

POLITECHNIKA WROCŁAWSKA
WYDZIAŁ ELEKTRONIKI

KIERUNEK: ELEKTRONIKA I TELEKOMUNIKACJA

SPECJALNOŚĆ: ZASTOSOWANIA INŻYNIERII KOMPUTEROWEJ

PRACA DYPLOMOWA
MAGISTERSKA

Analiza pomiarów spektroskopem masowym
SELDI-TOF-MS mająca na celu wyodrębnienie
próbek nowotworowych jako podstawa
internetowego serwisu wspomagającego pracę
lekarza specjalisty

SELDI-TOF-MS data analysis for cancer detection
as a base for online diagnostic software

AUTOR:
Marcin Radlak

PROWADZĄCY PRACĘ:
Dr inż. Ryszard Klemous

OPIEKUN:
Dr inż. Ryszard Klempous

OCENA PRACY:

Abstract

In this report, the process of detecting cancerous/healthy sample based on SELDI-TOF-MS data set is presented: from raw data, through pre-processing to classification. Methods and algorithms, their properties and suggested implementation ideas are presented. They are collected in an organized way and aim to present the state of the art over current research. Additionally, idea of 24h/day distributed work organization is introduced. This form of research and projects realization, which includes many participants that are not limited by geographical location or time, may provide an excellent opportunity to accelerate current work in this and many other fields. Finally, comparison of classifying methods is presented and dependency of preprocessing methods into classifier performance is evaluated using designed for the purpose of this project Mass Spectrometry Analysis Software software.

Streszczenie

Praca ta została napisana w celu przedstawienia procesu detekcji próbek nowotworowych / zdrowych z wykorzystaniem spektrometru masowego SELDI-TOF-MS: od surowych danych, poprzez metody przetwarzania kończąc na klasyfikacji. Metody i algorytmy, ich właściwości i sugerowane sposoby implementacji zostały opisane w usystematyzowany sposób mający na celu przedstawienie obecnego stanu wiedzy w tej dziedzinie. Dodatkowo, idea 24/7 została opisana. Ta forma współpracy pomiędzy grupami badawczymi, które nie są ograniczone pod względem lokalizacji czy czasu, może przyczynić się do przyspieszenia prac zarówno w tej jak i innych dziedzinach nauki i zarządzania projektem. Na koniec, wykorzystując napisane na potrzeby pracy oprogramowanie do analizy danych spektrometru masowego (MSAS), przedstawiono porównanie opisanych metod klasyfikacji oraz wpływ wstępnego przetwarzania danych wejściowych na ich wydajność.

Keywords: SELDI-TOF-MS, Mass Spectrometry, data analysis, cancer detection.

Acknowledgements

I would like to thank my supervisor, Dr Ryszard Klempous for advice and constant help during the process of preparation, literature review and planning. I also would like to thank my friends for patience and understanding.

Finally, I would like to thank my parents for education they gave me and for continuous support from their side which allowed me to achieve everything I have.

Contents

Introduction	1
1 Introduction	1
1.1 Project organization	2
2 Proteomics	4
2.1 Mass spectrometry	5
2.1.1 Electrospray Ionization	5
2.1.2 Matrix Assisted Laser Desorption Ionization	5
2.1.3 Surface-Enhanced Laser Desorption/Ionization	7
2.2 Experimental Steps	9
2.3 Theory conclusion	10
3 Sequence of data analysis	11
3.1 Preprocessing	14
3.1.1 Dimension reduction	14
3.1.2 Miscellaneous	15
3.1.3 Peaks manipulation	17
3.2 Data Analysis and Classification	19
3.2.1 Heuristic approach	21
3.2.2 Exact approach	23
3.3 Verification	24
4 Software Design Elements	26
4.1 Requirements for data analysis Tool	26
4.2 24/7 work organization	27
4.3 Steps of software design	30
4.3.1 Analysis	30
4.3.2 Design	30
4.3.3 Paper prototyping	31

4.3.4	Construction	31
4.4	Implementation of proposed solutions	31
5	Case study	33
5.1	Data set characteristics	33
5.2	Program requirements	33
5.3	Program demonstration	34
5.3.1	Step 1: Loading Data	34
5.3.2	Step 2: Preprocessing	36
5.3.3	Step 3: Analysis and Classification	39
5.3.4	Step 4: Results	39
5.4	Additional features	40
5.5	Classifier performance according to preprocessing methods	40
6	Results and conclusion	43
6.1	Strengths, weaknesses and future developments	44
6.2	Final word	44
	Acknowledgements	44
	Appendix A: Instruction for program execution	44
	Bibliography	45

List of Figures

2.1	Electrospray Ionization (ESI)	6
2.2	Matrix Assisted Laser Desorption Ionization (MALDI)	7
2.3	Sample Spectrogram	8
2.4	Spectrometry block diagram	8
2.5	Steps of sample processing for SELDI-TOF-MS	10
3.1	Sample Spectrogram	11
3.2	Example differences between samples	12
3.3	Data set presented in a Heat Map form	13
3.4	No baseline correction applied	15
3.5	Baseline correction applied	16
3.6	General Idea of Principal Component Analysis	17
3.7	Sample Mean Spectrogram	18
3.8	Sample Mean Spectrogram with detected Peaks	18
3.9	Classification example: a) input data (unclassified), b) two classes, c) six classes, d) four classes	20
3.10	Classes: a) Well separated, b) overlapping	21
3.11	Model of Artificial Neural Network (ANN)	22
3.12	Probability density of two class samples (after: Vitzthum et al. [17]	25
4.1	Time zones and example research centers arrangement	28
4.2	Schedule of work: a) one research center b) 3 research centers	29
5.1	Screen shoot of the Mass Spectrometry Data Analysis program	34
5.2	Zoomed unprocessed Spectra plotted as Standard Plot using Plot Tool	35
5.3	Loaded Spectra after baseline correction applied	36
5.4	Loaded Spectra not smoothed (left) and after smoothing (right)	38

Chapter 1

Introduction

There is no doubt that cancer is very serious disease which every year affects millions of people all over the world. To reduce this amount, extensive work by research centers, pharmaceutical companies and charity organizations is conducted to develop methods of prediction, prevention and treatment. The purpose of this project is to introduce the concept of cancer detection using data obtained from SELDI-TOF-MS. This type of Mass Spectrometry has been proposed, because of its ability to produce high resolution spectrogram of proteins content in an organic sample. Assuming, that cancerous cells consist of proteins which are usually absent in healthy tissue, there is a hope to develop a method being able to distinguish between those two states, giving a solid base for real diagnosis which doctors have to make.

Although, hardware is a breakthrough in the field it is still not precise enough to produce superior results expected from diagnosis equipment. Low repeatability of results, high noise and huge amount of data are only few of the difficulties encountered. Therefore, to supplement hardware deficiency, it is important to utilize more or less intelligent data analysis methods. Properly selected might, improve accuracy of the hardware. In this report, the full process of detecting cancerous/healthy sample is presented: from raw data, through pre-processing to classification. Methods and algorithms, their properties and suggested implementation ideas are presented. They are collected in an organized way and aim to present the state of the art over current research. Additionally, idea of 24h/day distributed work organization is introduced. This form of research and projects realization, which includes many participants that are not limited by geographical location or time, may provide an excellent opportunity to accelerate current work in this and many other fields. Finally, dependency of preprocessing methods into classifier performance is evaluated using, designed for the purpose of this project, analysis software. The project is organized as described in detail in the following section.

1.1 Project organization

The work is organized as follows:

- Chapter 2 - This is to introduce the concept of Mass Spectrometry, general terms and technology, to provide an overview of the hardware capabilities and deficiencies. The aim is to answer two main questions: "How is it possible to analyze organic samples digitally and what type of data the hardware produce. The technological process of deriving data is also described in this chapter.
- Chapter 3 - This chapter presents methods used for data manipulation. The sequence of analysis is described: from raw data derived from mass spectrometer, through preprocessing towards classification. Moreover, the methodology used for testing classifier is also described. This chapter is to provide a theoretical foundations of available methods, which developed Mass Spectrometry Analysis Software (MSAS) is based on.
- Chapter 4 - This chapter is to provide, overview of methodology for software design, particularly Graphical User Interface (GUI). Key points on usability and other factors crucial to produce useful application, aimed for data analysis. Furthermore, to accelerate future work, an idea of 24h work organization is presented: general introduction to this novel approach, advantages, disadvantages and implementation proposal for this specific Analysis Software is also a part of this chapter.
- Chapter 5 - In this chapter, case study on sample data for Ovarian Cancer freely available from National Cancer Institute ¹ is described. Step by step description of how developed analysis software (MSAS) can be used for classification and what results it provides. Furthermore, dependency of preprocessing methods for the classifier performance is analyzed. Achieved results, are the basis for discussion about future work in the field, giving guidelines about what should be done and what should be avoided.
- Chapter 6 - Final chapter, which consists of a summary of the work performed. Strengths and weaknesses, summary of conclusions and ideas for future development are suggested.

This project is mostly focused on software development and analysis of sample data. Thus sometimes just a brief theory is described because of the limited scope

¹<http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp>

of this project. However there is always a reference suggested which, in author's opinion, is great complement to this work and which describes theoretical issues in detail.

The reason why the focus has been put on algorithms of data analysis is because it is not possible to create well performing and useful online application addressed to community of doctors. The software should work as a support for diagnosis to be done. Therefore, instead of approaching problem with poor theoretical background of process of sample classification as diseased/nondiseased, pressure has been put to create solid foundation for further work.

Chapter 2

Proteomics

To be able to fully understand what steps are required to follow, how to prepare experiments and how to interpret their results it is important to have an understanding of fundamental definitions of corresponding terms. The most important are presented as follows, together with their brief explanation.

Genetics is a study of heredity. Genome is the total hereditary material - composition of genetic information of an individual. The size is usually defined as the number of base pairs. (i.e normal human genome consists of 3 billion base pairs of DNA). Ability to store information in genes lead to projects which main goal was to create directory of genes that can be used to answer questions such as what specific genes do and how they work. Because of this, branch of genetics called genomics has emerged.

Oxford English Dictionary [11] defines it as: "The scientific study of genomes, esp. of their organization and evolution, using nucleotide sequencing and gene mapping". It is - simply speaking - the study of how an individual's genes interact with each other and with the environment to create the complexity of life.

By the analogy to genome, term proteome was introduced. While the first one relates to information about genes, proteome relates to information about proteins produced from the information encoded in genes.

According to *Oxford English Dictionary* [11] proteomics is: "The study of the set of proteins expressed by an organism, its relationship to the genes coding for them, and to physiological and pathological processes". [PROTEOME n. + -ics, after GENOMICS n.]".

The question which would be perfect to ask now is how genome and proteome relate to each other? One gene can give instructions to about 100 proteins. The biggest difference is that while it is possible to find what genes functions are, this task is much harder with proteins. Different configurations of proteins create different

molecules. Therefore, instead of describing each single protein, an idea of looking for general patterns they form has been introduced - this is proposed solution for the problem that nobody knows which protein is "good" and which is "bad" one.

At this stage the difficulty of how to analyze proteins has arisen. The solution was mass spectrometry, which can measure masses of molecules (mass to charge ratio - M/Z) in an organic sample. How is it possible? Following sections introduce theoretical concept of this idea.

2.1 Mass spectrometry

Mass spectrometry is based on conversion of proteins in sample into ions, isolating them and detecting according to the ratio between mass and charge. This general definition introduce many technological problems with this approach of which the greatest one is ionization - how to ionize molecules to avoid splitting into smaller parts what could result in protein being detected as two smaller mass proteins.

According to Aebersold and Goodlett [1], during the decade of the 1990s, changes in MS instrumentation and techniques revolutionized protein chemistry and fundamentally changed the analysis of proteins. These changes were catalyzed by two technical breakthroughs in the late 1980s: the development of the two ionization methods electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI)

2.1.1 Electrospray Ionization

ESI uses high voltage to change the state of substance, from liquid to gas. A liquid is passed through nozzle (as presented on figure 2.1¹). By charging the liquid electrically to a very high voltage it becomes unstable because it has to hold more charge. Finally it reaches a critical point, when it can not hold any more charge and at the tip of the nozzle blows into a cloud of charged particles. They are all charged at the same potential what cause them not to stick together. This state is applied to a detector. By using ESI it is possible to analyze molecules between 50 - 80 000 Da.

