UPPSALA
UNIVERSITET

PROJEKTRAPPORT

Project #9

Brain region definition and comparison based on protein expression data

Wolmar Nyberg Åkerström
Moyen Mohammad Mustaquim

Report in Scientific Computing Advanced Course

May 2008

INSTITUTIONEN FÖR INFORMATIONSTEKNOLOGI

Abstract

The importance of applied biosciences and biotechnology in research and industry is rapidly increasing. Mathematical methods and concepts play a crucial role in these fields. As we are entering the post-genomic era, models-of-data, such as mining and filtering methods must be complemented with models-of-processes that explain relationships between genomic information and phenomena at biochemical and physiological levels.

Matrix-assisted laser desorption/ionization mass spectrometric imaging (MALDI-MSI) is an up-and-coming technology allowing true label-free molecular imaging of flat samples like biological tissue sections. A given MALDI-MSI data set containing the induced state of Parkinson's disease, initiates the interest in finding interesting markers of the disease. Manual exploration of such data is time consuming and requires a good idea on what to look for to be plausible. It should be possible to automate this process to some extent by using tools from the field of statistical analysis and thus instigate our project. Comparison of regions found in both halves of the brain to find interesting signatures related to the Parkinson's disease in the processed data of spectral peaks corresponding to protein expressions.

Feature selection was used to isolate protein expressions and reduce the dimension of the spectra. Further reduction into signatures representing regions was performed through a popular and a novel method of Projection Pursuit (PP) , Principal Component Analysis (PCA) and Independent Component Analysis (ICA). An attempt to match and improve these regions by removing asymmetries in the source data was also done.

Results reveal that a relatively small subset of the original MALDI-TOF spectra can be used to produce well defined regions using PCA and ICA. ICA will produce superior region definitions to PCA and it is possible to refine and analyze these regions to some extent without any use of spatial information in the algorithms.

Content

Content 2

1 Introduction 3
1.1 Supplied resources and assumptions 4
1.2 Data acquisition using MALDI-IMS 4
1.3 Protein expression signatures 5
1.4 MATLAB 5

2 Theory 5
2.1 Feature selection of protein expressions 5
2.1.1 MALDI-TOF sample spectra 6
2.1.2 Contiguous thresholding 7
2.1.3 Local maxima extraction 7
2.1.4 Wavelet smoothing 8
2.2 Feature extraction of protein signatures 8
2.2.1 Principal component analysis 9
2.2.2 Independent component analysis 9
2.3 Region definition by segmentation 10
2.3.1 Non-spatial segmentation and comparison 11
2.3.2 Successive symmetric region improvement 12
2.3.3 Regional comparison 12

3 Results 12
3.1 Feature Selection 12
3.2 Feature Extraction 16
3.3 Detected regions and symmetry 17

4 Implementation 20
4.1.1 PROPACK 20
4.1.2 FastICA 20
4.1.3 Rice Wavelet Toolbox 21

5 Discussion 21
5.1 Performance 21
5.2 Possible improvements 21
5.3 Sources of error 22
5.4 Extension to 3D 22

6 Conclusion 23

7 Acknowledgements 23

1 Introduction

This report was produced as a result of the work done during completion of Scientific Computing Advanced Course, directed by the Department of Information Technology at Uppsala University, first semester of 2008.

The problem under survey originates from the field of biology, where recent techniques allow detailed screening for proteins within tissue samples. These spectra of proteins can be investigated manually by an experienced scientist to identify important signatures. In this case sample sets taken consist of slices taken from rat brains with a state similar to Parkinson's induced in one of its halves. The process of investigating the full extent of these detailed spectra requires a large amount of work and has to be remade to some extent for every sample set. A way to automate this process to some extent would be desirable and this project is an attempt to explore this possibility.

The general idea of this implementation builds on using a definition of regions considered to be symmetric in structure over the two halves of the sample set to compare a list of protein intensities. As a product of this comparison a list of relative differences in intensities will be produced, which hopefully will aid in the exploration of the full spectra.

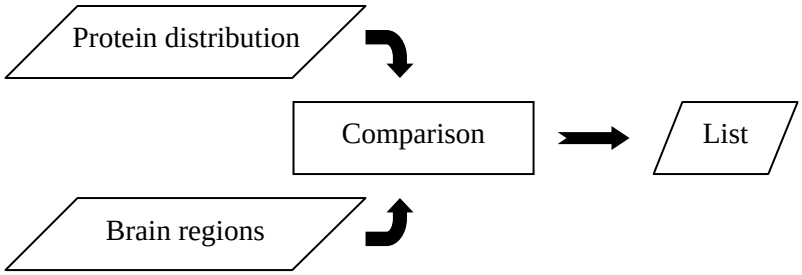

Fig 1: The basic components of the process.

Resources supplied for this project consist of data sets that have received a moderate amount of pre-processing, such as baseline adjustment. As no regions are given and the data still needed further processing to produce meaningful results using a reasonable amount of computational resources, the project was decided to involve a study of the data and acquisition process as well as statistical data processing to identify proteins and regions before comparison. The implementation is based on a commercial software for mathematical programming together with packages for well known and general methods.

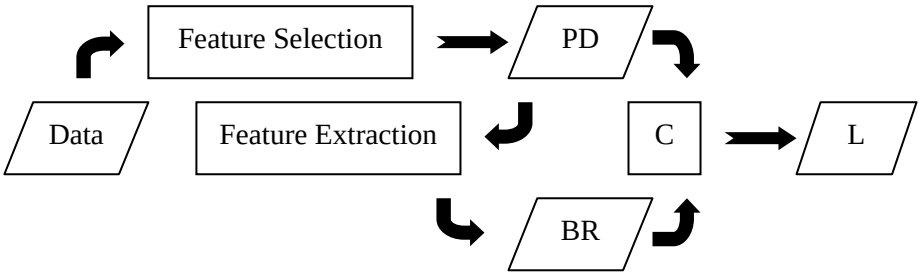

Fig 2: The basic approach attempted.

1.1 Supplied resources and assumptions

The supplied datasets were delivered in a binary format fulfilling the specifications of a, more or less, open format commonly used in this area of research (Analyze 7.5), preprocessing with baseline adjustment had already been made. Tissue samples consisted of a small slice from a part of a rats brain, more or less symmetric over the centre of the brain. Data was received from three diseased as well as three control rats, with a sample size of grids from 23 by 9 to 36 by 13 and a dimension of around 22000 to 32000 per sample.

Samples used consist of brain tissue from rats with a state similar to Parkinson's disease induced in only one of the half of the brain; the goal has been to provide a sample symmetric in size over the two halves. Control samples of rats with no disease have been provided as well.

1.2 Data acquisition using MALDI-IMS

The technique used to produce the protein spectra investigated in this project is called Matrix Assisted Laser Desorption/Ionization (MALDI) Imaging Mass Spectrometry (IMS). It is a technique often used for mass spectrometry of biological tissue samples and is capable of registering a wide spectra of masses through time-of-flight measurements (TOF).

Matrix of crystallized molecules is applied to the sample to protect the fragile bio molecules. As a laser pulse hit the sample the matrix will ionize while absorbing most of the energy. Bio molecules will thus be left intact. The ionized matrix will transfer charge to the bio molecules which allow them to be accelerated and later registered as they hit a sensor. Time-of-flight data is registered, which can be translated into mass ratio data.

