Morten Beck Rye

Image segmentation and multivariate analysis in two-dimensional gel electrophoresis

Thesis for the degree philosophiae doctor

Trondheim, November 2007

Norwegian University of Science and Technology
Faculty of Natural Sciences and Technology
Department of Chemistry

NTNU

NTNU
Norwegian University of Science and Technology

Thesis for the degree philosophiae doctor

Faculty of Natural Sciences and Technology
Department of Chemistry

ISBN 978-82-471-4400-8 (printed version)
SBN 978-82-471-4414-5 (electronic version)
ISSN 1503-8181

Doctoral theses at NTNU, 2007:200

Printed by NTNU-trykk

This thesis is dedicated to my girlfriend, Cecilia Larsen

Acknowledgments

First I would like to thank my supervisor, Bjørn K. Alsberg for funding, ideas, valuable advice and experienced evaluation of the work in this thesis. I would then like to thank my girlfriend, family, friends and colleagues for support and encouragement during my years of work, which has not always been smooth and easy. A special thanks to my colleagues Lars Gidskehaug, Arnar Flatberg and Bård Buttingsrud for valuable advice, discussion, and assistance on practical problems. Einar Ryeng is thanked for assistance on computer related problems, and all of the four above for providing the perfect company on the various conferences we have attended during this period. Also a very special thanks to Terje Bruvoll for kindly helping out with administrative issues, extra funding and necessary equipment when needed. The people at MATFORSK, Ås, especially Ellen Mosleth Færgestad, Harald Grove, Harald Martens and Xiaohong Jia, also deserves special thanks for fruitful cooperation and providing the data-material used throughout this work. I would also like to thank Beata Walczak in Katowice, Poland, for the cooperation, and for kindly providing one of the figures used in the thesis. I would finally like to thank various reviewers for valuable comments and suggested improvements on several of the presented manuscripts.

Summary

The topic of this thesis is data-analysis on images from two-dimensional electrophoretic gels. Because of the complexity of these images, there are numerous steps and approaches to such an analysis, and no “golden standard” has yet been established on how to produce the desired output. In this thesis focus is put on two essential fields concerning 2D-gel analysis; registration of images by segmentation and protein spot identification, and data-analysis on the output of such a registration by multivariate methods. Image segmentation is mainly concerned with the task of identifying individual protein spots in a gel-image. This has generally been the natural starting point of all methods and procedures developed since the introduction of 2D-gels in the mid-seventies, simply because this best reproduces the results created by a human analyst, who manually identify protein-spot entities. The amount of data produced in a 2D-gel experiment can be quite large, especially in multiple gels where the human analyst is dependent on additional statistical data-analytical tools to produce results. Because of the correlated nature of most gel-data, analysis by multivariate methods is natural choice, and are therefore adopted in this thesis. The goal of this thesis is to introduce the above mentioned procedures at different stages in the analysis pipeline where they are not yet fully exploited, rather than to improve already existing algorithms. In this way new insight and ideas on how to handle data from 2D-gel experiments are achieved. The thesis starts with a review of segmentation methodology, and introduces a selected procedure used to identify protein spots throughout. Output from the segmentation is then used to create a multivariate spot-filtering model, which aims to separate protein spots from noise and artifacts often creating problems in 2D-gel analysis. Lately the use of common spot boundaries in multiple gels have been the method of choice when gels are analysed. How such boundaries should be defined is an important subject of discussion, and thus a new method for defining common boundaries based on the individual segmentation of each gel is introduced. Segmentation may be a natural starting point when gels are analysed, but it is not necessarily the most correct. Often the introduction of fixed spot entities introduces restrictions to the data which cause problems at later stages in the analysis. Analysing pixels from multiple gels directly has no such restrictions, and it is shown in this thesis that the output of such an analysis based on multivariate methods can produce very useful results. It can also give insight to the data problematic to achieve with the spot boundary approach. At last in the thesis an improved pixel-based approach is introduced, where a less restricted segmentation is used to reduce and concentrate the amount of data analysed, improving the final output.

Contents

Chapter 1

Introduction

1.1 Background

The introduction of two-dimensional gel electrophoresis (2-DE) in the mid-seventies was a major change in the field of protein research. Before this, protein analysis was limited to the purification and investigation of, at best, a handful of proteins at a time. The connection of these few proteins to the larger biological system was ignored, simply because the means to study these relationships were not available. The emergence of two-dimensional protein-separations suddenly made it possible to look at thousands of proteins at the same time. Visions for the new methodology were soon put forth. One would now be able to estimate the number of proteins, made by any biological system, and describe the molecular anatomy of this system on the level of protein expression and modification patterns. Physiological states of whole cells could be monitored, and medical diagnostics could be performed. Because of the large amount of data produced from the 2-DE experiments, it was soon recognised that users of 2-DE needed aid of analytical tools to perform an efficient analysis, and in the late seventies the first computer based analysis systems for 2D-gels emerged. The creators of these systems soon became aware of the numerous challenges inherited in the complex protein patterns produced, many of which, after 30 years of software development, still remains unsolved. 2-DE is today recognised as a valuable tool in protein research, but the task of creating a fully automated and reliable computer software for identification, quantification and comparison of proteins in 2D-gels is still far ahead. Most researchers today recognise the limitation in both the gel methodology itself, and the software-based methods used to analyse them. Recent studies have also questioned whether the task of identifying and quantifying all proteins in a cell sample can be solved by 2-DE alone. Nevertheless, software solutions have been developed which make the life of a 2-DE analyst easier. These solutions are not fully automatic, but rather have the focus of improving visualisation and collection of data from the gels, helping the analyst to recognise the significant variations in the data, and make correct decisions based on these data. Such an approach is also the starting point for this thesis, where the overall task is to introduce new methods to improve the analysis of data from 2-DE.

1.2 Chapter outline

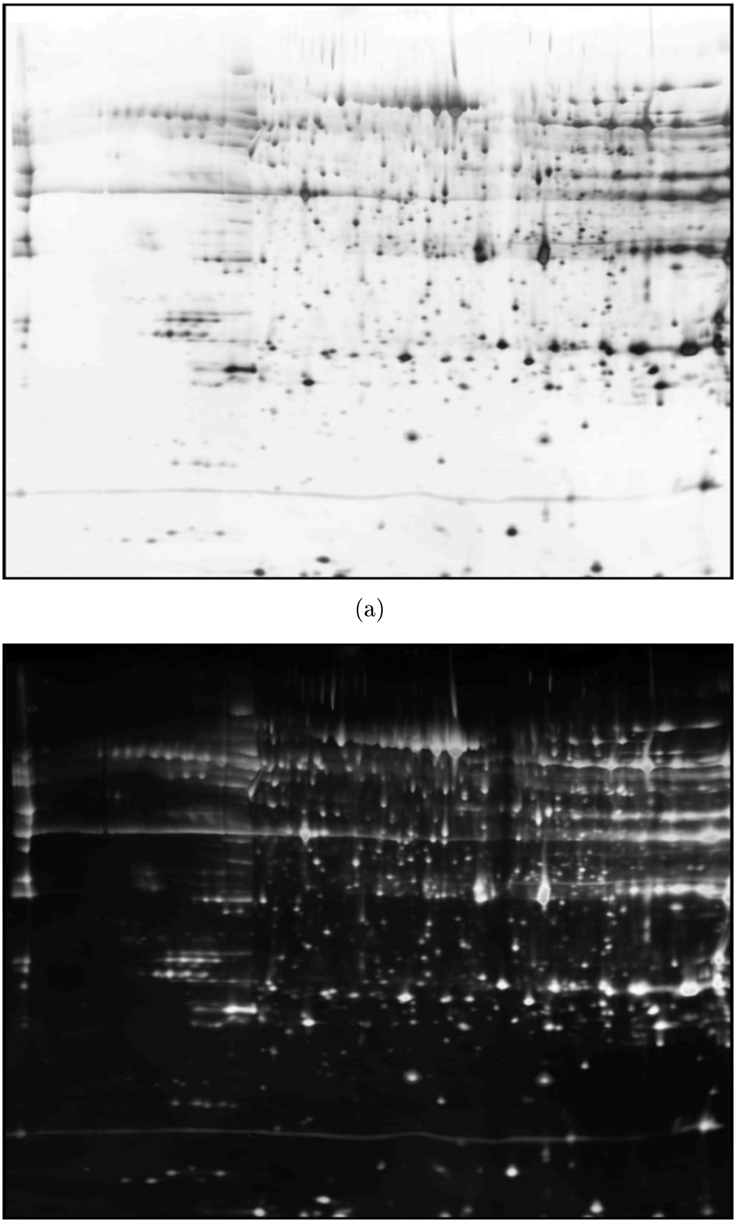
There are many steps involved in the analysis pipeline of 2D-gels, where each step has several suggested solutions. To develop and improve solutions to all these steps would be too much a task for this thesis. Thus a focus is put on one crucial step in the analysis pipeline: the segmentation of 2-DE images. Segmentation in 2-DE is concerned with identifying and separating proteins in a gel-image, and provides a fundamental basis for data representation and the final analysis. Another topic of this study is the use of multivariate methods in relation to the output of the image segmentation. The nature of data in 2-DE, where many proteins are correlated, makes analysis by multivariate approaches a natural choice. The contents of this thesis will thus include image segmentation in 2-DE, multivariate analysis of 2D-gels, and how multivariate analysis can be used in combination with output from the segmentation step to improve the overall analysis. The outline of the thesis is as follows: the introduction chapter describes the general methodology of 2-DE, automatic analysis of data from 2-DE, and comments on the current status of method development in the scientific community today in relation to commercial software. Focus is then put on the image segmentation part of the analysis. The most common segmentation methods in 2-DE are reviewed, and a selected segmentation pipeline used throughout the rest of this thesis is presented. Finally in the introduction common multivariate methods are described briefly. Chapter 2-5 contains the main work of this thesis, and each chapter is presented as a scientific article. A short summary of these four chapters will be given at the end of the introduction.

