	5	0.38	74.62
9	5	5	5	5	5	0.38	74.62
10	5	5	5	5	5	0.38	74.62
11	-	-	-	-	-	0.38	99.62

2.4 MALDI TOF MS

2 mg/ml α-Cyano-4-hydroxycinnamic acid (CHCA) was prepared freshly in aqueous solution containing 30% (v/v) Acetonitrile and 0.01% (v/v) Tri fluoro-actetic acid. CHCA was added to the MALDI target using sandwich model, spotting first 0.5 µl CHCA, followed by 1 µl sample and then 0.5 µl additional CHCA. Each sample was spotted one time on the plate except for Sample 11, which was spotted on five different spots (named BSA 1 -5). In addition five CHCA blanks were spotted in parallel with the samples (named matrix blank 1-5), and 8 spots containing peptides mass standards were also added to the plate. Samples were analyzed in an Applied Biosystems 4800 MALDI TOF TOF Mass spectrometer. Before analysis the plate's calibration settings were updated through external mass-calibration (m/z) using the peptide mass standards. The spotted samples were then analyzed, each one spot making rise to one spectrum. 1000 laser shots were fired on each spot, and the output of the analysis is the average signal of these shots over the mass range 700-4000 m/z. Each spectrum was then externally calibrated using the previously described plate calibration.

2.5 Data Preprocessing

The result from running the samples on the MALDI is a series of eleven spectra as well as a five BSA digest and one matrix reference spectra.

Several preprocessing steps are required to prepare this data for normalization. The first observation is that the lengths of the spectra vary slightly. Spectra 0-1 contain 124 489 data pairs (pairs of intensity and m/z values, see Table 1), while spectra 2-4 contain 124 488 pairs, and finally spectra 5-10 contain 124 487 pairs. The normalization methods are easier to implement if all spectra contain the same amount of pairs, and I solved this problem by simply

removing the last data pairs so that all spectra has 124 487 pairs. Since the last pairs of all spectra do not contain any information, this will not influence the further data analysis.

As can readily be seen, apart from noise, a spectrum consists of peaks, valleys and long stretches of no signal. Some normalization methods in the literature use the entire spectrum with all its features, while others operate solely on peaks. Peak detection can be performed by the use of specialized software, but in this work I decided to perform the peak detection by hand. The reason for this was the often non-optimal performance of peak detection algorithms, and also because of the fact that the spectra in this work consists of relatively few peaks which also appeared at the same place in each spectra. This facilitates manual peak detection.

Throughout this work I will use indexes of the intensity values of the spectra instead of the m/z values since the indexes are easier to directly compare across all spectra, and the spectra align better if compared on a per index basis as compared to m/z values (*Figure 8*). Indexes are also easier to access from a software development perspective.

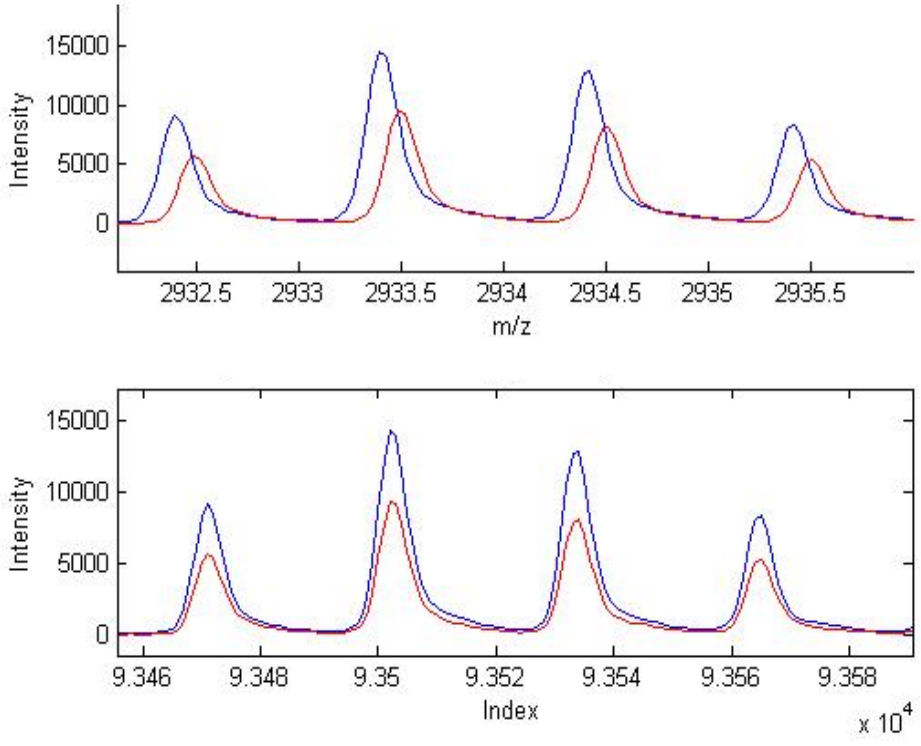


Figure 8: This figure shows how spectra 0 and 10 align better if they are plotted against their index (their position in the paired array) instead of m/z.

Overlaying several spectra on top of each other, I was able to easily determine peak regions and to single out 50 different peaks belonging to the peptides and the BSA background. Isotope peaks are separated from each other in this definition of a peak (*Figure 9*) so that the low intensity values in the valleys

are not calculated as part of the peak. The 50 peaks are using the same index values that describe the width of the peak at the baseline in all spectra even though the actual peak width at the baseline might vary across spectra. The reason for this is that that some of the normalization methods require peaks to be of the same width. The data for the peak shown in *Figure 9* looks like this in MATLAB:

```
p28 = [57071 57101; 57111 57142; 57150 57181; 57189 57215;
57230 57250; 57270 57280];
```

These represent the pairs of indexes of which in-between, intensity values are attributed to the peak. That is, the intensity values from index 57071-57101 are part of the peak, as are 57111-57142 and so on.

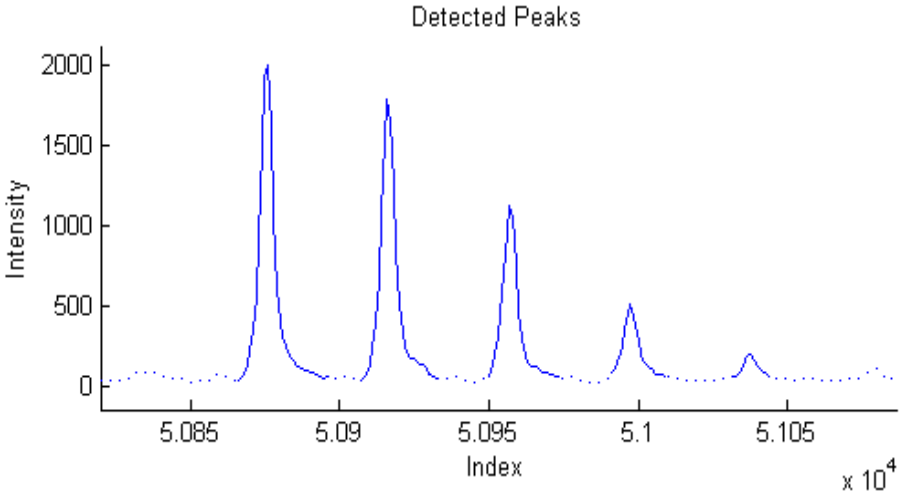


Figure 9 The image shows the isotopic pattern of one of the 50 detected peptide peaks, the isotope peaks are separated so that only actual peak data is included. The solid line represents the intensity values that are part of the peak definition.

Examining the peaks for the different peptides across the different spectra showed a rather large anomaly in one of the spectra. *Figure 10* shows this anomaly, namely that peptide C is hardly visible in spectrum 0, even though it is supposed to be of the same concentration as in the other visible spectra that is also shown in the picture. There is already a big spread in the data with the second lowest values being 2.6 times lower than the highest, but the values from spectra 0 are over 12 times as low as the second lowest, which is a far greater relative difference. It is also important to note that all the other peaks in the relevant spectra were of normal intensity. Because of this, spectrum 0 was removed from the batch. It is unfortunately not possible to disregard this particular peak since that would have an adverse effect on several of the normalization methods. This error most likely occurred in the sample preparation.

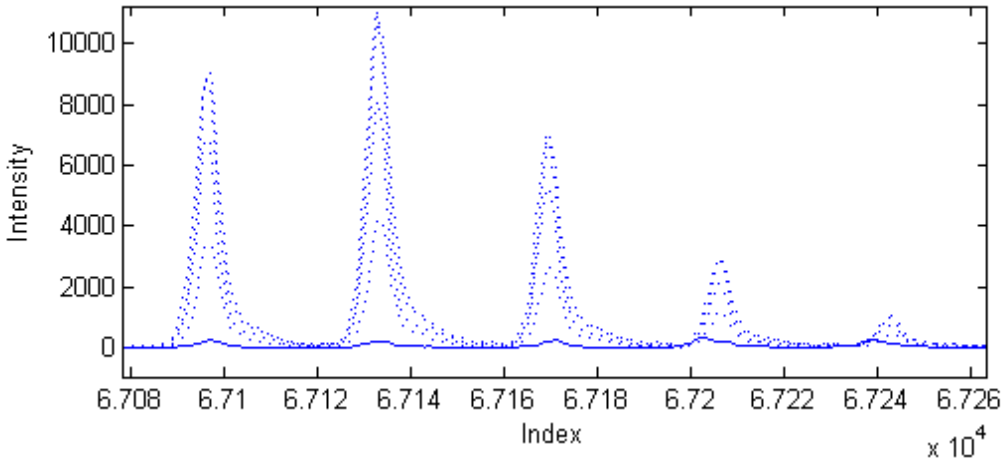


Figure 10 This image shows peptide C in spectra 0 (solid line) and 7-10 (dotted lines). The concentration of this peptide is supposed to be the same in all five spectra.

3 Normalization methods

Previous work showing comparisons between different normalizations are rare [5, 7], while articles showing new normalization methods are more common [8, 9, 10, 11], although these articles often contain comparisons with similar or competing techniques. In addition to methods designed specifically for mass spectrometric data, methods originally developed for other instruments like micro arrays are also relatively common in the literature [5, 4, 12].

This section of the thesis describes all of the normalization approaches used in the experiment. These methods are: Total ion current, Median, Internal standard, Standard deviation of noise, Cumulative intensity, Linear regression, LOWESS, Quantile and Top-L.

Almost all of the different normalization techniques in this experiment are tested in different variants for several reasons. First, it is not always stated in the references exactly how the normalization is performed, and hence I tried the different perceivable variants. There are also sometimes different variants of how to apply the normalization. Finally I found it interesting to see what would happen if the originally presented conditions of the normalization were tweaked to see how this would influence the results. Some of the methods may be applied in a way that they were not designed for and hence give erroneous results; this may, however, also give interesting outcomes, because it shows what not to do.

After each normalization described below has been performed, all spectra are scaled so that the highest intensity value in the entire set of spectra is 1. The number 1 is arbitrary, but the scaling allows for easier comparison between normalizations.