The sample sets acquired for this project originates from a thin slice of a brain-half symmetric region of a rat brain. This region includes the Substantia Nigra and has been prepared for MALDI-IMS according to recommended techniques [3]. Animals used as source for these samples consist of six rats out of

which three had a state similar to Parkinson's disease induced in one half of the brain.

The resulting dimension of each sample set consist of equally spaced rectangular grids with dimensions of 23 to 36 by 9 by 12 samples. Each sample contains a spectra of about 22500 to 32000 discrete measurements of m/z values.

Data produced by the MALDI-IMS does not directly translate into protein intensities which means some form of pre-processing has to undertaken to select and group valid parts of the spectra.

1.3 Protein expression signatures

Protein expression is a subcomponent of gene expression. It consists of the stages after DNA has been translated into amino acid chains, which are ultimately folded into proteins. Protein expression is commonly used by proteomics researchers to denote the measurement of the presence and abundance of one or more proteins in a particular cell or tissue. From the acquired data these are what one need to be investigating.

Traditional brain regions have not been marked within the sample sets. Matching regions will be required for later comparison. The assumption will be that there are different setups of proteins across regions, which could be interpreted as their signatures. These signatures would be used to identify regions. As these signatures has not been supplied for the sample sets used, each data set will have to be investigated to extract possible signatures.

1.4 MATLAB

MATLAB is a commercial software by The MathWorks, providing a numerical programming environment and high-level programming language. It provides many features for matrix manipulation as well as plotting and various popular algorithms are implemented to work with this software.

To reduce the overhead of setting up a programming environment capable of handling the wide range of methods required MATLAB has been chosen as the target of the implementation. It is available and coherent across a wide range of popular computing platforms and has been widely adopted by scientists and developers.

2 Theory

2.1 Feature selection of protein expressions

The spectrum of each sample contain a significantly larger amount of information than can be expected to be realistic, this additional information results from the nature of the acquisition device as well as examined samples. We will seek to

reduce the spectra into a subset of features that are more likely to be relevant in this study. This task constitutes the essence of feature selection.

Ideally every peak would be completely resolved at a single point in time. This however is not the case due to variations caused by distribution based characteristics of some variables related to the data acquisition. Noise must also be considered as a factor.

The two approaches attempted within the scope of this project are based on subset selection using filtering, which makes them relatively simple and computationally low in cost. Using a wrapper would arguably produce better results but would also require more assumptions on the data as well as more time in development.

2.1.1 MALDI-TOF sample spectra

As mentioned in earlier sections the resulting spectra from the data acquisition does not directly translate into distinct protein expression intensities, instead there are some uncertainties in the location of each protein expression. Ignoring the factor of detection errors in the device there are still uncertainties that result from the stochastic nature of the measurements such as the initial distribution of speeds of the induced ions and the existence of isotopes as well as differences in the amount of bindings with water molecules. The distributions associated with the first two sources has been investigated to be normal and binomial respectively [8].

Each isotope will produce a bell shaped feature in the spectra with a total intensity following a binomial distribution of the associated protein expression (Fig 3). In the provided sample sets the variance of initial ion speeds are large compared to the weight difference of isotopes, yielding overlapping bell shapes.

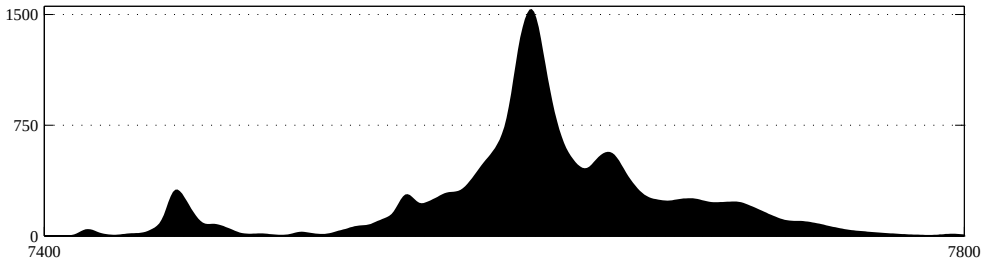

Fig 3: Close view of the largest peak of the spectra and its surroundings.

With the idea that noise will flatten out as the number of samples grow larger and the hope that the sample set will provide a large enough base of measurements to ensure a stable average and that individual peaks will take a solid shape, we hope to be able to identify peak locations by investigating the average spectra of the sample set.

Drift will produce smooth transition.

2.1.2 Contiguous thresholding

Looking at the shape of the spectra [fig3] one can argue that since the bell shaped peaks associated with one protein expression will be poorly resolved, thus not diving to a value close to zero between peaks, it is possible to define a subset of variables as the maxima of every contiguous region of values all larger than a certain threshold. Although the idea is naïve and not very general it should provide a decent result for at least the most prominent peaks, under the assumption that the data samples are of similar nature.

The optimization algorithm in this case is case is remarkably simple and is very easy to implement in a programming environment. D represent the index of the selected variables.

$$\max J(X_D) = \begin{cases} \sum \tilde{x}_d & a \notin C_b \\ 0 & b \neq a, a \in D \end{cases}$$

$$C_a = \left\{ i \,|\tilde{x}_i > T; i = k, k+1, ..., a, ..., l-1, l \right\}$$

For each variable selected the borders of the contiguous regions, k and l, are registered.

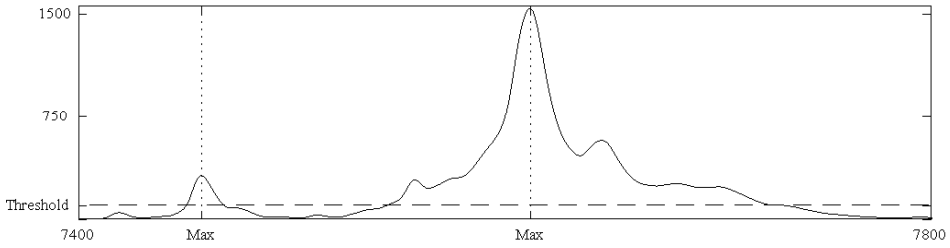

Fig 4: Threshold level marked with a dashed line.

2.1.3 Local maxima extraction

Another approach that would separate isotopes is to define the subset to be the maxima of each bell shaped peak in the spectra. The advantage would be that the method is not dependent on the spectrum being poorly resolved to perform as designed. This approach is hard to formalize but relatively straight forward to implement. To isolate the binomial forms of each protein expression the method can be reapplied to the feature selected set.

$$\max J(X_D) = \begin{cases} 1 & a \notin C_b \\ 0 & b \neq a, a \in D \end{cases}$$

$$C_a = \begin{array}{l} \left\{ k \,|\Delta\tilde{x}_{k-1} < 0; \tilde{x}_k > 0; \Delta\tilde{x}_i >= 0; i = k, k+1, ..., a \right\} \\ \cup \left\{ l \,|\Delta\tilde{x}_{l+1} > 0; \tilde{x}_k > 0; \Delta\tilde{x}_i <= 0; i = a, a+1, ..., l \right\} \end{array}$$

For each variable selected the borders of the peaks, k and l, are registered.