1.3 Creating gel images by 2-DE

Two dimensional polyacrylamide gel electrophoresis (2-D PAGE) is a method introduced by O'Farrel [1] in 1975. The method described a new way to separate proteins from biological samples in two independent dimensions, and is still the method of choice for the majority of differential protein expression studies. The basic principle for the technology is to separate the proteins in each dimension using two independent high resolution properties: isoelectric point (pI) and molecular mass (M_r). Before separation all proteins must be completely solubilized to break the interaction between proteins and to remove non-protein components. After this, separation by isoelectric point is carried out by a procedure called isoelectric focusing (IEF). A pH gradient is applied, and the proteins are allowed to migrate in the first dimension until their net charge is zero. The pH at which a protein has zero net charge is called the isoelectric point or pI-value. Since proteins have different charges, they will occupy different locations in the isoelectric dimension after the isoelectric focusing. However, several proteins also have similar pI-value, so the separation on isoelectric focusing alone is not sufficient to identify individual proteins in a sample under investigation. Thus a second orthogonal separation is applied, based on each proteins molecular mass. In the second dimension protein migration is caused by applying a second electric field in the presence of sodium dodecyl sulfate (SDS). When SDS is present, the electric field will cause the proteins to migrate to positions in the second dimension proportional to their molecular mass. Because few proteins have both identical pI and M_r value, it is possible to create a two dimensional map, where most individual proteins are located with unique (pi, M_r) coordinates. It should thus be possible to identify most proteins in a biological sample by the described procedure.

It is, however, not sufficient that the proteins are separated. They must also be made visible for analysis, either by a human analyst or a computer. Staining procedures are applied to make the proteins contained in a gel visible for the human eye. Common staining methods are silver staining, fluorescent staining and Coomassie Blue. Radioactive labelling has also been used, especially in the early years of 2-DE. The most widely used method is silver staining, and is considered the “golden standard” for 2D-gels. All gels used in this thesis were silver stained. After the staining process, the proteins in the gel will appear as dark spots on a transparent surface. A single gel can include a few hundred up to several thousands of spots depending on the sample analysed. In theory each spot is supposed to represent one isolated protein specie. When the gel-surface is viewed as a three-dimensional landscape, most protein spots will resemble Gaussian shaped depressions. The reason for the Gaussian shape, is that proteins are subject to a diffusion process during migration [2, 3].

After the creation of the original gels they are scanned to produce gel images. When scanned, the transparent part of the gel-surface becomes white, and the



(a)

(b)

Figure 1.1: (a): Gel image showing black protein spots on a white surface. (b): Inverted image, where spots are white and the background dark. The displayed image is one of the experimental gels used in chapter 3-5.

proteins appear as black spots on the white surface, as shown in figure 1.1(a). In gel-images the colours of the original gel are transferred to pixel intensities, also called greyvalues or greytones. Greytone images (that is, each pixel is associated with a single value, where black equals zero and white equals the maximum value depending on scale and depth) are usually sufficient to describe the intensity variations in the gel. Thus protein spots appear in the gel images as areas having pixel intensities approaching zero, while the background consist of pixels with intensities close to the maximum value. It is often convenient to invert the images, so the spots appears as peaks in a landscape, while the background appears black with intensities close to zero. In this thesis the latter image representation is preferred, meaning gel images will appear as a landscape with protein spot peaks, unless stated otherwise. Both representations are displayed in figure 1.1.

Different scanners produce different image resolutions. The gel images used in this thesis have a resolution between 1500 and 2000 pixels in each direction, and an 8 bit image depth, the latter meaning that each pixel can take an intensity between 0 and 255. Several scanners may produce images with both higher resolution and depth (12 or 16 bit), however, the current resolution and depth used for gel images in this thesis were sufficient for our purposes. Images of the size and depth presented are within the usual standards of gel images generally used in 2-DE research. The transition to higher resolution and depth, will also increase the computational time for analysis considerably, while the gain in knowledge and conclusions are limited.

1.4 Analysing gel images from 2-DE

1.4.1 Aim of the gel-analysis

The goal of 2-DE is primarily to identify differentially expressed proteins. This means that a 2-DE experiment usually includes several (multiple) gels. Motivation for such experiments is often to investigate differences in protein expression when comparing biological samples. For instance, such comparisons can be made between cell samples from a healthy and sick patient to identify proteins which are active during a disease, or several gels can be run on cell samples undergoing some change or participating in a cell-cycle to investigate which proteins that change in expression during the different stages. It thus becomes important to discover proteins spots that have disappeared, appeared or increased/decreased in size and intensity between the gels. The analysis and identification of these differentially expressed proteins are usually performed by a human expert, comparing spots on different gels by visual inspection with the aid of computer software.

1.4.2 Automation

The manual analysis and inspection of multiple gels are both time consuming and subject to ambiguities. Identifying thousands of spots, finding matches to each spot in several gels, and decide whether each spots intensity change significantly is a tedious procedure when done manually. Results achieved by a human observer are also difficult to reproduce because of the subjectivity introduced. Results from Prehm et al. [4] show that human analysts only agree on 60-90% of the spots identified on a single gel (depending on the complexity of the spot pattern), indicating that this source of error is significant. Keeping track of comparable spots in several gels are also difficult for a human observer, especially when the number if gels become large.