2.1.2 Matrix Assisted Laser Desorption Ionization

MALDI is a different method for creating ions from sample. In this case, laser beam is applied to a matrix which protects molecules from being destroyed by direct laser

¹after <http://www.newobjective.com>

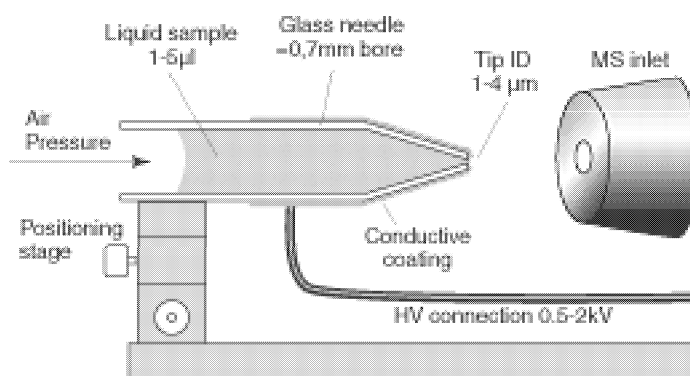


Figure 2.1: Electrospray Ionization (ESI)

application. It also allows ionization of more complex structures like peptides or proteins which are very fragile and lose its structure when ionized by traditional methods.

First, the sample is dissolved in solvent (for some proteins it can be water). Next step is addition of substance like trans-cinnamic acid or 2,5-dihydroxybenzoic acid. This substance has to absorb ultraviolet light. Resulting sample is placed on the probe in airtight chamber. Afterwards, the air is pumped out what creates vacuum in the chamber. It causes solvent to evaporate and only UV absorbing compound is left with some protein sample in it. This process is called desorption and is a basis for the process. Finally laser beam is applied (ultraviolet frequency). UV absorbing compound absorbs all the energy and passes some of it to binded molecules of proteins. If enough energy is passed, molecules start to evaporate and change their state to gas phase. The process is presented briefly on figure 2.2². The difference between ESI and MALDI is that the second one is able to ionize molecules of weight up to 1 000 000 Da

Last part of the process is detection. Proteins are in gas state in the air. They are charged so applying electric field cause them to move towards opposite charge electrode. According to Newtons Law

$$a = \frac{F}{m} \quad (2.1)$$

for constant force F acceleration will depend on mass of molecule m . If the mass is bigger, acceleration is lower so smaller molecules will get faster to detector. By recording amounts of molecules arriving to detector surface, mass spectrogram can be produced, i.e figure 2.3

²after <http://qbab.aber.ac.uk>

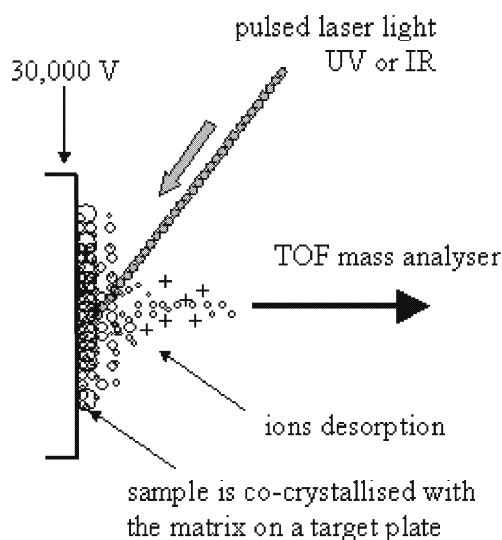


Figure 2.2: Matrix Assisted Laser Desorption Ionization (MALDI)

To summarize steps of proteins detection, which are performed by every mass spectrogram, figure 2.4 presents these as a blocks diagram. Different technologies are better at some stages of this process, but still those steps are performed:

- Sample introduction - the process of collecting tissue from probe into the system
- Ion source - mechanism to convert organic tissue into ions
- Ion analyzer - mechanism to organize samples in a way that detector will be able to distinguish differences between them
- Ions detector - process of converting ions into digital signal
- Data analysis - Hardware implemented procedures for data improvement

These steps are also implemented in recent advancement: SELDI-TOF-MS which produces data of the highest resolution compared to other two methods.

2.1.3 Surface-Enhanced Laser Desorption/Ionization

As every invention has to be improved, SELDI ionization method has been developed as an extension of MALDI. Major difference between those two methods have been described in application note of Ciphergen Protein Chip Vorderwölbecke et al. [18]. In both cases, proteins to be analyzed are cocrystallized with UV-absorbing compounds

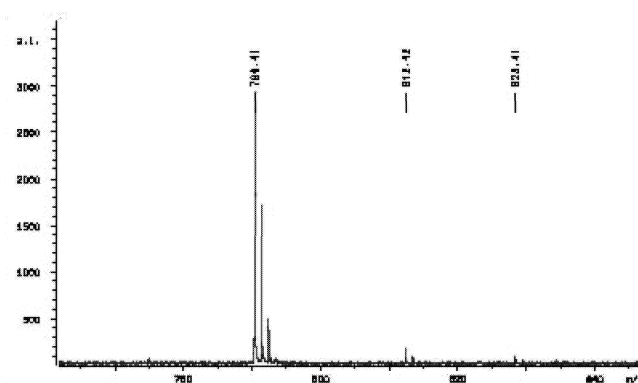


Figure 2.3: Sample Spectrogram

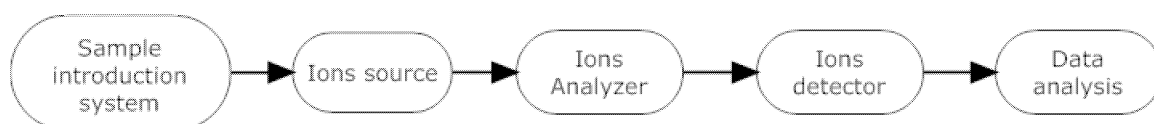


Figure 2.4: Spectrometry block diagram

and vaporized by a pulsed-UV laser beam. Ionized proteins are then accelerated in an electric field, and the mass to charge ratios of the different protein ion species can be deduced from their velocity. The differences between SELDI and MALDI are in the construction of the sample targets, the design of the analyzer and the software tools used to interpret the acquired data.

In the SELDI method, protein solutions are applied to the spots of ProteinChip Arrays, which have been derivatized with planar chromatographic chemistries. The proteins actively interact with the chromatographic array surface, and become sequestered according to their surface interaction potential as well as separated from salts and other sample contaminants by subsequent on-spot washing with appropriate buffer solutions. The chromatographic surfaces provide a very good support for the cocrystallization of matrix and target proteins, resulting in the formation of a homogenous layer on the spot, thereby delivering an ideal crystalline surface for the subsequent analysis.

Sample preparation for SELDI experiments is quite different than the process for MALDI. For MALDI analysis, protein solutions are typically premixed with the matrix and dried on a passive surface. With the exception of flash washing with cold distilled water, on-target purification is not possible, and pre-target deposition sample cleanup procedures must be applied to reduce chemical noise and ion

suppression. Also, on-target segregation of protein populations is not practical because the surface has only weak and unpredictable interaction properties. For these reasons, prefractionation using a variety of microtechniques is often used. Taken together, these sample preparation requirements complicate the MALDI analysis, often resulting in sample loss as well as artifactual qualitative and quantitative variances.

The analyzers used for SELDI and MALDI were designed with different purposes in mind. The ProteinChip Reader is especially adapted to achieve high-sensitivity quantification and good reproducibility. The ion source and detector are constructed to support very efficient ion transmission and ion detection over a wide mass range. The precise positioning of the laser beam is controlled by software both in manual and automatic mode. The process is visualized in a user-friendly format by a pixel raster map to facilitate the multiple analyses of the same sample spot, and software tools allow normalization of the resulting spectra to their total ion current for internal quantitative calibration. These features assure high precision and reproducibility even when great numbers of complex biological samples need to be comparatively analyzed. In contrast, MALDI devices are not designed for reliable quantitative precision over a wide mass range. They are a very good choice if high accuracy in the lower peptide range is needed without a requirement for high reproducibility of signal intensities. But if a good correlation between signal intensities and protein concentration is to be achieved over a wide mass and sample concentration range, the SELDI-TOF-MS-based ProteinChip Reader will always produce data with better reproducibility for hundreds of samples per day.

2.2 Experimental Steps

According to Vorderwulbecke et al. [18] this process consists of following steps (also presented on figure 2.5: after choosing an array from a selection of chromatographic and preactivated ProteinChip Arrays, samples are applied and incubated on the spots. On-spot washing ensures efficient sample cleanup, and the spot surfaces allow the formation of a homogenous layer of cocrystallized proteins and matrix compounds. In the ProteinChip Reader, a laser beam is directed on the spot causing desorption and ionization of the proteins. A defined laser beam raster is used to selectively cover the entire spot surface and allows repeated reading of a single spot without using the same positions twice. Multiple spectra from a statistically meaningful area are then averaged in a final spectrum in which the mass-to-charge ratios of the ionized proteins are given and a good correlation between signal intensities

and analyte concentration is achieved for the different peptides and proteins in the sample.

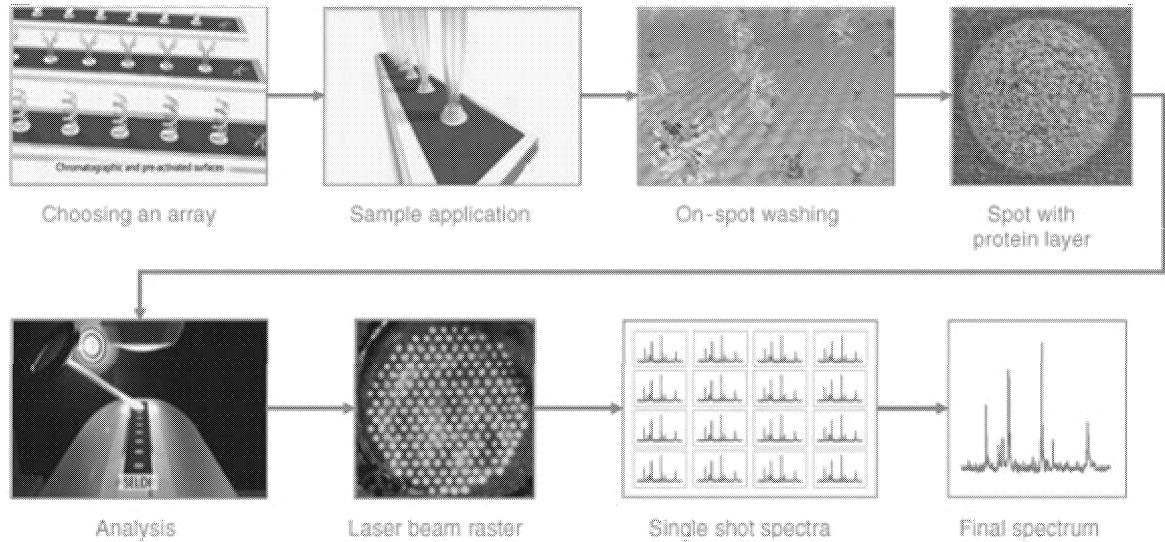


Figure 2.5: Steps of sample processing for SELDI-TOF-MS

2.3 Theory conclusion

After the technology process has been explained, it will be easier to understand what are actual problems and limitations after the raw data have been produced. It is important to remember, that MALDI allows high accuracy to be achieved whereas SELDI-TOF-MS was developed to achieve better reproducibility over many experiments. Problems with reproducibility were described in detail by Baggerly, Morris and Combes [3].

