

BRNO UNIVERSITY OF TECHNOLOGY  
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION  
DEPARTMENT OF BIOMEDICAL ENGINEERING

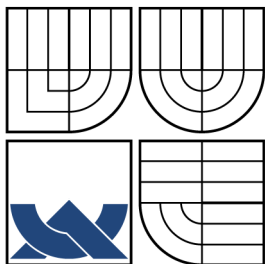
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ  
ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

DYNAMIC SPOT ANALYSIS IN THE 2D ELECTROPHORESIS GELS  
IMAGES

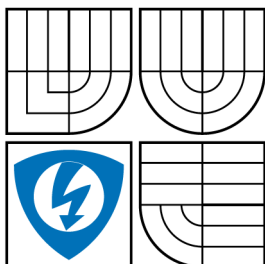
BACHELOR'S THESIS  
BAKALÁŘSKÁ PRÁCE

AUTHOR  
AUTOR PRÁCE

LENKA POLÁŠKOVÁ



BRNO UNIVERSITY OF TECHNOLOGY  
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ



FACULTY OF ELECTRICAL ENGINEERING AND  
COMMUNICATION  
DEPARTMENT OF BIOMEDICAL ENGINEERING



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH  
TECHNOLOGIÍ  
ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

## DYNAMIC SPOT ANALYSIS IN THE 2D ELECTROPHORESIS GELS IMAGES

DYNAMIKA SPOTU U OBRAZŮ 2D ELEKTROGRAFICKÝCH GELŮ

BACHELOR'S THESIS  
BAKALÁŘSKÁ PRÁCE

AUTHOR  
AUTOR PRÁCE

LENKA POLÁŠKOVÁ

SUPERVISOR  
VEDOUCÍ PRÁCE

Ing. JIŘÍ NEDVĚD

BRNO 2014

## **ABSTRACT**

The text briefly describes factors and parameters which influence the results of 2D electrophoresis focusing on image processing as one manner to reduce incorrect interpretation of its outputs. As dataset, image processing performance uses images from repeated execution of one experiment also known as multiplicates. Using multiplicates analysis it is possible to observe or lower the changes of one experiment and to compare them with outputs of other experiments. The aim of this work is to provide support for specialist who takes care about describing the character patterns located in electrophoretic images.

## **KEYWORDS**

electrophoresis, image processing, spot dynamics

## **ABSTRAKT**

Práce shrnuje faktory a parametry, které ovlivňují výsledky 2D elektroforézy, se zaměřením na zpracování obrazu jako jeden ze způsobů snížení nesprávné interpretace jejích výstupů. Proces zpracování obrazu využívá jako zdroj dat především obrazů z opakovaných provedení téhož pokusu, neboli víceplik. Pomocí analýzy obrazů víceplik je možno pozorovat nebo korigovat změny jednoho pokusu a také porovnávat je s výstupy jiných pokusů. Cílem práce je poskytnout podporu specialistovi, který má na starosti popsat vlastnosti struktur nacházejících se v elektroforetických obrazech.

## **KLÍČOVÁ SLOVA**

elektroforéza, zpracování obrazu, dynamika spotu

## DECLARATION

I declare that I have elaborated my bachelor's thesis on the theme of "Dynamic spot analysis in the 2D electrophoresis gels images" independently, under the supervision of the bachelor's thesis supervisor and with the use of technical literature and other sources of information which are all quoted in the thesis and detailed in the list of literature at the end of the thesis.

As the author of the bachelor's thesis I furthermore declare that, concerning the creation of this bachelor's thesis, I have not infringed any copyright. In particular, I have not unlawfully encroached on anyone's personal copyright and I am fully aware of the consequences in the case of breaking Regulation § 11 and the following of the Copyright Act No 121/2000 Vol., including the possible consequences of criminal law resulted from Regulation § 152 of Criminal Act No 140/1961 Vol.

Brno .....

.....

(author's signature)

## ACKNOWLEDGEMENT

I would like to express a special appreciation and thanks to my supervisor Ing. Jiří Nedvěd for a kind and valuable support as my consultant, advisor and teacher. I would also thank to my other consultant, RNDr. Lochmanová, for willing cooperation.

Brno .....

.....  
(author's signature)

# CONTENTS

<b>Introduction</b>	<b>7</b>
<b>1 Theoretical part</b>	<b>8</b>
1.1 Genome, proteome, electrophoresis . . . . .	8
1.2 Samples for electrophoresis . . . . .	8
1.3 2D electrophoresis . . . . .	9
1.4 Variability of the output in 2D electrophoresis . . . . .	10
1.5 Image processing . . . . .	11
1.5.1 Image aquisition . . . . .	11
1.5.2 Image adjusting . . . . .	11
1.5.3 Segmentation . . . . .	13
1.5.4 Spot detection . . . . .	15
1.5.5 Spot analysis . . . . .	16
1.5.6 Analysis of electrophoretic map . . . . .	16
<b>2 Practical part</b>	<b>18</b>
2.1 Core of the program . . . . .	18
2.1.1 Image contrast adjusting . . . . .	19
2.1.2 Edge repairing . . . . .	19
2.1.3 Crucial step - spots detection . . . . .	19
2.1.4 Spot selection . . . . .	20
2.2 Graphical user interface . . . . .	21
2.3 Building standalone application . . . . .	22
<b>3 Results of bachelor work</b>	<b>24</b>
3.1 Final application . . . . .	24
3.2 Future prospects . . . . .	25
<b>4 Conclusion</b>	<b>28</b>
<b>Bibliography</b>	<b>29</b>
<b>List of symbols, physical constants and abbreviations</b>	<b>31</b>
<b>5 Appendix</b>	<b>33</b>
5.1 The content of DVD . . . . .	33

## LIST OF FIGURES

1.1	Image shown in 1.1a has lower mean gray value than other images in the same set of multiplicates. Adjusted image is shown in 1.1b. . . . .	12
1.2	Erosion and dilatation, original matrix in red boundary, structure element in blue boundary and resulting matrix in yellow boundary. . . . .	12
1.3	Two separated spots are represented as two catchment basins (created in MATLAB) . . . . .	14
1.4	Segmentation of two overlapping objects is based on computation of inter-distance. . . . .	14
1.5	Imersion approach in watershed algorithm, from: [1] . . . . .	15
2.1	General scheme of application's algorithm . . . . .	18
2.2	Distorted edges in 2.2a and straightened edges in 2.2b . . . . .	20
2.3	MATLAB objects' hierarchy (redrawn partially from: [2]) . . . . .	22
3.1	Ask dialog . . . . .	25
3.2	View of the region of interest . . . . .	25
3.3	Detected spots and the control panel with settings . . . . .	26
3.4	Application's graphical output . . . . .	26
5.1	Electrophoretic gel, black spots represent clusters of proteins . . . . .	33

## INTRODUCTION

The research of genome and proteome undoubtedly covers an important field in molecular biology and biomedical engineering. The continual development in these branches allows not only to construct the list of proteins coded by genes, map proteins on the chromosomes, examine their expression and conditions in which this expression is executed, but also helps to explain the influence of related components to the proteins' production. Results provided by the research will be used for data mining in order to gain information leading to improve the disease therapy. There exist multiple approaches to the given task. Gel electrophoresis belongs among these as an easy, quick, visual method, which have also several modifications. In this publication, the electrophoresis will be described in details with aim to explain the reason of using multiplicates during electrophoretic result's analysis.



# 1 THEORETICAL PART

## 1.1 Genome, proteome, electrophoresis

At the beginning of this work, let us first explain the terms genome and proteome. Genom is a complex of all DNA contained in an organism, however genetic information is encoded only by a small part of it. Genes are protein-coding sequences of DNA. Prokaryotes store its DNA in cytoplasm as opposed to eukaryotes, which have their DNA stored in cell nucleus and a small part of it in mitochondria. Proteome is a complex of all proteins appearing in an organism at a given time [3]. Genes are translated to a chain of amino acids, which binds together a protein. This whole mechanism is generally called gene expression. Expression is invoked by circumstances of internal or external origin and thus is strongly dependent on signals, which come from the cellular environment itself or from the whole organism. Expression is also influenced by non-coding regions of DNA - regulatory sequences (i.e. promoters, silencers, enhancers etc.). The content of proteome is thus dependent on its own expression cycle. To finish, the most important information concerning expression's process and regulation is retrieved by its examination. [4]

Electrophoresis is a widely used technique of separating particles with respect to their mobility in electrical field. The main targets of the electrophoresis in bioengineering are: proteins, nucleic acids and DNA fragments. However, it is also possible to apply it to sugars, amino acids or ions [5]. As mentioned in introduction of this work, the method of electrophoresis will be described as a convenient tool of protein separation.