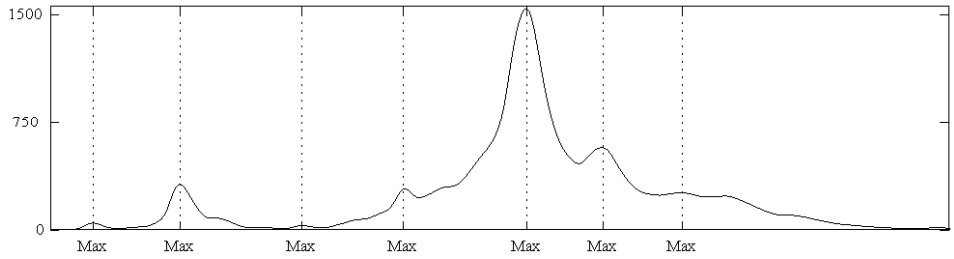

Fig 5: Every maxima marked with dotted lines.

The methods presented in this section are very sensitive to error caused by noise, drift in the samples as well as the very stochastical nature of the peaks the methods rely on. To remedy this uncertainty to some extent it is possible to use more samples from the same distribution, instead of selecting a distinct spectra within a peak we will seek the average or area under the peak. The resulting variables are no longer a subset of the original variables, instead it is a transformation of them which is the essence of Feature Extraction.

2.1.4 Wavelet smoothing

Due to the stochastic nature of the spectra small variations in the shape of the peaks are a natural occurrence, this poses a problem to the local maxima extraction approach as it will yield many false features. To remedy this problem a wavelet decomposition can be used to filter out the components of the highest frequency without distorting the overall shape of the spectra.

2.2 Feature extraction of protein signatures

In general, within a sample set, the number of identified features using the feature selection is still much larger than the number of regions that can actually be distinguished. To narrow down the number of features even further one can apply a transform of the space spanned by the data set and again apply a selection of sub features that characterize the full spectrum of the data. This process of using the full spectra to produce a representation is called feature extraction.

There are different approaches to feature extraction which affords a vast spectra of different results. We have chosen to apply relatively general methods such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA) as they will produce understandable results for most types of data.

2.2.1 Principal component analysis

PCA in its essence is a linear transform of a dataset, which will produce an uncorrelated set of new variables. These new variable will be associated with a set of vectors that together represent an orthogonal basis for the dataset. Another important property of this transform is the fact that the variables will be ranked based on how much of the original dataset they can represent.

PCA belongs to a class of methods based on Projection Pursuit (PP), which searches for lower dimensional subspaces onto which data is projected. In the case of PCA one will seek the to project the data onto the direction of the most variance, these vectors will be called the principal components. This can be done using a least squares based regression line search or by finding the eigenvalues and vectors of the covariance matrix. The latter approach is the key of the Hotelling transform ([1], pp 319-329).

In essence PCA consists of finding the transform:

$$\vec{\xi} = A^T \vec{x}$$

Where $\vec{\xi}$ is the principal components and $\vec{x}$ is a vector whose elements are the variables subject to the transform. The matrix A consist of the eigenvectors of the estimated covariance matrix, these correspond to the stationary values that maximize the variance of $\vec{\xi}$.

The eigenvectors of the covariance matrix can be obtained through Singular Value Decomposition (SVD) ([1], pp 326-327), where the right singular vectors will correspond to the eigenvectors of the estimated covariance matrix ($X^T X$). Where X is the normalized, data matrix adjusted to have zero mean.

2.2.2 Independent component analysis

ICA belongs to a class of blind source separation (BSS) methods for separating additive data components into its underlying source signals. Just as PCA it is based on PP but instead of projecting the data onto the direction of the most variance one wishes to find the directions of maximum statistical independency. Thus the method is based on the assumption that the “source signals” are statistically independent. In summary, one wish to find the set of estimated signal components that can represent the original data while maximizing independence.

A common measurement of independence is to estimate how far the projected data is from a normal (Gaussian) distribution, non-Gaussian. This is motivated by the central limit theorem and is likely a good choice in cases where no special assumptions can be made regarding the data. Non-Gaussian has been the choice of optimization criteria for the implementation used in this project.

Pre-processing the data, such as whitening the variables are used to reduce the complexity of the algorithm and can be performed using PCA.

Thus given a sequence of observations of a set of variables, estimate the “unmixing matrix” and the corresponding original source signals. Based on maximizing the non-Gaussianness of the sources.

$$\vec{s} = W\vec{x}$$

One approach consist of finding w such that $\max J(w^T x)$ with $\|w\| = 1$, where J is essentially a measure of non-Gaussianness (negentrophy). More information on the subject can be found in [7].

2.3 Region definition by segmentation

By producing intensity maps over the contribution of each protein signature over the sample set one can get an idea of how the samples relate to one another. Under the assumption that identifiable regions in the brain have a similar protein structure one can search for spatially and intensity map related samples.

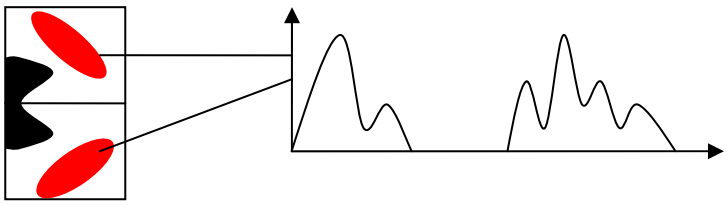

Fig 6: A symmetric region with associated spectra.

The number of regions is assumed to be small compared to the dimension of protein expressions. This fact motivates the use only a small subset of the results of the feature extraction methods suggested in the previous section.

To group the spatially distributed samples into regions, assumption on their nature must be taken. Each regions is assumed to be spatially connected, more or less homogenous and be large enough to be accurately resolved by the data acquisition process. Ideally each region would be uniquely defined by its own signature, this however cannot be certain to hold as more than one region can have the same protein structure. Instead an intensity map targeted for region definition would be desired to consist of fully homogenous regions with jumps at the border toward the neighbouring regions. Defining regions from this can be done by using a sliding threshold or a modified watershed [2], creating a hierarchy of merging regions.

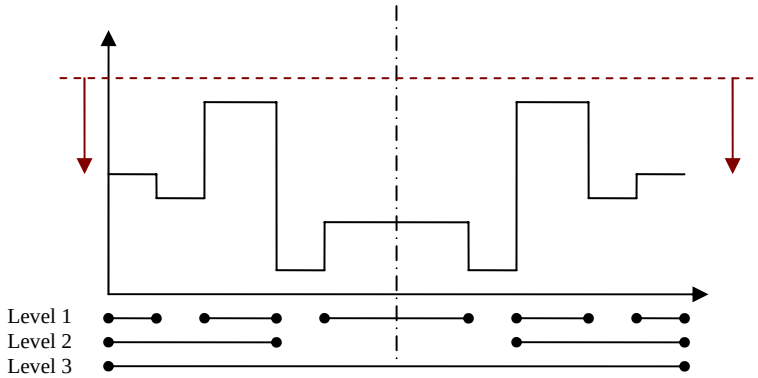

Fig 7: Sliding threshold levels.

To be able to compare the data across the two halves of the brain the datasets need to be partitioned into corresponding sets. Parameters to compare across these sets consist of: Size of the region, intensity distribution, shape and relative location. The last two parameters being hard to consider due to the limited information provided about the possible natural/induced deformations and differences as well as noise and loss of information from discretizing the grid.

The approach used is a rough simplification of the above. The threshold used will be static and located at zero intensity, symbolizing the deficiency/excess of a protein signature across the sample set. Identification of the subsets corresponding to the two halves of the brain will be done manually.

The regions will be matched across the two halves of the brain. Regions that display symmetric qualities will be used to determine a deviation plot over the peak spectra of the samples they correspond to.