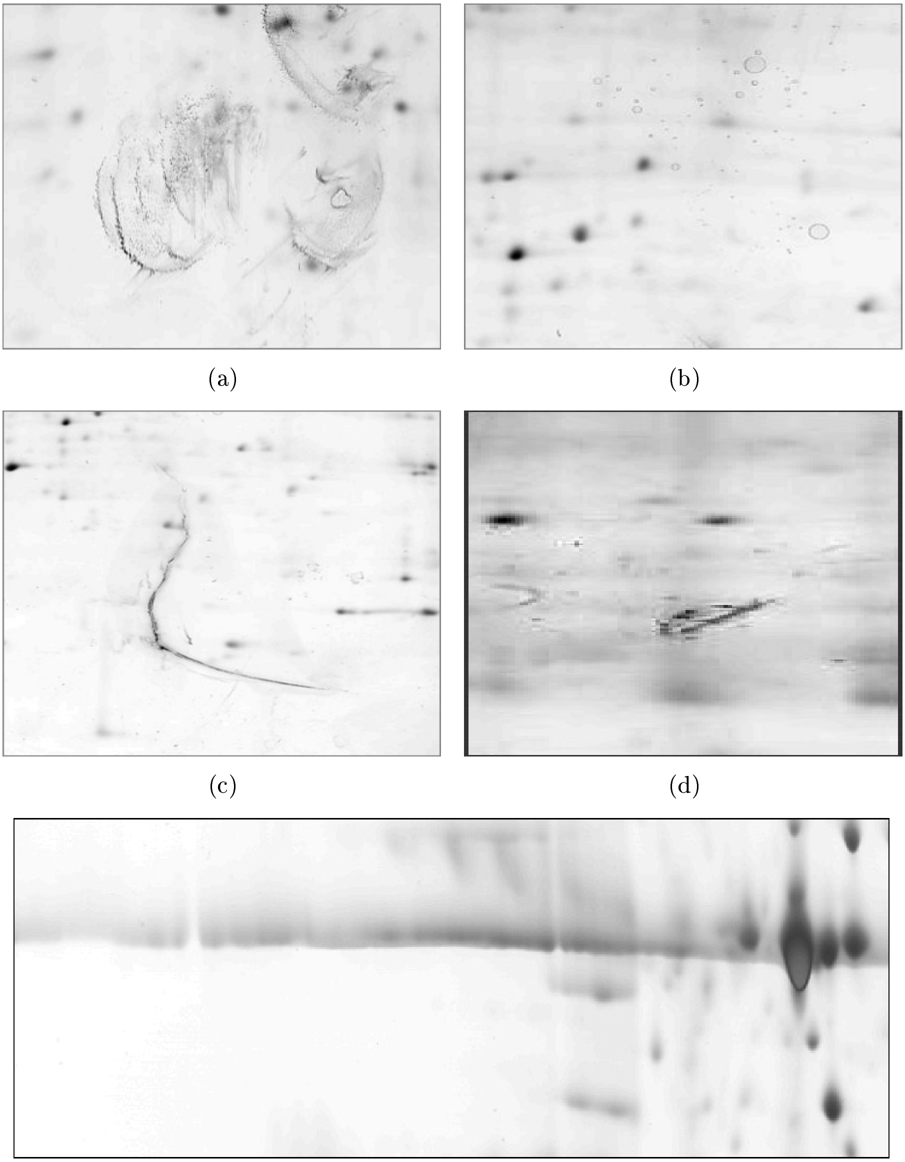
It would be of great advantage if an automatic standard procedure could produce reliable results from a 2-DE experiment. The benefits on both speed, reproducibility, and applicability would be considerable, and in the first years after 2D-gels appeared in 1975, several hardware and software solutions trying to automatise the gel analysis procedure were published [5–12]. These included both single standing methods and more user friendly program packages. Despite the effort of finding a reliable automation procedure for analysing gels from 2-DE, several challenges encountered in these early reports still remains unsolved today. The most important ones are listed in the following section, and possible solutions to several of these challenges are suggested in the articles presented in chapter 2-5 in this thesis.

1.4.3 Challenges for automatic analysis

Reproducibility is an important issue in 2-DE. Two gels run from the same sample should ideally produce two identical gels, with spots located at identical positions with equal size and intensity. Unfortunately, in reality this is far from the case. Due to geometrical distortions the position and shape of the spots are shifted between the gels subject to analysis. These distortion can be applied both globally for the whole gel, and locally in specific regions. Dowsey et al. [13] lists four main factors contributing to these distortion: Differences in the structure of the media (the polyacrylamide net), characteristics of the transporting solute (SDS), the solvent conditions and variations in the electric field applied. The last factor is well described by the *current leakage model* published by Gustafsson et al. [14]. The above factors explains variations within the gel material itself and the chemical composition of the material used in creating the gel. In addition variations are found depending on which laboratory the gels are run in, the equipment and procedure used in this laboratory and by which laboratory assistants that conducted the gel creation procedure. At last we also have biological variations among the samples. Two cell samples from the same biological tissue do not necessarily produce identical spots in the gel. A thorough discussion on the reproducibility and the error sources present when creating 2D-gels are beyond the scope of this thesis, and a good overview can be found in [15]. Nevertheless, all of the above factors mean that great care must be taken when designing experiments in 2-DE to create reliable and conclusive results.

Other disturbing factors influencing the results from automatic 2-DE analysis are the presence of noise, dust particles, fingerprints, cracks in the gel surface and other artifacts not related to protein content. A human observer may in most cases distinguish these artifacts from true protein spots, but to make a computer perform the same separation is not trivial. It is often observed that these unwanted artifacts are treated as proteins in the automatic analysis, thus corrupting the final output. In addition proteins themselves can be sources of artifacts, such as streaks and tails. Examples of common artifacts are shown in figure 1.2. Streaks and tails are well handled by most analysis software. A multivariate model for filtering other artifacts are developed in chapter 2. Background intensity is another problem encountered in 2-DE. The background intensity is not uniform, and usually increases with the density of spots in a local area. This means that a simple threshold value is not sufficient to separate pixels belonging to protein spots from background pixels, and more sophisticated methods for background correction must be applied. The effect of applying a single threshold to a region with large background variation is shown in figure 1.3.

A different but serious challenge is the large difference in expression of proteins analysed by 2-DE. The dynamic range between the smallest and the largest concentrations of individual proteins can be up to 10^6 for cells and tissues and as much as 10^{12} in body fluids, as stated by [13, 16]. A normal image of a silver stained gel has a maximum dynamic range of only 10^4, which obviously creates problems when proteins are to be identified and their concentrations calculated. At the lower end of the dynamic range, many proteins are hardly visible in the gel because of the small amount present in the sample. Extremely sensitive algorithms are necessary to identify areas where these proteins are present, and at the same time distinguish the signals caused by these faint protein spots



(a)

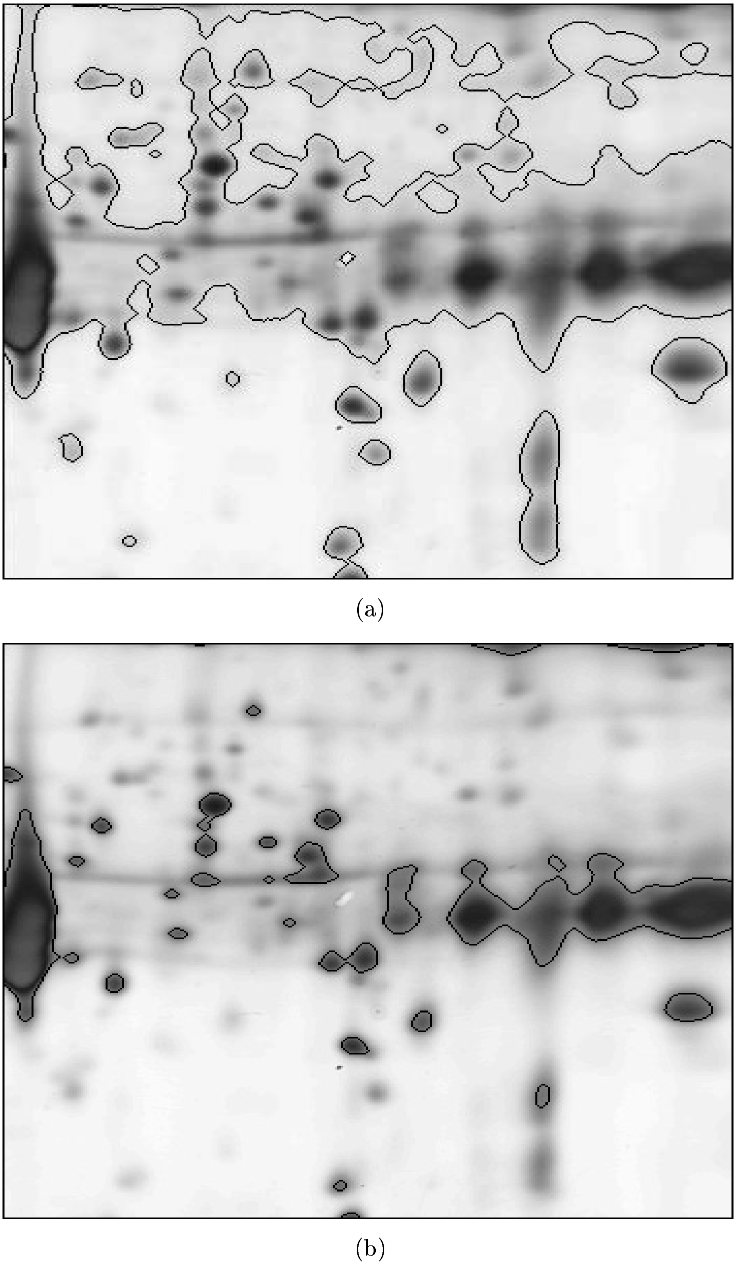
(b)

(c)

(d)

(e)

Figure 1.2: Examples of disturbing factors influencing the results from 2-DE analysis. (a): Fingerprints. (b): Dust and other small pollutions. (c). Crack in the gel-surface. (d): Artifact not resulting from proteins. (e): Horizontally directed streak.