Protein mobility in electric field is given by its size, form and charge. Different electrophoretic modifications use the dependency of these parameters in different ways. Other factors influencing the electrophoretic process are the applied voltage, gel viscosity, ionic forces and others [6]. The general principle of electrophoresis is to dip a mixture of proteins into wells made in agarose or polyacrylamide gel carrier situated in a plastic platform. Whole platform is placed into a chamber full of conductive liquid and the opposite sides are connected to a direct voltage supply set to a value in order of tens of volts (the voltage shouldn't exceed 125 V). Then the voltage is applied for tens of minutes or units of hours [7]. Current flowing through the gel is stable during this experiment (the source acts as a constant current source). Electrophoretic conductivity of proteins is thus proportional to the applied voltage. Using a constant voltage source is also a common way. [5]

## 1.2 Samples for electrophoresis

Intracellular, interstitial and transcellular fluids represent sources of sample for electrophoresis. Interstitial fluid (blood) is a part of extracellular environment. Consequently, blood serum is retrieved from this source by blood clotting. Serum proteins are then determined by a one-dimensional electrophoresis. Collecting transcellular fluid (saliva, phlegm, products of glands) don't provide whole proteome, which could be useful thanks to the

fact, that sample doesn't contain non-required proteins. Typically hormonal and enzymatic levels are evaluated using this approach.

Body fluids are important source of disease markers but their revelation is limited as proteins interfere with other components (i.e. salts). Intracellular fluid is collected from tissue cells by centrifugation. In that case it is required to define what kind of cells were used for the collection. [3]

Before the own electrophoresis process can begin, many other operations follow after sample extraction from the source. Special protocols have been developed for these purposes and following operations require skilled experts and modern technology.

### 1.3 2D electrophoresis

Two-dimensional electrophoresis provides possibility to resolve the proteins according to two parameters - according to their isoelectric point  $pI$  and according to their molecular mass  $M_r$  [3]. Proteins are first resolved in one dimension according to first parameter and this step is so-called isoelectric focusing (IEF). This method is based on different proteins' charge. The net charge is a sum of the charges of amido acid side chains. Theoretically, all acidic ends of amino acid are dissociated and all basic ends of amino acids are protonated at defined pH value <sup>1</sup> [3]. This value of pH is called the isoelectric point. The principle of IEF is thus to create a pH gradient in direction from cathode to anode. Originally this gradient was established on polyacrylamid rods, which were formed inside glass or plastic tubes. This method was found to be nonstable, irreproducible and hard to work with. Nowadays, the tubes are replaced by strips of gels on a plastic film, where carrier ampholytes are attached with gels molecules and thus a stable pH gradient can be established [8][3]. This concept is a base for immobilised pH gradient method (IPG) which is used for two-dimensional electrophoresis as a separation technique in the first dimension. IPG removed all mentioned weaknesses of IEF [8]. After electric field is applied, proteins start to move from the cathodic end to the anodic one and stay fixed in place where their total charge equals zero [9]. For greater resolution, more strips are produced and placed next each other inside one cooled platform. Each strip has its own range of pH. Sample is then dipped into every single strip [8]. Strip pH range is chosen to be 3-10 or 4-7, other ranges are also applicable according to what proteins are to be determined [3][8]. Building more strips into one platform also allows to choose the pH range narrower which is convenient for separation of proteins with similar values of  $pI$ . That proteins are thus spread out over greater physical distances [8]. These gels are also known as „zoom gels“. Gel strips have also another important advantage, which is the possibility of loading greater amount of a sample making the running of preparative gels possible for later characterisation analysis. [3]

After IEF is done, in the next step each strip is built into upper side of another square platform where their charges are unificated in sodium dodecyl sulfate (SDS-PAGE). Now

---

<sup>1</sup>practically nonachievable

the proteins will be resolved according to the second parameter - molecular mass.

Resulting spots are composed from clusters of proteins. They are visualised by either staining with Coomassie blue dye, silver staining, SYPRO Ruby or by labeling such as radiolabeling, immunological labeling or labeling by fluoro/phosphorimagers etc. [10]. It is important to know that different staining techniques stain different proteins. There are proteins that are not stained by CBB but that do with silver staining and vice versa. The electrophoretic output is then converted to the digital data using scanning techniques and stored in a computer. The example of such image is shown in 5.1. Following step of electrophoretic gel analysis is the image processing and pieces of gels can be removed to be used again with another analysis (i.e. mass spectrometry analysis [10]). Other techniques combining mass spectrometry with liquid chromatography or immunochemical approaches are executed in parallel with the same sample as well. These methods usually provide better results, however, all of these methods have some limits. Using 2D electrophoresis, the analysis resolution of up to 10 000 proteins is available and 2000 of them are resolved usually. [3]

## 1.4 Variability of the output in 2D electrophoresis

The outputs of the same electrophoretic experiments don't look identically. Structures intended for electrophoresis are very sensible to even small changes in input parameters and interactions during the sample preparation, their own separation or gel visualisation after electrophoretic separation. During sample preparation, proteins interact with substances in solution. Very hard fault is caused by a mechanical gel damage during manipulation with the sample or gel. Non-required interactions are mainly caused by agents such as salts, DNA, stainings, proteases or proteins that interact with each other.

Incorrect manipulation may cause insufficient gel solubilisation or hydration or even it can make proteins enter into the wells incompletely. It is not always possible to avoid these situations. According to [11], most common source of variability are current inhomogeneities inside of the gel. These aspects result in changes of proteins migration and thus it creates a drift of spots against the control sample. It is preferred to reduce variability experimentally during sample preparation, to reduce it by image processing techniques. [11]

One way how to examine or reduce the influence of input parameters variability on resulting spot layout is to run the same process with the same sample multiple times in parallel. The same sample is filled into several platforms where electrophoresis is being done. The outputs of these experiments are thus called multiplicates. The analysis of the same spots in each multiplicate is averaged for achieving the highest likelihood of final analysis result.

## 1.5 Image processing

### 1.5.1 Image aquisition

Depending on the method of proteins staining or labelling, there are specified methods of image aquisition. If Coomassie Blue or silver staining is used, special scanners are required. Laser densitometers serve for visualisation of radiolabelled spots and fluoro/phosphorescent scanners provide visualisation of fluoro/phospholabelled images. Lasers with different wavelength are combined with different filters for storage phosphor screen or fluorescent dyers. CCD camera is also a convenient tool, its resolution can compete with the resolution of scanners [3]. Image must be obtained as .TIFF file with at least 16-bit intensity. The acquisition set must be calibrated for the measurements to be accurate.

### 1.5.2 Image adjusting

Geometric image distortion is often caused by current leakage over platform boundaries that are not completely isolated. Current characteristics may be drawn as equipotential lines given by linear interpolation of values of  $pI$  and exponential interpolation of values of  $M_r$  representing positions of known protein markers. Equipotential lines are straight in homogenous gel whereas in case of current leakage, they are curved. Position of each distorted pixel then can be remapped using vertical and horizontal rotations or translations. This reparation process is generally called image-warping. For these purposes, so-called „current leakage“ model was established. This model assumes original coordinates regarding to the voltage applied on the cathode, gel length, velocity of proteins in gel and current leakage out of boundary, which is set as solution of differential equations. [9]

The intensity correction in the image and normalization between multiplicates is reached by level transformation function adjusting in each image. It is convenient to use thresholding to enhance contrast of spots against the background. The intensity correction can be done using histogram equalisation or adjusting level transformation function. [12]

#### 1.5.2.1 Opening and closing

Structuring element is a small matrix (usually 3x3 or 5x5) which is compared with the image in every positions. For comparing, a small matrix of the same dimensions as structure element matrix is extracted from image. This small image matrix is called a window. If the positions of value in both matrices equals then central value in resulting matrix is set to one, otherwise it is set to zero. This is called dilatation 1.2a. If any two corresponding positions are of value one in both matrices, then the central pixel in resulting matrix is set to one, otherwise it is set to zero. This is called erosion 1.2b.