2.3.1 Non-spatial segmentation and comparison

As suggested above, the method of choice for segmenting the intensity maps into regions is based on one threshold. This leaves two segments, regions, for every intensity map. This is a low number but for a small sample set containing only a few regions it is a viable choice.

The size of the region will be specified by the number of samples on either side of the segmentation threshold. The intensity distribution will be based on the distance to the threshold.

These regions do not have any spatial restrictions attached and as such they cannot be differentiated from one another by shape or relative locations. This is a large drawback and may result in false positives when matching regions.

If one can assume most spectra in a data set to be symmetric however an additional condition for evaluating symmetry can be found. Using the footprint of the region in the domain of the sample set one can compare the weighted average intensity of every spectra. If the number of spectra that differ under these weights is very large the region can be assumed to be asymmetric.

Thus, comparison can be based on: The size, intensity distribution and the footprint.

The threshold used for this project has been fixed to zero due to the nature of the intensity maps of the extracted features. The distance from zero ymbolizing the deficiency/excess of a protein signature across the sample set.

2.3.2 Successive symmetric region improvement

The extracted protein signatures depend directly on the original subset of the spectra. Since we wish to find regions that display symmetric qualities across the two halves of the brain

It is likely that the feature extraction process would produce signatures tending more towards symmetric regions if asymmetric qualities could be removed from the original variables. Spectra that from the beginning show major differences in average intensity level across the two halves can be removed, but the effect is limited as the average tend to hide local differences.

To remedy the problem of local differences slipping through an initial region definition can be done using feature extraction on the slightly improved dataset. This initial definition will hopefully reveal a set of intensity maps with region-like symmetric qualities. Using the set of symmetric regions extracted from the intensity maps one can compare the weighted average intensity of every spectra over the associated samples. The spectra that differ greatly over any of the symmetric regions will be removed from the dataset.

Repeating this procedure until no more spectra can be removed or until the limit of what is an acceptable dataset is reached.

2.3.3 Regional comparison

When the region definition is satisfactory, the these regions can be used to evaluate the deviation of the original spectra of protein expressions based on the differences across the region’s footprint over the two halves of the brain. The resulting quotas can be used to create a deviation plots conveying information on spectra that might be of interest for further study.

3 Results

3.1 Feature Selection

The methods developed for this task takes parameters corresponding to low-pass filtering of the dataset, setting a minimum intensity of a maxima to be recognized as well as a threshold value for the total area of a peak. Before applying this implementation an estimate of identifiable peaks were taken from 24[3] as around 1400 distinct peaks, attempts was made to get values close to this measure as well as explore more restrictive choices of parameters.

The spectra of each sample set is in the scales of over 20000 parameters (Fig 8), exploring every one of them would be very hard. In this case however there is a clear spatial dependence among the samples making it possible to review a spectra over the full sample set as an image. As an example the highlighted section of Fig 8 is depicted as images (intensity maps) of every successive spectra in Fig 9. Many of these images look the very similar, which is a product of the fact that the spectra doesn't directly reflect the protein expressions.

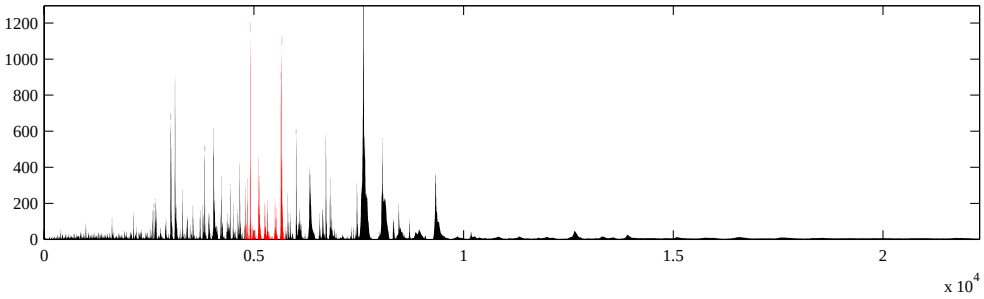

Fig 8: Full MALDI-TOF spectra, highlighted region used for closer study.

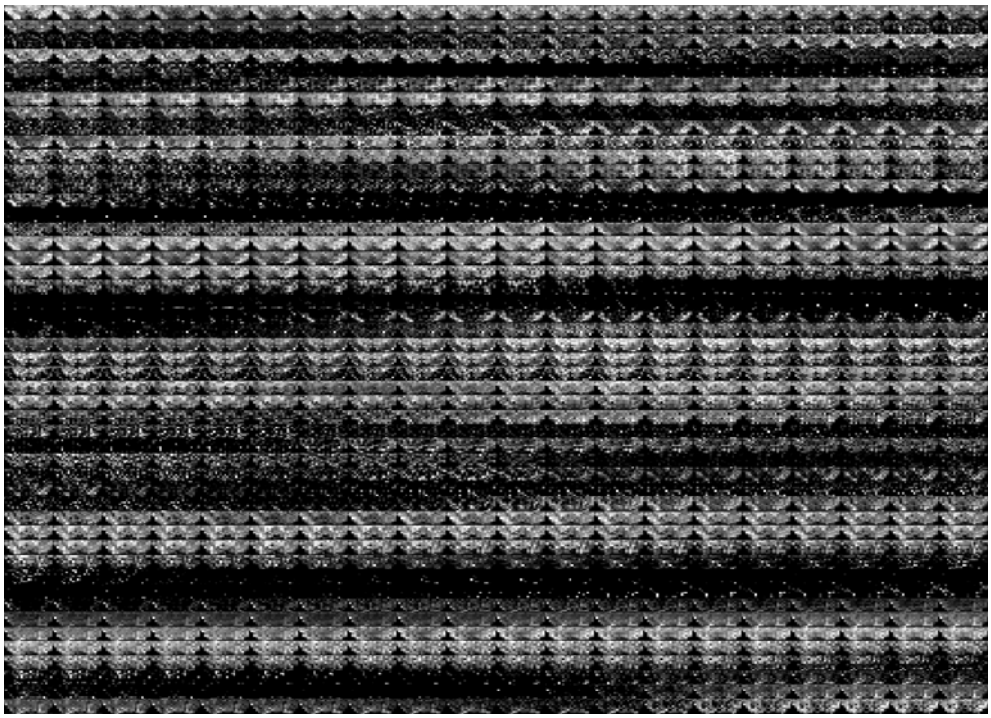

Fig 9: Intensity maps assiciated with every spectra higlighted in Fig 8.

Applying the least restrictive approach of feature selection attempted in this report will greatly reduce the redundancy in the set of features. The peaks identified using local maxima extraction on the highlighted region of Fig 8 are marked in Fig 10. The resulting data set has been reduced from over 800 spectra to around 50, while still retaining the important classes of images as can be seen in Fig 11. There is still some minor tendencies to redundancy in this set however.

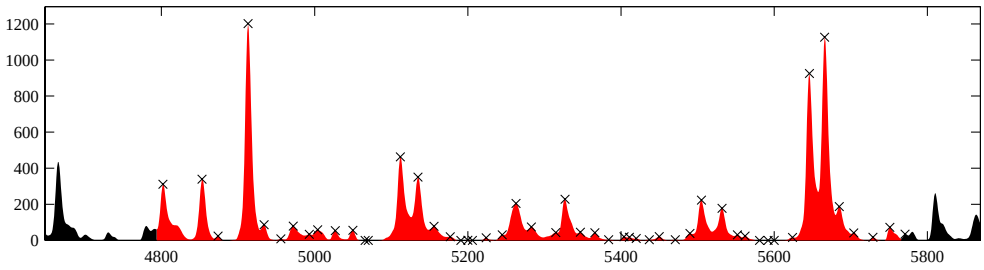

Fig 10: Extracted features, local maxima, marked with x.