3.1 Spectral subsets

All normalizations tested in this study are performed on a subset of the raw spectra. The subset is the part of the spectra that the normalization is performed on. This thesis uses the three different subsets explained below.

3.1.1 Entire spectrum

“Entire spectrum” means, as the name suggests, that all data points, both those in the peaks and in the noise, in a spectrum are used for calculating the normalization coefficient. For the spectra used in the current study, this means 124 487 data points, which is the length of the entire raw spectra.

3.1.2 Peak regions

Normalization using only the peaks regions that were extracted in the pre-processing step (2.5 Data Preprocessing) means that instead of calculating the normalization coefficient by using the entire spectrum, only the subset of the spectrum that includes the different peaks regions is used. The subset used here brings down the number of data points to 5 000.

3.1.3 Peak heights

The final subset of the spectrum used for normalization is to only use the max intensity value for each isotope peak. The final subset consists of 202 data points.

3.2 Total Ion Current

"Total area under the curve", or "Total Ion Current" (TIC) as it is most often called, is probably the most common normalization technique for MALDI data. Among its benefits is that it is both simple to understand, simple to implement and fast. The underlying assumption for TIC is that "we assume that on average, the number of proteins that are being over expressed is approximately equal to the number of proteins being under expressed and that the number of proteins whose expression levels change is few relative to the total number of proteins bound to the surface"[13]. This means that the total intensity of the spectrum should be proportional to the total concentration in the sample, and that if there is an equal amount of sample, then the summed intensity in the spectrum should be the same across all spectra. If it is not already equal, then the method multiplies each individual spectrum with a normalization factor so that they all end up with the same total intensity.

The variants listed below are the different approaches to TIC normalization tested in this work.

TIC1 – TIC intensity normalization on the entire spectrum
TIC2 – TIC intensity normalization on peak regions
TIC3 – TIC intensity normalization on the peak heights
TIC4 – TIC intensity normalization on peak regions also normalizing on total peptide amount in spectra.
Looking at the experimental setup we can easily see that for this experiment not all samples have the same total amount of peptides. In spectrum 4 there are 2.83 amounts of non BSA peptides, while in spectrum 5 there are only 1.83 amounts. If intensity is directly related to concentration then there should be a higher total intensity in spectrum 4 than in spectrum 5 (by a factor of 1.55). This approach tests this by dividing each spectrum by their non BSA peptide amount.
TIC5 – TIC intensity normalization on non peptide region. TIC5 normalizes on the subset of the spectrum consisting of everything apart from peptide A-E. The only activity in this region of the spectra is from the BSA, and since BSA is constant across all samples this total intensity should be equal across spectra.

Example 1:

There are two spectra, each consisting of 8 intensity values: Example Spectrum 1 = [2 1 3 1 3 80 130 90] and Example Spectrum 2 = [1 2 2 1 1 150 250 200]. Position 5-8 in each of the two spectra has prior to the normalization been classified as a peak region. Spectrum 1 has a TIC value (sum of all intensities) of 310 in the entire spectrum and 300 in the peak region, while Spectrum 2 has a TIC value of 607 in the entire spectrum and 600 in the peak region.

Using the previously classified TIC2 normalization (normalization on peak regions) on these spectra yield the following:
The TIC value for example spectrum 1 (TIC_{s1}) = 300
The TIC value for example spectrum 2 (TIC_{s2}) = 600 (since we are normalizing only on the peak regions, the TIC value is calculated only from the intensities within that area of the spectrum)

$2*TIC_{s1} = TIC_{s2}$
Multiplying all intensity values in Spectrum 1 with the value 2,0 results in both spectra having the same total intensity in the subset of the spectra that the normalization is basing its coefficient on. Note that even though only the peak regions are used to compute the normalization coefficient the entire spectrum is affected by the normalization. The result of the normalization is:
Normalized Example Spectra 1 = 2[2 1 3 1 3 80 130 90] = [4 2 6 2 6 160 260 180].*
Normalized Example Spectra 2 = 1[1 2 2 1 1 150 250 200].*

3.3 Median

Median normalization is similar to TIC normalization not only in that it is also a global normalization. If the number of data values is equal in all spectra, then setting the means of all spectra to the same value will produce the same results as setting the sum of all intensities to the same value. In median normalization the same kind of operation is performed as with TIC normalization, but instead of using the sum or mean, the median is used.

The variants listed below are the different approaches to Median normalization tested in this experiment.

Median1 – Median intensity normalization on the entire spectrum
Median2 – Median intensity normalization on peak regions
Median3 – Median intensity normalization on the peak heights

Example 2:

In this example we are using the same Spectra as in Example 1. That means Spectrum 1 = [2 1 3 1 3 80 130 90] and Spectrum 2 = [1 2 2 1 1 150 250 200]. Spectrum 1 has a median value of 3, and Spectrum 2 has a median value of 2.

Using Median1 normalization (on the entire spectrum) on these spectra yields the following:
The median value for example spectra 1 ($Median_{s1}$) = 3
The median value for example spectra 2 ($Median_{s2}$) = 2
$Median_{s1} = 1,5*Median_{s2}$

Multiplying the intensity values in Spectrum 2 with 1,5 results in the same median values for both spectra. The result of the normalization is:
Normalized Example Spectra 1 = [2 1 3 1 3 80 130 90]
Normalized Example Spectra 2 = 1,5[1 2 2 1 1 150 250 200]= [1,5 3 3 1,5 1,5 225 375 300].*

3.4 Internal Standard Normalization

The idea of internal standard normalization is that if you have a substance of known and equal concentration in each sample, then by setting the intensity of the peak corresponding to this substance as equal in all spectra then the other peaks should also be directly comparable.

In our examples the BSA background is constant in all experiments, and hence it is possible to use one of the BSA peaks as an internal standard.

Example 3:

In this example there are two spectra, S_1 = [1 3 2 5] and S_2 = [2 3 4 10]. The 3rd intensity value is the value representing the pre-determined internal standard that is of the same concentration in both of the samples that the spectra are derived from.

The 3rd value in the first spectra is 2 and in the second spectra 4. We can easily see that 4/2 = 2 so by multiplying S_1 with 2 the result is the normalized spectrum 1 (SN_1) = [2 6 4 10].

By comparing SN1 and S2 we can see that the 1st and 4th values are from substances with the same concentration in both S_1 and S_2 while the substance related to the 2nd value is of only of half the concentration.

In the internal standard normalizations in this work, the internal standards are the sum of all intensity values within a decided peak region.

The variants listed below are the different approaches to the internal standard normalization tested in this experiment. The different regions are selected by having relatively high intensity values in the BSA spectra.

Internal1 – Internal standard normalization using BSA peak at intensity index 40530-40720 (1480 m/z)
Internal2 – Internal standard normalization using BSA peak at intensity index 50866-51044 (1727 m/z)
Internal3 – Internal standard normalization using BSA peak at intensity index 57071-57280 (1883 m/z)

3.5 Standard Deviation of Noise Normalization

The Standard Deviation of Noise normalization is to our knowledge not previously published. The idea is that the standard deviation of the noise should be of equal size in all spectra. In that regard it shares a lot of its reasoning with the global type normalizations, but instead of normalizing on the actual peaks of interest, we normalize on the noise.

The first task was finding a large region of the spectra with only noise that lacked any peaks (*Figure 11*). A suitable region in this case was 2603 m/z to 2661 m/z. The standard deviation (using n-1) was calculated from the intensity values in that region for all spectra. Normalization was then achieved by dividing all intensities in the entire spectrum by the standard deviation of that spectrum. Hence all spectra then had the same standard deviation of 1 in the selected noise region.

Example 4:

*In this example there are two new spectra, S_1 = [1000 3000 5000 2 1 3 2 1] and S_2 = S_1*2 = [2000 6000 10000 4 2 6 4 2].*

The pre-defined noise region for these spectra are indexes 4-8, and the standard deviation for these regions are as follows:
The standard deviation for S1 (Std_{S1}) = std(2 1 3 2 1) = 0.84
The standard deviation for S2 (Std_{S2}) = std(4 2 6 4 2) = 1.67

In order to normalize the spectra we divide the full spectra by their standard deviations.
The normalized spectra 1 (N_{S1}) becomes S_1/Std_{S1} = [1195 3586 5976 2.4 1.2 3.6 2.4 1.2]
The normalized spectra 2 (N_{S2}) becomes S_2/Std_{S2} = [1195 3586 5976 2.4 1.2 3.6 2.4 1.2]
So by comparing the standard deviation of the noise regions we have been able to normalize these two simple spectra.

The variant listed below is the only approach to Standard deviation of noise normalization tested in this experiment.
STD1 – Standard Deviation normalization on the noise region of the spectra between 2603 and 2661 m/z.

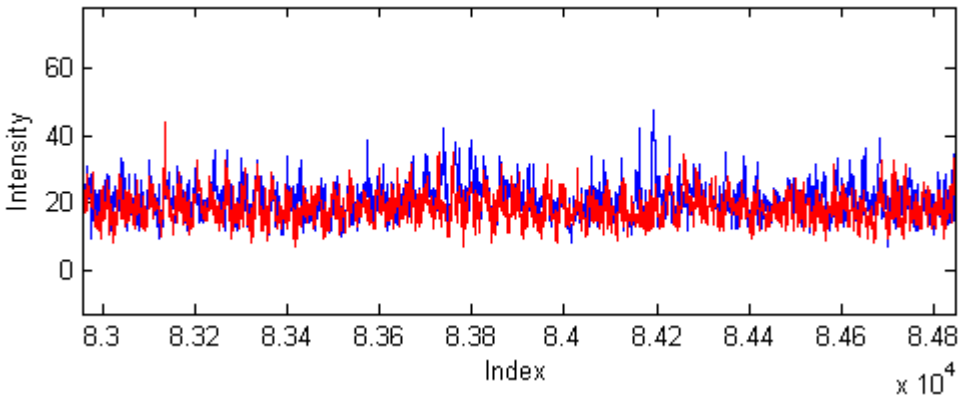


Figure 11: This is the area from 2603 m/z to 2661 m/z in spectra 9 and 10 which is the region chosen for the noise normalization . The overall intensity is very low with the highest peaks being less than 50 when BSA and peptide peaks in the same spectra are of several thousand intensity units. In addition none of the "peaks" from one spectrum is of similar height in the other.

3.6 Cumulative Intensity Normalization

Cumulative Intensity Normalization is described in the paper Quality Assessment of Tandem Mass Spectra Based on Cumulative Intensity Normalization by Paek and Na [9] and is a technique that is different from any other normalization technique mentioned here. The idea is that the method "considers both the magnitude of individual fragment ion peaks and their ranking in raw intensities" [9].