Opening is defined as erosion followed by dilatation and closing is defined, as opposite process - as dilatation followed by erosion. If a small structure element is used for opening, random details in image are removed and thus the background is subtracted. Remaining

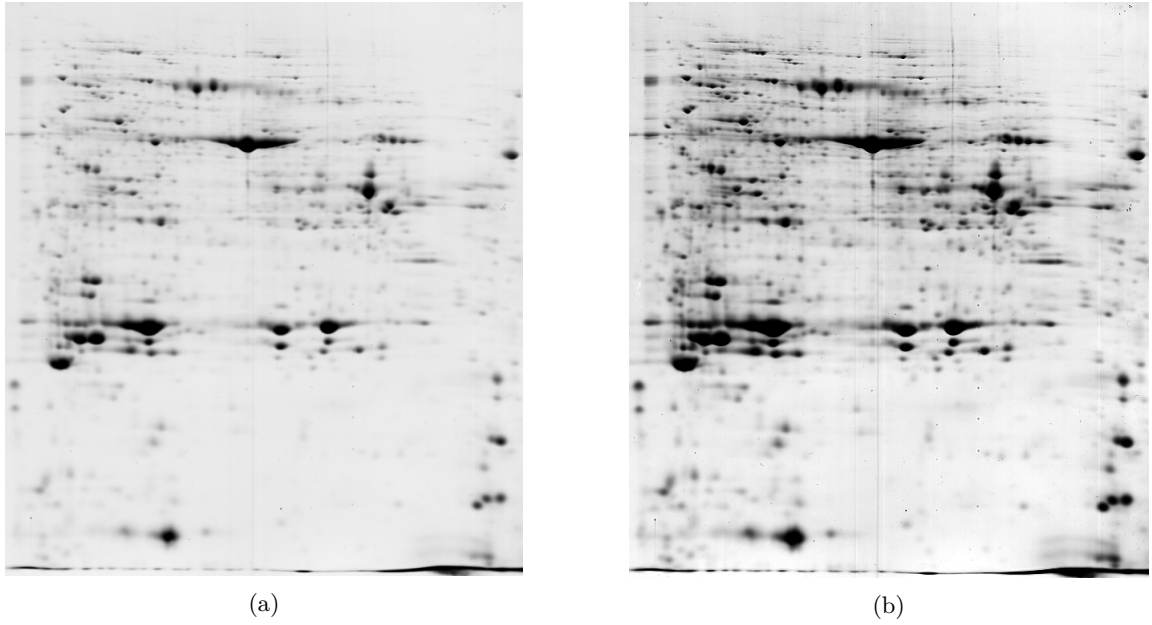


Fig. 1.1: Image shown in 1.1a has lower mean gray value than other images in the same set of multiplicates. Adjusted image is shown in 1.1b.

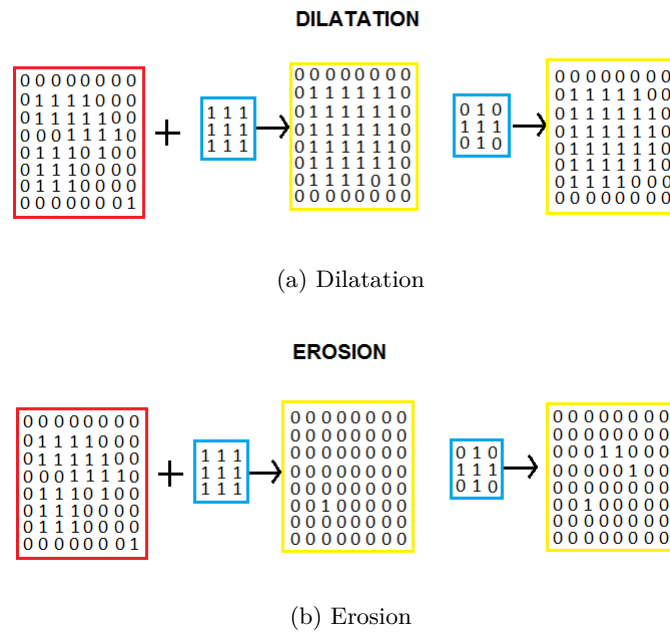


Fig. 1.2: Erosion and dilatation, original matrix in red boundary, structure element in blue boundary and resulting matrix in yellow boundary.

spots are emphasized by closing. As structure element sphere „the rolling ball“ is used and is greater than the greatest spot in image [12].

### 1.5.3 Segmentation

It is convenient to use some segmentation method before spot detection itself. This enables one to distinguish objects in the image and even in conditions where the objects are overlapping. Segmentation is indeed an important step in spot detection.

#### 1.5.3.1 Thresholding

The simplest way of segmentation is thresholding. Thresholded image is then in binary form and is defined by:

$$g(x, y) = \begin{cases} 1 & \text{if } f(x, y) \geq T \\ 0 & \text{if } f(x, y) < T \end{cases} \quad (1.1)$$

This approach can become a partial step of sophisticated methods.

#### 1.5.3.2 Edge operators

Alternative method for image segmentation represents edge enhancing using so-called edge operators. Edge is defined as a region where level transformation function escalates. Edge operators' work is based on convolution with convenient kernel or finding the biggest second-order derivative. Many operators were developed, the most common are i.e. Roberts operator, Prewitt operator, Canny operator and Sobel operator. [13]

#### 1.5.3.3 Watershed

The most popular approach to segmentation is watershed transform [13]. The image is represented as landscape or topographic relief which is flooded by water. Pixels with the least intensity are connected by curves representing rivers. Regions splitted by these curves are called catchment basins. Drawn curves split objects in image 1.3b (original extract is in Fig. 1.3a). Several modifications of this method were developed.

The simplest way is based on computing distances between pixels. Image is first thresholded then the distance values are computed. The distance is calculated from every pixel to the nearest nonzero-valued pixel. The matrix of distances' values is then viewed as an intensity image so this new image undergo watershed transform. An example is shown in Fig. 1.4. At the left side there is an image of two overlapping objects. Distance values were computed from its negative and the resulting intensity image is drawn at the center. At right lower side is situated the result of watershed transform of previous image. So this way enables to resolve partly overlapping objects.

The gradient method is another approach used to preprocess an image prior to using watershed transform. First step in this approach is to enhance the edges using some edge

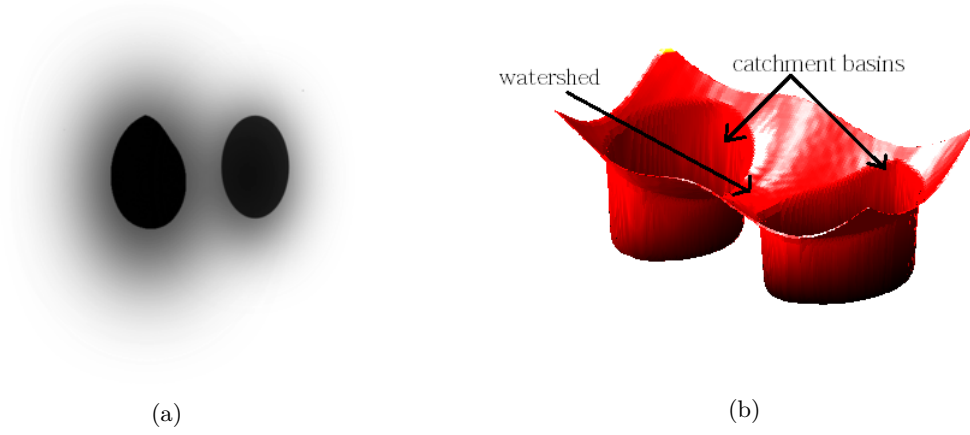


Fig. 1.3: Two separated spots are represented as two catchment basins (created in MATLAB)

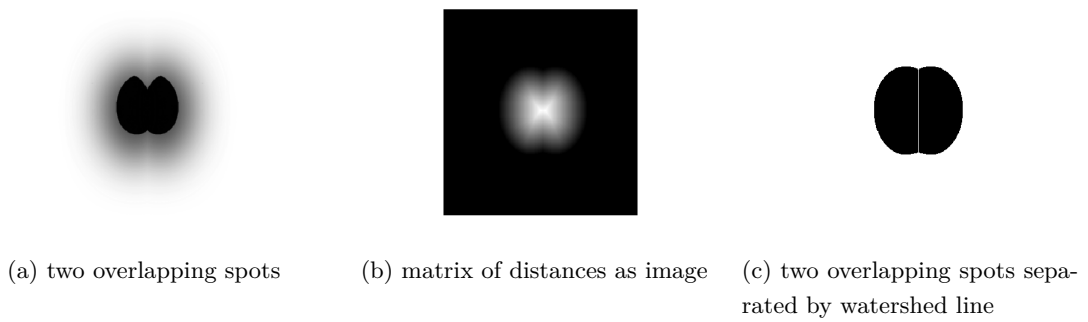


Fig. 1.4: Segmentation of two overlapping objects is based on computation of inter-distance.