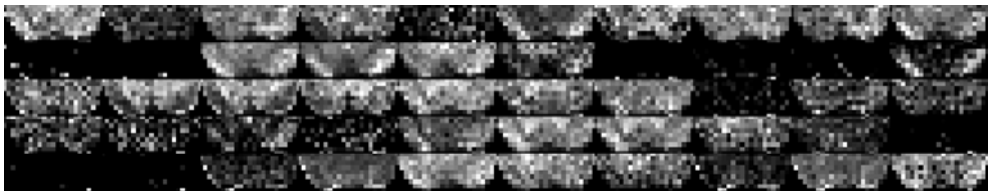

Fig 11: Intensity maps associated with extracted features of Fig 10.

Exploring the simplest approach to the feature selection on the same set as in the above discussion we can see that the number of selected spectra is lower (Fig 12). The contiguous threshold will not identify the smallest peaks as they fall below the threshold value and will group peaks that do not resolve well enough to break the condition of contiguous regions. In this case the result is somewhat better than local maxima as the smaller peaks sensitive to noise are removed and redundancy is very slightly improved (Fig 13).

As some peaks identified by the approach can consist of several poorly resolved sub-peaks one can apply a transform to the feature set by averaging over the span of the peak. This transform would ideally improve the quality of the image by using a larger set of measurements from the same statistical distribution. Improvements are minor at best in this case as seen in Fig 14.

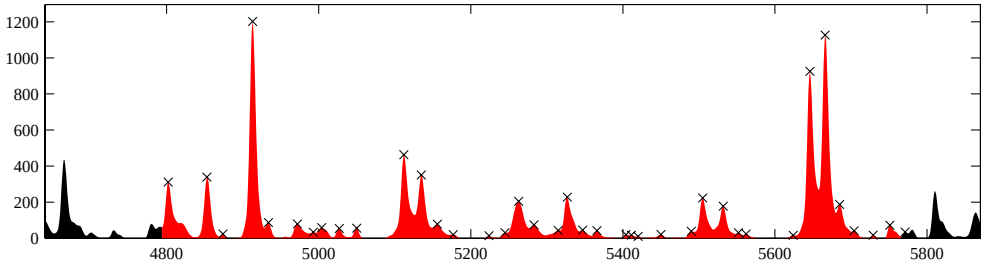

Fig 12: Extracted features, contiguous threshold, marked with x.

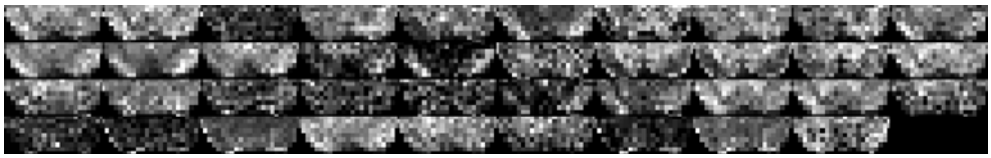

Fig 13: Intensity maps associated with extracted features of Fig 12.

Fig 14: Intensity maps of the average, transformed, intensities of the contiguous threshold.

The most restrictive approach used is the local maxima selection applied twice in succession to isolate the peaks of every cluster of sub-peaks. As can be seen the number of isolated features are reduced considerably and the redundancy of the data is as good as gone. This is at a cost however as can be seen in Fig 16 as some features meld together, the somewhat clear definition of substantia nigra is lost. However this feature only corresponded to one isolated peak of Fig 11 which may indicate it is either insignificant or misplaced and may appear somewhere else in the spectra as a better resolved peak. Using a transformation on this feature set is more rewarding than the previous attempt as can be seen in Fig 17.

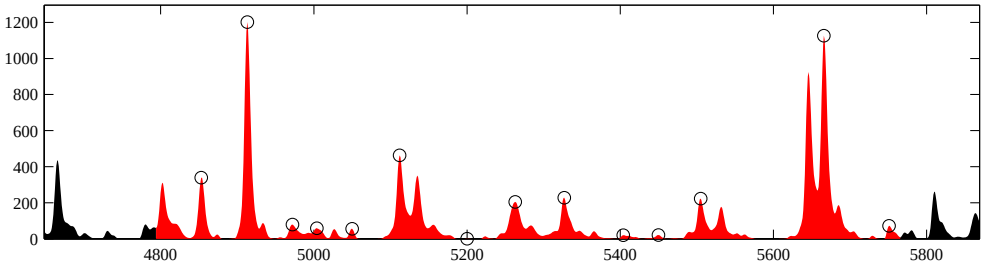

Fig 15: Features extracted, double-pass local maxima, marked with o.

Fig 16: Intensity maps associated with extracted features of Fig 15.

Fig 17: Intensity maps of the average, transformed, intensities of double-pass local maxima.

To add additional constraints to the methods investigated threshold values for identifiable peak sizes have been implemented, which allow filtering of too small peaks in both maximum height and transformed size. Low pass filtering is also applied to filter out smaller variations within the peaks. Depending on the choice of parameters and method the number of features can be chosen in a range of about 2000 to 200 which is a reduction of between 90% and 99% of the original spectra.

The overview conducted in this section suggest that a the more restrictive approach may exclude features that can be identified with less restriction but noise levels and the number of replicated patterns are lower. Before deciding on a set of parameters the resulting feature set should be carefully reviewed.

Careful investigation of the extracted features over the full spectra suggest that a large number of the identified peaks found using local maxima extraction are located in the upper end of the range. This part of the spectra is dominated by low intensities which are more sensitive to noise as well as drift among the samples. The intensity images in this region are also hard to visually interpret due to seemingly random variations which suggest that classification errors will occur.

3.2 Feature Extraction

All the data used in this section has been normalized and had its mean adjusted to zero to produce good results using both PCA and ICA.

PCA will have a natural order of it's components based on the amount of the total variance they represent. This makes it very useful for finding a reduced dimension representation of the data, as can be seen in Fig 18 and Fig 19 the components projected in the lower end do not display any understandable features. Fig 18 display the result of using the full spectra (22000 spectra) of a sample set to produce the principal components. Compared to Fig 19, which display the result of PCA on a very restrictive peak selection (117 peaks), there are actually about the same amount of visually distinguishable regions.

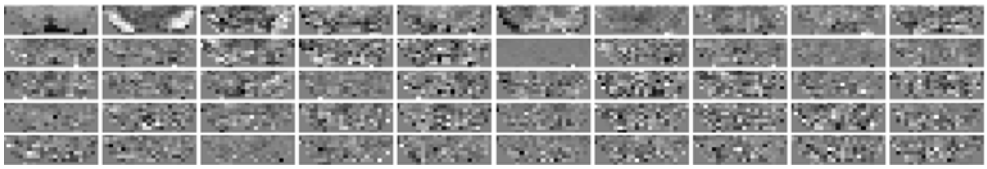

Fig 18: PCA applied to normalized data, corresponding to the unreduced original spectra.