(a)

(b)

Figure 1.3: Effect of applying two different single thresholds to a gel-image with non-uniform image background. (a): Threshold at 0.2. Spots in the region with low background intensity are identified, but no separation is possible in the region with high background intensity (b): Threshold at 0.5. Proteins in the region with high background intensity are separated, but spots with lower intensities are not detected. Pixel intensities are or on a 0-1 scale.

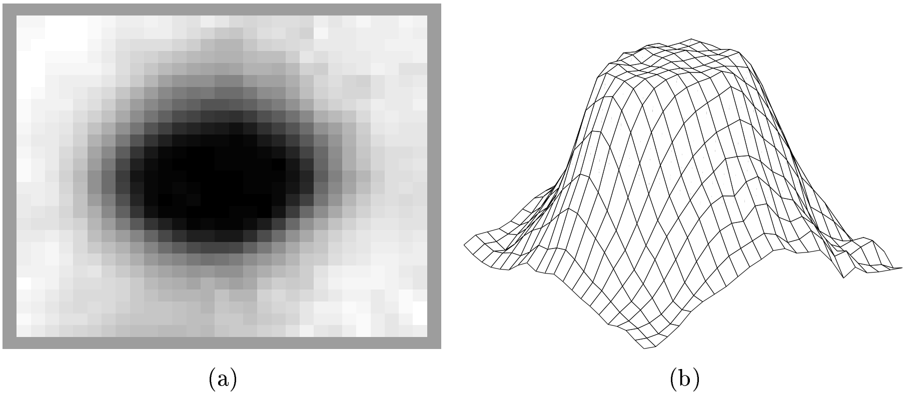


Figure 1.4: Saturated spot with a “flat” uniform intensity surface. (a): Intensity image. (b): Landscape representation.

from noise. At the other end of the dynamic range, there is the problem with protein spot saturation. The image scanning device is capable of capturing colour differences in the protein spots up to a certain point, where it can no longer distinguish the darkest colours. This point becomes the saturation value in the gel image, and is usually the maximum intensity value. Saturated protein spots are characterised by a flat surface around the spot-centre with uniform intensity equal to the maximum image greyvalue. All spots should ideally display a Gaussian shape, however, because of the high concentration of these proteins the image scanner are unable to detect the Gaussian peak, and it is replaced by a uniform intensity surface as shown in figure 1.4. It follows from this that it is impossible to determine the exact concentration of all proteins from the spot intensity measures. It is thus more beneficial to look for relative differences in protein spot intensities, or the presence/absence of spots when analysing gels from 2-DE.

It should be noted that the direct pixel intensity conversions made by a scanner are not linearly proportional to the protein concentrations. To achieve a scale that is reflects the concentration (or absorbanse) of each protein, a logarithmic or similar transformation of the image is necessary. But even with a logarithmic transformation, the relationship between concentration and intensity is rarely linear. One way to identify this relationship is to include reference proteins with known concentration differences, and make adjustments according to the changes in the spot intensities for these proteins. However, because of the presented difficulties, the aim of a gel-analysis is often to identify relative changes and presence/absence of protein spots rather than estimating the exact

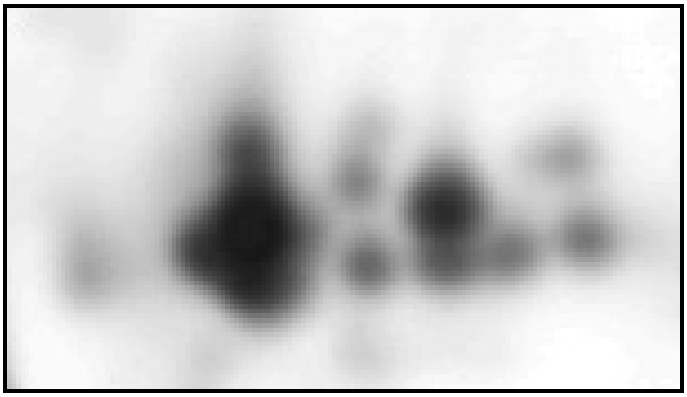


Figure 1.5: Complex spot region with several partly overlapping protein spots.

Table 1.1: Summary of results from Campostrini et al. [17]. Spot singlets are by far the rarest specie when the total number of spots reach a few hundred.

Number of spots	Number of singlets
74	71%
639	32%
1500	27%
3000	10%

concentrations.

The large amount of spots present in a gel means that different proteins have a tendency to occupy the same area on the gel surface. This leads to so-called complex regions, where several partly overlapping spots create spot-clusters which are very difficult to analyse. An example of such an area i shown in figure 1.5. To identify individual protein boundaries in such regions is a major challenge for automated spot identification software. Combined with the poor reproducibility of 2D-gels, the consequences are that multiple gels may show large variations in the number of identified spots and their boundaries in the complex regions. This introduces large errors when spots are compared automatically, with wrong estimates of both the total intensity of a spot and the absence/presence of spots. Such errors have a large effect on the final statistical analysis, and is described more thoroughly in chapter 4. Some ways to circumvent the problem are topics of this thesis, and are presented in chapter 4 and 5.

Later studies [17, 18] have suggested that the problem regarding multiple spots (spots consisting of more than one protein) is substantial. Results from one of these studies are summarised in table 1.1. Table 1.1 shows that treating single isolated proteins spots as consisting of only one protein is incorrect in most cases. Most spots actually consist of two or more proteins, even if they appear as singlets. When this is the case, 2-DE alone is not sufficient to identify all proteins in a sample, and additional analysis tools are needed. One possibility is to use Mass Spectrometry (MS) on each spot to resolve multiple proteins. MS has been widely used for protein identification in the last years [16, 19]. These latter findings also suggests that the methods introduced in chapter 4 and 5, where gels are analysed without making assumptions on isolated spots, are useful alternatives to the usual spot-volume approaches.

1.4.4 Commercial software

Several commercial software packages dealing with the automatic analysis of gels from 2-DE exist on the market today. The ones found available pr. May 2007 are listed in table 1.2. The table is based on an overview published by Marengo et al. [20] in 2005. The field of 2D-gel analysis today is greatly dominated by commercial software packages. The main reasons for this is that many of the methods used in these packages outperform methods published in the literature, and that the whole analysis can be performed in a user-friendly environment. Though the existence of such program packages has made life easier for the general gel analyst who wants to identify the important proteins in his or her experiment, it is the authors view that the domination of these packages has some unfortunate consequences. Especially this is true in the research field of method development. Since the software packages are commercial, the included methods, routines and algorithms used to identify, compare and analyse the 2D-gels are hidden from the user, and is thus not attainable by the general researcher. This makes the development of a golden standard for analysis very difficult. Since each program package uses its own self-made pipeline of methods, with its own defined set of parameters for these methods, the output of the analysis may differ significantly according to the selected package [21]. Several studies have compared software packages [22–26] with the aim to investigate which ones that perform best under different conditions. Though such studies are of interest from a user point-of-view, the black box approaches these software packages offer make it impossible for a researcher to inspect in detail the methods used and to reproduce the results on his/her own computer. Thus a systematic comparison of which methods that perform best in the different stages of the analysis procedure has yet to be

Table 1.2: List of commercial software packages available per May 2007.