First, all intensities are converted to relative intensities, which means they are divided by the total summed intensity value of the entire spectral region used for the normalization. Starting with the smallest intensity value, all values are

replaced with the sum of all intensity values that are of lower rank (smaller value) than the current value, *i.e.* the smallest value (lowest rank) r1 is converted to r1, the second to smallest value r2 is converted to r1+r2 and so on. The highest value will be the sum of all the relative intensities in the spectrum and will as such sum to 1. The formula for the normalized intensity of the nth value is as such:

Equation 1

$$NIn = \frac{\sum\{I_{raw}(x) | Rank(x) \leq n\}}{TIC}$$

Where NIn is the normalized intensity of the nth highest peak, $I_{raw}(x)$ is the raw intensity of intensity value x, TIC is the total ion intensity of a spectrum and Rank(x) represents the order of a value at x when sorted by magnitude of raw intensities in ascending order.

*Example 5 (*Figure 12*):*

In the spectrum in this example there are 4 intensity values, [1, 3, 2, 10]. Since the sum of the intensity values in this spectrum is 16, the relative intensities to the total intensity of this spectra are [1/16, 3/16, 2/16, 10/16].

Starting with the lowest value (rank) we perform the normalization as it is described above. First we sort the relative intensity spectrum from lowest to highest value [1/16, 2/16, 3/16, 10/16]

We now calculate the new ranks:
r1 = 1/16
r2 = 1/16+2/16 = 3/16
r3 = 3/16+3/16 = 6/16
r4 = 6/16 + 10/16 = 16/16 = 1
The ranks are placed in a list [r1 r2 r3 r4] = [1/16, 3/16, 6/16, 1]
*This list represents the sorted normalized spectrum. To obtain our normalized spectrum we need to undo the sorting and reorganize the spectrum so it has its original order. The result of that procedure is the final normalized spectra: [1/16, 6/16, 3/16, 1] (*Figure 11*)*

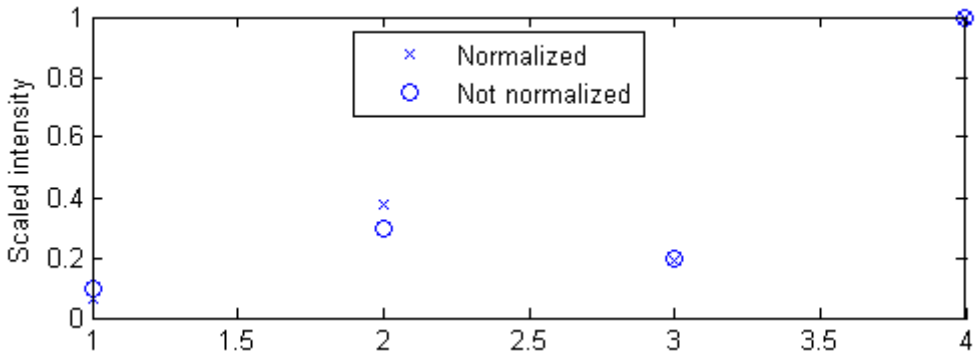


Figure 12 This picture shows the effects of cumulative intensity normalization in this very simple case. The non normalized data is scaled so that the highest value equals 1

This normalization is designed to be applied to fragment ion peaks, but it could be interesting to see the results of it being applied on all three data selections defined in 3.1 in addition to a subset that instead of taking the highest intensity values per peak takes the summed intensity of that peak.

The variants listed below are the different approaches to Cumulative normalization tested in this work.

Cumu1 – Cumulative intensity normalization on the entire spectrum
Cumu2 – Cumulative intensity normalization on peaks regions
Cumu3 – Cumulative intensity normalization on peak heights
Cumu4 – Cumulative intensity normalization on summed area under peak regions

3.7 Linear Regression Normalization

Linear Regression Normalization is usually used for normalizing DNA or RNA arrays but has seen some use in mass spectrometry [5].

The first thing that is done is creating a spectrum, S_a, that is the spectrum of the means of the intensity values at each index in all the different spectra. Then for each spectrum to be normalized, S_b, we create a Minus versus Average (MA) plot with
$M = log_2(S_a)-log_2(S_b) = log_2(S_a/S_b)$ and
$A = log_2(S_a)+log_2(S_b) = log_2(S_a*S_b)$
where M is a list of the difference of the logarithmic intensity values for all indexes in the spectrum to be normalized, S_b, and the average, S_a, spectra, and A is a list of the sum of the logarithms of the intensity values for the indices in the same spectra. (This is just a linear transformation of the average.)

An MA plot (*Figure 13*) is a plot that shows the intensity difference between two points in the same location of the spectrum (M) plotted against the intensity of the same two points (A). If there is no systematic bias the MA plot would look similar to *Figure 14* in which the mean spectrum S_a has been replaced with the S_b spectrum but with added normally distributed noise. In the ideal case the normalization performed on the plot shown in *Figure 13* would change the distribution so that it similar to *Figure 14* with all data points symmetrically spaced around the x axis.

After all the points are shown in the MA plot, a linear regression step is performed on this plot, which in this case is a least squares regression. The regression line is plotted (not shown), and for each data point the deviation from zero of the regression line is added to the logarithm transformed values of S_b. These new values are then retransformed back into the normal scale and are the normalized intensities.

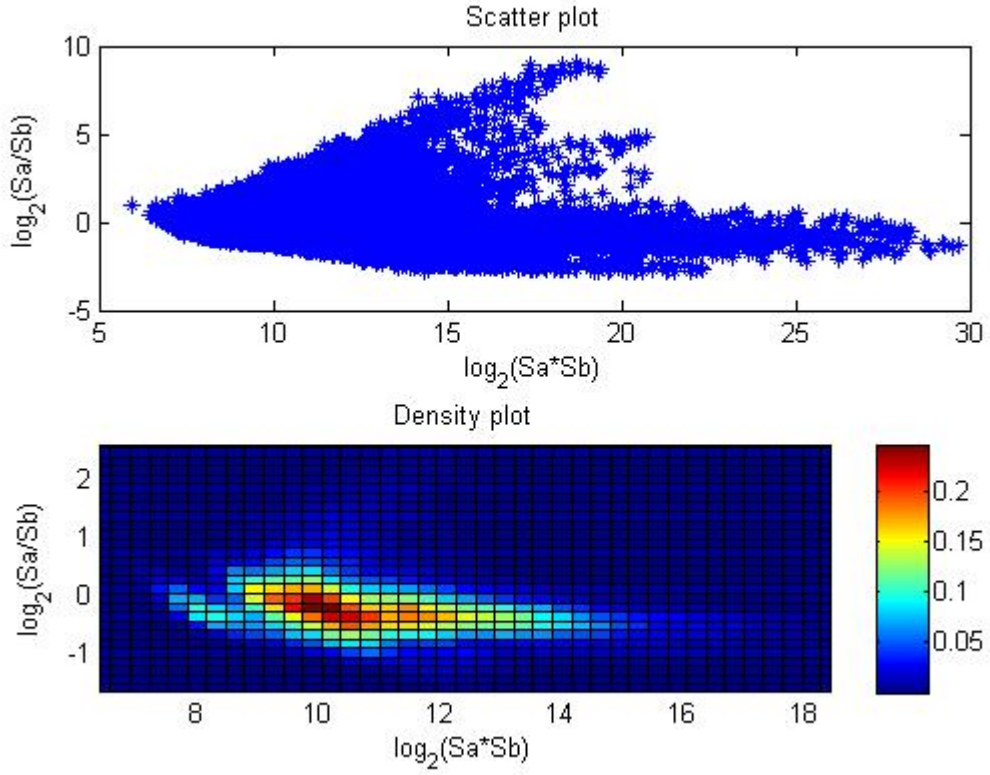


Figure 13: This illustration shows two MA plots of the same spectrum. I will use the density plot in further illustrations instead of the standard scatter plot, as in the scatter plot a lot of data is obfuscated by outliers, and it is hard to see where most of the data points lie. Paying attention to the scaling of the two plots, it seems like data is very spread in the scatter plo,t but in the density plot we can see that it is not nearly as spread as the scatter plot makes an impression of.

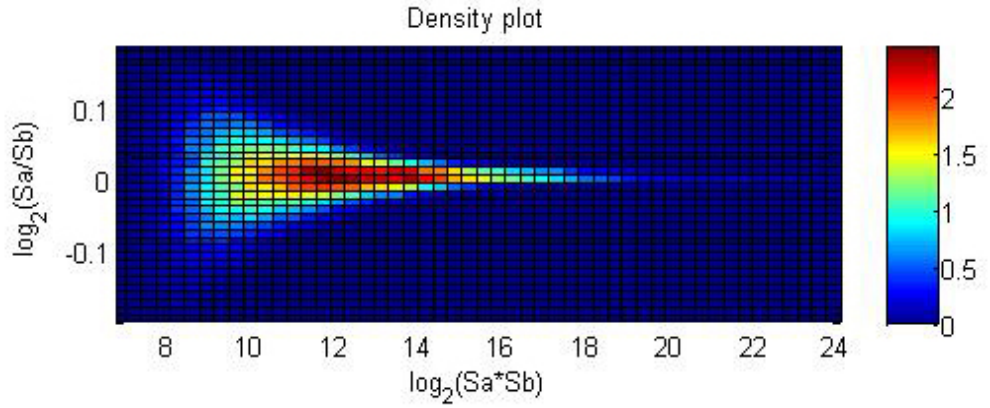


Figure 14: This density plot shows an MA plot where S_a is the same spectra as S_b only with added normally distributed noise. The result is a uniform distribution around the x-axis.

Example 6:
In this example there are two spectra, S_a = [1 3 2 5 4] which is the reference spectrum that is used to normalize against and $S_b = S_a^{1.2}$+small random number = [0.80 3.86 2.40 7.24 5.40]. This means that S_b is essentially S_a but has both a bias that grows exponentially with increasing intensity and a small random normally distributed error.

First we create the MA plot by calculating the M and A variables.
$M = \log_2(S_a)-\log_2(S_b)$ = [0.3258 -0.3637 -0.2624 -0.5335 -0.432]
$A = \log_2(S_a)+\log_2(S_b)$ = [-0.3258 3.5336 2.2624 5.1773 4.432]

*Using the MA plot we create from these values we are able to perform a linear regression with the degree of our choice, which in this case is 1 (*Figure 14*) (linear regressions do not necessarily have to be of the first order).*

*The least squares regression r = kx + i is calculated, and the value of r in each point is kM + i = -0.15*M+0.2 = [0.2556 -0.332 -0.1384 -0.5822 -0.4687]. Adding this value of r to the logarithm of the intensity value of S_b gives the logarithm of the normalized value for S_b.*
$LNS_b = \log_2(S_b)+r$ = [-0.0702 1.6167 1.124 2.2732 1.9632], where LNS_b is the logarithm normalized intensities for S_b.

*If we plot these values instead of S_b in our MA plot, we can see them distributed roughly evenly across the zero with the systematic bias gone (*Figure 16*).*

Finally we remove the logarithm from LNS_b so that we are left with the normalized intensities (NS_b).
$NS_b = 2^{LNSb}$ = [0.9525 3.0666 2.1795 4.8339 3.8993]

Because of the random error in S_b the normalized result is not a perfect match of S_a but the systematic bias resulting from the exponential factor added to intensity has been completely removed.