<table><tr><td>3</td><td>2</td><td>2</td></tr><tr><td>3</td><td>1</td><td>1</td></tr><tr><td>0</td><td>1</td><td>0</td></tr></table>	3	2	2	3	1	1	0	1	0	<table><tr><td>3</td><td>2</td><td>2</td></tr><tr><td>3</td><td>1</td><td>1</td></tr><tr><td>A</td><td>1</td><td>B</td></tr></table>	3	2	2	3	1	1	A	1	B	<table><tr><td>3</td><td>2</td><td>2</td></tr><tr><td>3</td><td>W</td><td>B</td></tr><tr><td>A</td><td>W</td><td>B</td></tr></table>	3	2	2	3	W	B	A	W	B	<table><tr><td>3</td><td>B</td><td>B</td></tr><tr><td>3</td><td>B</td><td>B</td></tr><tr><td>A</td><td>W</td><td>B</td></tr></table>	3	B	B	3	B	B	A	W	B	<table><tr><td>B</td><td>B</td><td>B</td></tr><tr><td>W</td><td>B</td><td>B</td></tr><tr><td>A</td><td>W</td><td>B</td></tr></table>	B	B	B	W	B	B	A	W	B
3	2	2																																															
3	1	1																																															
0	1	0																																															
3	2	2																																															
3	1	1																																															
A	1	B																																															
3	2	2																																															
3	W	B																																															
A	W	B																																															
3	B	B																																															
3	B	B																																															
A	W	B																																															
B	B	B																																															
W	B	B																																															
A	W	B																																															
(a)	(b) $h = 0$	(c) $h = 1$	(d) $h = 2$	(e) $h = 3$																																													

Fig. 1.5: Imersion approach in watershed algorithm, from: [1]

operator (i.e. Sobel operator). If watershed transform was applied now, an oversegmentation would occur and so many lines would be drawn without resolution fulfillment. So before watershed is applied, image has to be smoothed by for example yet mentioned opening and closing.

The immersion approach seems to be the most promising method of preprocessing before constructing watershed transform. This algorithmic definition of watershed transform was given by Vincent and Soille. First, pixels in the image are sorted ascendently according to their intensity values. Starting at the minimum gray-level value, neighboring pixels with the same gray-level value are markeded as one class and pixels neighboring different values at each side are markeded as divide in each step. [1] This approach is shown in Fig. 1.5.

#### 1.5.4 Spot detection

Spot detection requires implementation of methods of object recognition. Resolving objects in an image is based on determining their common patterns - so called pattern recognition. Two main approaches stands for object recognition [12]

- decision-theoretic methods: quantitative descriptors are used for pattern description, such as surface, length, texture etc.
- structural methods: symbolic information is coding pattern description, such as label 2 and relationships among the objects <sup>2</sup> [12]

Objects are classified. Objects with common patterns are assigned to the same class (having the same label). For object recognition it is important to determine their common relationship and similarities and therefore the methods of machine learning are used, such as neuron network, decision tree, cluster analysis, fuzzy logic etc.

##### 1.5.4.1 Laplacian-Gaussian method

Laplacian-Gaussian (LoG) approaches are based on detecting zero value between two edges. Edges are computed by convolution with second-order derivative of Gaussian function (Laplacian of Gaussian function) ref. to subsection 1.5.3.2. [10], [12]

<sup>2</sup> label is the string assigned to the object for its designation



#### 1.5.4.2 Objects similarity

Objects similarity is determined by finding common patterns using their analysis in region of interest. This approach belongs to decision-theoretical approaches. Some patterns may be removed from participation in decision process by i.e. estimation or principal component analysis to lower the probability of getting false results. Every remaining patterns influence the decision with given weights. This method will be used in following bachelor work.

#### 1.5.5 Spot analysis

The most important properties for spot analysis are the following:

- area - total spot area given as multiple of number of pixels in a spot and pixel surface value,
- centroid x - horizontal coordinate of a point which is given as geometric center of spot,
- centroid y - vertical coordinate of a point which is given as geometric center of spot,
- integrated density - sum of intensity of each pixel in given spot, for purposes of this work, this value is the most important output,
- boundary - coordinates of pixels around a spot determining spot boundary.

Other features may be defined as following:

- width - double of semi-major axis of an ellipse, if spot is viewed as an ellipse,
- height - double of semi-minor axis of an ellipse, if spot is viewed as an ellipse.

#### 1.5.6 Analysis of electrophoretic map

Electrophoretic map analysis is an important step since it provides statistical data about proteome, which are then used for database mining (i.e. estimation of randomness in spot distribution mentioned further). Many methods that provide these information were developed, two of them will be described in details: SMO method and an autocovariance function method. Both methods consider spot distribution in both vertical and horizontal directions in gel as random variable determined by Poisson's distribution and a maximum between 4-6 pH and 20-60 kDa. This hypothesis was tested on map constructed experimentally from values of  $pI$  and  $\log(M_r)$  selected from SWISS-2D-PAGE databasis. Histogram plotted in both dimensions corresponded to Poisson's distribution and hypothesis was confirmed also by chi-square test (probability level of 0.5). [10]

SMO method includes computations of: total expression, degree of spot order (in range of absolutely disordered to absolutely ordered) and degree of spot overlapping. Autocovariant function method provides information about degree of spot order with aim to identify ordered clusters, so-called spot trains.[10]

##### 1.5.6.1 SMO method

Statistical Model of Peak Overlapping (SMO) is a method, which allows to compute the amount of proteins in a single spot and its order of overlapping. Image matrix is divided

into many strips. In each strip, the amount of spots is calculated as a number of so-called critical interdistance values ( $x_0$ ) which is the smallest distance by which the centers of two adjacent non-overlapping spots can be separated. This value depends on the highest physical resolution of the gel (ref. zoom gels). If two or more spots fall inside the same interdistance, they will be counted as one spot formed by two or more proteins. Total area of every spot is defined as observed area (*observed peak area*  $y_{obs}$ ). Choosing increasing  $x_0$  interdistance, a decreasing number of spots is counted yielding increasing  $y_{obs}$  values. For the simplest case, a linear relationship exists between experimental values of observed area  $y_{obs}$  and interdistance  $x_0$  and theoretical values: amount of proteins in cluster and spot mean intensity. The values of pairs  $\{y_{obs}, x_0\}$  can be used twice: once, from histogram plotted from distances of these pair values, a degree of order of structures is estimated and the pair values are interspaced by line to derive statistical estimate of amount of proteins. Computation of this value for each strip gives the total expression. Consequently, it is possible to use amount of proteins to relate an order of spots overlapping. This information can be useful for studying an influence of different experimental conditions (i.e. size of gel, dimensions of strips, acquisition system performance) in process of electrophoresis to its results. Same spots that participate in SMO estimations are processed by the mass spectrometry analysis for comparison. Thanks to its good results SMO method is considered as a strong tool. [10]

#### 1.5.6.2 Autocovariant function analysis

Correlation approaches use template of region containing required structures to find similar structures in regions inside the image [12]. The higher the value between template and compared region, the higher the value of correlation. If the template is chosen from the image itself it is called the autocorrelation. Covariance differs from correlation in fact that it excepts the mean value of image function.

2D autocovariant function (ACVF) tracks distances among spots. If equal distances occur repeatedly, the covariance value increases. Image matrix consists of gridded surface where all nodes are equally spaced, so the size of one field is given by  $N_x, N_y$  and in every field a value of autocovariant function is computed by 1.2 (from: [10]).

$$2D ACVF = \frac{1}{N_x N_y} \sum_{i=1}^{N_x-1} \sum_{j=1}^{N_y-l} (f_{i,j} - \bar{f})(f_{i+k,j+l} - \bar{f}) \quad (1.2)$$

This method allows to detect already mentioned spot trains, which can be related with some specific structures of proteins and specific changes caused by posttranslated modification. [10]

## 2 PRACTICAL PART

The aim of practical part is to apply the methods of image processing described above for spot detection and its features' measurement. Multiplicates will play a significant role to limit mentioned variability of input parameters. Designed graphical interface will provide the possibility to load more multiplicates for analysis. The measurements will be plotted to one graph to show a comparison of the results on selected images. General description of algorithm is given in Fig. 2.1.

The methods of image processing will be applied to obtain result of analysis of 2D electrophoresis using multiplicates. Consequently, the general results of work will be put together and presented and so a conclusion will be made.

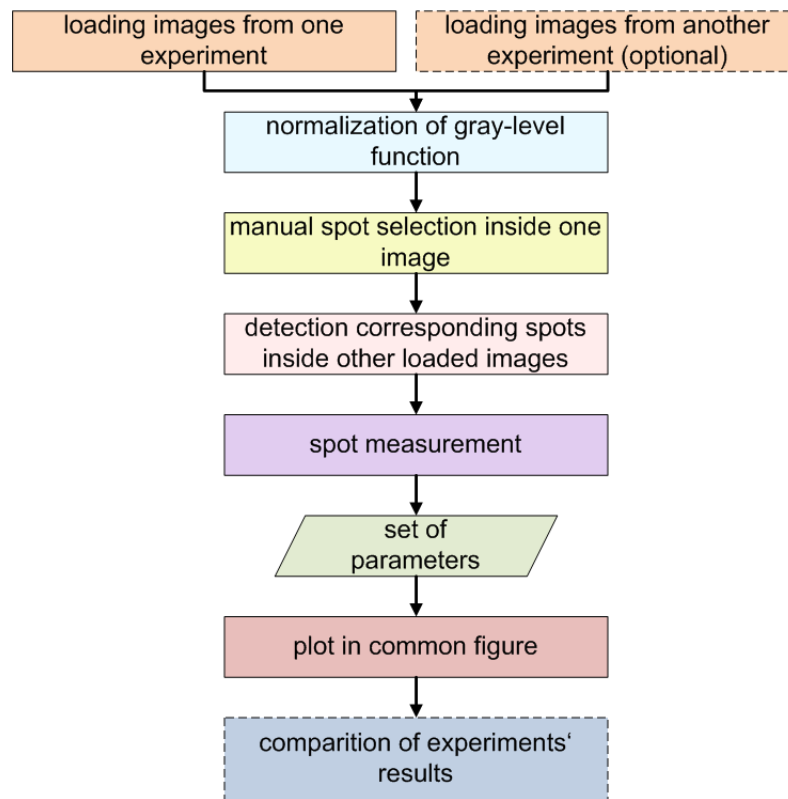


Fig. 2.1: General scheme of application's algorithm

### 2.1 Core of the program

The core of the program includes algorithms for spot detection and its measurement. In each image a small area (100x100 or 200x200 pixels) is selected as a working area where next steps (section 2.1.1, 2.1.3, 2.1.4) are performed.