Name	Company
Delta 2D	DECODON
GELLAB II+	Scanalytics
Melanie/ImageMaster 2D Platinum	GeneBio/GE Healthcare
Progenesis SameSpots	Nonlinear Dynamics
Alpha GelFox 2D	Alpha Innotech

established. A consequence of the diversity of approaches is that the variation caused by the subjectivity factor among human analysts, is now replaced by a just as serious ambiguity caused by the variation among software packages. In spite of the existing differences, there are still some recognisable standard steps in an automatic analysis pipeline. These steps are described more thoroughly in the following chapter. A flow chart of a standard analysis pipeline is given in figure 1.6.

1.4.5 Automated gel analysis pipeline

Pre-processing of the gel images is a natural first step in the analysis pipeline. By pre-processing is here meant procedures that improve the general quality of the image, with respect to enhancing the separation between protein spots and other parts of the image. The most important step in the pre-processing stage is usually noise removal and, if several gels are compared, normalisation of intensity values over all gels. For general noise removal, smoothing masks are applied to the image. These masks may represent both Gaussian, Polynomial and Adaptive smoothing [27], and the degree of smoothing is adjusted by the size of the smoothing mask or the number of successive smoothing iterations. Lately more sophisticated denoising methods based on wavelets have also been used to remove noise in gel images [28]. Spikes is a special type of single noise-pixels with an extremely high intensity value compared to its neighbouring pixels which is often seen in 2D-gel images. Such spikes are removed with a median filter. How accurate the denoising procedure need to be to produce an satisfactory image segmentation is dependent on the choice of segmentation procedure.

If measuring protein concentrations is the goal of the analysis, a logarithmic transformation is necessary (section 1.3.3), which is also regarded as a pre-processing step.

When an experiment consist of several gels, the intensity values have to be adjusted on each gel in order for the gels to be comparable. The usual way to correct the intensity values is to adjust for the total amount of staining colour added to the image, which is regarded as identical to the total intensity of the image (the sum of all pixel intensities). Normalisation is performed by dividing each pixel intensity by the total image intensity. Normalising has the same effect as calculating relative volumes and intensities in the data acquisition and quantification step described later in this section.

The next step after pre-processing is image alignment. The goal of the alignment step is primarily to adjust for the geometric distortions described in section 1.3.3, and is performed by warping and deforming the images in such a way that each pixel at a specific position in one image, is comparable to all pixels at an identical position in all other images. There are several routines, both in commercial software and published in the literature that performs image alignment [14, 27, 29–35]. Alignment procedures usually make use of some pre-defined landmark spots which are used as anchors for an optimisation. The optimisation problem is often formulated as minimising a correlation function based on the corresponding pixel intensity differences between all images. Not all reports perform image aligning before the rest of the steps in the analysis pipeline. Sometimes the aligning is incorporated into the matching procedure described later. In these cases the optimisation is based on pattern-similarities between sets of protein spots, rather than the original greytone correspondence in the image. There are in general several variations to both landmark selecting and alignment optimisation reported. It is the authors opinion, especially after observing alignment results in commercial software, that satisfactory alignments are produced. This step is thus not considered the bottleneck in the analysis pipeline. Image alignments in this thesis are performed using the commercial software SameSpots (Nonlinear Dynamics). Typical results from the aligning step are illustrated in figure 1.7. It should also be mentioned here that alignment should be performed after the normalisation step, because warping of images my influence the total intensity in an image.

After the alignment, identification of isolated protein spots on each gel is carried out. This procedure is also often referred to as segmentation. Segmentation of gel images is the foundation of much work in this thesis, and a lot of effort was put into creating a reliable segmentation procedure. The details and the development of this segmentation procedure are described in section 1.5. For now it is sufficient to say that the spot identification step produces protein spots, each with a corresponding spot boundary. Each image segment

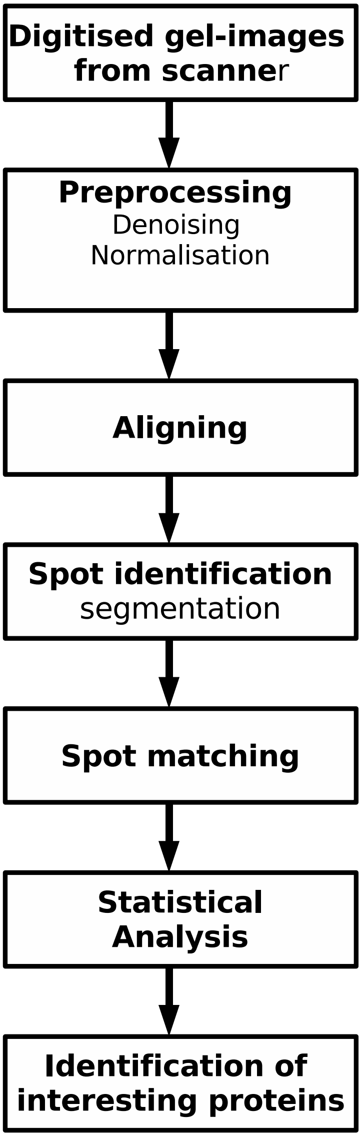


Figure 1.6: Flow chart of a general analysis pipeline. Sometimes the aligning step is removed, and incorporated into the spot matching procedure.

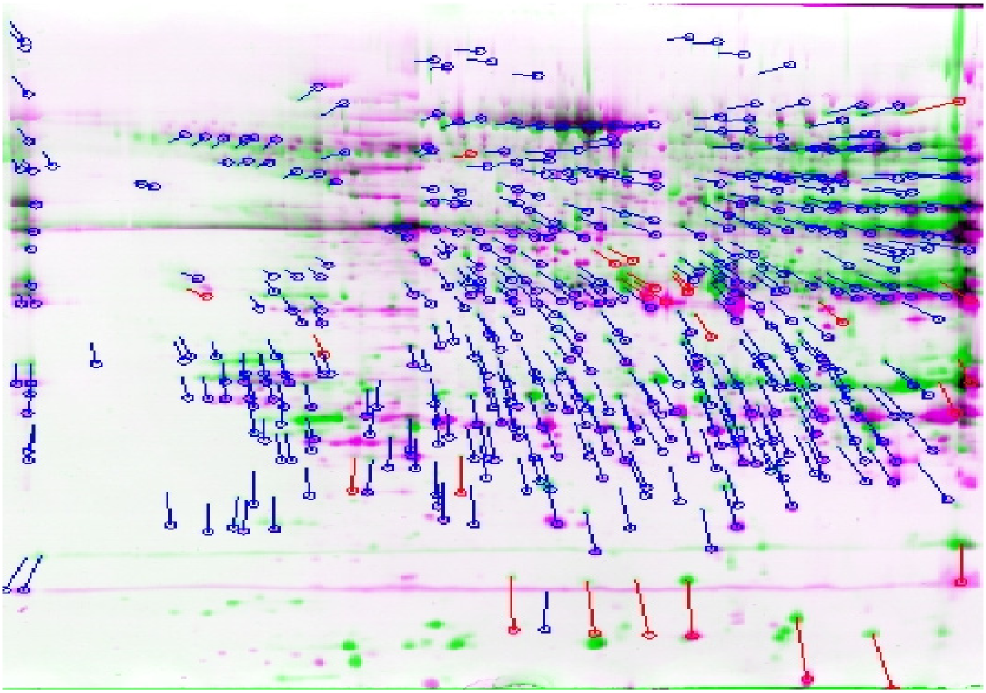


Figure 1.7: Results from the image alignment in the analysis pipeline. The arrows indicate how one image is transformed according to the reference, so all pixels in the two images are directly comparable. Courtesy of Ellen Færgestad.