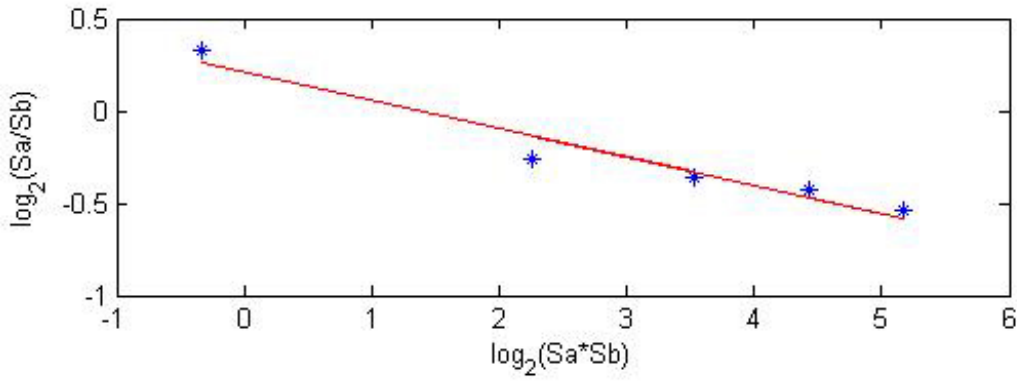


Figure 15: This image shows the MA plot of the data in example 6 (blue stars) and the least squares linear regression performed on this data. We can easily see that data is biased since the data points are not scattered across zero.

The variants listed below are the different approaches to Linear regression normalization tested in this experiment.

Linreg1 – Linear regression intensity normalization of the first order on the entire spectrum
Linreg2 – Linear regression intensity normalization of the second order on the entire spectrum

Linreg3 – Linear regression intensity normalization of the first order on peak regions
Linreg4 – Linear regression intensity normalization of the second order on peak regions
Linreg5 – Linear regression intensity normalization of the first order on peak height
Linreg6 – Linear regression intensity normalization of the second order on peak height

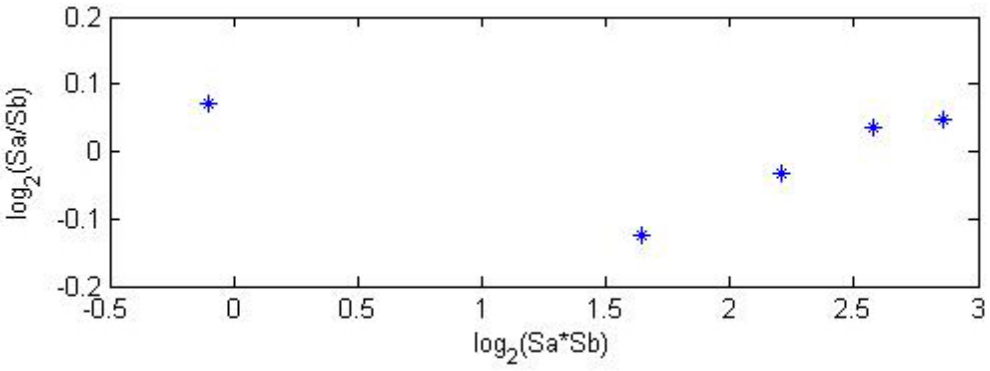


Figure 16: MA plot of data in example 6 after normalization has been performed. There is now no clear bias, and data is scattered somewhat evenly across the zero.

3.8 LOWESS Normalization

Locally Weighted Scatter plot Smoothing (LOWESS) Normalization is a method that shares many similarities with the linear regression technique. The general methodology is the same, but instead of single linear regression curve being the basis for the normalization step, the MA plot is subdivided into smaller sections, each with a regression curve that are fit together.

A program called MAANOVA 2.0 [14] originally developed for micro-array data had its code slightly modified in this work to fit the MALDI data. Details about the original code can be found in the reference [14]. The normalizations are run with the default settings for this program of 3 iterations and a smoother span of 0.2. After the LOWESS line has been calculated, the normalization process is exactly like in the linear regression case (*Figure 17*).

The variants listed below are the different approaches to LOWESS normalization tested in this experiment.

Lowess1 – Locally Weighted Scatter plot Smoothing normalization on entire spectrum
Lowess2 - Locally Weighted Scatter plot Smoothing normalization on peak regions

Lowess3 - Locally Weighted Scatter plot Smoothing normalization peak height

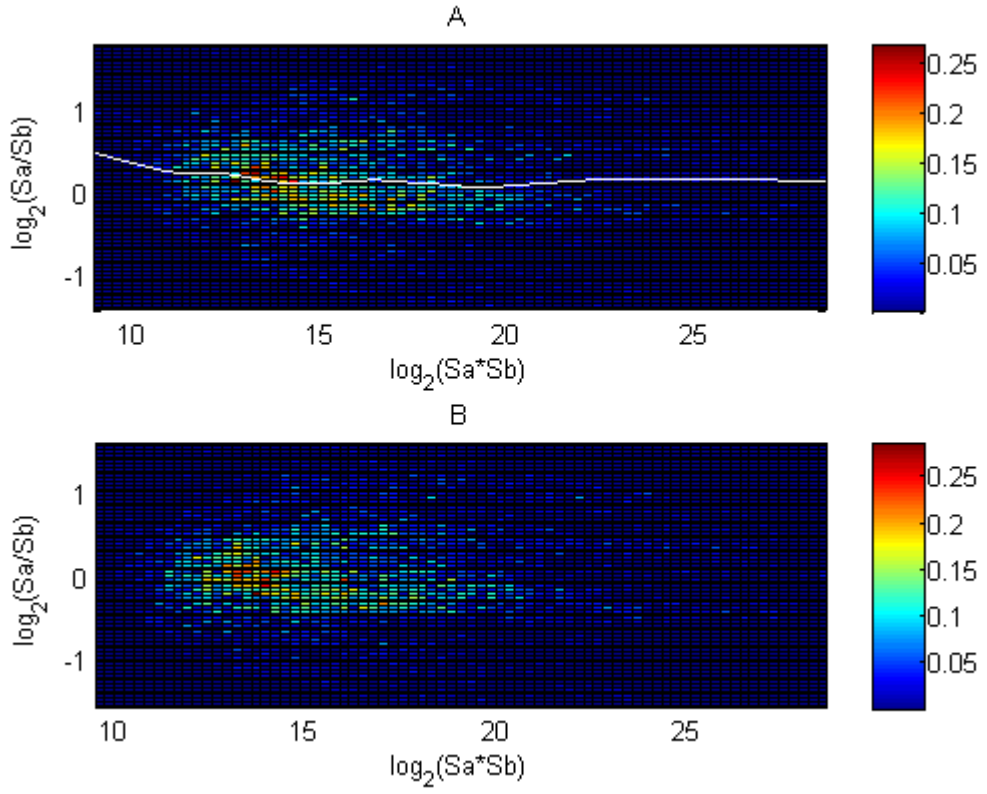


Figure 17: Picture A shows the lowess-line (white) on the non normalized MA plot from one of the spectra, the noticeable change in curvature of the line shows that it is not a normal polynomial regression. Picture B shows the normalized MA plot of the same data.

3.9 Quantile Normalization

Quantile normalization was originally developed for multiple high-density arrays. The underlying assumption for the method is that peptide abundances in different samples are expected to be roughly similar [5].

Quantile normalization consists of two steps as outlined by Ballman et.al [8] (Array is interchangeable with spectrum)

1. *Create a mapping between ranks and values. For rank 1 find the n values, one per array, that are the smallest value on the array, and save their averages. Similarly for rank 2 and the second smallest values, and on up to the n largest values, one per array.*
2. *For each array, replace the actual values with these averages.*

This makes the distributions of the values identical in the different spectra.

Example 7:

In this example there are two spectra, S_1 = [1 6 3 7 9] and S_2 = [2 8 4 7 11].

First the spectra are sorted: S_{S1} = [1(1) 3(3) 6(2) 7(4) 9(5)] and S_{S2} = [2(1) 4(3) 7(4) 8(2) 11(5)] with the figures in the parenthesis being these values original position in the S_1 and S_2 spectra. The average array for the sorted spectra is calculated: AS = [1.5 3.5 6.5 7.5 10]. These averages are now inserted into the sorted spectra: S_{S1} = [1.5(1) 3.5(3) 6.5(2) 7.5(4) 10(5)] and S_{S2} = [1.5(1) 3.5(3) 6.5(4) 7.5(2) 10(5)].

Finally the S_{S1} and S_{S2} spectra are reverted to their unsorted order by using the numbers in the parenthesis, and the result is S_1 = [1.5 6.5 3.5 7.5 10] as well as S_2 = [1.5 7.5 3.5 6.5 10]. We can see that the distribution is now equal, but the spectra are not because 6.5 and 7.5 were of different ranks in the individual spectra.

The variants listed below are the different approaches to Quantile normalization tested in this experiment.

Quantile1 – Quantile normalization on entire spectrum
Quantile2 – Quantile normalization on peak regions
Quantile3 – Quantile normalization on peak height

3.10 Top L-Ordered Statistics Normalization

Top L-Ordered statistics normalization [10] is aimed at avoiding bias that may be caused by non-random missing events in the spectra, which are peaks falling below what is detectable in that spectrum.

The mathematical definition of the normalization that can be found in the paper by Wang *et.al* [10] is as follows:

Equation 2

For the case of K(K>2) samples, denote the intensity measurements of the k_{th} sample as $X^k = (x_1^k, x_2^k, \ldots, x_{n\,k}^k)$. For a given number $L(L < min(\{n_k\}_{k=1}^K))$, define the population median as $\mu_0 = \frac{1}{K}\sum_k median\left(x_{(1)}^k, x_{(2)}^k, \ldots, x_{(L)}^k\right)$
Then the scaling coefficient for the k_{th} sample is $\lambda^k = \frac{1}{\mu 0}\, median(x_{(1)}^k, x_{(2)}^k, \ldots, x_{(L)}^k)$

More intuitively what is done is that first the L highest intensities in each spectrum are saved, L must be lower than the lowest number of intensity values in any of the spectra. The averages of the median values for the L highest intensities in each spectrum are calculated. The median of the L highest intensities for each spectrum are divided by this value, and the result is the scaling coefficient (λ^k) that we multiply the spectrum with.

A problem with the approach used by Top L-Ordered normalization is that changing the value of L can drastically alter the results. It was immediately clear that in order for this method to be of any use there would have to be a way to find a good value of L. It was also important that this value of L was not derived from the dataset used for the experiment, as that would compromise the validity of the experiment by risking over fitting L to perfectly match our particular data. Instead I decided to calibrate the value of L on a

dataset containing only BSA and no other peptides that was created at the same time as my experimental data. This is reasonable, because the BSA spectra would be similar but not identical, which is a situation that seems likely to be the case for how this technique would be calibrated in a real setting.