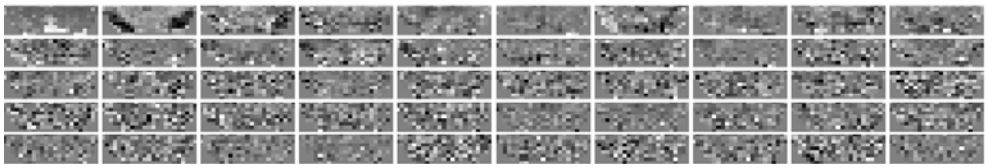

Fig 19: PCA applied to a very restrictive set of selected features.

After using the set of PCA components to whiten the data, ICA can be used to find the statistically independent features. These components do not have a natural measurement of importance but instead tend to produce well defined regions over the full set of reduced variables. In Fig 20 the result of PCA displayed in Fig 18 is used to whiten the data and create a set of independent components. Regions are clear and some regions not apparent through PCA are well defined using ICA When using the very restrictive dataset as a basis for ICA one can see that the difference is slight here as well.

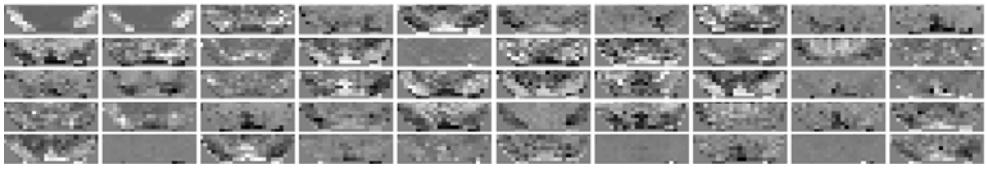

Fig 20: ICA used to improve the output of PCA, full original spectra.

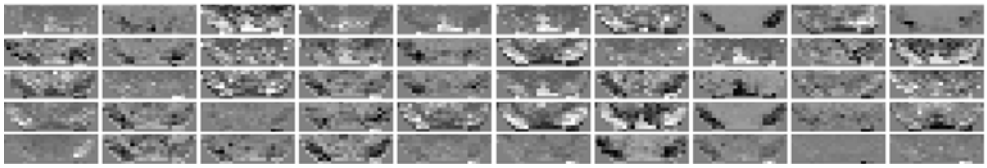

Fig 21: ICA applied to a very restrictive set of selected features.

To conclude this section one can point out that while PCA will provide natural ranking of components, the number of features the method can successfully extract is very limited. ICA would be the natural choice if some tendencies towards redundancy are acceptable. Another interesting conclusion is the fact that a very small subset of the original spectra is capable of producing results almost indistinguishable from the original full spectra. While the protein expressions might need to be selected with less restriction to produce meaningful results to the scientist in the end, the basis for the region definition could be made up by a much smaller subset, which would speed up the algorithm considerably.

3.3 Detected regions and symmetry

As the basis of regional separation was chosen to be the zero, two sets of regions will be identified for each feature. Many parameter choices has been implemented for this algorithm due to the difficulty in defining a strict maximization criteria from the limited amount of available information.

As described in the theory section symmetric regions can successively refined by removing asymmetric spectra through repeated feature extraction and regional comparison. The method will successively remove peaks that differ among matching regions, in the end resulting in an asymmetric and symmetric subset of the spectra. In this section one run through the algorithm will be presented and analyzed.

The original data set consist of the 117 peaks partially displayed in Fig 23 (asymmetric) and Fig 24 (symmetric). This run make use of ICA with a limit of 5 components. The regions resulting from the first iteration of the algorithm can be observed in Fig 22, which already display quite symmetric qualities.

After 10 iterations the algorithm terminates and 17 of the 117 peaks have been removed. In this case no symmetric regions were misclassified, but due to the inexact approach used there are cases when a few symmetric peaks will be lost. As can be seen by reviewing the symmetric spectra some asymmetric features still remain, but in this case it is usually better to chose the parameters to be restrictive to avoid a total collapse with almost every peak being removed.

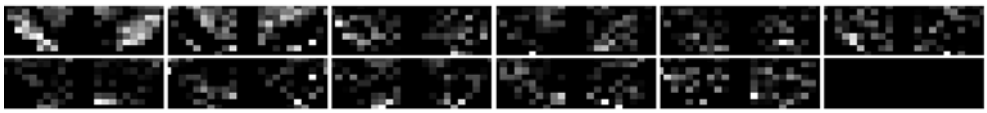

Fig 22: Symmetric regions at the first iteration.

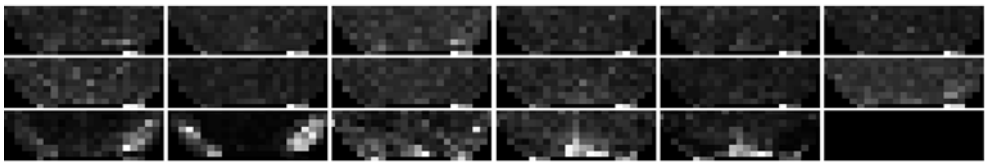

Fig 23: Intensity maps of spectra classified as assymmetric.

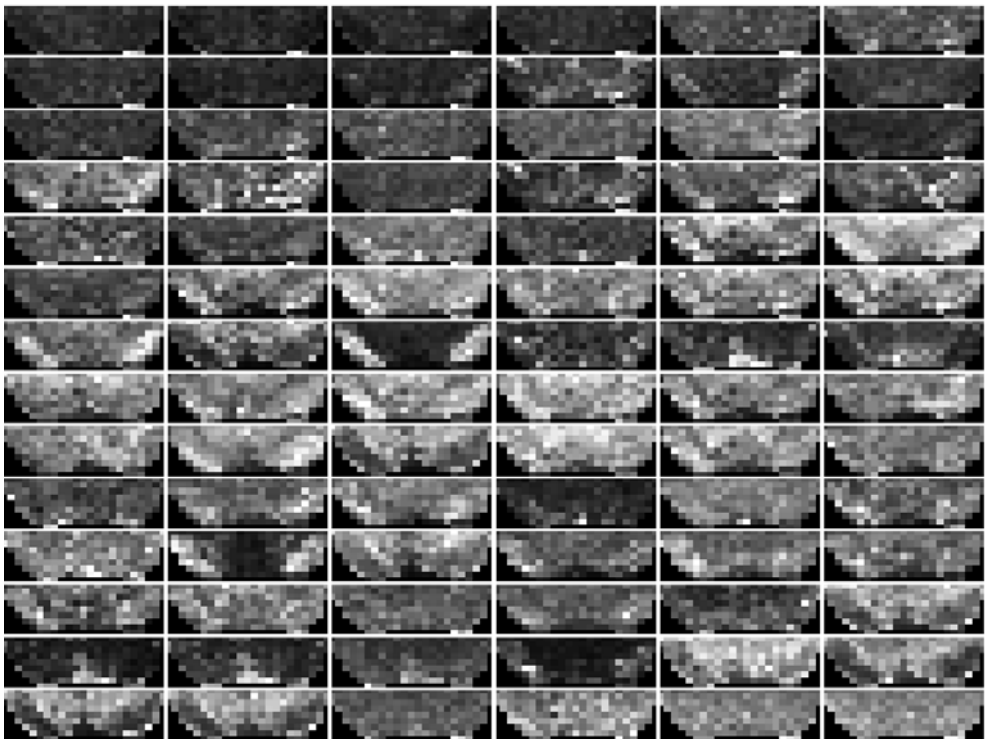

Fig 24: Intensity maps of spectra classified as symmetric.