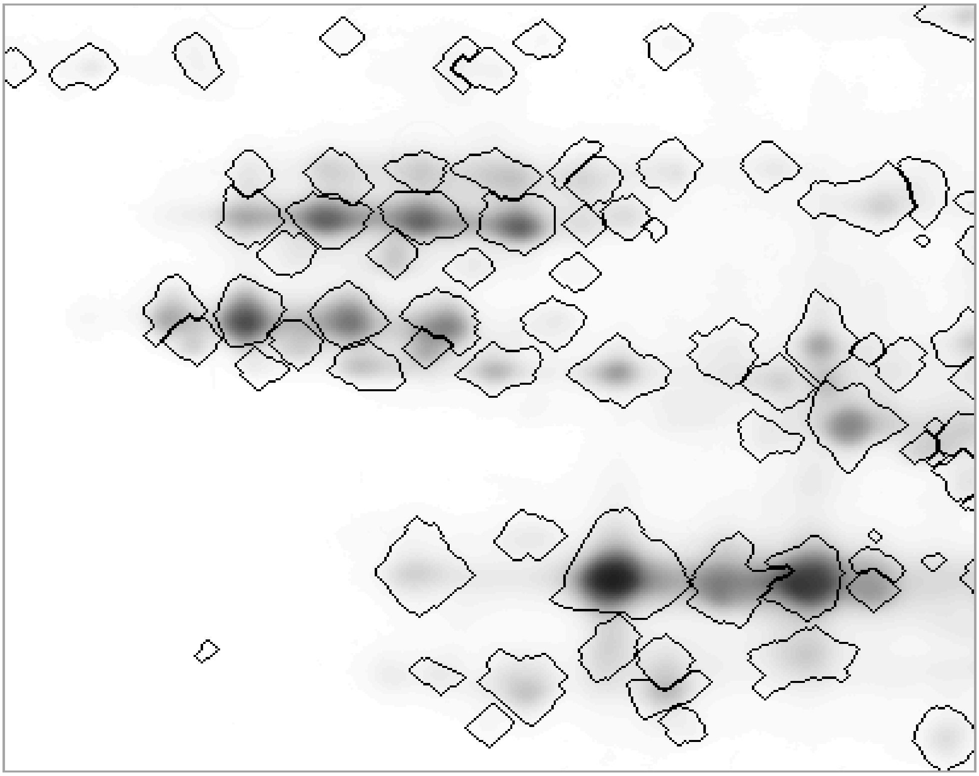


Figure 1.8: Typical results from the image segmentation procedure.

produced by the spot identification procedure ideally represent a single protein. The image segments are the basis for the following quantification and spot matching steps. Typical results after image segmentation are shown in figure 1.8.

The data acquisition and quantification calculates data and values for each of the image segments created in the segmentation procedure. Numerous variables and parameters can be calculated, the most usual being spot centre, volume, mean or maximum intensity. The volume of the spot is sum of all pixel intensities inside a spot boundary, while the optical density is the maximum intensity value within the same boundary. Relative volume and optical density are often also used to account for the difference in total intensity on each gel. When the relative values are used, the volume and intensity is divided by the total intensity in the image, and thus produces relative amounts similar to results from the normalisation procedure described earlier. Spot centres are usually placed at a spots centre of mass, using the following formula:

$$COM_x = \frac{\sum_x \sum_y \; xI(x,y)}{\sum_x \sum_y I(x,y)} \tag{1.1}$$

$$COM_y = \frac{\sum_x \sum_y yI(x,y)}{\sum_x \sum_y I(x,y)} \tag{1.2}$$

Here COM_x and COM_y are the centre of mass coordinate in vertical and horizontal direction respectively, and $I(x, y)$ is the pixel intensity at coordinate (x, y). The sum is taken over all pixels constituting a protein spot. In general, at this stage, any kind of parameters can be calculated based on the intensity values of each spot, the shape of the spot boundary and the position and orientation of the spot in the gel. Gaussian spot models [8, 27, 36] are also common to create at this stage in the analysis pipeline. Based on the calculated parameters, or the closeness to Gaussian shape, one can create spot filters, giving a score indicating how close each spot segment actually resembles the expected protein spot shape. Such filters are described and developed more thoroughly in chapter 2.

Before the quantified values can be compared between spots in multiple gels, corresponding spots have to be identified in all gels. The procedure of finding corresponding spots is often referred to as spot matching. Several methods for spot matching are described in the literature [9, 11, 27, 29, 32, 37, 38]. The general solution to the spot matching step relies on using the similarity of geometrical features extracted from the gels under investigation. Usually these features are point patterns, where each point represents the centre coordinates for a single spot. Point Pattern matching (PPM) are then applied for mapping the geometric point patterns from the different gels on to each other, minimising the influence of spot displacements. Performed in this way, the gel alignment described above is sometimes inherent in the matching procedure. However, due to difficulties in reproducing protein spot segments from gel to gel, producing a complete match between gels segmented individually has proved to be very difficult. Match tables including comparable spots from several gels are thus subject to several errors which corrupts the final comparison and statistical analysis [39, 40]. The consequences of erroneous values in match tables are described more thoroughly in chapter 4. Because of this, the importance of common spot boundaries has lately been stressed by both users and producers by commercial software (DELTA-2D and Progenesis SameSpots).

This issue has also been targeted earlier in the literature [41, 42]. Common spot boundaries means defining a single set of boundaries which are used on all the gels in the analysis. Thus the match between the gels are always 100%. For the common spot boundaries approach to work, it is crucial that the gels are properly aligned before analysis. In chapter 3 a new method for defining common spot boundaries is presented. Yet another approach to the matching problem, is to perform analysis without making assumptions on spot boundaries at all. This idea has not been presented earlier in the 2-DE community, and the results presented in chapter 4 and 5 show that this approach can be very useful.

After all spots are identified and matches have been found between the experimental gels, the significant differences in intensities can be investigated using statistics or multivariate analysis. Common statistical measures like mean, standard deviations and histograms can be calculated for a specific group of spot matches, and used to remove spurious spots and decide whether the differential expression between several groups of spots are significant. The significance is performed using confidence intervals and probability values. Because of the large amount of data usually produced in 2-DE experiments, multivariate analysis has also been common to results from 2D-gels [26, 35, 43, 44]. Both factor analysis, heuristic clustering and chemometric methods like Principal Component Analysis (PCA) and Partial Least Squares Regression (PLSR) can be applied. A more thorough description of the applied chemometric methods is given in section 1.6.

It should be noted that the presented steps for the analysis pipeline are based on a general description. The nature of the methods, and the order in which some of them are performed, may differ from system to system. Several systems also use combinations of alignment, segmentation and matching iteratively to achieve the desired output.

1.5 Segmentation and spot identification in 2-DE

Segmentation in 2-DE is generally referred to as the process of separating areas in the image related to protein spots from image background, noise and other artifacts not related to proteins. The segmentation procedure produces a set of image segments consisting of connected neighbouring pixels enclosed by a spot boundary. Ideally each image segment represents the spot of one single and isolated protein.

1.5.1 Spot identification and background intensity correction

It is important to distinguish between the process of spot identification and background intensity correction. Non-uniform background intensity is often regarded as the main reason why spots in gel-images can not be identified by a single threshold. By single threshold is here meant labelling all pixels above a certain intensity value as belonging to proteins, while all pixels below this threshold is regarded as background. By subtracting the non-uniform background it is often believed that a single threshold on the new, background corrected image will produce the desired protein spot shapes. Experience has shown, however, that more sophisticated methods are necessary to reproduce the circular protein spots that would be naturally identified by a human observer. One should thus distinguish between the methods that aim to reproduce the spot-shapes as viewed by a human observer, and methods that subtracts background to produce more correct spot intensities for the final analysis. Images produced for segmentation purposes should not be viewed as reflecting true intensity values, while the background subtracted images would generally not be suitable for segmentation on their own. It is also a question whether the background subtracted image actually reflect the true protein content of a protein spot area. This, of course, depends on both the source of the background intensity and the procedure used for background estimation. If background intensity is introduced e.g. during the scanning procedure, it would be correct to remove this intensity. However, background intensity may also results from protein-content that is not properly concentrated into protein spots, but distributed more evenly in areas around spots. The evidence of this is that the background intensity are higher in areas where the density of protein spots is large. To actually identify proteins in these local areas is too difficult in both automatic and manual analysis, and areas like these should be regarded as background and not included in the final analysis. It is, however, not clear whether the actual protein spots in these areas are placed on top of the background or merged in the background as illustrated in figure 1.9. It is thus not obvious whether a general background intensity correction gives more correct spot-intensities in general. It should be mentioned that for the study introduced in chapter 4 and 5, using background correction produced slightly better results.