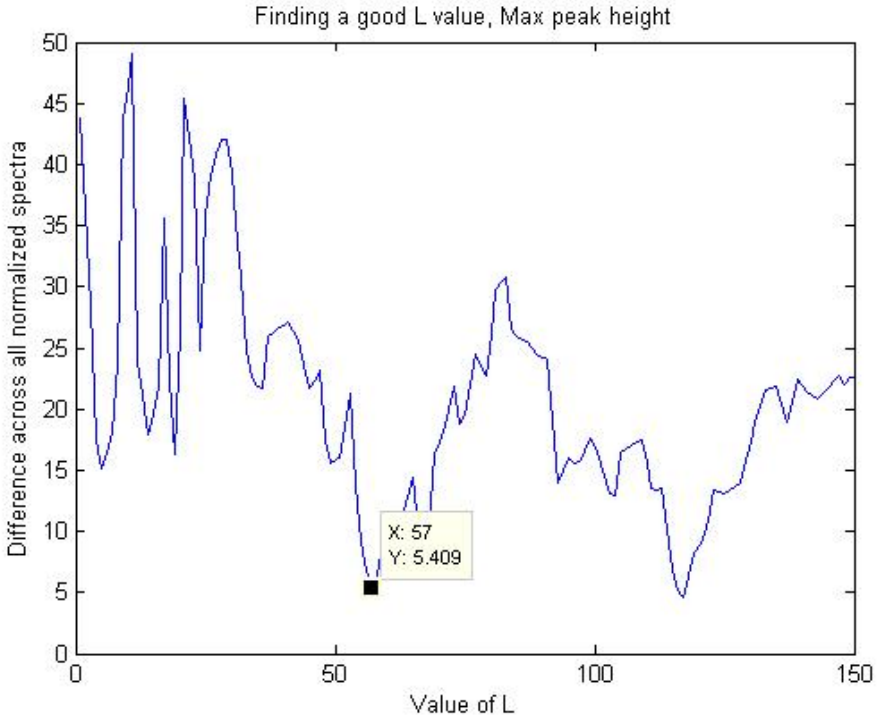


Figure 18: The image shows the plot of the difference in the sum of BSA peak intensity across normalized BSA spectra using different values of L with the max height per peak methodology. The same kind of plot was used for both TopL1 and TopL2.

All the BSA spectra have the same concentration, and if it is possible to find the value for L that minimizes the normalized differences in the BSA peak regions between the different spectra, then that would be the best value for L to use in the normalization.

By testing different values of L against this difference and plotting it, the optimal values for L (*Figure 18*) were found.

Example 8:

In this example there are K = 3 spectra and these are, S_1= [1 3 2 10 5 4], S_2 = [2 5 3 18 9 7] and S_3 = [1 2 4 8 3 5] in addition the value for L has been chosen to 4.

The 4 (L) highest intensities in each spectrum are sorted into sorted spectra (S_{Sx}): S_{S1} = [10 5 4 3], S_{S2} = [18 9 7 5], S_{S3} = [8 5 4 3].

The average of the medians of S_{S1}, S_{S2} and S_{S3} is calculated.
μ_0 = mean(4.5 8 4.5) = 5.67.

The median of the 4 highest intensities in each spectrum is scaled by $\frac{1}{\mu 0}$, and the result is the normalization scale factor.
*Scale1 = 1/5.67 * 4.5 = 0.79*
*Scale2 = 1/5.67 * 8 = 1.41*
*Scale3 = 1/5.67 * 4.5 = 0.79*

Finally we multiply the spectra by this scaling factor, and the result is the normalized spectra (N_{Sx}).
*N_{S1} = S_1*Scale1 = [0.79 2.38 1.59 7.94 3.97 3.18]*
*N_{S2} = S_2*Scale2 = [2.82 7.06 4.24 25.41 12.71 9.88]*
*N_{S3} = S_3*Scale3 = [0.79 1.59 3.18 6.35 2.38 3.97]*

The variants listed below are the different approaches to Top L normalization tested in this experiment.

TopL1 – Top L-Ordered statistics on entire spectrum with L=400
TopL2 – Top L-Ordered statistics on peak height with L=57

4 Evaluation

To determine the efficiency of the different normalization approaches outlined in the previous paragraphs, I need to convert the now normalized spectra into concentrations for the different peptides. This will enable comparison of the calculated and known concentrations and enable the analysis of the different normalization approaches performance. Further, I need to perform a statistical analysis to see if the results are significant or not; insignificant results will be discarded.

4.1 From normalized spectra to peptide intensities

Since I know the m/z values of each peptide, it was easy to locate the peptide peaks in the spectra. For each set of peptide isotope peaks in each spectrum, the intensity values within the peak regions were summed, here called SoPI (Sum of Peak Intensities). This measurement denotes the intensity value for each peak (*Figure 19*). This is done for all normalized spectra yielded from all the different normalization techniques. The result is a series of tables showing normalized summed intensity values for the 5 different peaks, corresponding to the selected peptides, in the 10 different spectra (*Table 5*).

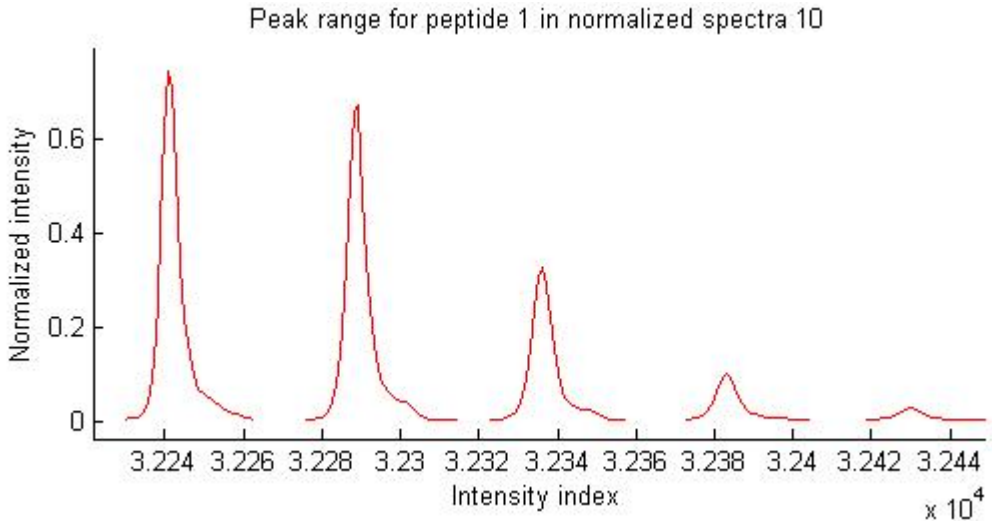


Figure 19: The image shows spectra 10 normalized with peak mean TIC normalization and the isotopic pattern of peptide A. Summing up all intensities showed in this image yields the total normalized SoPI value for this peptide: 12.98.

Table 5: This table shows the normalized SoPI values for the different peptides in the different spectra for the peak mean TIC normalization approach (TIC2). This table is comparable to Table 3 but showing the normalized SoPI values instead of the measured values.

	Peptide (SoPI)				
Spectra	A	B	C	D	E
1	1.8375	7.0095	8.4328	19.4811	11.2229
2	5.731	14.0323	0.1996	24.2135	4.4092
3	12.4572	0.7154	8.3902	21.3883	1.219
4	5.2967	10.3014	6.5991	23.1782	0.0901
5	13.725	3.2359	1.6765	11.6708	5.2812
6	15.9682	11.5508	4.7024	6.0667	8.5728

4.2 Calculating concentrations

The intensity values in *Table 5* can be seen as predictions of the concentration of the different peptides; however, they are by themselves not very useful predictions, and we need to convert them into real concentrations. For every peptide it is possible to create a plot which shows the predicted intensity plotted against the measured concentration. From this plot it is then possible to create a concentration curve using linear regression. Existing theory assumes a linear relationship between concentration and intensity, and, although that might not always be readily apparent in this data, we assume any deviation from linearity is because of measurement error, static interference or other types of interference.

It would be possible to just create a least squares regression line on the SoPI/measured intensity plot and then immediately pick out the predicted intensities from this regression; however, when making models it is always important to try and prevent over fitting the model to data. Over fitting a model means that the parameters that make up the model are too tightly tuned to the existing data, and because of this, the model performs poorly on new data. For a model to have much value it needs to maintain its predictive capabilities.

The technique used in this work to prevent over fitting is called *leave one out, cross validation*. Cross validation means that the data set is split into groups. The predictions for a group is calculated using the data for all other groups but not the own group. The number of groups can be everything from two groups splitting the data in half to as many groups as there are data values; this last variant is *leave one out, cross validation (loocv).*

The advantage of the loocv technique is that it is possible to use it on small data sets such as the one used for this study. Other techniques may reduce over fitting more than loocv does, but they would leave us with too few data points for any meaningful analysis, and because of this, loocv is ideal for this study.

Using the loocv methodology, we make a unique concentration curve for each data point that is to be predicted. The data point the prediction is made for is removed from the calculation of the curve. After the equation of the concentration curve has been calculated, it is possible to predict the concentration values for the data point. This is then repeated for all data points in the data set.

Example 9: Peptide C in TIC5 have intensity values [8.4 0.2 8.4 6.6 1.7 4.7].

Let us calculate the predicted concentration value for concentration number 6 in the intensity array.

We remove value 6 (4.7) from the array, which leaves [8.4 0.2 8.4 6.6 1.7], and calculate the concentration curve by linear regression of the first degree. The result is a curve $y = ax + b$ where in this case $a = 8.5$ and $b = 0.5$. Using these numbers and the value of $y = 4.7$, we can calculate the concentration $x = (y-b)/a = (4.7-0.5)/8.5 = 0.49$ which can be compared to what we know is the measured concentration: 0.333.

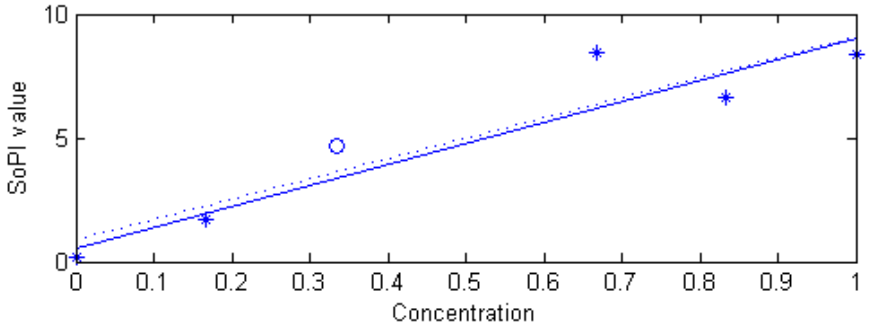


Figure 20: This image shows the concentration curves and the data points they are based on from example 9. Stars represent data points that are used to calculate the curve in the regression the circles is the data point left out (the concentration that is to be determined). The dotted line shows how the linear regression would look if all points were used in the calculation and the solid line shows the regression line for the case where one sample point is excluded as in loocv.