### 2.1.1 Image contrast adjusting

The image intensity value should be first adjusted. Adjusting means preparation of level transform function based on mean value and variance that are calculated over the entire image. Let  $r$  denote a discrete random variable representing discrete gray-levels in the range  $[0, L - 1]$  and let  $p(r_i)$  denote the normalized histogram component corresponding to the  $i$ -th value of  $r$ .  $P(r_i)$  could be viewed as estimate of the probability of occurrence of gray level  $r_i$ . Thus mean value is defined by:

$$m = \sum_{i=0}^{L-1} (r_i \cdot p(r_i)) \quad (2.1)$$

Variance is defined by:

$$\mu = \sum_{i=0}^{L-1} ((r_i - m)^2 \cdot p(r_i)) \quad (2.2)$$

As all loaded images were built with same sample it is supposed that the structure layout inside of them will be the same. So image adjusting is applied to each image separately and resulting images will have united mean gray value. This step is crucial because the intensity value of spots in each image is computed later and compared to each other. [12]

### 2.1.2 Edge repairing

As discussed in sec. 1.4, one of the source of the output's variability is image's geometric distortion. Some of samples of gel images have geometric distortion along the edges due to inappropriate handling during image scanning. For small distortions along one dimension a reparation may be undertaken. Let us reconsider the image as a matrix of pixels. There are missing pixel values in the image along the distorted edge. The number of missing pixels varies in each row/column of the image matrix. Reparation technique uses interpolation between inner pixels to fulfill the missing pixels with the convenient values. This is a similar process to the one discussed in sec. 1.5.2. The difference is in the fact, that correct pixel values are not given from the database but they are defined by the user. For reparation, the user should select the pixels along the distorted edge and run the reparation process.

### 2.1.3 Crucial step - spots detection

As image is adjusted, its negative is stored in some variable for later use. Then the image is auto-thresholded and computation of distances (quasi-euclidean algorithm) among pixels is applied. The distance values are then represented as image intensity values, thus distance mask image is given. Let it be denoted as *mask1*. In negative image, the structures are dilatated with the disk structure element at value of intensity of 10. Then the maximas are extended to fit from the value of intensity of 20. This image contains enhanced spots, but also a high amount of information noise. Let this image be denoted as *mask2*. The

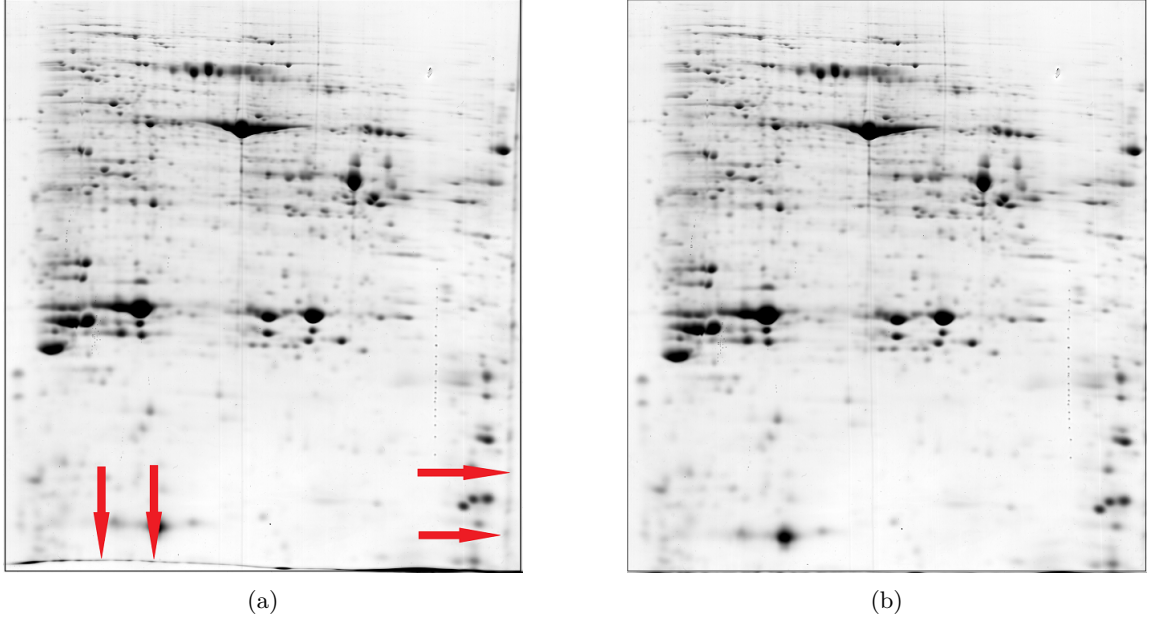


Fig. 2.2: Distorted edges in 2.2a and straightened edges in 2.2b

structures in this *mask2* are then labeled using Matlab `bwlabel` function. This function assigns a unique value to every pixel belonging to each of insulated structures in the binary image. Now, the implication will take part. If the structure in *mask1* is overlapped by the structure in *mask2*, a valid spot will be detected. *Mask1* now undergo the watershed transform, the result is an image of ridge lines. The image of the ridge lines is merged with the image of validated spot by logical OR to recognize closing or overlapping objects. This final image is labeled again.

Now the original image is inverted and a sum of intensity value is calculated from labeled objects (inverting original image removes the influence of background on the calculation of integral intensity). The result is a vector of integral intensity values of each object in the selected area.

As mentioned in section 1.5.5 other object attributes are also obtained from labeled object - an x coordinate of center of mass and y coordinate of center of mass of each object in selected area.

#### 2.1.4 Spot selection

The term spot selection in this context means choosing required spot from all detected spots in each loaded image. First, the user selects one image of loaded multiplicates and finds inside it the required spot, then clicks on it. Let  $x_{click}, y_{click}$  are the coordinates expressing where mouse motion down was registered in the image axes. In the spots' detection step, all spots inside a predefined area are found and their integrated densities and coordinates are evaluated. Let denote **IntDen** as a vector of integrated densities, **x** as a vector of x-coordinates and **y** as a vector of y-coordinates belonging to every single

spot in selected area. The aim now is to handle which of these objects is the required spot. This is done in the following step - in a decision. The decision rule is based on a simple cluster analysis, where, generally, euclidean distance of all three attributes is calculated. The calculation of n-th euclidean distance for the first image is given by:

$$edist_n = \sqrt{W_1 \cdot (x_n - x_{click})^2 + W_2 \cdot (y_n - y_{click})^2 + W_3 \cdot (IntDen_n - 0)^2} \quad (2.3)$$

where  $W_1$ ,  $W_2$ ,  $W_3$  belongs to the  $\mathbf{W}$  which is a vector of weights assigned to attributes.

The attributes having the least value of euclidean distance represent the attributes of required spot. Let the variables  $x_{ref}$ ,  $y_{ref}$  be denominated as the coordinates of required spot and  $IntDen_{ref}$  is the integrated density value of required spot. After first spot have been found, the task is to find corresponding spots in each one of other images. For the following images the calculation of euclidean distance is similar:

$$edist_n = \sqrt{W_1 \cdot (x_n - x_{ref})^2 + W_2 \cdot (y_n - y_{ref})^2 + W_3 \cdot (IntDen_n - IntDen_{ref})^2} \quad (2.4)$$

In this case the required spot is also evaluated as the one, whose attributes have the least euclidean distance value.