The final set of regions look much the same as the first iteration but slightly improved as one of the not fully symmetric regions has been removed.

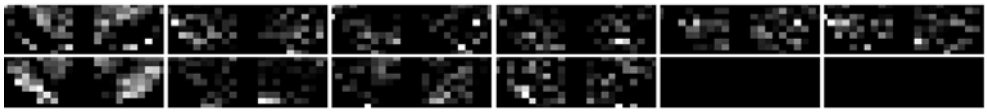

Fig 25: Symmetric regions at the last iteration.

Using one of the more well refined regions the spectra in its footprint can be compared. The result is displayed in Fig 26 and Table 1.

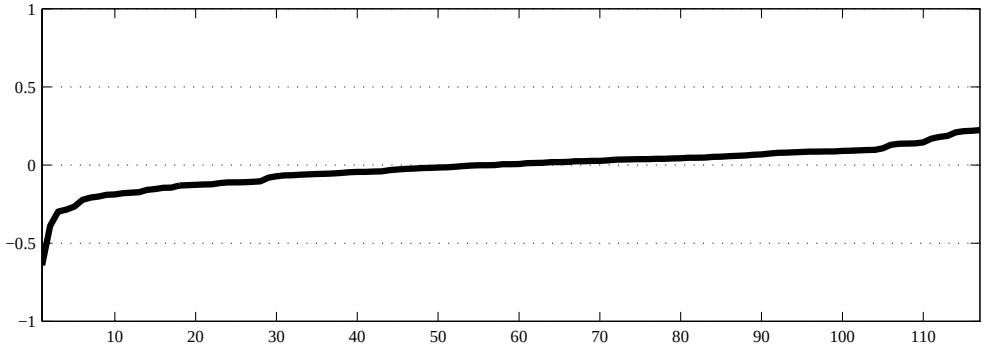

Fig 26: Deviations in (restrictive) spectra found using a symmetric region.

Peak index	Deviance	Original index
1	-0.071	186
2	-0.057	325
3	-0.066	387
4	-0.065	442
5	-0.015	506
6	0.087	581
7	-0.061	628
8	0.019	715
9	-0.004	767
10	0.057	813
11	0.006	862
12	-0.124	948
13	-0.054	995
14	-0.208	1055
15	0.034	1122
16	0.040	1191
17	0.020	1239
18	0.019	1289
19	0.006	1363
20	-0.014	1456
21	-0.115	1567
22	-0.110	1617
23	0.082	1650
24	-0.046	1765
25	-0.285	1841
26	0.048	1917
27	-0.056	1956
28	0.079	2053
29	-0.081	2131
30	0.036	2196
31	-0.222	2258
32	0.169	2288
33	0.087	2321
34	0.139	2423
35	-0.027	2526
36	-0.123	2648
37	-0.032	2676
38	-0.059	2730
39	0.092	2767
40	0.078	2906
41	0.013	2962
42	-0.159	3017
43	-0.128	3121
44	0.005	3259
45	0.038	3302
46	-0.266	3409
47	0.130	3479
48	0.187	3547
49	0.094	3784
50	-0.025	3828
51	0.096	3921
52	-0.173	4004
53	0.137	4042
54	0.107	4233
55	-0.041	4304
56	0.090	4362
57	0.042	4435
58	0.136	4505
59	0.074	4554
60	0.014	4618
61	-0.104	4665
62	0.060	4780
63	-0.007	4853
64	0.180	4913
65	0.053	5004
66	-0.018	5112
67	0.025	5263
68	0.209	5327
69	-0.126	5505
70	0.027	5666
71	-0.019	5751
72	0.047	5810
73	0.218	5864
74	0.040	6008
75	0.013	6086
76	0.097	6218
77	-0.180	6327
78	-0.154	6558
79	0.219	6643
80	0.038	6718
81	0.145	6816
82	0.027	6858
83	-0.298	6917
84	0.024	6958
85	-0.050	7114
86	-0.041	7323
87	0.068	7458
88	-0.001	7612
89	-0.188	8063
90	0.086	8114
91	0.043	8319
92	-0.177	8455
93	0.086	8709
94	0.062	8851
95	-0.044	8943
96	-0.111	9334
97	0.224	9406
98	-0.641	9836
99	-0.000	9859
100	-0.108	10178
101	0.066	10266
102	-0.144	10822
103	-0.131	11336
104	-0.022	11968
105	0.053	12030
106	-0.146	12647
107	0.037	13313
108	-0.011	13559
109	0.047	13916
110	-0.001	15094
111	-0.191	15754
112	-0.201	16491
113	-0.111	16574
114	0.084	16614
115	-0.044	17569
116	0.030	17629
117	-0.390	17661

Table 1: Deviance and original index associated to a peak index.

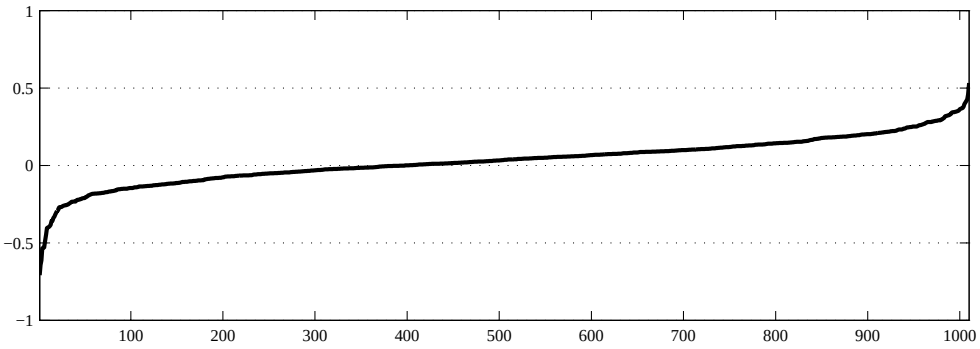

Fig 27: Deviations associated with a more realistic spectra than Fig 26..

4 Implementation

As explained in the introduction the platform of choice is the programming environment provided by MathWorks' MATLAB. The implementation consist of a collection of MATLAB scripts, m-files, that make use of the tools available from the environment in combination with a few additional packages that contain effective implementations of well-known algorithms. In particular, packages are used for particular implementationa of PCA, ICA and wavelet smoothing.

The over all functionality structure for the collection scripts can be summarized into the following categories: Data incorporation; reading and partitioning the binary data sets. Dimensionality reduction; from peak selection and protein expression identification. Region definition/analysis; finding symmetries, defining and improving regions. And the final comparison and visualization.

4.1.1 PROPACK

PROPACK is a software package with a set of functions for computing the singular value decomposition of large and sparse or structured matrices. The interesting routines to this project are the implementation for finding eigenvalues and vectors. Eigen values and eigenvectors of a hermitian matrix

4.1.2 FastICA

One widely used implementation of ICA is the FastICA algorithm. It is relatively fast, memory efficient and self-sufficient, meaning it has the possibility to take care of the necessary pre-processing of the data. FastICA uses a fixed-point optimization scheme based on Newton-iteration which has been tested to be several magnitudes faster than conventional gradient descent based methods. FastICA can search for the independent components one at a time or all at once. Its performance can also be tuned somewhat by choosing from a range of nonlinearities..