4.3 Root Mean Squared Error of Cross Validation

The result of performing loocv on all of these data points for all the different normalizations is a series of tables for each normalization. These tables contain the predicted concentration of all the peptides.

The goal of being able to compare the different normalization approaches is now within reach. By comparing the predicted concentrations with the measured quantities, it is now possible to see if the different normalization methods have yielded more accurate results than if no normalization had been performed at all.

By subtracting the measured concentrations from the predicted concentrations it is possible to see the error of prediction for every peptide in every sample. This error can be both positive and negative, but since we are only interested in how large the errors are, a good measurement tool is the Root Mean Squared Error of Cross Validation (1)

$$RMSECV = \sqrt{\frac{\sum(P-M)^2}{n}} \qquad (1)$$

Where P is the predicted concentration value, M is the measured concentration value and n is the number of data points.

There are two different values of RMSECV that are of interest - first of all, the total RMSECV value for the normalization procedure of all samples over all peptides. This value will represent the overall effectiveness of the normalization. Secondly the RMSECV values of the different peptides. These values will show differences in how the normalization performs for different peptides. The smaller the RMSECV values, the better the normalization.

4.4 Wilcoxon Ranked Sign-Test

The results of all these RMSECV calculations are a great number of values that should show if it is better to normalize by a particular method or not. However, as with all empirical data it is important to find out if variations in the results are because of actual differences or because of statistical variation. The way to determine this is by using statistical significance tests.

One way of testing if normalization achieves anything is to compare the absolute value of the prediction error between normalized and non normalized results. If the error reduction is significant, then it is reasonable to assume that this difference is the result of a successful normalization. What is compared is the distribution of absolute values of the errors between normalized and non normalized data for all the peptide concentrations.

The error (*predicted concentration - measured concentration)* is normally distributed, and a suitable test would be a student's t-test. The absolute value of the error (|predicted concentration - measured concentration|) on the other hand is not normally distributed (*Figure 21*), and as such a non parametric approach is required: the Wilcoxon ranked sign-test. This test compares two symmetric continuous distributions of random variables by comparing median values. The benefit of the sign-test is that there is no assumption of the distribution of the data, which is important as performing a test that assumes a certain distribution on data of another distribution is very likely to yield erroneous results.

The test that is performed is a paired, two-sided ranked Wilcoxon sign test of the null hypothesis that the data in the vector of the differences between absolute errors for normalization and absolute errors for no normalization (Absolute error for normalization – absolute error for no normalization), comes from a continuous and symmetric distribution with zero median, against the alternative that it does not have zero median [15]. If the null hypothesis is false, then it is reasonable to assume that the difference in error is because of the normalization and not because of random factors.

The Wilcoxon ranked sign-test is a more advanced and sensitive version of the normal sign-test. Its only downside is that it requires the two compared distributions to be symmetric. It may be that this is not the case for some of the normalized results in this work; however, as we are testing the null hypothesis that both distributions are equal, it is of little consequence since for this hypothesis to be true the distributions must be symmetrical (equal distributions are by default symmetrical).

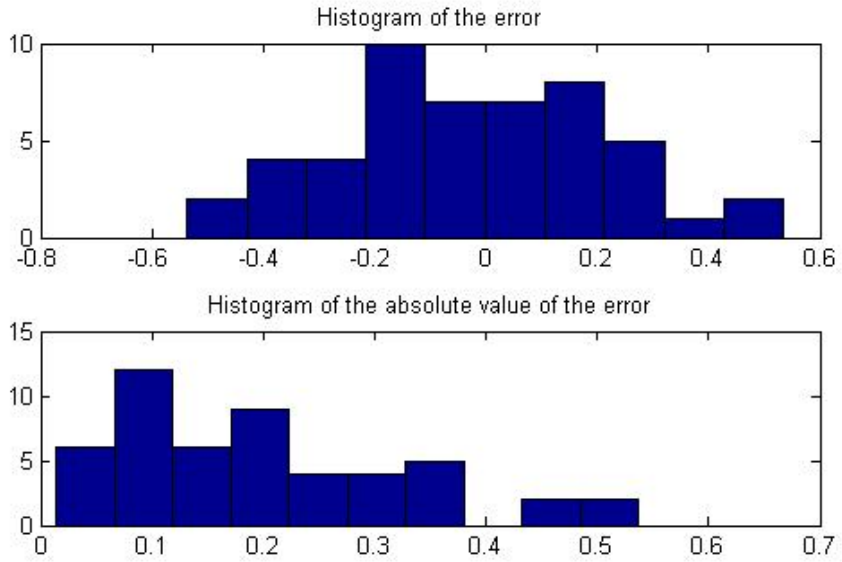


Figure 21: The normally distributed error changes to another distribution when taken as the absolute value, and it is no longer possible to use a t-test.

4.5 Repeatability study

There were four spectra from samples composed of peptides with the same concentration. These are normalized just like the other spectra; however, no cross validation is performed on this data, and instead the coefficient of variation (2) is calculated. This number represents the spread in intensity value from sources that should be of the same concentration. A small *CV* means that the intensity values are grouped tightly together while a large value means that they are spread over a large area.

$$CV = \frac{\sigma}{\mu} \qquad (2)$$

Where CV is the coefficient of variation, σ is the standard deviation and μ is the mean. These calculations are performed on a per peptide basis. This value can be expressed as a percentage.

5 Results

As mentioned earlier in the thesis, the output from these calculations is the Root Mean Squared Errors of Cross Validation for both individual peptides and for all peptides for all normalizations (Table 9 Appendix 1). The RMSECV values are measurements of the size of the errors of the measured and predicted concentration.

In order to be able to see the relative improvement of a normalization approach compared to the case with no normalization, each RMSECV value was divided with the value for the non normalized case:
(*new value = old value/Not normalized value*).

This results in a value for each normalization approach of how the relative size of the error compares to the non-normalized case. In this new relative scale, a small number is better. A value of 0,5 would mean that the error between predicted and measured concentration is 50% of that with no normalization, while a value of 2,0 would mean that the error have grown to twice the size of the no normalization case (*Table 10* Appendix 2).

Most normalizations seem to perform better than the no normalization case for all of the peptides. However, the difference in performance varies greatly between different peptides as can be seen in *Table 6*. That different peptides respond differently to mass spectrometry is no surprise, and as such it is no big stretch of imagination that that response is also transferred to the normalization.

Table 6: This table shows the mean, median and standard deviation of the relative improvements for all normalizations for the different peptides.

Overall relative performance					
	Pep A	Pep B	Pep C	Pep D	Pep E
Mean	1,1100	0,788	0,6837	1,1307	0,3679
Median	0,8935	0,6297	0,6231	0,9375	0,2899
Std	0,4946	0,6247	0,3123	0,8371	0,3094

One risk when trying to draw conclusions from data like (*Table 10* Appendix 2) is that what is observed might not be because of inherent differences but rather random errors and that what seems to be a good approach is the result of these errors lining up in a favourable way. This is the reason for statistical tests, in this case the Wilcoxon ranked sign-test described in section 4.4. The *p* value that is displayed here is the probability of obtaining a result at least as extreme as the one measured even if the null hypothesis is true (in this case that median of the absolute value of the error for *normalized – non normalized* data is zero). Hence a small value of *p* makes it more likely that the improvement or worsening of the result of a normalization is because of the

normalization itself rather than random chance. The relative improvements of the normalizations with significant *p* values are shown in *Figure 22* and *Table 7*.

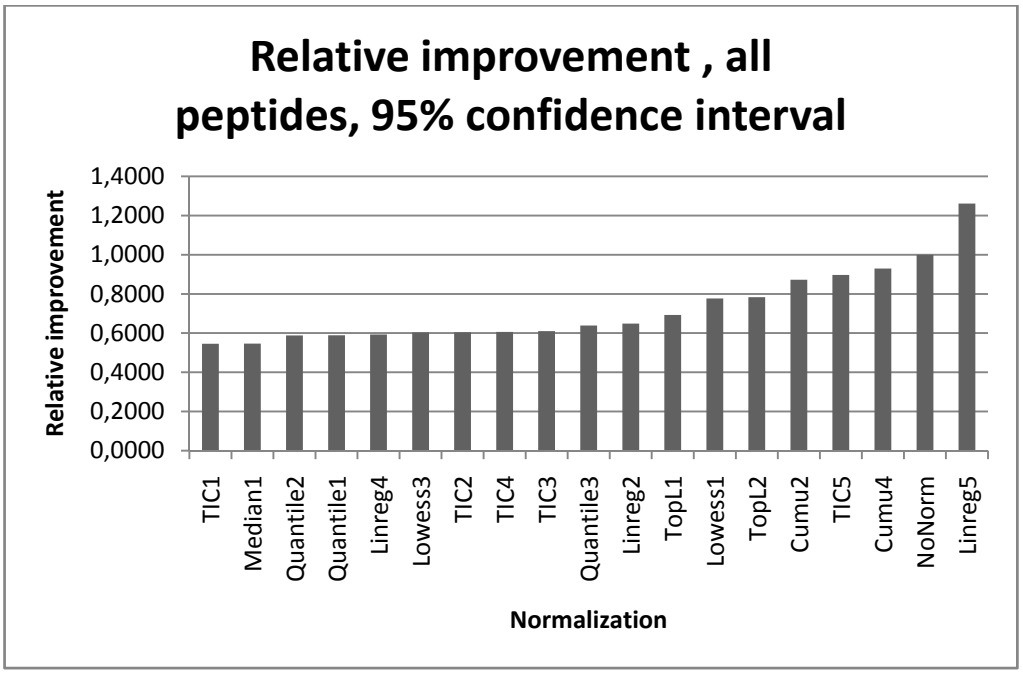


Figure 22: This graph shows the relative improvements of the different normalizations. Only the normalizations that were shown as significant by the statistical test are displayed.

Table 7: This table shows the relative RMSECV values with the normalizations sorted by best to worst normalization. In addition to relative RMSECV values, this table also shows the p *value from the paired Wilcoxon ranked sign test performed on the difference of the absolute value of the errors between normalized and non-normalized errors.*

Relative improvement			Relative improvement		
Normalization	All Peptides	Ranked sign, p	Normalization	All Peptides	Ranked sign, p
TIC1	0,5457	0,0021	Linreg2	0,6485	0,0016
Median1	0,5461	0,0125	TopL1	0,6929	0,027
Quantile2	0,5883	0,0057	Lowess1	0,7764	0,0407
Quantile1	0,5886	0,0087	TopL2	0,7832	0,0008
Linreg4	0,5932	0,0166	Cumu2	0,8720	0,0285
Lowess3	0,6037	0,0093	TIC5	0,8965	0,0111
TIC2	0,6041	0,0004	Cumu4	0,9292	0,0032
TIC4	0,6063	0,0068	NoNorm	1,0000	1
TIC3	0,6105	0,0057	Linreg5	1,2612	0,0104
Quantile3	0,6391	0,0104			

6 Discussion

This section contains a discussion about the results of the normalization procedures as well as the repeatability study. I also discuss why we choose to work with peptides, why some results are omitted, and I reflect on if some of the principal differences between normalizations affect the results.