Chapter 3

Sequence of data analysis

Chapter 2 introduced foundations of the technology used to produce raw data. In this project the biggest interest is to be able to distinguish between normal or cancer sample. SELDI-TOF-MS produce data which present structure of a sample. Intuitively, if a tissue is infected by a disease, it should contain different, specific to a disease, proteins. The idea is to investigate samples collected and processed by mass spectrometer with respect to their spectrograms. Two randomly selected samples spectra are presented on the figure 3.1

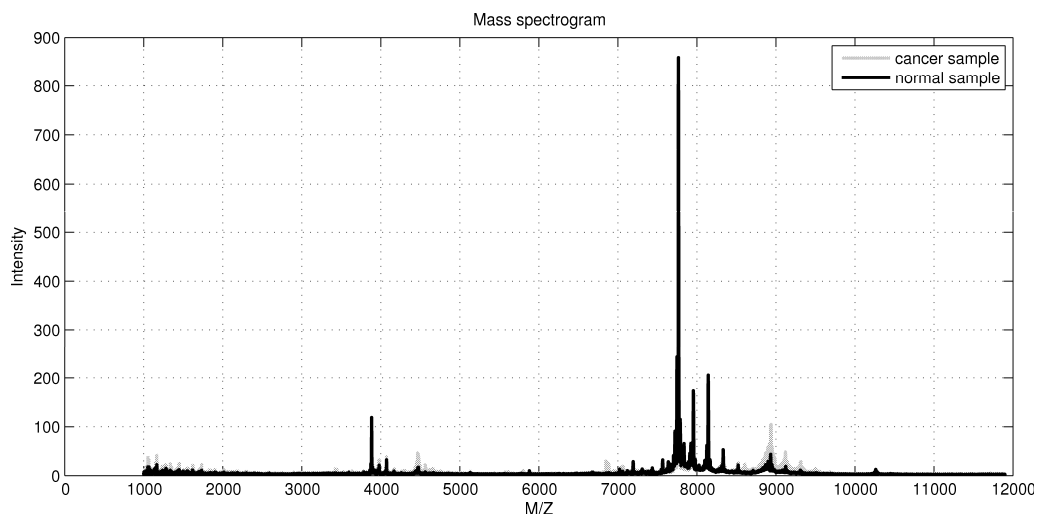


Figure 3.1: Sample Spectrogram

When those samples are analyzed in detail, it is clearly noticeable that spectrograms for cancer and normal samples differ significantly. This may confirm above claim, that it is indeed a good idea to use this tool to solve the problem.

Figure 3.2 presents zoomed view of the samples from figure 3.1. In the range

5790-5815Da there is undoubtedly significant difference between samples. By identifying areas similar to this one - classification algorithm for detection would be easy to construct, based on the peak presence i.e. at 5800Da.

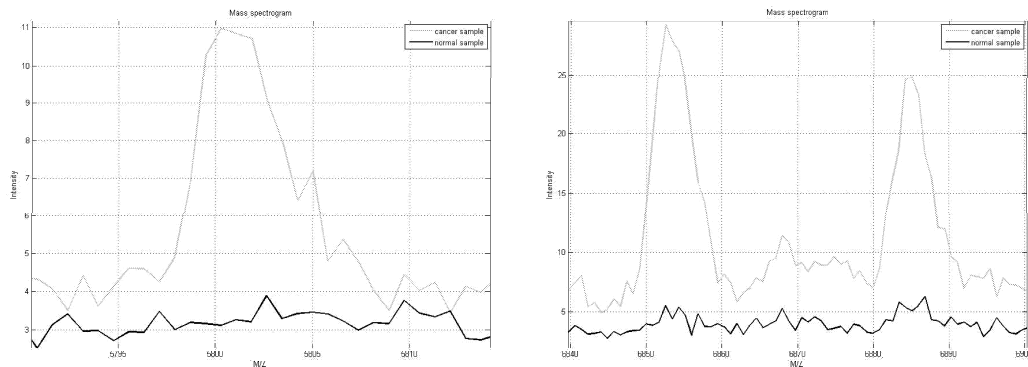


Figure 3.2: Example differences between samples

At first, the task seems to be very easy, but in reality it is very difficult and there is no method which can produce 100% accuracy. Number of problems are encountered, which are the factors decreasing quality of the results.

First of them is that it is very difficult to define which value defines major peak and which does only minor. Every sample is different and if a wider analysis is done over many samples, conclusion is that there are many samples, but the peaks are either much smaller or are not present at all. It is therefore very difficult to classify samples. What makes it even harder task is that generally, there are only few areas for which cancerous samples differ from normal over all available input spectrograms.

There is also problem of overlapping samples. What is meant by this, is a situation, when two samples belong to different classes (cancerous or healthy) but are not different at areas, where most of the samples in a data set differ.

There are also many other difficulties when approaching this problem, i.e. presence of noise, shifts in values (vertical and horizontal), lack of normalization and many more with high dimensionality at the end. Figure 3.3 briefly presents all of the above. That is why before detection can be performed, samples have to be preprocessed to be able to predict unknown sample by the classifier. Some of the methods, most common, are presented in following sections. Most of them were implemented in the Mass Spectrometry Analysis Software (MSAS) designed in Matlab environment for the purpose of this project. It was decided to use Matlab as it provides a great framework for developing and testing algorithms. There are also many sta-

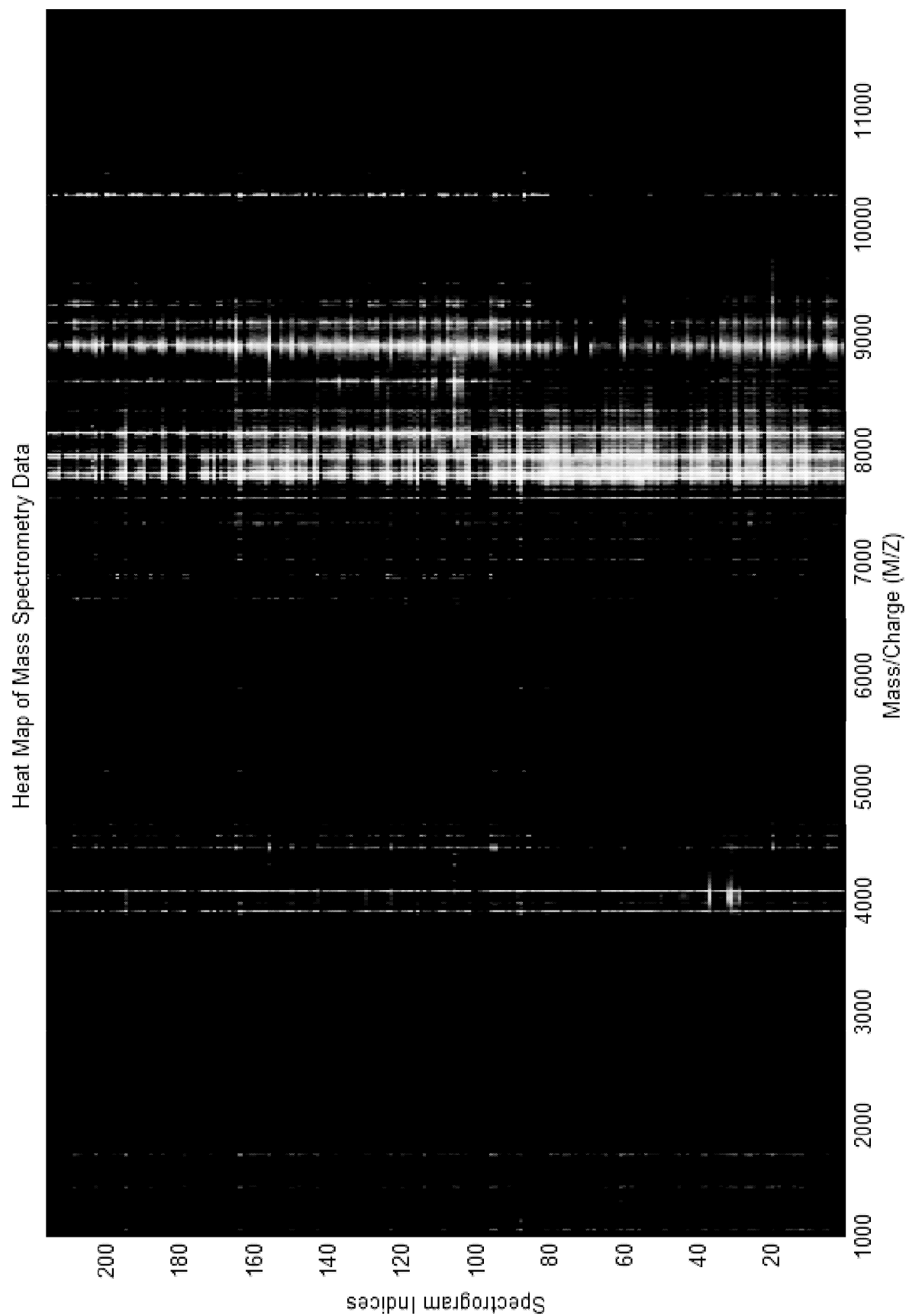


Figure 3.3: Data set presented in a Heat Map form

tistical tools available on the market, i.e. SAS, which provides a platform for data analysis. It has extensive statistical capabilities. But there is one major disadvantage of this kind of tools - its price. Therefore, MATLAB has been chosen, because it is widely available through academic community. By developing software on a platform which is easily accessible, there is greater chance for further development, thus a chance for improvement.

In this section, theoretical description of methods commonly used in Mass Data Mining is presented.

3.1 Preprocessing

To be able to compare different samples to each other, it is important to prepare them for this process. Different factors have to be eliminated to prepare data for analysis and classification. If the samples were not preprocessed, classifiers may detect normal sample as cancer, or cancer as normal what, if it was medical tool could result in very dramatic consequences. Person who is healthy could be mistakenly diagnosed as having cancer and would be directed for unnecessary treatment which is costly and very depressing. On the other side, ill person diagnosed as being healthy, may lose chance of being cured if the disease state is advanced. Following subsections will outline common methods for data preprocessing.

3.1.1 Dimension reduction

One of the biggest problems with data set is its size. Every sample, from the data set used for the purpose of this paper, consists of approx. 350 000 points (M/Z values) from the range of 1000mz to 12000mz. For example, in this data set, each file represent one sample with its M/Z and intensity values. It is a size of approx 5MB in computer memory. Multiplying it by 216 (number of all samples provided) results in more than 1GB of memory required only to load the data. Many methods used in later analysis require duplication of the data so even more memory is required. Finally, operating on such a big data set is time consuming. Therefore, size reduction should allow elimination of insignificant areas of spectra, leaving important ones for further analysis. Furthermore, this large data set has to be optimized when loaded into memory so it will be accessed quicker. Mechanism for this kind of manipulation is available by operating system. Matlab, used as a Development Environment (IDE) is also highly optimized, when operating on large data sets. Idea implemented in this project to reduce the size of the data, thus calculation can be done in reasonable time, is to re-sample data set. Instead of

loading 350 000 values, data has been reduced to 15 000 what significantly increased speed of calculations, but as pre-tests proved, no significant loss of information was encountered.

3.1.2 Miscellaneous

Baseline correction All the samples are not at the same level. Some of them are shifted with respect to M/Z axe. This is clearly illustrated on figure 3.4.

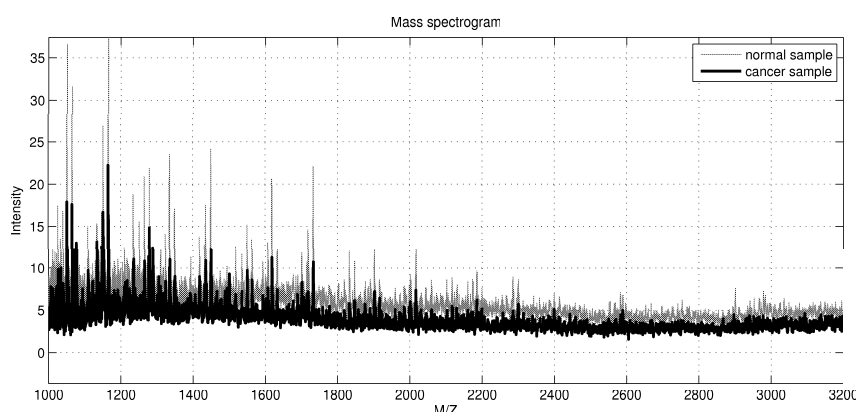


Figure 3.4: No baseline correction applied

Cancer sample is shifted up according to normal sample. To allow equal comparison the baseline correction should be applied so only significant differences will be presented. Baseline correction provides flatter baselines and averages the baseline to zero. This improves the accuracy of classifier, the appearance of the spectrum, and the quality of a result from subtracting one spectrum from another. All intensity values are also shifted up, so it could be with advantage for the future analysis if the samples were shifted to the 0 value.

Figure 3.5 presents the results of baseline correction. Intensities are shifted to begin from 0 but intensity values were kept unchanged. This step may increase comparison of samples, thus performance of classifier.