4.1.3 Rice Wavelet Toolbox

The rice wavelet toolbox (RWT) is a combined collection of MATLAB M files along with compiled C MEX files which is used for design and analysis of 1D and 2D wavelet. It is open source software and distributed free or charge. The toolbox offers tools for de-noising and interfaces directly with MATLAB code.

To interface with this toolbox a script developed at University of Texas that was developed as part of a research project dedicated to investigating MALDI-TOF spectra [6].

5 Discussion

5.1 Performance

As discussed in the previous section the region definition does not improve dramatically by performing the feature selection on the spectra, worth mentioning however is that the computational demand of the ICA will be greatly reduced, speeding up the process by a large factor. This would be increasingly beneficial when expanding the sample set by increasing resolution, spatial extension of the tissue sample or extension to include consecutive slices in an attempt to produce three dimensional regions.

Another issue related to the remark about increasing the size sample set is that the cost of forming the covariance matrix to determine the principal components will increase by a power of two. This could lead to issues regarding memory efficiency and using SVD instead of direct eigenvector calculation would be beneficial.

An improvement that might be seen by expanding the sample sets however may be that the peaks will have a larger number of statistical observations to rely on, which can lead to a more accurate feature selection.

The largest limitation of this work is the inability of the region definition procedure to use subsets of the full sample set or match regions spatially. This will lead the algorithm to be very likely to collapse when expanded to include sample sets containing a larger number of regions.

5.2 Possible improvements

Coupling the spectrum based comparison of the two halves of the brain with an algorithm that uses more spatial information as well as some known data of regions could improve the performance a lot. Although for this to be possible, problems of noise, deformations, natural differences and loss of information due to the small resolution of the sample sets must be addressed in detail. Analyzing the sample procedure, the biological structure of the brain as well as using manually identified and labelled regions to assist in finding relevant parameters and their limits. Fitting known regions to the data might be helpful to extract the relevant differences.

The feature extraction procedure used to produce the regions may be improved by imposing known limitations of the transformation. The ability to retrieve strictly positive results would for example be beneficial in this case as negative densities are not physically plausible. This is not guaranteed in the case of PCA nor the implementation of ICA used for this report.

5.3 Sources of error

As in any analysis that rely on measured quantities the accuracy of the samples are of utmost importance, by following well supported guidelines and a good understanding of the nature of the measurements the errors can be minimized. The most important possible sources of error that are not implementation specific consist of:

1. Calibration error of the measurement equipment, may produce drift or scaling problems in the data causing mismatched spectra and in the end deformed intensity maps.
2. The possibility of deformation of the tissue when isolating the relevant regions. This would lead to deformations in the sample sets as well and make the region matching harder.
3. Uneven sample preparation, different thickness of the slices or uneven distribution of matrix across the tissue. As a result there may be variations in scaling of the sample spectra, displayed as variations in the intensity maps.
4. Spatial discretization of the sample may result in smaller features to be lost, such as line segments (and plane regions in 3D).
5. The discretization of the mass spectra due to the time interval used as well as the energy transferred from the laser pulse, will make individual spectra harder to resolve.
6. The signal must be strong compared to the noise level in order to avoid false positives, the stochastical nature of the variables requires a certain number of observations to be able to discriminate them from noise. May produce false positive identification of peaks.

5.4 Extension to 3D

The demands of extending the algorithm to work with sample sets connected in three dimensions would drastically increase the amount of processed data and thus increase the need for memory efficiency.

The large amount of data may make it hard to keep the sampling device well calibrated through every set of samples and means for basic synchronization of peaks between slices would be advisable, both for scaling and drift.

There would also be need for a spatial adjustment to correct for variations in deformations caused by the individual cuts. As well as slight differences in the orientation of the grid over the tissue sample.

The methods as they are should be easy to extend to 3 dimensions, but as long as spatial information is not used the number of regions one expect to find within the sample sets must remain low.

6 Conclusion

Mass spectrometry data obtained from a brain-half-symmetric tissue samples was analyzed to extract region definitions and corresponding protein distributions. This was implemented using different feature selection approaches such as subset selection and linear transformations.

The regions were then examined for symmetry and compared across the centre axis of the brain, projecting a constraint of symmetry across the protein spectra over the regions. The level of symmetry of protein expression was measured by the relative deviation from the weighted average intensity.

Results show that restrictive approaches to feature selection across the spectra perform equally well compared to the full spectra when used as a base for feature extraction based on PCA and ICA. ICA will produce superior results to PCA when it comes to the number of identifiable regions and general quality of the produced intensity maps. The region comparison approach attempted should however be complemented with spatial information based on the nature of the sample sets and sample preparations procedure.

7 Acknowledgements

The accomplishment of this project work has been one of the many significant academic challenges we have faced so far. Without the support, patient and guidance of the few people, this study would not have been completed. It is to them we owe our deepest gratitude. First, we thank our supervisor Carl Nettelblad, for his continuous support in this project course. Carl was always there to listen and to give advice. He taught us how to ask questions and express our ideas. He showed us different ways to approach a research problem and the need to be persistent to accomplish any goal. We give our heartiest gratitude to our project architect and adviser Malin Andersson who is responsible for involving us in this project in the first place by creating an opportunity to work further from her previous research work. We would like to thank Stefan Pålson, study adviser in Scientific Computing Department, Uppsala Universitet for setting us up as a team and allowing us to work on this course. Finally a very special appreciation and thanks must go to Maya Neytcheva, Scientific Computing Department, Uppsala Universitet for her all kind of support, continuous follow up of the project and allowing us working on a divisive project in the field of biology to relate with computational mathematics which actually worked out as a super motivation for us to accomplish the given task productively.

References

[1] A. Webb, *Statistical Pattern Recognition* (Second Edition), Wiley, Chichester, 2002

[2] R. C. Gonzalez & R. E. Woods, *Digital Image Processing* (Second International Edition), New Jersey, Prentice-Hall, 2002

[3] M. Andersson, M. R. Groseclose, A. Y. Deutch & R. M. Caprioli, *Imaging mass spectrometry of proteins and peptides: 3D volume reconstruction, Nature Methods*, Vol. 5 No. 1 pp.101-108, 2008

[4] D. S. Cornett, M. L. Reyzner, P. Chaurand & R. M. Caprioli, *MALDI imaging mass spectrometry: molecular snapshots of biochemical systems*, Nature Methods, Vol. 4 No. 10 pp. 828-833, 2007

[5] R. Van de Plas, F. Ojeda, M. Dewil, L. Den Bosch, Bart De Moor & E. Waelkens, *Prospective exploration of biochemical tissue composition via imaging mass spectrometry guided by Principal Component Analysis*, Pacific Symposium on Biocomputing 12 pp. 458-469, 2007

[6] J. S. Morris, K. R. Coombes, J. Koomen, K. A. Baggerly & R. Kobayashi, *Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum*, Oxford University Press Bioinformatics, Vol. 21 no. 9 pp. 1764-1775, 2005

[7] A. Hyvärinen & E. Oja, *Independent Component Analysis: A Tutorial* [1], <http://www.cis.hut.fi/aapo/papers/IJCNN99_tutorialweb/>, 2008-06-15.

[8] K. R. Coombes, J. Koomen, K. A. Baggerly, J. S. Morris & R. Kobayashi, *Understanding the Characteristics of Mass Spectrometry Data Through the Use of Simulation*, Cancer Informatics, 1.1, 2005.

[1] Revised version*: Independent Component Analysis: Algorithms and Applications*, Neural Networks, 13(4-5):411-430, 2000