6.1 Peptide and concentration selection

The sample mixture needed to contain several substances that I could measure, and we decided that using peptides would make the measurement process easier, since these would be visible in the spectra as an isotopic pattern in contrast to proteins that would be broken down by digestation and show up as multiple peaks in the spectra. Using multiple proteins in the mix could also cause overlap in certain locations of the spectra, and this interference would decrease the accuracy of measurements and as such decrease the validity of any results. However, the most important reason for using peptides is that it gave us complete control over the concentrations. The enzymes used to digest proteins are not always 100% effective, and this would increase variability and make concentration determinations more uncertain.

Using only these peptides would create an unrealistic situation, as there are almost always other substances other than the one being measured in a sample. To simulate this we decided to add a protein called BSA at equal concentration in all samples that we would not measure but that would only act as background. Adding this protein means that the spectra is a more realistic representative of a real situation because of the background matrix produced by this protein, while still leaving us with peptide peaks of known concentration.

6.2 Individual vs. group normalization of spectra

The concept of a normalization being based on individual or groups of spectra is an important distinction. Certain normalizations only normalize on a single spectrum basis, and other techniques normalize on a group of spectra where features of all the spectra in this group influence the normalization. A normalization that is independent of other spectra has the advantage that new spectra can be normalized and added to an already existing pool of spectra without the initial normalization having to be redone.

Total ion current, median, internal standard, standard deviation of noise and cumulative intensity normalization are all individual normalizations. Linear regression, LOWESS, quantile and Top L ordered statistics, on the other hand, are grouped normalizations where it is not possible to add a spectrum for normalization without having to redo the entire procedure.

This distinction becomes important in validation. If the sample “left out” in the validation is normalized with the others, and where the data from this sample affects the normalization of the other samples, then it is no longer independent. If the sample is not independent of the other samples, then the results from this validation are likely to be skewed, and the model over fitted.

Individual normalizations are hence preferable over grouped normalizations whenever any kind of validation is used.

6.3 Normalization evaluation

All significant results can be seen in *Table 7*, and the results from all normalizations are found in section 2 in the appendix.

The reason for showing all normalizations in the appendix is that twelve of the normalizations have *p* values that are higher than 0.05, and as such they do not pass the 95% confidence interval. The discarded normalizations may very well be potent and functional normalizations (RMSECV values ranged from highest to lowest); however, as they are not significant, we cannot trust them, and therefore it is not possible to draw any conclusions from that part of the data.

The total ion current and quantile normalizations (see appendix 4 for a compilation of normalization names and what they do) seem to be the most robust. They all give significant results, and they mostly score closely together (0.55-0.64 relative improvement) regardless of which subset of the spectrum that the normalization was applied to, the only exception being the total ion current approach that normalized on non-peptide regions (TIC5). This feature of knowing that the normalization will perform regardless of what type of subset it is applied to is a nice benefit for these two normalization approaches. When doing validation one must however keep in mind that TIC is an individual normalization and Quantile is grouped.

Median1 is an approach that mathematically is very similar to TIC1, using median instead of mean, and they also score very closely together. However, in practical terms, there is quite a big difference in that the median on the total spectrum is only a measurement of the size of the noise (due to noise being more common than signal) while the total ion current approach takes both noise and signal into consideration. Another difference is that, unlike total ion current, it is only the median approach on the entire spectrum that yields significant results, so overall TIC is probably the preferred approach since they are otherwise so similar.

The regression based normalizations (linear regression and LOWESS) have a wide performance spread from being the 5th best normalization to the worst. Unlike total ion current and quantile, these seem to be very dependent on implementation and what the normalization is applied to. The linear regression variants using a second order regression line are scoring well. Unfortunately only one of three regressions using a first order regression is significant, but the results are divergent enough that I think it is valid to say that a second order regression should be utilized for this normalization. LOWESS, being a significantly more complicated normalization that also requires more computing power, does not produce better results than the ordinary linear regression, and as such it is hard to recommend this approach even though the results are not bad.

The most consistent underperforming normalization in the group is the cumulative intensity normalizations. While the two normalizations that yielded significant results still give better result than not doing any normalization, there are far easier and more effective methods to use than these two.

Another underperforming normalization is the Top-L, and while it performs better than cumulative intensity, it is from an implementation perspective a worse overall normalization. The problem is that it is unnecessarily complicated for the results it is able to produce. The normalization procedure itself is not very complicated, but finding a suitable value for the L variable can be a major problem. When faced with the choice of using this normalization or total ion current, it is easy to see why the latter is so popular.

Performance-wise, it does not seem possible to separate normalizations that were performed on subsections of the spectra from normalizations that were performed on the entire spectra. For example, half of the top six scoring normalizations were performed on the entire spectra, while the other half was performed on different subsets. Instead the important part is that the data feed into the normalization is adapted to it. Some normalizations works best with peak regions and some with the entire spectra, and care must be taken to choose the correct subset of spectra for the particular normalization that is being employed.

As was mentioned in section 1.4, normalization can be subdivided into two categories, global and local normalization. However, here it is also hard to separate these two methods by performance. There are several normalizations from both groups that perform among the best, and there seems to be no bias towards one group or the other.

6.4 Repeatability

As can be seen by the repeatability study (*Table 8* and *Table 11*, Appendix 3), there is a large spread in data with coefficient of variations (CV) mostly varying between 30-40%.

This CV value is not reduced by normalization unless for particular methods such as Cumulative intensity normalization where the entire intensity distribution is dramatically altered. The large uncertainty in the measurements makes it harder to interpret the results of individual normalizations in the study; however, it does not affect the overreaching results. There are still big hurdles to overcome though, as in many cases the large repeatability errors, some of which are lab problems, make the entire procedure unreliable.

Table 8: This table shows the Coefficient of Variance of all peptides for all normalizations

Relative standard deviation of repetability study					
Normalization	CV	Normalization	CV	Normalization	CV
TIC1	0,33	STD1	0,33	Linreg6	0,34
TIC2	0,33	Cumu1	0,06	Lowess1	0,32
TIC3	0,34	Cumu2	0,20	Lowess2	0,33
TIC4	0,35	Cumu3	0,20	Lowess3	0,32
TIC5	0,37	Cumu4	0,19	Quantile1	0,33
Median1	0,33	Linreg1	0,32	Quantile2	0,34
Median2	0,33	Linreg2	0,33	Quantile3	0,33
Median3	0,33	Linreg3	0,34	TopL1	0,34
Internal1	0,45	Linreg4	0,33	TopL2	0,36
Internal2	0,36	Linreg5	0,33	NoNorm	0,33
Internal3	0,50				

6.5 Concluding Remarks

A conclusion from the relatively large repeatability errors is that when using MALDI for quantitative purposes, relatively large differences in intensities are required for conclusions about differential expression. An unfortunate conclusion of this is of course that, because of the measurement errors, for it to be possible to use MALDI as a screening tool, there needs to be a large concentration differences between studied groups.

That said, normalization certainly does help towards being able to use MALDI as a screening technique as it generally improves quantitative accuracy.

7 References

[1] MARTIN et al. *Identification of serum biomarkers for colon cancer by proteomic analysis*. British Journal of Cancer, 2006, 94, 1898-1905.

[2] RAY et al. *Early Alzheimer's disease defined by patterns of cellular communication factors in plasma.* Nature Medicine 2007

[3] WIKIPEDIA – MALDI-MS-TOF
http://en.wikipedia.org/wiki/Maldi
Last visited 24th of February 2008

[4] LISTGARTEN & EMILI. *Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry*. Molecular & Cellular Proteomics, 2005, 4.4, 419-434

[5] CALLISTER et al. *Normalization approaches for removing systematic scientific biases associated with mass spectrometry and label-free proteomics*. Journal of Proteome Research 2006, 5, 277-286.

[6] Bio-Rad ProteinChip buffers and reagents product description
http://www.bio-rad.com/B2B/BioRad/product/
br_category.jsp?BV_SessionID=@@@@@1396985540.1190209798@@@@@&BV_EngineID=
cccgaddmdelihkmcfngcfkmdhkkdfll.0&categoryPath=%2fCatalogs%2fLife+Science+Researc
h%2fSELDI+Technology%2fProteinChip+SELDI+System%2fProteinChip+Buffers+and+Rea
gents&catLevel=5&divName=Corporate&loggedIn=false&lang=English&country=HQ&catO
ID=-36407&isPA=false&serviceLevel=Lit+Request
Last visited 28th of September 2007

[7] NORRIS et al. *Preparing MALDI TOF mass spectra for statistical analysis: A practical approach*. www.biodesix.com/Documents/A052621.pdf

[8] BALLMAN et al. Faster *cyclic loess: normalization RNA arrays via linear models*. Bioinformatics, 2004, vol 20, issue 16, 2778-2786.

[9] PAEK & NA. *Quality assessment of tandem mass spectra based on cumulative intensity normalization*. Journal of Proteome Research 2006, 5, 3241-3248.

[10] WANG et al. *Normalization regarding non-random missing values in high-throughput mass spectrometry data*. Pacific Symposium on Biocomputing, 11:315-326(2006)

[11] ZIEN et al. *Centralization: a new method for the normalization of gene expression data*. Bioinformatics, 2001, vol 17, 323-331.

[12] SAUVE & SPEED. *Normalization, baseline correction and alignment of high-throughput mass spectrometry data*. *Proceedings of the Genomic Signal Processing and Statistics workshop, Baltimore, MD, USA.*, May 26-27, 2004.

[13] FUNG & ENDERWICK. *ProteinChip clinical proteomics: Computational challenges and solutions*. Computational Proteomics Supplement 2002, 32, 34-41.

[14] Jackson Laboratory MAANOVA 2.0 download page
http://www.jax.org/staff/churchill/labsite/software/anova/index.html
Last visited 28th of September 2007

[15] Matlab (R2007a) Function description for signrank

Appendix

Appendix 1 - RMSECV values for normalizations

Table 9: This table shows the RMSECV values for the different normalization approaches and for the difference peptides as well as for all peptides at once. The smaller the value the more accurate the normalization has been.