This method belongs to the decision-theoretic approaches because the attributes are represented by quantitative descriptors (ref. section 1.5.4)

## 2.2 Graphical user interface

A graphical user interface (GUI) is created to mediate program functions through graphical windows, components and icons. GUI facilitates the user interaction with the program, the user does not have to create a script or type commands at the command line to accomplish the tasks. Unlike coding programs to accomplish tasks, the user of GUI need not understand the details of how the tasks are performed. GUI is, for purposes of this bachelor work necessary, because the user has no possibility to run the program in MATLAB command line. [2]

The components hierarchy of MATLAB graphical user interface is well shown in Fig. 2.3. M-file that implements graphical output is assigned at least one figure window. The developer can insert one or more UI-objects (push buttons, list views, text views,...), annotation objects or axes objects inside one figure and also within one m-file, the developer can plot another independent figures. The axes are GUI objects enabling to display graphics such as graphs and images. Thus every single image is shown inside its own axes object. The axes don't use the same coordinated system as figures, therefore x-y coordinates of current point returned by axes don't match the ones returned by figure and thus mapping x-y coordinations of axes on the figure must be done using predefined conversion (function `ds2fnu`<sup>1</sup>). MATLAB GUI can be build in two ways

---

<sup>1</sup>this function is not a standart library function

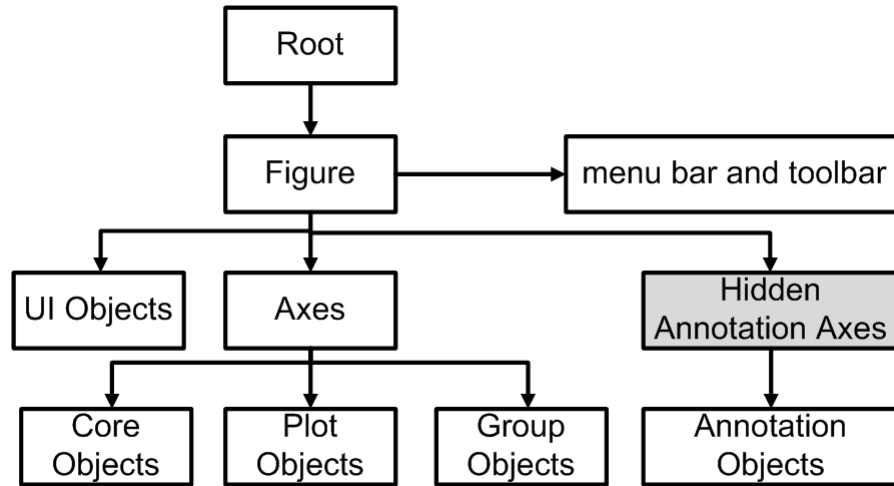


Fig. 2.3: MATLAB objects' hierarchy (redrawn partially from: [2])

- using GUIDE tool (GUI Development Enviroment)
- programming source codes, functions are used to create layout and place components explicitly
- combination of GUIDE tool for creating layout and placing the components and source codes for specifying the properties and behaviour of the GUI

For purposes of this work, every single image is plotted inside its own axes, all axes are placed in one figure. The main challenge is that all objects inside figures (and the figures themselves) must be created dynamically, because the number of images to be loaded is a priori unkown. So the second way of building GUI was used. Programming GUI also means to program all tools, which user may need and which are not automatically available in executable version (i.e. tool for deleting a group of objects by mouse drag). Developer must also assure visual interpretations of all program exceptions through popup error or warning logs.[2]

## 2.3 Building standalone application

MATLAB Compiler™ enables to share MATLAB® programs as standalone applications or shared libraries (generally called apps). The user may run the executable application at any platform using MATLAB Runtime Compiler (MCR) without MATLAB editor enviroment itself being installed.

There are two possibilities:

- compile with Lcc C and package
- compile with Microsoft Visual C/C+ and package
- package with .mlappinstall

Lcc is set as a standard compiler. MATLAB requires a compiler supporting ANSI C or C++ to be installed on the computer, where the compilation runs. Compiling is indu-

ced using `deploytool` command typed to the command window. Then developer selects which `.m` file would be the main application and selects other files to be included in package. Developer also may assign MCRInstaller, which is to be run once on user's target machine. On Windows, MCRInstaller is a self-extracting executable that installs the necessary components to run the application. On UNIX, MCRInstaller is a ZIP file. MCR is version-specific. MathWorks provides MCR tool as trial 30-days version, or full version at their web page. As MATLAB Runtime Compiler automatically includes all toolboxes (can be changed in MATLAB preferences) to the application package, including all standard functions is conditioned by having MATLAB licence. Otherwise some settings will not be available at standalone application, which influences the program functionality (as it will be described later). [14] "MATLAB does not guarantee to find every dependent file." [2]

To share a program as an application, the program doesn't have to be compiled. MATLAB app installer file `.mlappinstall` is an archive file for sharing a MATLAB GUI as an app. "A single app installer file contains everything necessary to install and run an app, including source code, supporting data information (such as a product dependencies) and the app icon." Sharing source codes is not necessary (maybe even unadvisable) for purposes of this work.[2]





## 3 RESULTS OF BACHELOR WORK

### 3.1 Final application

Program was principally designed as described at chapters 1.5.3.3 and 1.5.4.

The application starts after launching the executable file. First window, which appears is a question dialog, where the user specifies how many images are to be loaded and their paths in the system (Fig. 3.1). Then images are loaded one next to each other in one figure. Zooming at one image automatically induces zooming with the same region and scale in each other images. Spot detection is triggered by user clicking inside an axes at the requiring spot. The same spots in each other images are automatically found and all detected spots are analyzed and measured. User can use a command *View roi* from *Menu* to see how the detected spots look like after the detection algorithm execution (Fig. 3.2). That may be useful if some spot is not detected well, so the user can check if it was at least recognized by the detection algorithm. The user can also view all detected spots since the program's launch by command *Spots* from *Menu*. The spots will be labeled by a square magenta marker.

The main result is a graphical expression of spots' integrated densities, which are plotted as peaks in a common plot. The figure must adapt it's size for new values and also must have a tool for spot peak removal , adding some text label inside a plot **T** and moving the text once placed into a figure  <sup>1 2</sup>. The figure also contains the reference line which enables the user to visually compare the levels of the values in the plot. The example of how a graphical output may look like is at Fig. 3.4, where there are three proteins detected in each of three multiplicates. For purposes of this example, the seventh spot was considered as invalid, it was removed from the plot.

The user may intervene to detection algorithm in a limited way. For this purposes, the user open the control panel. The control panel (Fig. 3.3) contains the settings of:

- values of decision weigths (in range of 1-20)
- position of working area (in pixels)
- manual or automatic detection mode

All these parameters influence the detection of spots in the same way in all images except the first one. It is expected that the user clicks right at the required spot in the first selected image, so settings of parameters for this image is not needed. In automatic detection mode (check *Automatic*), the detection of the same spot executes in all images automatically. In manual mode (check *Manual*), only selected spot is detected and measured every time. At Fig. 3.3 an example of how to use the control panel is shown. The user tries to detect the appropriate spot in the image situated at the right side of the paper using the control panel. As the appropriate spot is laying higher, the working area should be held down

---

<sup>1</sup>As noted in 2.3, these possibilities are available in plot-edit mode (developing), but are not show automatically after compiling. Developer must enclose them programatically

<sup>2</sup>original icons provided by MathWorks were used because MATLAB compiler refused to compile another images as icons placed at the toolbar. Creating own icons from images is a little bit tricky.