Normalization This will be the next step to prepare data that are similar in general, so only important differences can be revealed. Group of samples can be normalized by standardizing the area under the curve to the group median. It would also be helpful if the normalization process could recalculate the intensities to be enclosed between desired values as every spectra intensity is very different to each other. Generally speaking, normalization is a process applied to all data in a

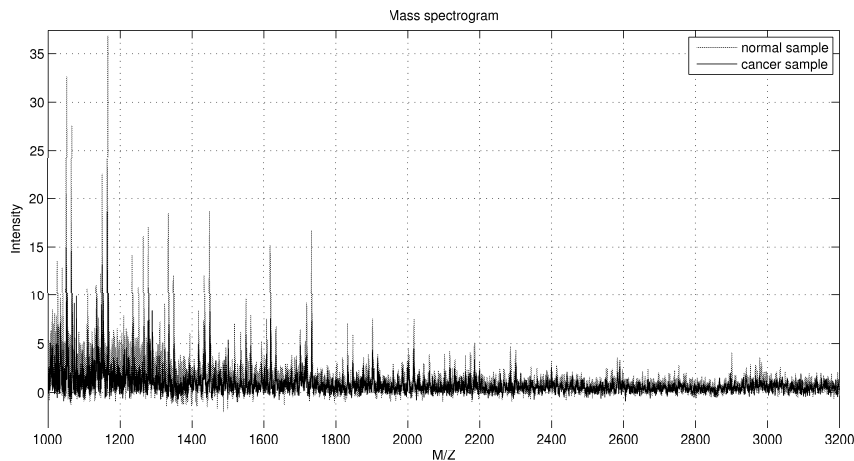


Figure 3.5: Baseline correction applied

set that produce some statistical property. This step may (or may not) influence performance of classifier what will be investigated further in detail.

Denoising - Smoothing As every digital data, mass spectra contain noise, which is impossible to avoid with every type of electronic device. There are many techniques for removing noise and many research has been done towards denoising this type of data i.e. Coombes et al. [8] or Satten et al. [15]. As the scope of this project is limited, they are not discussed in detail. Having the mass spectra standardized and normalized, removing noise can sometimes significantly improve the performance of classifier.

Principal Components Analysis (PCA) PCA is a way of identification of patterns in data. It presents the data in a form, that similar properties and differences are emphasized. This tool is very often used for dimension reduction as it can group data in clusters of much smaller size, but with properties exactly the same as raw data with very small loss of information. The idea of PCA is presented on figure 3.6. Algorithm of PCA as described by Smith [16] is following:

- 1) Data preparation
- 2) Mean subtraction
- 3) Calculation of covariance matrix
- 4) Calculation of eigenvectors and eigenvalues of covariance matrix
- 5) Choosing components and forming a feature vector

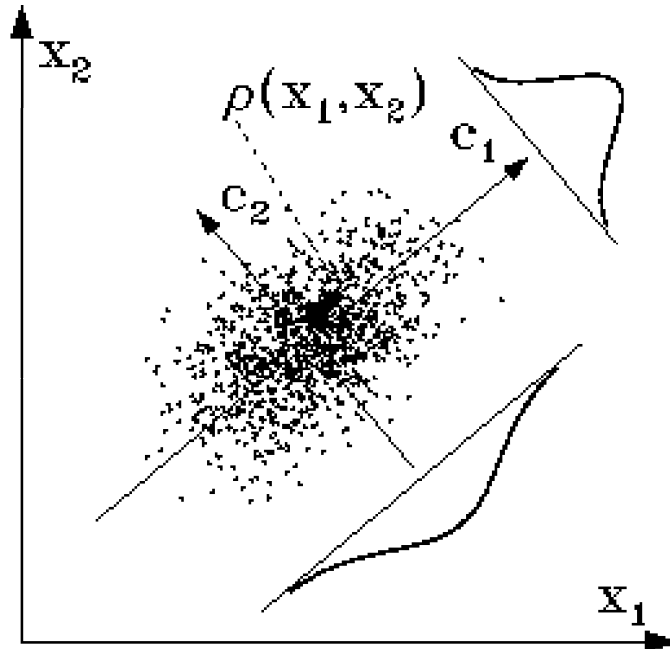


Figure 3.6: General Idea of Principal Component Analysis

By applying above steps, data are classified into groups of similar properties. In this way dimension is much reduced so less computing power is required for further analysis. PCA can be used as classification method on its own but research shows, i.e. Lilien et al. [10], that it is better to use it as preprocessing method for another classifier, i.e. Linear Discriminant Analysis.

3.1.3 Peaks manipulation

Peaks detection As a method of peak detection an algorithm developed for the purpose of this project has been used. Example spectrogram is presented on figure 3.7. Some random peaks are marked. there is no "good" or "bad" method for choosing the peaks. On the other hand, peak detection is mostly required as an input for the peak alignment algorithm. The idea was to use the whole spectrogram to detect peaks. Steps of algorithm I have designed and implemented are presented as follows:

- 1) Set the number of maximum peaks mp to detect
- 2) Calculate mean across every sample, for all Mass/Charge ratio
- 3) Calculate mi = maximum intensity in the mean vector
- 4) Choose MZ ratios for which Intensity is over m

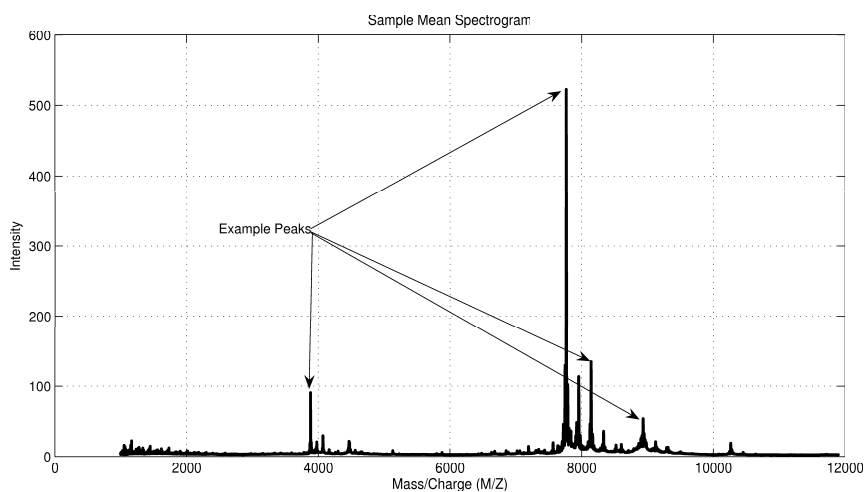


Figure 3.7: Sample Mean Spectrogram

- 5) If number of values is too small, divide maximum intensity by 2 and repeat from step 4
- 6) Calculate differences (approximate derivatives) of vector, and choose MZ ratios for which sign of differences changes from positive to negative
- 7) Repeat step 6 until number of peaks reduced to desired maximum number of peaks to be detected

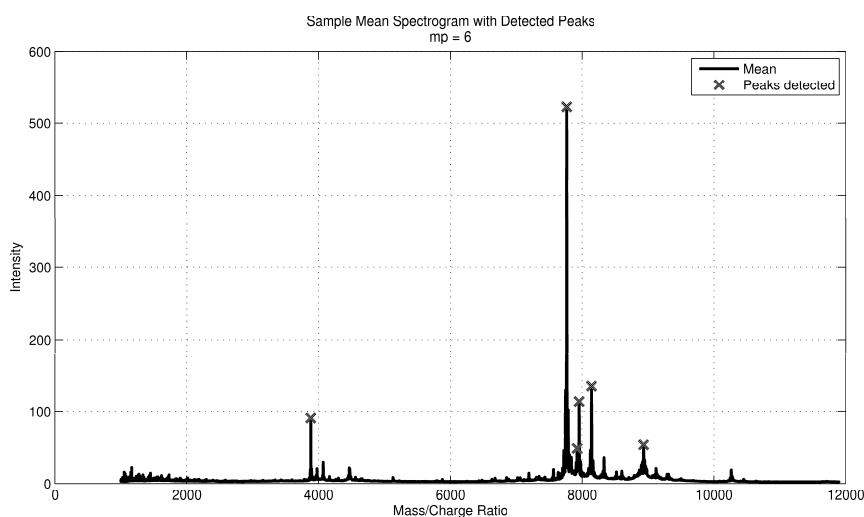


Figure 3.8: Sample Mean Spectrogram with detected Peaks

Algorithm has been tested on different sets of mass spectrometry data and it works properly - detects peaks according to desired number of detected peaks. Re-

sults for spectrogram presented on figure 3.7 are shown on figure 3.8 with maximum values of peaks to detect equal to $mp = 6$.

The importance set for this part of preprocessing is to detect few highest peaks and it is not very important which are these - they should be the highest of all the spectra.

This step is a part of a next one which will align peaks according to detected ones.

Peaks alignment While baseline correction was used to shift samples vertically, it is also important to align them horizontally. There are different sources of shift for the samples which may result in peak presence at different MZ values. It could lead to misinterpretation and while peaks should be the same ones, they could be detected as two different ones. Peaks alignment algorithm will be able to rescale data horizontally in a way, that every peak which is the same for every spectra is evenly placed.

3.2 Data Analysis and Classification

To develop a software which can be used for cancer diagnosis based on mass spectrometry, classifying algorithm is required. Previous steps were employed to prepare raw data, thus common differences in particular samples were removed and classifier will receive data which differ at significant areas because of their belonging to the class rather than because of measurement process.

With this requirement in mind many classifying algorithms have been developed and all the research work is focused on developing new ones, which will be applicable to Mass Spectrometry specific data set. Further sections provide brief description of different approaches.

The main idea of classification is to be able to extract areas of similar properties and group them into classes. By creating those areas of similarities (particularly in unsupervised classification) it is possible to build a model which generalizes the properties of new data to be classified into one of the classes. There are two main terms, although sometimes used interchangeably, differ much: clustering and classification.

Clustering Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset share some common properties - often proximity according to some

defined distance measure. Example of clustering is presented on figure 3.9

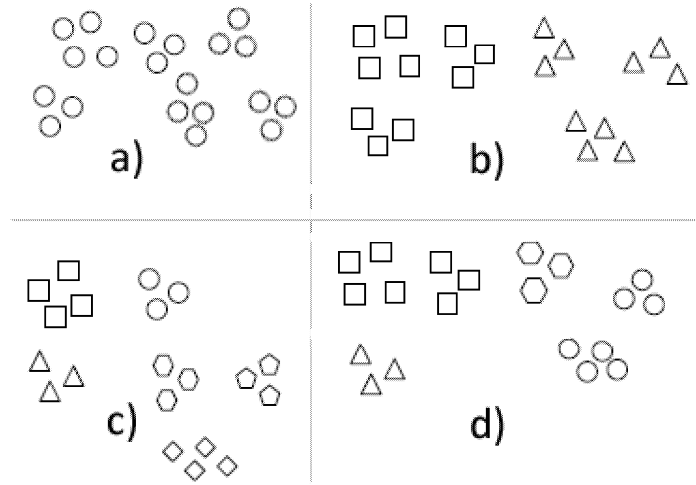


Figure 3.9: Classification example: a) input data (unclassified), b) two classes, c) six classes, d) four classes

It is important to state at this point, that clustering is usually unsupervised type of learning. That means that algorithm organizes data into classes, which are not previously defined. There is also no opportunity to test the goodness of the output.

Classification Classification, on the other hand, is a supervised learning algorithm. What this term means is that algorithm is trained to classify data into predefined classes. Performance of classification is measured as percentage of correctly classified samples which will be described further in detail. It is easy to test the performance, because for each input, in training set, there is an output assigned. By minimizing the error between desired and actual output, performance may be improved.

There are many factors which influence how classifying algorithm performs. One of the reason of poor classification is when data are overlapping themselves - classes are not easily separable. This is caused when data sets have similar parameters, so it can not be straightforward classified to one of the classes. Problem is presented on figure 3.10 Therefore it is important to use a classifier, which will be able to extract hidden information in data set, distinguish differences and classify samples correctly.