1.5.2 Challenges for segmentation in 2-DE

There are obstacles that makes the segmentation of 2-DE images a difficult task. One important issue is the non-uniform background introduced in the

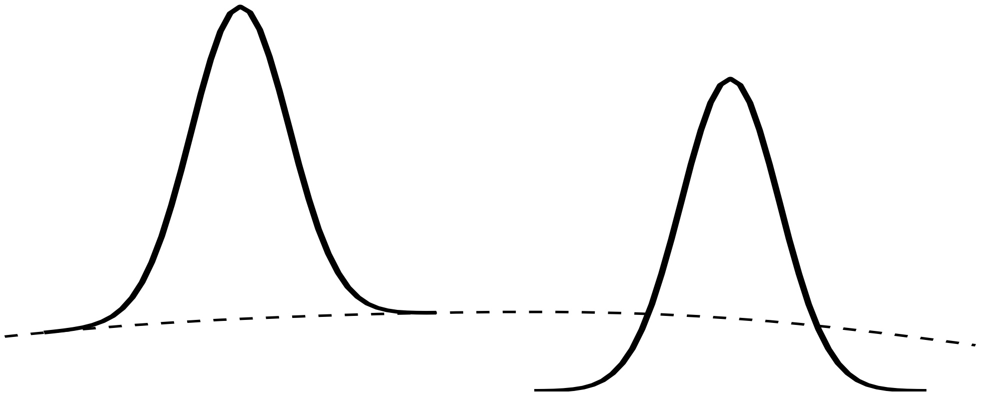


Figure 1.9: Intensity distortions caused by image background (dotted line). The spot to the left is placed on top of the background, and a correct spot-intensity is obtained by subtracting the background. The spot on the left is merged with the background. Performing background subtraction in this situation will result in erroneously reduced spot intensities.

previous section. An even more challenging problem is the highly overlapping spot clusters. As mentioned earlier, the number of overlapping spots in 2-DE is quite large. Apart from the results in table 1.1, which states that even protein spots that appear as singlets might actually be doublets or even triplets, the number of protein spots with a high degree of visual overlap is substantial. In highly overlapping spot-clusters similar to the one displayed in figure 1.5, the challenge of resolving the individual spots becomes close to impossible. Several methods to resolve overlapping spots have been proposed, and some of them are included in the following overview. Spots that appear as shoulders of larger spots, and have no distinct peak themselves, are especially problematic as shown in figure 1.10.

One proposal to identify these shoulders is to model the larger spots by Gaussian functions, and subtract the Gaussian model from the original image [5, 45]. This should reveal the hidden spots shoulders. This, however, touches upon another problem in 2-DE spot-identification, namely that protein spots often display considerable deviations from the ideal Gaussian shape [2, 3, 5]. This goes for both flat-surfaced saturated spots, and regular intensity spots with a non-elliptical or irregular contour. Subtracting a Gaussian modelled spot in these cases would introduce unwanted artifacts to the analysis. The problem of handling shoulders in overlapping proteins can be overcome by using a pixel based analysis as the one presented in chapter 4 and 5. Artifacts and noise

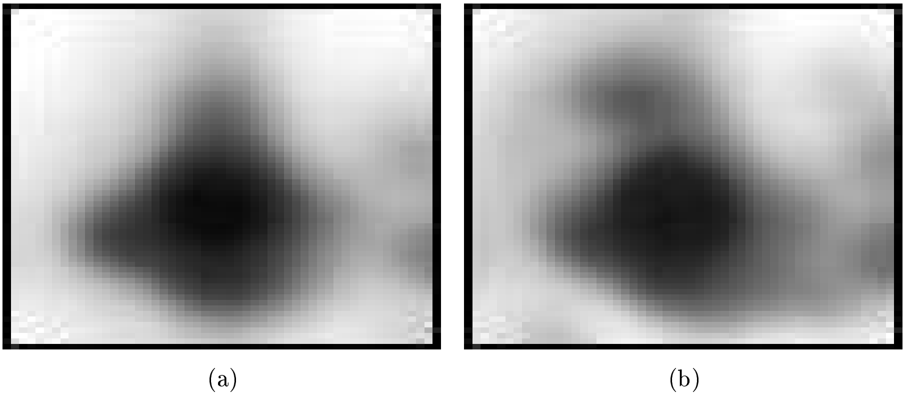


Figure 1.10: (a): Small spot as a shoulder above a larger spot showing no distinct peak.(b): The same spot in a replicate gel, but this time showing a clear distinct peak.

not related to proteins also constitute a major problem in 2-DE segmentation. Most segmentation routines are not able to distinguish artifacts resulting from dust, fingerprints, cracks in the gel surfaces and other sources of pollutions from true protein spots. The presence of such artifacts will certainly introduce problems in the final protein spot analysis.

Generally a substantial number of methods and approaches have been suggested in the literature to accomplish the segmentation task during the 30-year existence of 2-DE. For spot identification, four types of approaches are common: Stepwise threshold, second derivatives, image morphology and watersheds. These approaches are reviewed in the following chapters, together with some basic methods for background intensity correction. Finally the selected segmentation pipeline used in general throughout this book is presented.

1.5.3 Stepwise Threshold

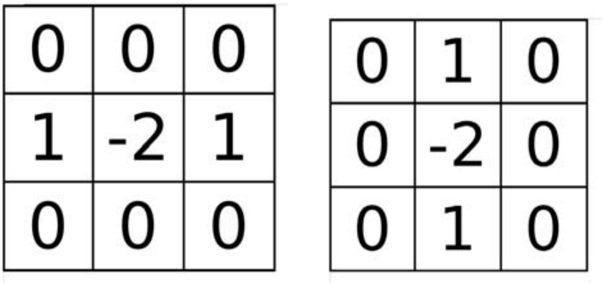
As already mentioned a single threshold is not sufficient to separate protein spots from background in 2-DE. The basic stepwise thresholding procedure [7, 9] compensates for this by considering thresholds at several intensity levels. The idea is to begin at the lowest intensity level considered to be a possible signal (protein spot), and collect pixels in connected areas at this level. After this initial threshold, increasing intensity levels are gradually considered for each connected area. At a certain intensity level the connected area may be split into two or more connected areas. If so, these new areas may again be split at a yet higher intensity level. This procedure continues for each connected area until no more splits are possible, or the areas created by the split does not satisfy some pre-defined criteria. Such criteria can be based on size, shape and maximum intensity of the area. When no more splits are identified or accepted, the resulting image segment is defined as the area created in the last split, and is said to constitute a protein spot. Slightly modified versions of the stepwise thresholding approach has been presented. Cutler et al. [46] turned the procedure upside down, starting at the highest intensity level and working downwards. At each new intensity level, new pixels were added to the connected areas defined at the previous level. In this way, instead of splitting connected areas, the areas were allowed to grow and merge at each intensity level until they met some specific criteria. Another, similar approach was introduced by Tyson et al. [36]. Pixel intensities above a certain threshold were classified as belonging to regions of major protein staining. These regions were then allowed to grow, and pixels that could be found in a monotonically decreasing path from these were assigned to the regions. The major advantage of the stepwise threshold approach is its ability to resolve all potential intensity variations into distinct connected areas. Its disadvantage is its sensitivity to noise and artifacts, and some criteria is usually needed to accept or reject the connected areas at the final stage. Another problem with this approach is its inability to detect shoulders in overlapping protein spots which does not have an distinct peak.