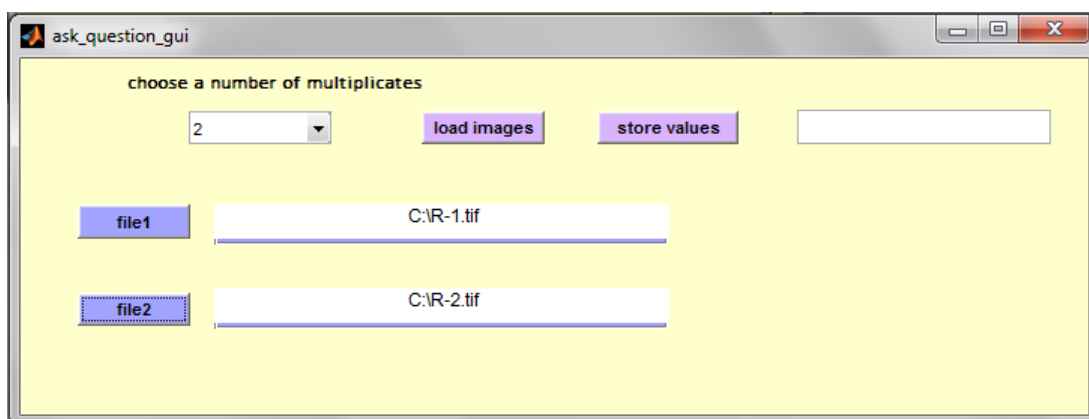


Fig. 3.1: Ask dialog

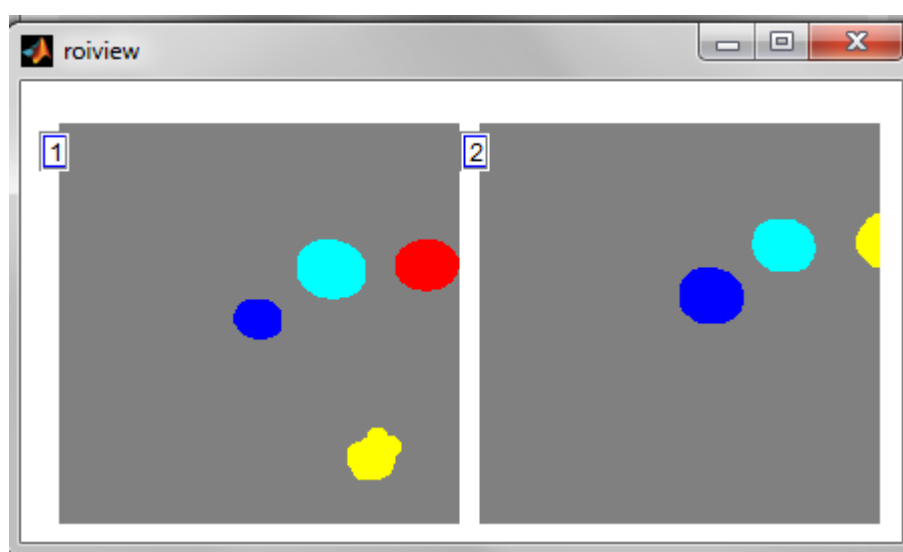


Fig. 3.2: View of the region of interest

about 20 px, and as it is closed to the other density - similar spots' x-y weights should be risen up.

As geometric distortion of a gel image was discussed in chap. 1.5.2 it is often related with negligence during scanning procedure. For purposes of straightening flexed edges, a special tool is provided by the application. The technique of edge repairing was discussed in 2.1.2. This tool must be used carefully, because it always produces a new distortion in the image.

## 3.2 Future prospects

The main part of the work has been done now but it might not be the final lap - the application may be extended.

One way how to extend this application is to enrich it's detection capabilities. Since

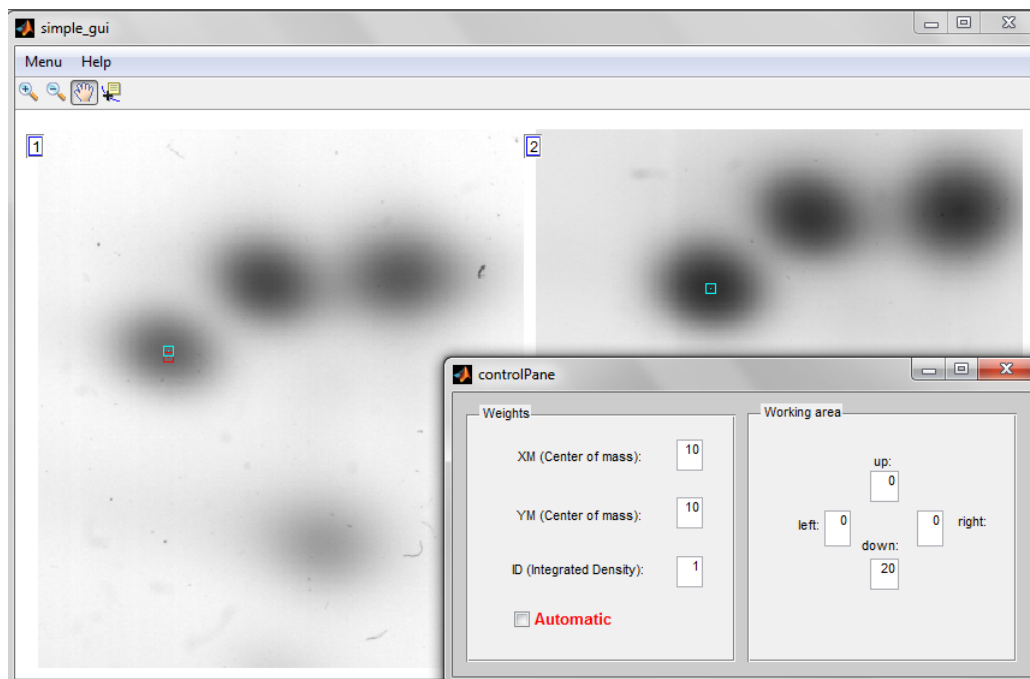


Fig. 3.3: Detected spots and the control panel with settings

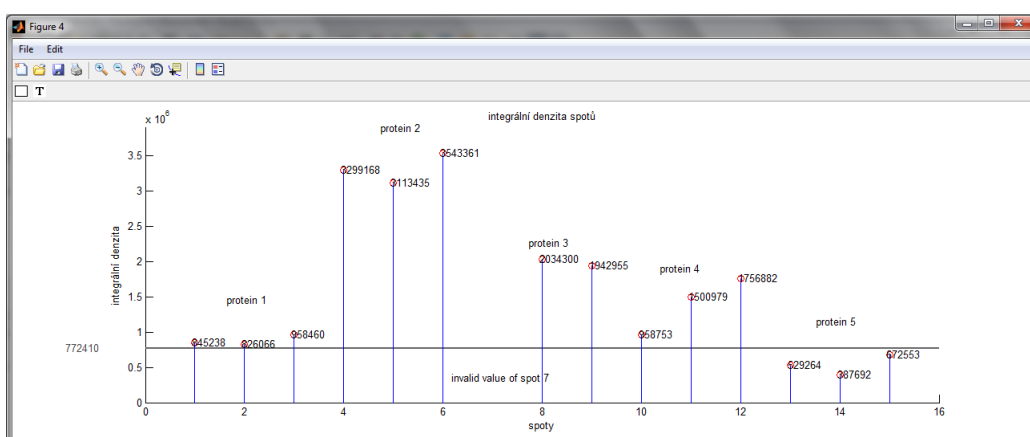


Fig. 3.4: Application's graphical output

only one algorithm has been implemented now the detection potencial is limited. These limits can be well demonstrated on the following example. If two or more weak spots are close one to another the algorithm does not resolve them. This is, in contrast to spot dynamics, an unavoidable error of the method. Also very weak spots are removed from the result view during the detection method as they are seen as a background. These and other errors may be lowered in another algorithm appended to the contemporary one. One possibility is to overwrite the same detection method with regard to the local region where the detection executes and thus parameters as local gray minima and maxima can be handled or either another different method can be used. For example, as typical approach in object detection the Hough transformation can be mentioned. In the image, Hough transform enables to find the objects that can be described by some analytic equation. Thus lines, circles or ellipses are recognized.[15] That may be useful for purposes of searching spots in the image.

Other way is to add a possibility of getting some statistical information about the detected objects. The scope of this extension would depend on user (expert) needs. As mentioned in theoretical part of this paper, the gel can be seen globally as a map of proteins and there are some statistical measurements providing additional informations about the sample. SMO method deals with total expression and autocovariant function concerns the discovering of similar patterns which is used in target research of spot trains.

## 4 CONCLUSION

The electrophoresis stays as one of the main technique providing proteome and genome analysis and it will be undoubtedly upside of the ladder for the future. One electrophoresis experiment costs plenty of effort and care. The results of electrophoresis are very important because they may help to improve medical research or to construct patient's diagnosis. With the effort to obtain precise outputs the multiplicates are made as a set of outputs sourcing from one sample. Principally multiplicates can be used to reduce the non-required changes in the spot dynamics due the variability in the electrophoresis trial or they also can be used as a tool for monitoring the changes. As the working enviroment MATLAB was chosen thanks to its sophisticated image processing tool.

The main goal of this bachelor work was to produce a program that handles a set of images (not only multiplicates) and provides spot detection and analysis of required spot in each of multiplicates following by graphical output of the analysis result. These tasks were accomplished after studying the reasons and sources of variability in previous semestral work. All these tasks have been realized and program will be offered to RNDr. Lochmanová for purposes of research.