Classification algorithms can be divided into two major groups as proposed by Lilien et al. [10]. First, those algorithms which depend on the data and on experiment conditions - return different results. With those algorithms, often population of solutions is initialized, at performance of these solutions is improved for number of

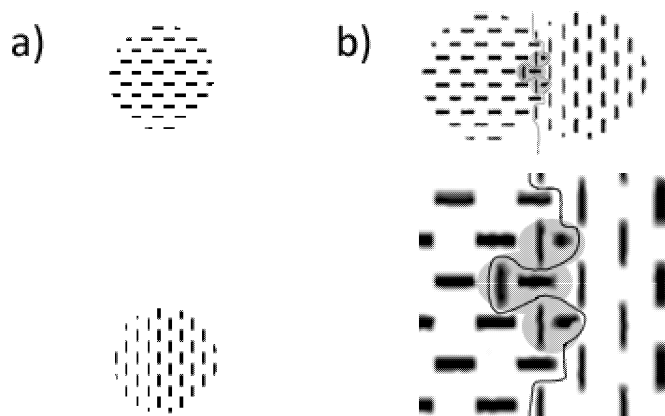


Figure 3.10: Classes: a) Well separated, b) overlapping

generations, i.e. Neural Networks or Evolutionary Computations. Furthermore, initial conditions are often the reason why method performs better during one run and worse during another. Results produced are always dependant on initial conditions thus output is not deterministic. These methods are called heuristic.

Second group of algorithms are mathematical models. Those methods have strictly determined steps to follow therefore for specific input, always return the same output, even if the initial conditions are different each time. These methods are called deterministic or exact models, i.e. PCA, LDA, k-Nearest Neighbors.

An overview of the methods, which can be applied to mass spectrometry data has also been described by Bensmail and Haoudi [5]. In further sections, a general overview of methods used in this project will be presented.

3.2.1 Heuristic approach

Neural Networks (NN). Artificial Neural Networks (ANN) are computing tool based on rules of human brain functioning. Observing real brain neuron, scientist noticed that it is connected with many similar neurons. By generalizing processes in brain and describing it in mathematical notation, very powerful analysis tool has emerged. Example model of Artificial Neuron Network is presented on figure 3.11.

As real neurons, artificial ones (simplified) have inputs and outputs. Through inputs they collect outputs from neurons they are connected to. Inputs are multiplied by according weight and summed. Next, the value is passed through nonlinear activation function. If the input is strong enough to activate the neuron, it "fires" - produces output for the neuron it is connected to. By combining neurons in multi layer networks, it is possible to solve many difficult nonlinear problems, which are impossible to solve by traditional mathematics i.e they tend to approximate func-

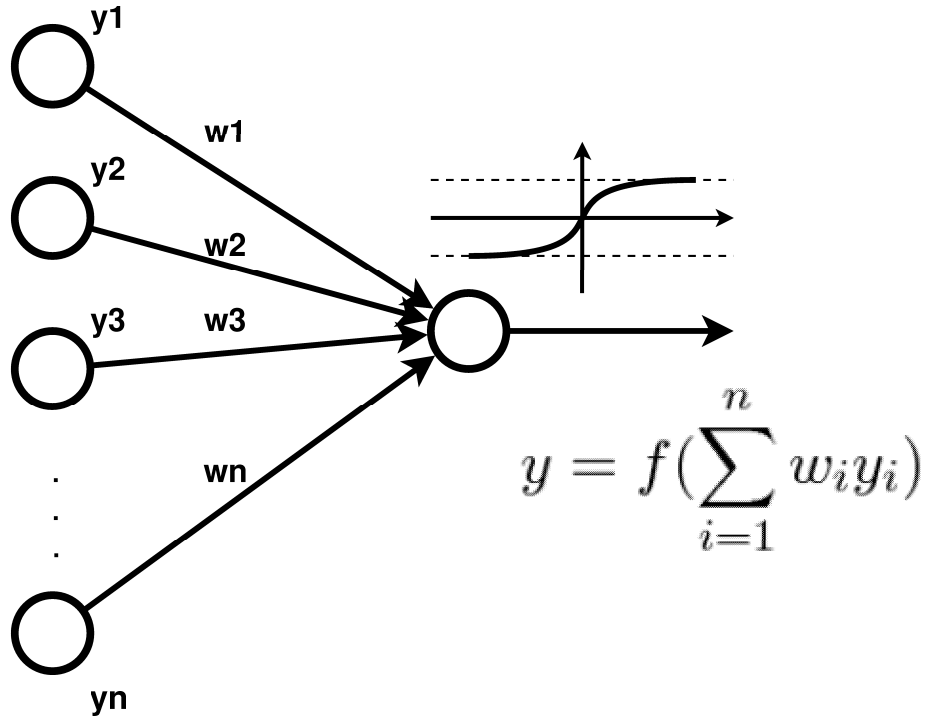


Figure 3.11: Model of Artificial Neural Network (ANN)

tions very well with reasonably small computation power requirements. Some of ANN's applications include: speech and pattern recognition, image recognition, financial prediction and many more. They also perform well in proteomics and the results are described in following papers Ball et al. [4], Zhou et al. [21] or Yu and Chen [20]. Problem with Neural Networks is that for different data, some architectures perform much better than the others and it is very laborious task to achieve very good results also because of large number of parameters to adjust, i.e. number of nodes, number of hidden layers, types of activation functions, connection between nodes, etc. Therefore an idea of using Evolutionary Computations to approximate parameters of Neural Networks has been introduced but it will not be discussed in this project.

Evolutionary Computation (EC) Evolutionary Computation methods may be grouped into following subgroups: Evolutionary Programming (EP), Genetic Algorithms (GA) and Genetic Programming. In general, Evolutionary methods work according to general algorithm:

1. Generate population of solutions
2. Evaluate each individual of population according to fitness function

3. Apply operators, i.e. selection, crossover or mutation to create offsprings
4. Evaluate offsprings according to fitness function
5. Select individuals for next population
6. If stopping criteria not met - return to 3.

By applying different operators, Genetic Algorithms explore intelligently the search space defined by Fitness Function which is a measure of goodness of the solution. As it is optimization method, when fitness function is properly defined, GA can be used as a classifier. Problem will be defined as a set of parameters to adjust and the goal is to minimize (or maximize) objective function which could be, i.e. minimum error of the classifier. Evolutionary Computation is a wide area and which has many applications, also in mass spectrometry analysis, i.e. Petricoin et al. [12] used Genetic Algorithm to identify biomarkers in mass spectrometry data. However, the scope of this project is limited thus EC will not be implemented.

3.2.2 Exact approach

While heuristic approach is sometimes the only way of solving difficult problem, it is very common that results differs even if input data has not been changed, because there is a small addition of randomness in these algorithms. Therefore, whenever it is possible, exact algorithm algorithm should be used, which will always produce determined result, for the same input data. There are many tools available, but the problem with these methods is that they tend to be very data specific. Therefore, only few of them can be applied for mass spectrometry analysis. They are described as follows

Linear Discriminant Analysis (LDA) Implemented in so called Q5 algorithm by Lilien et al. [10]. The dimension has to be reduced before LDA is applied so Principal Component Analysis is performed to accomplish it. LDA, like PCA, looks for linear combination of variables which best describe the data. But the difference is that LDA also models differences between them whereas PCA does not include it. As defined, LDA approaches the problem by assuming that the probability density functions $p(\vec{x}|y = 1)$ and $p(\vec{x}|y = 0)$ are both normally distributed, with identical full-rank covariances $\Sigma_{y=0} = \Sigma_{y=1} = \Sigma$. It can be shown that the required probability $p(y|\vec{x})$ depends only on the dot product $\vec{w} \cdot \vec{x}$ where $\vec{w} = \Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_0)$

That is, the probability of an input x being in a class y is purely a function of this linear combination of the known observations.

When LDA classifier is trained, mean and covariance is calculated because those parameters are not known. But training in this case is shorter than time consuming heuristic methods. Performance of this classifier has been shown to be much better than any other classifier.

k-Nearest Neighbors k-Nearest Neighbors belongs to a class of supervised learning algorithm. It is another method used for classification. The algorithm examine data and divide them into a predefined number of classes. Those classes contains categories of parameters which are derived from data during training process. After algorithm is trained, when a test sample is applied, classifier finds the k nearest neighbors and assigns their label names to the training data set. Importance of each neighbor is weighted by its rank presented in terms of the distance to test sample.

Because of classification ability, k-NN algorithm was implemented to show its accuracy for Mass spectrometry data analysis. As it will be presented further, it's performance is lower than other algorithms.

3.3 Verification

Every classifier has to be verified to show how well does it perform for input data. Data set is divided into 2 subsets: training set and verification set and it is important to perform this process properly. For mass spectrometry data freely available on the Internet, number of samples is limited thus algorithms have been proposed to increase the size of it. But the easiest way is to permute data set and choose desired number of randomly selected samples for training and use remaining set for testing. Another methods used for the small data set is boosting or bagging (Donald et al. [9]).

When classifying system is ready - it is important to perform statistical tests to evaluate it's performance. It is crucial to test it on a data which have not been used for training. Performance should be tested for some number, i.e. 50 independent runs and mean and standard deviation should be calculated. This will create statistically significant sample and if mean is high there is high probability that this is indeed true performance of the classifier.

In general, performance of classifier can be expressed as a percentage of correctly classified samples. But this is not enough. Often, classifier may perform very well detecting existence disease, but poorly when rejecting its existence. Other classification algorithm may perform well when rejecting existence of disease but poorly detecting its existence. In medicine, more harmful would be one, which under per-

forms while detecting diseased samples. The reason is because it is better to false predict that the person is diseased, because then further investigation can be conducted. If, on the other side classifier would classify diseased person as healthy, and no further investigation would be conducted, person would lose its early chance to be cured. And time is the key factor when it comes to cancer disease. If diagnosed at early stage, chances for person to be cured are very high.

Therefore, instead of measuring only percentage of correctly classified samples, two other values are also examined. They are called PPV ¹ and NPV ².

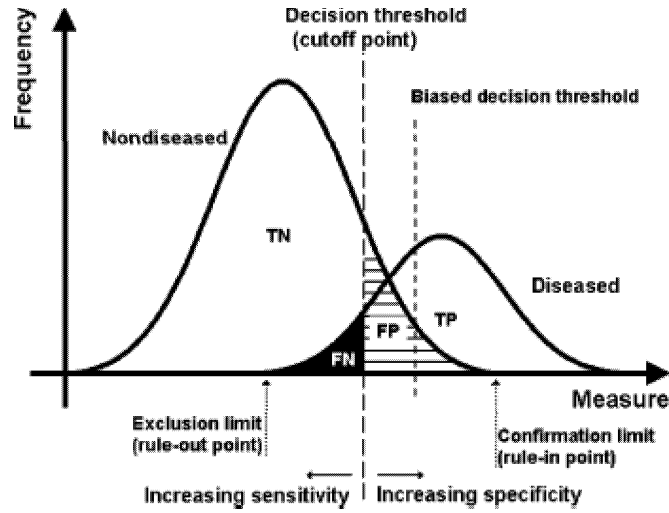


Figure 3.12: Probability density of two class samples (after: Vitzthum et al. [17])

According to 3.12 they are calculated as follows:

- 1) PPV - indicate the percentage that in case of a positive test, patient indeed has the specified disease;

$$PPV = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} = \frac{TP}{TP + FN}$$

- 2) NPV - indicate the percentage that in case of negative test, patient indeed is healthy;

$$NPV = \frac{\text{true negative}}{\text{true negative} + \text{false positive}} = \frac{TN}{TN + FP}$$

Further details about validation requirements of classifier were described in work described by Vitzthum et al. [17]

¹Positive Predictive Value

²Negative Predictive Value

Chapter 4

Software Design Elements

Important part of the project was to develop a software which will be used to for classification of raw data produced by SELDI-TOF-MS spectrogram. This was not an easy task because many factors had to be considered while writing the diagnosis software. In this chapter, problems and solutions will be described to show brief methodology applied when software was being developed. Additionally, 24/7 work organization has been proposed, which can significantly increase further software development.

4.1 Requirements for data analysis Tool

Methods which allow noninvasive or low invasive diagnosis play very important role in present medicine. They show the state of a patient without making him feel any discomfort while extracting data. These methods allow better diagnosis made by doctor thus they increased survivability and comfort of population these days. That is the reason why so much research is done to develop new methods which will allow to examine patient noninvasively and further increase a comfort of prevention or treatment process. Vitzthum et al. [17] in his article described the requirements of diagnostic application. He mentions three most important expectations of this kind of tool which are:

- 1) it is essential to provide diagnostic tests that allow for definite and reliable diagnosis tied to a decision on intervention (prevention, treatment or non-treatment)
- 2) it is essential to meet stringent performance characteristics for each analyte (in particular: test accuracy, including both precision of measurement and trueness of the measurement)

- 3) provide adequate diagnostic accuracy (i.e., diagnostic sensitivity and diagnostic specificity, determined by the desired positive and negative predictive values which depend on disease frequency).