	RMSECV Values					
Normalization	Pep A	Pep B	Pep C	Pep D	Pep E	All Peptides
TIC1	0,0913	0,1286	0,2299	0,1949	0,1018	0,145
TIC2	0,0912	0,1127	0,2218	0,2796	0,079	0,1605
TIC3	0,1114	0,1007	0,205	0,2982	0,0658	0,1622
TIC4	0,1841	0,1354	0,237	0,1854	0,1137	0,1611
TIC5	0,1418	0,1878	0,3547	0,3934	0,0674	0,2382
Median1	0,0977	0,1366	0,2229	0,1574	0,1538	0,1451
Median2	0,0786	0,1425	0,2392	0,2627	0,1109	0,1658
Median3	0,0895	0,1181	0,255	0,3484	0,0631	0,1881
Internal1	0,3363	0,1415	0,1307	0,2487	0,156	0,1985
Internal2	0,2558	0,0971	0,1296	0,1095	0,0976	0,1373
Internal3	0,1867	0,1643	0,1676	0,3015	0,1299	0,1815
STD1	0,0896	0,1833	0,2069	0,161	0,2455	0,1686
Cumu1	0,2494	0,7895	0,746	0,8946	0,6413	0,6395
Cumu2	0,2185	0,241	0,2771	0,3356	0,1643	0,2317
Cumu3	0,2227	0,1748	0,2564	0,3084	0,1648	0,2114
Cumu4	0,2173	0,2537	0,3075	0,3548	0,1838	0,2469
Linreg1	0,1432	0,3963	0,5221	1,5026	0,1321	0,674
Linreg2	0,188	0,129	0,1853	0,2952	0,0685	0,1723
Linreg3	0,115	0,2026	0,2494	0,482	0,0905	0,2439
Linreg4	0,0885	0,0841	0,2357	0,2662	0,0881	0,1576
Linreg5	0,1098	0,2265	0,2839	0,7248	0,0656	0,3351
Linreg6	0,1157	0,1104	0,2372	0,1524	0,0617	0,1347
Lowess1	0,1169	0,1226	0,2655	0,3893	0,067	0,2063
Lowess2	0,1237	0,0732	0,2619	0,2908	0,0567	0,1717
Lowess3	0,1177	0,0537	0,2017	0,3003	0,0816	0,1604
Quantile1	0,1137	0,0983	0,2226	0,2572	0,0925	0,1564
Quantile2	0,1123	0,099	0,2179	0,259	0,0979	0,1563
Quantile3	0,1017	0,0951	0,1807	0,3192	0,1382	0,1698
TopL1	0,0856	0,1385	0,2122	0,3481	0,1035	0,1841
TopL2	0,1075	0,1692	0,1998	0,4065	0,1201	0,2081
NoNorm	0,1291	0,216	0,3736	0,3192	0,3445	0,2657

Appendix 2 - Relative improvements and p-values for sign-test

Table 10: This shows the relative RMSECV value compared to the case of no normalization. A value under 1 means that the normalization has an error that is smaller than that for no normalization (0.5 means the error is half of that of the no norm case) and a value over 1 means the error has increased.

	Relative Improvement						Sign test,
Normalization	Pep A	Pep B	Pep C	Pep D	Pep E	All Peptides	p value
TIC1	0,7072	0,5954	0,6154	0,6106	0,2955	0,5457	0,0021
TIC2	0,7064	0,5218	0,5937	0,8759	0,2293	0,6041	0,0004
TIC3	0,8629	0,4662	0,5487	0,9342	0,1910	0,6105	0,0057
TIC4	1,4260	0,6269	0,6344	0,5808	0,3300	0,6063	0,0068
TIC5	1,0984	0,8694	0,9494	1,2325	0,1956	0,8965	0,0111
Median1	0,7568	0,6324	0,5966	0,4931	0,4464	0,5461	0,0125
Median2	0,6088	0,6597	0,6403	0,8230	0,3219	0,6240	0,102
Median3	0,6933	0,5468	0,6825	1,0915	0,1832	0,7079	0,0656
Internal1	2,6050	0,6551	0,3498	0,7791	0,4528	0,7471	0,165
Internal2	1,9814	0,4495	0,3469	0,3430	0,2833	0,5167	0,3389
Internal3	1,4462	0,7606	0,4486	0,9445	0,3771	0,6831	0,2059
STD1	0,6940	0,8486	0,5538	0,5044	0,7126	0,6346	0,5304
Cumu1	1,9318	3,6551	1,9968	2,8026	1,8615	2,4068	0,5999
Cumu2	1,6925	1,1157	0,7417	1,0514	0,4769	0,8720	0,0285
Cumu3	1,7250	0,8093	0,6863	0,9662	0,4784	0,7956	0,0752
Cumu4	1,6832	1,1745	0,8231	1,1115	0,5335	0,9292	0,0032
Linreg1	1,1092	1,8347	1,3975	4,7074	0,3835	2,5367	0,36
Linreg2	1,4562	0,5972	0,4960	0,9248	0,1988	0,6485	0,0016
Linreg3	0,8908	0,9380	0,6676	1,5100	0,2627	0,9180	0,0519
Linreg4	0,6855	0,3894	0,6309	0,8340	0,2557	0,5932	0,0166
Linreg5	0,8505	1,0486	0,7599	2,2707	0,1904	1,2612	0,0104
Linreg6	0,8962	0,5111	0,6349	0,4774	0,1791	0,5070	0,1779
Lowess1	0,9055	0,5676	0,7107	1,2196	0,1945	0,7764	0,0407
Lowess2	0,9582	0,3389	0,7010	0,9110	0,1646	0,6462	0,1529
Lowess3	0,9117	0,2486	0,5399	0,9408	0,2369	0,6037	0,0093
Quantile1	0,8807	0,4551	0,5958	0,8058	0,2685	0,5886	0,0087
Quantile2	0,8699	0,4583	0,5832	0,8114	0,2842	0,5883	0,0057
Quantile3	0,7878	0,4403	0,4837	1,0000	0,4012	0,6391	0,0104
TopL1	0,6631	0,6412	0,5680	1,0905	0,3004	0,6929	0,027
TopL2	0,8327	0,7833	0,5348	1,2735	0,3486	0,7832	0,0008
NoNorm	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1

Appendix 3 - Coefficient of variation of the repeatability study

Table 11: This table shows the coefficient of variation for the measurements of the normalized peptide intensities for the repeatability study. As well as showing the relative improvement of the CV value compared to the non-normalized case.

Coefficient of variation of the repeatability study							Relative improvement All Peptides
Normalization	Pep A	Pep B	Pep C	Pep D	Pep E	All Peptides	
TIC1	0,3418	0,3309	0,3607	0,3143	0,2791	0,3254	0,9873
TIC2	0,2656	0,2539	0,4427	0,3785	0,3261	0,3334	1,0115
TIC3	0,2365	0,225	0,4828	0,4119	0,3564	0,3425	1,0391
TIC4	0,3267	0,3167	0,4462	0,3734	0,3027	0,3532	1,0716
TIC5	0,6289	0,622	0,1384	0,2338	0,2216	0,3689	1,1192
Median1	0,381	0,3707	0,3269	0,2894	0,2619	0,326	0,9891
Median2	0,2381	0,2257	0,457	0,3937	0,3442	0,3317	1,0064
Median3	0,165	0,151	0,5138	0,4465	0,3966	0,3345	1,0149
Internal1	0,1007	0,1177	0,7586	0,6744	0,6057	0,4514	1,3695
Internal2	0,0036	0,0175	0,6655	0,5863	0,5236	0,3593	1,0901
Internal3	0,1468	0,1664	0,8078	0,7235	0,655	0,4999	1,5167
STD1	0,3323	0,3213	0,3801	0,3284	0,3008	0,3326	1,0091
Cumu1	0,0387	0,0281	0,0967	0,0684	0,0831	0,063	0,1911
Cumu2	0,1455	0,1402	0,2836	0,2067	0,2394	0,2031	0,6162
Cumu3	0,1405	0,1658	0,2784	0,1886	0,2263	0,1999	0,6065
Cumu4	0,1621	0,1665	0,2446	0,1702	0,1938	0,1874	0,5686
Linreg1	0,2878	0,2813	0,3968	0,3538	0,2984	0,3236	0,9818
Linreg2	0,2603	0,2601	0,4353	0,3982	0,3106	0,3329	1,0100
Linreg3	0,25	0,2383	0,4597	0,3982	0,3348	0,3362	1,0200
Linreg4	0,2775	0,259	0,4403	0,3660	0,3316	0,3349	1,0161
Linreg5	0,2199	0,2192	0,4702	0,3717	0,3464	0,3255	0,9876
Linreg6	0,2699	0,2544	0,4738	0,3609	0,3535	0,3425	1,0391
Lowess1	0,2025	0,2019	0,4342	0,4203	0,3439	0,3206	0,9727
Lowess2	0,2438	0,2256	0,4521	0,3814	0,3484	0,3303	1,0021
Lowess3	0,1849	0,1876	0,4717	0,3569	0,3775	0,3157	0,9578
Quantile1	0,2894	0,2651	0,4317	0,3558	0,3192	0,3323	1,0082
Quantile2	0,281	0,2534	0,4477	0,3605	0,3377	0,3361	1,0197
Quantile3	0,2578	0,2138	0,4698	0,3271	0,3626	0,3262	0,9897
TopL1	0,2726	0,2613	0,4447	0,3786	0,3231	0,336	1,0194
TopL2	0,2617	0,2515	0,4931	0,4185	0,3499	0,355	1,0771
NoNorm	0,3138	0,3022	0,3742	0,3396	0,3178	0,3296	1,0000

Appendix 4 - Compilation of normalizations

TIC1 – TIC intensity normalization on the entire spectrum
TIC2 – TIC intensity normalization on peak regions
TIC3 – TIC intensity normalization on the peak heights
TIC4 – TIC intensity normalization on peak regions also normalizing on peptide concentration in spectra.
TIC5 – TIC intensity normalization on non peptide region
Median1 – Median intensity normalization on the entire spectrum
Median2 – Median intensity normalization on peak regions
Median3 – Median intensity normalization on the peak heights
Internal1 – Internal standard normalization using BSA peak at intensity index 40530-40720 (1480 m/z)
Internal2 – Internal standard normalization using BSA peak at intensity index 50866-51044 (1727 m/z)
Internal3 – Internal standard normalization using BSA peak at intensity index 57071-57280 (1883 m/z)
STD1 – Standard Deviation normalization on the noise region of the spectra between 2603 and 2661 m/z.

Cumu1 – Cumulative intensity normalization on the entire spectrum
Cumu2 – Cumulative intensity normalization on peaks regions
Cumu3 – Cumulative intensity normalization on peak heights
Cumu4 – Cumulative intensity normalization on summed area under peak
Linreg1 – Linear regression intensity normalization of the first order on the entire spectrum
Linreg2 – Linear regression intensity normalization of the second order on the entire spectrum
Linreg3 – Linear regression intensity normalization of the first order on peak regions
Linreg4 – Linear regression intensity normalization of the second order on peak regions
Linreg5 – Linear regression intensity normalization of the first order on peak height
Linreg6 – Linear regression intensity normalization of the second order on peak height
Lowess1 – Locally Weighted Scatter plot Smoothing normalization on entire spectrum
Lowess2 - Locally Weighted Scatter plot Smoothing normalization on peak regions
Lowess3 - Locally Weighted Scatter plot Smoothing normalization peak height
Quantile1 – Quantile normalization on entire spectrum
Quantile2 – Quantile normalization on peak regions
Quantile3 – Quantile normalization on peak height
TopL1 – Top L-Ordered statistics on entire spectrum with L=400
TopL2 – Top L-Ordered statistics on peak height with L=57

TRITA-CSC-E 2008:038
ISRN-KTH/CSC/E--08/038--SE
ISSN-1653-5715

www.kth.se