## BIBLIOGRAPHY

- [1] J. B. T. M. ROERDIGN and A. MEIJSTER, “The Watershed Transform: Definitions, Algorithms and Parallelization Strategies,” *Fundamenta Informaticae*, pp. 187–228, 2001.
- [2] MathWorks, “Creating Graphical User Interfaces,” The MathWorks, Inc., Tech. Rep., 2014.
- [3] J. L. LÓPEZ, “Two-dimensional electrophoresis in proteome expression analysis.” *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences*, vol. 849, no. 1-2, pp. 190–202, 2007, [cit. 2.12. 2012]. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/17188947>
- [4] J. KOOLMAN and K.-H. RÖHM, *Barevný atlas biochemie*, 4th ed. Praha: Grada Publishing, a.s., 2012, [cit. 16. 12. 2013].
- [5] Amersham Biosciences, “Protein Electrophoresis,” 1999, [cit. 16. 12. 2013].
- [6] P. ŠIMAN, “Elektroforéza,” Ph.D. dissertation, Lékařská fakulta v Hradci Králové, 2013, [cit. 16. 12. 2013].
- [7] I. EDVOTEK, “Principles and Practice of Agarose Gel Electrophoresis,” The Biotechnology Education Company, Tech. Rep., 2003, [cit. 16. 12. 2013].
- [8] H. ISSAQ and T. VEENSTRA, “Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE): advances and perspectives.” *BioTechniques*, vol. 44, no. 5, pp. 697–8, 700, 2008, [cit. 2.12. 2012]. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/18474047>
- [9] A. W. DOWSEY, M. J. DUNN, and G.-Z. YANG, “The role of bioinformatics in two-dimensional gel electrophoresis.” *Proteomics*, vol. 3, no. 8, pp. 1567–96, 2003, [cit. 1.12. 2012]. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/12923783>
- [10] M. C. PIETROGRANDE, N. MARCHETTI, F. DONDI, and P. G. RIGHETTI, “Decoding 2D-PAGE complex maps: relevance to proteomics.” *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences*, vol. 833, no. 1, pp. 51–62, 2006, [cit. 2.12. 2012]. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/16448866>
- [11] A. M. BLAND, L. R. D’EUGENICO, M. A. DUGAN, M. G. JANECH, J. S. ALMEIDA, M. R. ZILE, and J. M. ARTHUR, “Comparison of Variability Associated with Sample Preparation in Two-Dimensional Gel Electrophoresis of Cardiac Tissue,” *Journal of Biomolecular techniques*, pp. 195–199. [cit. 1.12. 2012]. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2291783/>

- [12] C. R. GONZALEZ, E. R. WOODS, and L. E. STEVEN, *Digital image processing using MATLAB*, 2nd ed. Gatesmark Publishing, 2009, [cit. 5.12. 2012]. [Online]. Available: [http://www.imageprocessingplace.com/downloads\\_V3/dipum2e\\_downloads/dipum2e\\_sample\\_book\\_material\\_downloads/DIPUM2E\\_Chapter02.Pgs\\_13-50.pdf](http://www.imageprocessingplace.com/downloads_V3/dipum2e_downloads/dipum2e_sample_book_material_downloads/DIPUM2E_Chapter02.Pgs_13-50.pdf)
- [13] V. HLAVÁČ and M. SEDLÁČEK, “Zpracování signálu a obrazu Pracovní verze skriptu v tisku pro studenty FEL ČVUT,” Ph.D. dissertation, České vysoké učení technické v Praze, 1999, [cit. 13. 12. 2013].
- [14] MathWorks, “MATLAB Compiler,” 2013.
- [15] J. VLACH, “Hledání úseček a kružnic s využitím Houghovy transformace při zpracování obrazů v LabView,” *snímače a měřicí technika*, pp. 42–44, 2011.
- [16] S. EDDINS and MathWorks, “The Watershed Transform: Strategies for Image Segmentation,” The MathWorks, Inc., Tech. Rep., 2002, [cit. 29.11. 2012]. [Online]. Available: <http://www.mathworks.com/company/newsletters/articles/the-watershed-transform-strategies-for-image-segmentation.html>

## LIST OF SYMBOLS, PHYSICAL CONSTANTS AND ABBREVIATIONS

$pI$	value fo isoelectric point
$M_r$	molecular mass
$f_{i,j}$	intensity of pixel at coordinates i,j
$\bar{f}$	average intenzity of pixel
$N_x, N_y$	size of surfaces of subregions in direction of x and y coordinates
$k, l$	values in range of $pI$ and $\log(M_r)$
$f, g$	transformation function
$T$	treshold
$x_0$	minimal value of intercritical value between two spots
$y_{obs}$	area of one cluster (consisting of one or more proteins)
$[0, L - 1]$	range of discrete gray levels
$r$	gray level from range $[0, L - 1]$
$p(r_i)$	$i - th$ value of $r$
$m$	mean gray value
$\mu$	variance of mean gray value
$mask1$	image where intensity values are represented by distance values
$mask2$	image's negative containing enhanced spots
$x, y, x_{click}, y_{click}$	coordinates of mouse click inside the axis
<b>IntDen</b>	vector of integrated densities over all spots in selected image
<b>x</b>	vector of x-coordinates over all spots in selected image
<b>y</b>	vector of y-coordinates over all spots in selected image
$W_1$	weight of 1.st attribute in decision (x-coordinate)
$W_2$	weight of 2.nd attribute in decision (y-coordinate)
$W_3$	weight of 3.th attribute in decision (integrated density)
<b>W</b>	vector of weights in decision rule



$x_n$       x-coordinate of n-th spot in selected image  
 $y_n$       y-coordinate of n-th spot in selected image  
 $IntDen_n$  integrated density value of n-th spot in selected image  
 $edist_n$  n-th euclidean distance value between attributies of spot manually selected and  
                  spot selected by decision algorithm  
 $x_{ref}, y_{ref}$  coordinates of required spot in the manually selected image  
 $IntDen_{ref}$  integrated density value of required spot in manually selected image

## 5 APPENDIX

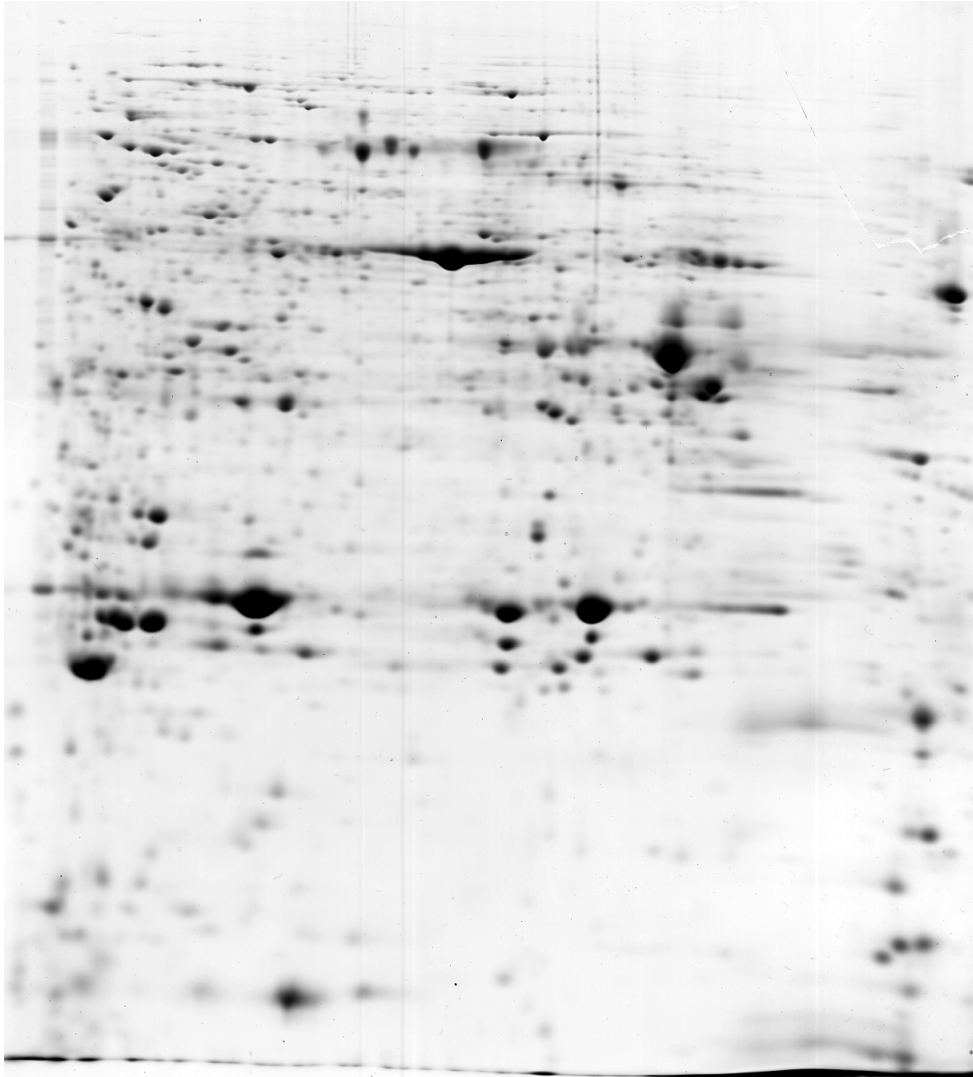


Fig. 5.1: Electrophoretic gel, black spots represent clusters of proteins

### 5.1 The content of DVD

- complete version of bachelor thesis in pdf format,
- complete version of program including: standard program launcher as an executable file (S2DESA.exe), package file of project (S2DESA\_pkg.exe - run only once), MCRInstaller (run only once), readme.txt,
- source codes of program packaged in .zip file.