SELDI-TOF-MS combined with appropriate data analysis software could possibly be a tool to meet desired requirements. But there is still a lot of work that needs to be done in order to achieve satisfactory results. But as there are many tools available, low mass spectrometry reproducibility (which is the biggest problem at the moment) can be supplemented by good analysis software which will be able to correct deficiency of the hardware. This is why this work has been conducted: to make a step towards better analysis software. Conclusions drawn from the results MSAS provides may be very helpful for future developments of similar software. It is therefore worth putting an effort into this field of research, and methods should be developed to speed up development of this research. One idea is to encourage research centers all over the world into larger groups, working towards the same goal. This is the fundamental aim of 24/7 - round the clock work organization.

4.2 24/7 work organization

The demand for diagnostic software has always been very high and together with increasing standard of living, it is even larger these days. 24/7 system is an approach to software development which is based on cooperation of few research groups located in different time zones.

The need for distributed project realization. There is no doubt that researchers, project, managers, programmers - people work at their best during day-time. It is very very well known property of human brain to be active during a day and regenerate over night. Therefore, a single group working on a project do it effectively only through 8 hours a day and sometimes even less. Thus if task requires 320 hours - it would take 40 days to accomplish it. No speed up is possible because of two major factors:

1. workers would have to extend their working day
2. more people would have to be employed.

First solution would only "virtually" boost the speed of a project, because people would have to work when their brain activity is lower, thus producibility is lower. This solution is only recommended for very short period of time, i.e. when a deadline approaches.

Second solution might be difficult to evaluate: it may take a long time to recruit required amount of people. Additionally, this process requires additional financial expenses which are not the case here.

Therefore, a solution is 24/7 work organization.

Detailed description, advantages and disadvantages 24/7 is a solution for described problems regarding traditional process of software development. It is a framework of effective project management which is intended to significantly speed up the software development. It is based on few research groups cooperating together, for ease of explanation 3 different research centers. The key idea is that they are located in different time zones all over the world, i.e. Wroclaw (Poland), Tuscon (USA) and Sydney (Australia) as proposed by Chaczko et al. [6]. This situation is illustrated on figure 4.1

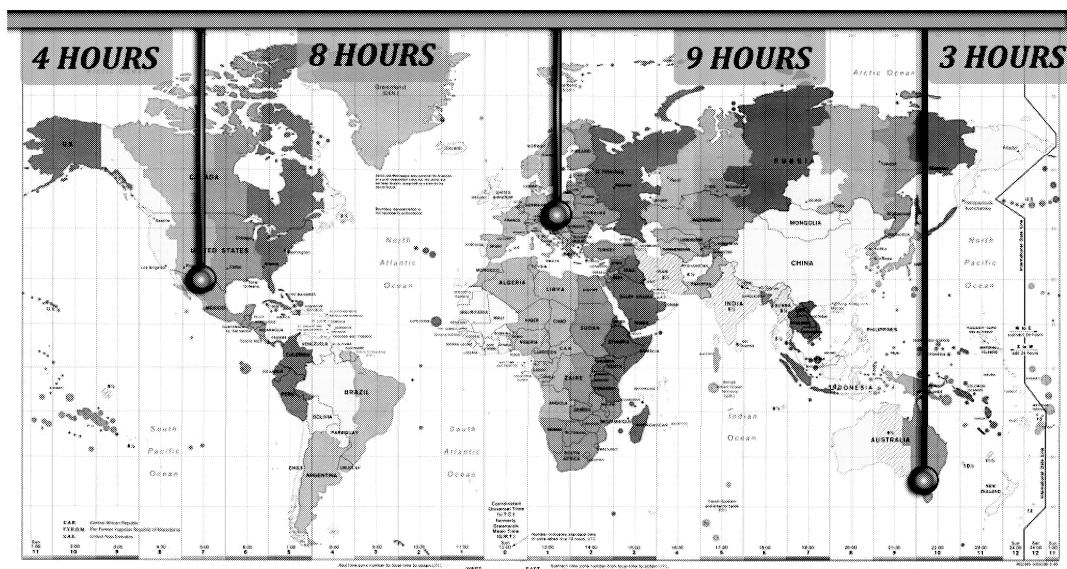


Figure 4.1: Time zones and example research centers arrangement

Preferably, as figure 4.1 presents, those research centers should be placed within reasonable time zones, not just any randomly chosen. This requirement exists to fulfill the aim, that only one group works at its highest activity time (working day) and after this working day is finished, pass the results to the following group, located in different time zone (i.e. 8 hours difference) where the working day has just begun. Process is repeated further as long as the project duration. To illustrate the process just described, figure 4.2 is provided.

Advantages and disadvantages. Advantages of proposed solution is clearly presented on figure 4.2. If 48 hours were considered, by the time single research center

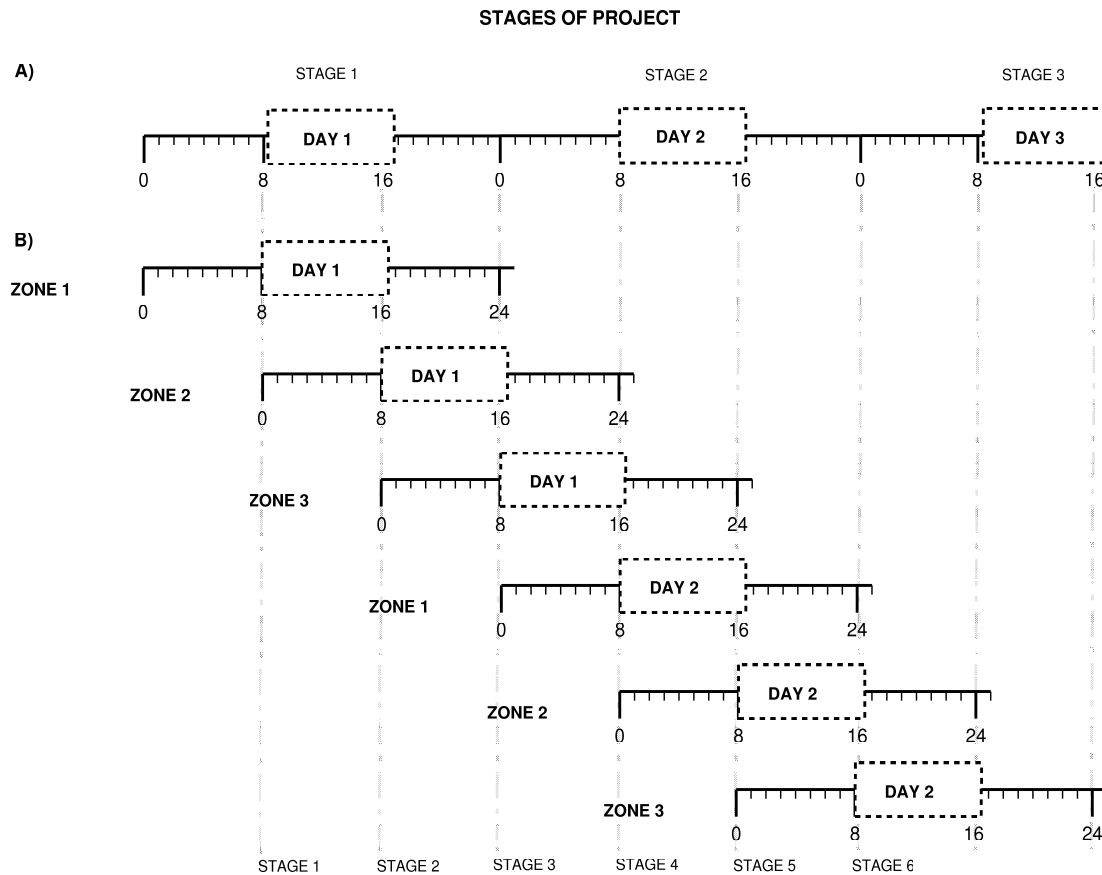


Figure 4.2: Schedule of work: a) one research center b) 3 research centers

finish their second stage of the project, 3 research centers will accomplish 6 stages. This is the most significant advantage of 24/7 work organization. Projects could be realized *almost* three times faster. The work "almost" means that it is not exactly 3 times faster due to additional processes which have to be performed when using 24/7 system. One of them is synchronization of the work between the change of shifts. It may take another several minutes for a group which intercepts results of previous group to fully understand what has been done within last few hours. Another difficulty with 24/7 approach is task sequencing. It is similar to a pipeline processing, but a group is not tightened to any goal that have to be achieved. If they do not finish what was assigned to them, they simply explain following group what was the difficulty and the work is carried on. It may even be with advantage for the project: when one cannot make a progress with a difficult part, fresh look of the following group may provide some new ideas and move the project further probably further increasing the speed of works. Finally, this kind of work organization may introduce kind of healthy competition when each of the group would like to progress more then the others.

By knowing the framework, following sections will propose ideas of employing 24/7 into Mass Spectrometry research as it provides opportunities for this to be used, but first of all, key factors of software design will be described.

4.3 Steps of software design

Most of researchers who work in the field of Proteomics are with Computer Science, Mathematics, Biology background. They develop and improve algorithms which in future may be applied for clinical diagnosis. By having knowledge of software programming, it is common that designed programs are difficult to use for someone without this knowledge. This is the reason why it is very important to consider requirements of end user, because it is the person who will be using software the most. In general, it will be addressed to doctors with no high proficiency in programming. Good interface must comply with following rules: simple, intuitive, nice presentation, ergonomic - and finally, should make process of analysis easier. Some very useful tips on GUI design were explained by Weinschenk et al. [19]

4.3.1 Analysis

Designing a software is very laborious task. It can be even more time consuming if there is no planning included. That is why very important aspect is to spend more time on the process which will allow very good preparation. Analysis should include the following steps, and if performed properly can save huge amount of time:

- 1) decide who will be the end user of software
- 2) identify how the software is going to be used
- 3) be aware of computer limitations
- 4) plan the software in a way that it will be easy to upgrade

4.3.2 Design

With the first step, analysis properly performed, many important aspects should be noticed which will help to avoid faults during design process. When designing, there are also many subcomponents of the process, which must be considered. First of all, it is worth remembering that main goals of GUI are usefulness, reliability and that it should make work easier. As the software being designed for this project is for data processing - information search to be more specific - GUI should provide great

flexibility. It also should not overwhelm the user and states where user do not know what is happening should be avoided. When i.e. the function is launched which requires long time to accomplish, user should be informed how much more time will it take (in this case progress bars should be considered). Analysis software should reduce the demand on the user so so he can focus on performed experiment rather than on technical issues of programming.

4.3.3 Paper prototyping

After an effort has been made to analyze the requirements, with design advice in mind, it is a moment to prototype an interface on paper. This vary fast technique will give a brief overview of the ideas will look on the screen. It is much better to see how everything looks like, before it is put into computer. Changing layout on a paper is also much easier and faster than after everything is programmed. It is more efficient to program correctly from beginning rather than having to reprogram parts of the software. This part of the process, if performed properly, will save huge amount of time in the future.

4.3.4 Construction

When all the previous steps have been performed with success, the GUI can be programmed. It will be much easier, time efficient and as will result in the software which is useful, and which will help to perform tasks it is designed for.

4.4 Implementation of proposed solutions

With 24/7 work organization scheme, jobs have to be scheduled. In traditional distributed task assignment process, everything had to be planned carefully. If tasks were partitioned improperly, some groups were inactive, due to delay of a team which results they were dependant on. Usually, to manage many groups, one centralized unit had to be created. Exchange of an information was a long process, thus gain achieved by distributing tasks was much smaller than expected. 24/7 system help to avoid many disadvantages of traditional distributed work scheduling.

With Mass Spectrometry research, it is possible to make advantage of 24/7. As described in chapter 3 - Steps of data manipulation - it is clear that this process is very structured and can be simply divided into modules, i.e. data preprocessing, data manipulation or classification. Those are only 3 general steps of data manipulation, but can be further divided into smaller modules. This is excellent property

to apply distributed scheduling. Traditional distributed tasks assignment could be one choice: each group would have to work on different module. But to avoid disadvantages of traditional scheme, 24/7 is proposed to be used. Jobs can be assigned much more loosely so there is no requirement that a group has to finish day goals to not to block another group. It is based on great cooperation between teams, which may encourage new ideas.

Chapter 5

Case study

The main goal of this project was to prepare a mass spectrometry data analysis software for data classification. As a result of a work through all semester, the goal has been achieved - working program has been written and used for testing preprocessing and classifying algorithms. Some of the ways of using the software will be presented in this chapter.

5.1 Data set characteristics

Example data set was obtained from National Cancer Institute ¹. There are different sets to choose from, but the most recent one has been chosen: High Resolution SELDI-TOF Study Set. It contains 216 samples out of which 121 were derived from population with ovarian cancer, 95 were derived from healthy population. This data set has not been preprocessed so it is a good input to test preprocessing methods and classifier. More on the data set has been written by Conrads et al. [7] and Baggerly, Edmonson, Morris and Coombes [2].

5.2 Program requirements


Mass Spectrometry Analysis Program has been developed using Matlab 7.1 environment. It requires standard Matlab 7, Bioinformatics Toolbox and Statistics Toolbox, which are supplied with standard edition. It does not require any other software, but the better hardware used, the faster and more pleasant will be the work with program. The most computing power consuming process is loading data which are stored in separate text file for every sample. I.e it takes approximately 10 minutes to load all samples on the Intel Centrino 1.5GHz with 1024MB of RAM memory.

¹<http://home.ccr.cancer.gov/ncifdaproteomics/OvarianCD.PostQAQC.zip>

There are no significant delays on any of the functions which were used in program so Pentium IV 1.4 GHz processor with 1024MB of memory will be minimum required hardware.

5.3 Program demonstration

When the program is launched, it looks as presented on figure 5.1. Core structure of the software is intuitive and follows the sequence of a trial in 4 steps: Step 1 - Loading data, step 2 - Preprocessing, step 3 - Analysis and classification and finally step 4 - Results. On the right hand side there are few options which allow to perform additional tasks. Those are: Saving current data (this can be done at any time), plotting current data and setting global parameters.



Politechnika Wroclawska

MASS SPECTROMETRY DATA ANALYSIS

by Martin Radlak

STEP 1 *LOADING DATA*

Data status: **loaded** Resolution:

Samples: 216 Normal: 95 Cancer: 121 Low cutoff:

High cutoff:

STEP 2 *PREPROCESSING*

<input type="checkbox"/> Baseline Correction	<input type="button" value="Settings"/>	<input type="button" value="Launch"/>
<input type="checkbox"/> Peaks Detection	<input type="button" value="Settings"/>	<input type="button" value="Launch"/>
<input type="checkbox"/> Peaks Alignment	<input type="button" value="Settings"/>	<input type="button" value="Launch"/>
<input type="checkbox"/> Normalization	<input type="button" value="Settings"/>	<input type="button" value="Launch"/>
<input type="checkbox"/> Smooth	<input type="button" value="Settings"/>	<input type="button" value="Launch"/>
<input type="checkbox"/> PCA (dim. reduction)	<input type="button" value="Settings"/>	<input type="button" value="Launch"/>
<input type="checkbox"/> 2-way t-test	<input type="button" value="Settings"/>	<input type="button" value="Launch"/>

STEP 3 *ANALYSIS AND CLASSIFICATION*

<input type="radio"/> k-Nearest Neighbor	<input type="button" value="Launch"/>
<input type="radio"/> LDA	<input type="button" value="Launch"/>
<input type="radio"/> Perceptron (NN)	<input type="button" value="Launch"/>
<input type="radio"/> FF Neural Network	<input type="button" value="Settings"/> <input type="button" value="Launch"/>

STEP 4 *RESULTS*

	Current	Mean	St.Dev	Mn-NPV	St-NPV	Mn-PPV	St-PPV
k-Nearest Neighbor	90.8257	90.2141	2.8028	91.1793	2.0404	89.6471	3.962
LDA	96.3303	97.2477	0.91743	97.9148	2.0423	96.7989	1.5478
Perceptron (NN)	95.4128	95.4128	0.91743	97.777	0.049411	93.7653	1.4653
FF Neural Network	92.6606	91.7431	3.3078	86.7725	1.8329	96.6667	5.7735
Current Iteration	3		<input type="button" value="Global Launch"/>				

Preprocessing step is used to prepare samples in a way that their parameters are adjusted to be similar. It is mostly used to make signal similar at common points and emphasize important differences

Analysis and Classification use preprocessed data to divide samples into two classes: normal or cancer

Figure 5.1: Screen shoot of the Mass Spectrometry Data Analysis program

5.3.1 Step 1: Loading Data

First step to perform is to load the data. Program allows to load new samples from separate files, stored in two catalogues: normal and cancer respectively. There is

possibility to adjust the size of samples by using following parameters:

- 1) resolution - how many points per each sample to load. This allows reduction of size of the sample without significantly changing sample properties. Default value is 15000 points.
- 2) low cutoff - defines which point should be the first one of the spectra. Default value is 1000 M/Z
- 3) high cutoff - defines which point should be the last one of the spectra. Default value is 11900 M/Z

If other than default values have to be used, they can be changed in corresponding edit boxes.

Loaded data can be stored which is highly recommended, as samples will be saved as mat file. This matlab data file allows faster data import than standard "load new data" command. Data status always display info, whether samples are

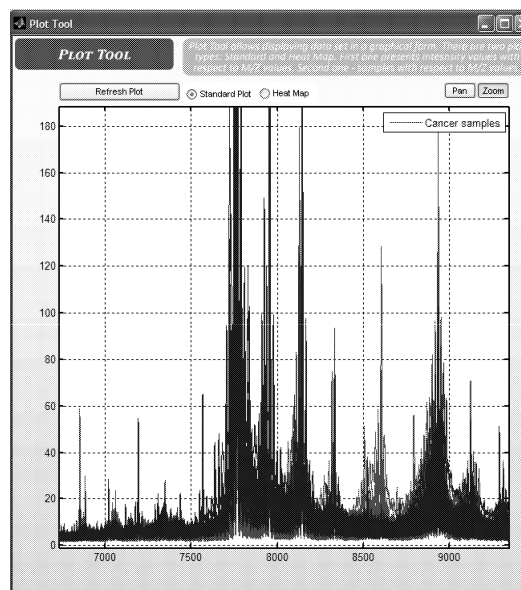


Figure 5.2: Zoomed unprocessed Spectra plotted as Standard Plot using Plot Tool

loaded or not, giving also quick summary of how many samples were loaded:

- 1) samples - number of all different samples
- 2) normal - number of all different healthy samples
- 3) cancer - number of all different cancerous samples

Loaded data can be plotted using "Plot Tools" This is presented on figure 5.2 for Standard Plot.

5.3.2 Step 2: Preprocessing

There are number of preprocessing algorithms implemented in the software. There are no limits to how many of them can be applied, so before classification is done: none, all or only part of preprocessing algorithms may be applied. Following is short description of settings which can be adjusted according to requirements.

Some of the functions were used from Matlab Bioinformatic Toolbox so description of parameters written according to Matlab Help.

Baseline correction

This algorithm adjusts the variable baseline of raw mass spectrum by following steps:

- 1) Estimates the baseline within multiple shifted windows of defined width.
- 2) Regresses the varying baseline to the window points using defined approximation method.
- 3) Adjusts the baseline of the spectrum (Y)

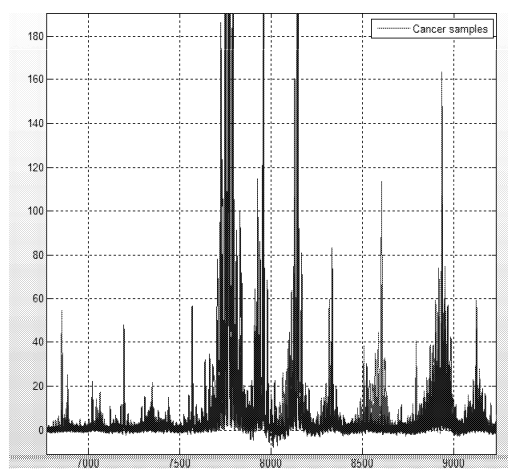


Figure 5.3: Loaded Spectra after baseline correction applied

The result of this algorithm depends on carefully choosing the window size and the step size. Width of peaks in the spectrum and the presence of possible drifts should also be considered. If there are wider peaks towards the end of the spectrum, it is a good idea to use variable parameters. Following are settings that may be adjusted:

- a) Windows size - specifies the width for the shifting window. This option is useful for cases where the resolution of the signal is dissimilar at different regions of the spectrogram.

- b) Step size - specifies the steps for the window
- c) Regression method - specifies the method to regress the window estimated points to a soft curve
- d) Smooth method - specifies the method for smoothing the curve of estimated points and eliminating the effects of possible outliers.
- e) Height preserve - sets the baseline subtraction mode to preserve the height of the tallest peak in the signal
- f) Show plot - plots the baseline estimated points, the regressed baseline, and the original spectrum

More information is available by typing "doc msbackadj" in Matlab command window.

Sample plot after algorithm was applied is presented on figure 5.3

Peaks detection

This command is used to detect some peaks in mass spectra. Using approximates derivatives it returns a vector of m/z values. This may be used by Peak Align algorithm. Following are settings that may be adjusted:

- a) Max peaks to detect - specifies how many peaks should be detected at most.

Peaks align

Align peaks in mass spectrum to reference peaks. This algorithm works best with 3 - 5 reference peaks. Following are settings that may be adjusted:

- a) Reference peaks -reference mass vector with a list of known masses in the sample spectrum.
- b) Weights - specifies the relative weights for every mass in the reference mass vector.
- c) Show Plot - shows plot of original and aligned spectra as a heat map

More information is available by typing "doc msalign" in Matlab command window.

Normalize

Normalizes a group of mass spectra by standardizing the area under the curve to the group median. Following settings may be adjusted:

- a) Low cutoff, high cutoff - range of normalization points.
- b) Max value - normalized points can be scaled according to an overall maximum intensity.

More information is available by typing "doc msnorm" in Matlab command window

Smooth

Smooth mass spectrum using nonparametric method. Following settings can be adjusted.

- a) Fit order - specifies the order of the smoother.
- b) Window size - specifies the window size for the smoothing kernel.
- c) Weighting function - selects the function for weighting the observed ion intensities. Samples close to the MZ location being smoothed have the most weight in determining the estimate.
- d) Show Plot - shows plot of smoothed spectrum over the original one

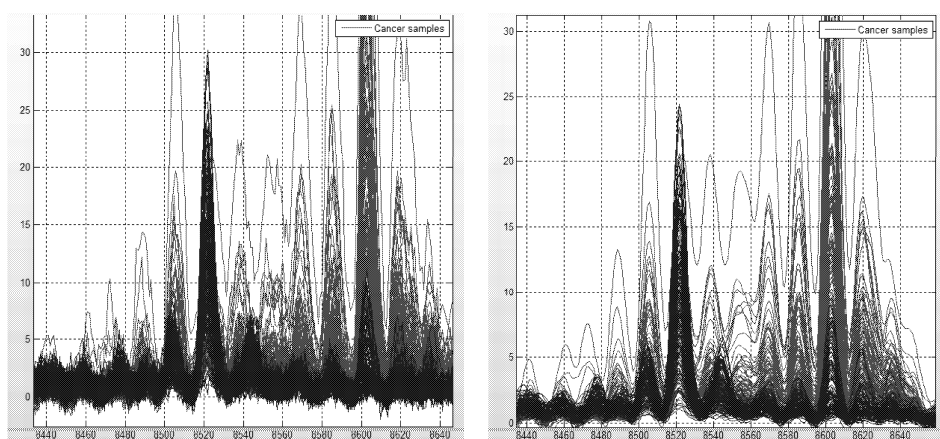


Figure 5.4: Loaded Spectra not smoothed (left) and after smoothing (right)

More information is available by typing "doc msloess" in Matlab command Window. Differences between smoothed not smoothed spectra are presented on figure 5.4

5.3.3 Step 3: Analysis and Classification

Four algorithms are available in this section: k-Nearest Neighbors, simplified LDA algorithm (Lilien et al. [10]), perceptron and 3 Layer Artificial Neural Network.

k-Nearest Neighbors

This classifier is launched with default settings of 3 nearest neighbors with the distance metric set to one minus the sample correlation between points.

Linear Discriminant Analysis

This LDA classifier also needs data dimension reduction what is done by use of PCA, as a process before it is applied. Every time it is launched, vector of training and testing samples is calculated.

Perceptron

Consist of a single layer with Hard Limit transfer function. It is used because can classify data into two classes.

3-Layer Artificial Neural Network

Before Neural Networks are used, Principal Component Analysis (PCA) is used to reduce dimension of data. Every time it is launched, vector of training and testing samples is calculated and PCA is applied every time, because of the data change. This classifier can be adjusted according to the following setting:

- a) Layer size - defines how many neurons each layer contains
- b) Training epochs - maximum number of epochs the network will be trained for.
- c) Training goal - Value of Mean Square Error, which is satisfying to stop training process
- d) Minimum gradient - Value of the gradient which force the network to stop training

5.3.4 Step 4: Results

Results are presented according to which classifying method has been used. It is a percentage of correctly classified samples. There are few different possibilities to see performance of a classifier. First option is to try a single run. Results are presented

in column "Current". If statistical sample was required, it can be done using "Global Launch". This evaluate selected preprocessing methods and performs classification for every classifier - 50 times. Mean and standard deviation is then calculated and displayed in following columns: "Percentage of correctly classified", "Standard Deviation of percentage of correctly classified", "NPV", "Standard Deviation of NPV", "PPV" and "Standard Deviation of PPV".

5.4 Additional features

As an additional feature, plotting tool is available to view samples at any time. There is standard plot or heat map available, together with zooming and panning options. Additionally, it is possible to set a ration of training data to evaluation data in "Global Settings".

5.5 Classifier performance according to preprocessing methods

In this part Mass Spectrometry Analysis Software has been used to examine efficiency of the following methods: k-Nearest Neighbors (KNN), PCA+LDA (PCA), Perceptron (PER) and Feed-Forward Neural Network (FFN) trained with Back-Propagation algorithm. Therefore, dependency of preprocessing methods in final classifier performance has been examined. Sample data [?] have been equally divided into training and evaluation set. Training and classification has been repeated 50 times to create statistically significant results. Training and evaluation data were randomly selected for each run to avoid repetition of the same set. Results are presented in Table 5.1. It is not straightforward to say that applying preprocessing methods will increase performance of a classifier. Some of the methods decrease its performance. I.e When only normalization has been applied, performance of LDA has been increased. Also, standard deviation has been lowered indicating that classifier produce much more repeatable output. On the other side, when three different preprocessing methods were applied together - Baseline Correction, Normalization and Smoothing - apart from k-Nearest Neighbors, all classifiers performed worse than if no preprocessing was applied. This leads to following conclusion: at first - preprocessing methods should be carefully chosen to not loose important information which raw data contain, secondly - performance of the classifier might be increase by applying correct preprocessing methods, but it is important to notice,

Table 5.1: Classifier performance based on preprocessing method

	KNN	LDA	PER	FFN
No preprocessing				
% Correct	91.74	98.28	96.44	92.33
St. Dev.	3.00	1.12	1.80	4.40
NPV (Mean)	89.24	98.66	98.02	93.32
NPV (StDev)	4.75	1.84	2.67	7.18
PPV (Mean)	92.06	98.07	95.46	92.79
PPV (StDev)	3.66	1.90	2.82	5.76
Baseline Corrected				
% Correct	91.25	98.35	96.50	93.19
St. Dev.	3.16	1.20	1.78	3.50
NPV (Mean)	90.83	98.41	97.84	95.948
NPV (StDev)	5.12	1.88	2.35	4.93
PPV (Mean)	91.89	98.37	95.68	92.22
PPV (StDev)	3.24	1.72	2.94	6.00
Normalization				
% Correct	90.81	98.68	97.27	94.24
St. Dev.	2.97	0.89	2.01	4.24
NPV (Mean)	89.08	99.10	98.20	94.47
NPV (StDev)	5.08	1.43	2.39	7.71
PPV (Mean)	92.60	98.49	96.74	95.01
PPV (StDev)	3.25	1.49	3.11	4.10
Smoothing				
% Correct	90.81	97.96	95.65	93.28
St. Dev.	2.94	1.40	2.14	3.63
NPV (Mean)	88.87	98.62	97.33	94.81
NPV (StDev)	4.74	1.42	2.93	5.78
PPV (Mean)	92.71	97.53	94.66	93.02
PPV (StDev)	3.01	2.20	3.27	5.29
Baseline Corr. + Normalization + Smoothing				
% Correct	92.09	98.28	95.95	91.27
St. Dev.	2.94	1.17	2.39	13.67
NPV (Mean)	91.19	98.11	94.66	88.17
NPV (StDev)	4.75	2.27	4.40	14.33
PPV (Mean)	93.12	98.49	97.28	94.82
PPV (StDev)	3.24	1.57	2.48	14.17

that if the classifier is well constructed, it can be superior, even if no preprocessing is applied.

At this point it is also important to comment on results achieved by using Feed-Forward Neural Network. As observed in Table 5.1 FFN performs poorly. It is better than KNN but is characterized by highest standard deviation of the results. The reason for it is probably the way of using Neural Network in this work. Tests were performed for 3 fully connected layers consisting of 9-5-2 hidden units. Network has been trained using simple back-propagation algorithm. Even though settings were not adjusted to the data, still produced results were satisfying. Therefore, it is possible to achieve increased performance, if a better Neural Network model has been prepared, i.e. different activation functions, architecture optimized using Evolutionary Computation or more stable training algorithm.

Chapter 6

Results and conclusion

As a result of intensive work, Data Analysis Software for Mass Spectrometry has been developed. As a main goal set for this project was to develop data analysis software and to analyze its performance - this has been achieved. But the process of preparing the software consisted not only of coding. Before that could be done, some technical background had to be gained about technology available and requirements for this type of software. By researching literature where many ideas have been proposed, few of the most common were chosen and planned in detail to include in the software. This fulfilled one of the objectives which was to do adequate literature review. Secondly, it allowed to understand relationship between cancer and mass spectrometry, what also has been described in the previous chapters.

By describing methods for data preprocessing and classification, second objective has been fulfilled, which was to review data analysis and classification methods. Theoretical description of these has been described in chapter 3.

By describing implementation of 24/7 work organization scheme, subsequent goal has been achieved - to propose a system which will allow to increase the speed of research.

Finally, by implementing data analysis and classification algorithms into a working program, another objective have also been fulfilled.

Considering this work with respect to objectives and aims, it can be assumed that all requirements have been met. The program is fully working piece of software which, even at this stage, can serve user with varying important information, i.e. performance of methods chosen, plot of the data and many more, described in detail in chapter 5.

6.1 Strengths, weaknesses and future developments

There are many strengths of work performed. Firstly, the software is based on Matlab environment which is a cross field development platform. Not only programmers use it, but also biologists, chemists or physicists. By encouraging these groups to create communities which will cooperate using, i.e. proposed 24/7 system. Researchers from different areas of science should be encouraged. This is required, as i.e. biologist know details which computer scientist know nothing about (i.e. structure of organic tissue - biologist, programming skills - computer scientists).

Additionally, many algorithms and general framework for cancer detection has been presented. This is laid solid foundations for future work which should be directed towards developing an online application: accessible by anyone - everywhere. This online system, could possibly collect data about patients, i.e. their medical history, which would provide additional knowledge. Furthermore, based on these data, new data mining algorithms could be employed to analyze these information and possibly find the causes of cancer disease.

6.2 Final word

This project introduced mechanisms and ideas which are not sole work. Information provided here are a part of large research. By introducing author's point of view, possibly new ideas will be emerge through further researchers.

The summary of this project has been presented in articles [14] and [13].

Appendix A: Instruction for program execution

Program source code is available for download from website:

<http://projects.radlak.com>.

Zip file should be downloaded and unpacked into separate folder, which should be added to matlab path. By typing: `miniProject1`, program screen is displayed as presented on figure 5.1.

The software is optimized for data available from National Cancer Institute: http://home.ccr.cancer.gov/ncifdaproteomics/OvarianCD_PostQAQC.zip. To load the new data, they should be unzipped to a folder, which contains two sub-folder: cancer and normal. Samples should be stored on a disk according to a folder which they belong to. All settings have been described in chapter 5.

Bibliography

- [1] Aebersold, R. and Goodlett, D. R. [2001], ‘Mass spectrometry in proteomics’, *Chemical Reviews* **101**.
- [2] Baggerly, K. A., Edmonson, S. R., Morris, J. S. and Coombes, K. R. [2004], ‘High-resolution serum proteomic patterns for ovarian cancer detection’, *Endocrine-Related Cancer* **11**.
- [3] Baggerly, K. A., Morris, J. S. and Combes, K. R. [2004], ‘Reproducibility of seldi-tof protein patterns in serum: comparing datasets from different experiments’, *Bioinformatics* **20**(5).
- [4] Ball, G., Mian, S., Holding, F., Allibone, R. O., Lowe, J., Ali, S., Li, G., McCardle, S., Ellis, I. O., Creaser, C. and Rees, R. C. [2002], ‘An integrated approach utilizing artificial neural networks and seldi mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers’, *Bioinformatics* **18**(3).
- [5] Bensmail, H. and Haoudi, A. [n.d.], ‘Postgenomics: Proteomics and bioinformatics in cancer research’, *Journal of Biomedicine and Biotechnology* .
- [6] Chaczko, Z., Klempous, R., Nikodem, J. and Rozenblit, J. [2006], 24/7 software development in virtual student exchange groups: Redefining the work and study week, in ‘ITHET 7th Annual International Conference, Sydney, Australia’.
- [7] Conrads, T. P., Fusaro, V. A., Ross, S., Johann, D., Rajapakse, V., Hitt, B. A., Steinberg, S. M., Kohn, E. C., Fishman, D. A., Whiteley, G., Barrett, J. C., Liotta, L. A., III, E. F. P. and Veenstra, T. D. [2004], ‘High-resolution serum proteomic features for ovarian cancer detection’, *Endocrine-Related Cancer* **11**.
- [8] Coombes, K. R., Tsavachidis, S., Morris, J. S., Baggerly, K. A., Hung, M.-C. and Kuerer, H. M. [2005], ‘Improved peak detection and quantification of mass

- spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform', *Proteomics* **5**(16).
- [9] Donald, D., Hancock, T., Coomans, D. and Everingham, Y. [2005], 'Bagged super wavelet reduction for boosted prostate cancer classification of seldi-tof mass spectral serum profiles', *Chemometries and Intelligent Laboratory Systems* **82**.
- [10] Lilien, R. H., Farid, H. and Donald, B. R. [2003], 'Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum', *Journal of Computational Biology* **10**(6).
- [11] *Oxford English Dictionary* [2006], Oxford University Press.
- [12] Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C. and Liotta, L. A. [2002], 'Use of proteomic patterns in serum to identify ovarian cancer', *The Lancet* **16**(2).
- [13] Radlak, M. and Klempous, R. [2007a], 'Seldi-tof-ms data manipulation with designed software for methods benchmarks'.
- [14] Radlak, M. and Klempous, R. [2007b], 'Seldi-tof-ms pattern analysis for cancer detection as a base for diagnostic software'.
- [15] Satten, G. A., Datta, S., Moura, H., Woolfitt, A. R., da G. Carvalho, M., Carlone, G. M., De, B. K., Pavlopoulos, A. and Barr1, J. R. [2004], 'Standardization and denoising algorithms for mass spectra to classify whole-organism bacterial specimens', *Bioinformatics* **20**(17).
- [16] Smith, L. I. [2002], 'A tutorial on principal components analysis'.
- [17] Vitzthum, F., Behrens, F., Anderson, A. N. and Shaw, J. H. [2005], 'Proteomics: From basic research to diagnostic application a review of requirements and needs', *Journal of Proteome* **4**.
- [18] Vorderwylbecke, S., Cleverley, S., Weinberger, S. R. and Wiesner, A. [2005], 'Protein quantification by the seldi-tof-ms-based proteinchip system', *Nature Methods* **2**.
- [19] Weinschenk, S., Jamar, P., Yeo, S. C. and Jamar, P. [1997], *GUI Design Essentials*, Wiley.

- [20] Yu, J. and Chen, X.-W. [2005], ‘Bayesian neural network approaches to ovarian cancer identification from high-resolution mass spectrometry data’, *Bioinformatics* **1**.
- [21] Zhou, Z.-H., IEEE, S. M. and Liu, X.-Y. [2005], ‘Training cost-sensitive neural networks with methods addressing the class imbalance problem’, *IEEE Transactions on Knowledge and Data Engineering* .