1.5.4 Second Derivatives

Another approach to identify protein-spots in a 2-DE images is the use of second derivative filters [5, 10, 27, 29, 47]. The second derivative is defined in the horizontal, x, and vertical, y, direction by the following formulae:

$$\frac{\partial^2 I(x,y)}{\partial^2 x} = 0, \quad and \quad \frac{\partial^2 I(x,y)}{\partial^2 y} = 0 \tag{1.3}$$

where $I(x,y)$ is the intensity at image coordinate (x,y). In Appel et al. [27] the zeros is replaced by a threshold in each direction, and the value of the derivative at image coordinate (x,y), is taken as the minimum of the two calculated values. When used on images, the second derivative is usually transformed into filter masks which are used on each pixel on the image. The simplest second derivative masks are shown for the horizontal and vertical direction in figure 1.11.



0	0	0
1	-2	1
0	0	0

(a)

0	1	0
0	-2	0
0	1	0

(b)

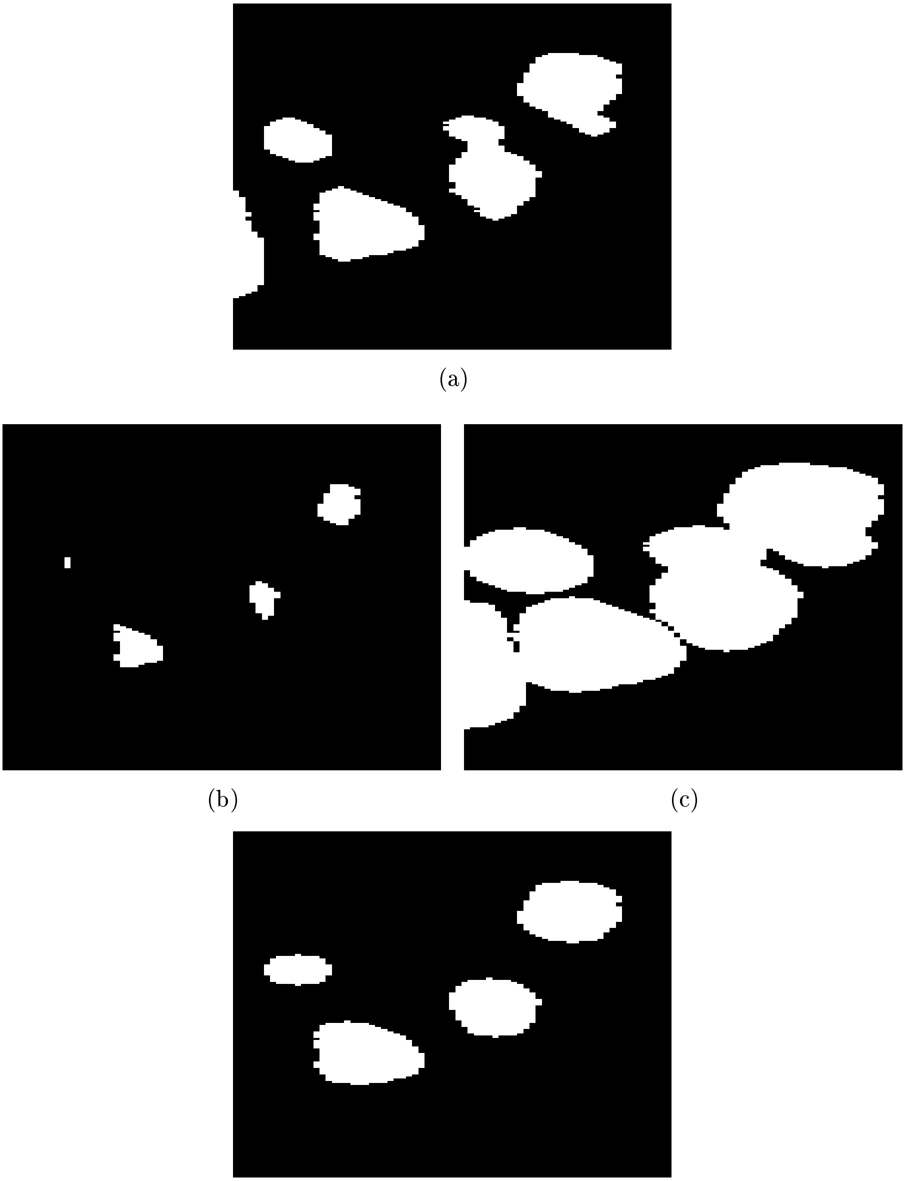
Figure 1.11: Masks used for calculating second derivatives in images. a): Vertical direction. b): Horizontal direction.

When using second derivatives, the gel image is viewed as a landscape where the protein spots are hills and peaks rising from the surrounding landscape constituting the image background. The second derivatives zero crossings identifies the locations of the hill-slopes where the curvature of a peak changes from concave to convex. In the convex region the sign of the second derivative will be negative, and thus the pixels which have negative second derivative are collected into connected regions and labelled as protein spots. The main advantage of this approach is, because of the high sensitivity of the second derivative, the ability to detect shoulders which are not detected by the stepwise thresholding approach (figure 1.10). This sensitivity is exploited by Olson et al. [47], who use the second derivative approach combined with the stepwise threshold to achieve the desired segmentation. However, high sensitivity can also become a drawback, because noise is also easily detected, and additional methods are needed to separate the true protein spots from noisy features. To compensate for this, it is common to perform some smoothing or denoising on the images before segmentation, to reduce the impact of noise. Polynomial smoothing [9, 27, 36] is widely used, but also Gaussian and adaptive smoothing [27] have been suggested. Generally these smoothing templates have a tendency to distort the original signal in such a way that segmentation becomes difficult. Peaks are lowered, so weak spots can no longer be distinguished from the background, and the degree of overlap in spot clusters increases, complicating the resolution of such regions into individual spots. Based on this, more sophisticated denoising methods, like wavelets [28], have been introduced which do not distort the original signal two the same degree. It is generally agreed that the second derivative approach is dependent on proper denoising methods to perform satisfactorily. Another drawback with this approach is that the number of pixels in the convex area are generally much smaller than what is normally said to constitute a protein spot. This is because the zero-crossing of the second derivative is associated with the steepest part of a peaks slope, rather than the beginning of a peak. Thus the second derivative approach is sometimes combined with region growing methods [10] similar to those described in the previous section to increase the number of pixels associated with each spot.

1.5.5 Image Morphology

Image Morphology [48] is method for identifying features with a specific size and shape in an image. The features are identified by constructing a structural element with a size and shape similar to the features one wants to identify. This structural element is then used on the image with an operation referred to as "morphological opening". The advantage of the morphological approach, is its ability to identify other features than just protein spots. One such feature is the previously mentioned streaks, which often causes problem in 2-DE analysis. Streaks are identified by selecting a horizontal or vertical line of a certain length as structural element, while spots are identified by a disk of a specified radius. The size of the structural element should generally be chosen as the smallest shape that will not fit into the structures of interest in the image. Morphology has been used in 2-DE by [8, 29, 49–51], for identification of protein spots or streaks. The general procedure for morphological identification of protein spots and streaks is as follows. The morphological opening of an image is the erosion of the image with the structural element, followed by dilation with the same structural element. Erosion and dilation are both fundamental morphological operations, and consist of replacing each pixel value in the image with the minimum and maximum value of its neighbourhood respectively. The size and shape of the neighbourhood is decided by the structural element. The effect of erosion and dilation on a binary image is illustrated in figure 1.12(b) and 1.12(c) respectively.

When erosion and dilation are used successively in an image it is called opening, which is illustrated in figure 1.12(d) for a binary image. Opening has a similar effect on images with more than two greyvalues. The opening of a general greyvalue image by a circular disk with a specified radius, can be viewed as a "rolling ball" with the same radius rolling along the underside of the greytone landscape. Depending on the radius, the ball will fit into certain structures, but will be too large to fit into other structures. In the case of the disk, such features are typically circular peaks (protein spots) with a radius smaller than the disk.



(a)

(b)

(c)

(d)

Figure 1.12: Erosion, dilation and opening using a circular disk with radius of five pixels as structural element. (a): Original image. (b): Result of erosion. (c): Result of dilation. (d): Result of opening. It can be seen that structures which do not fit into the structural element are removed, while other structures remain intact when opening is performed.

After closing is performed on each pixel in the original image, a new image is created where the features the ball did not fit into are removed. What we are really interested in, however, is the features which were removed by the rolling ball, which we obtain by subtracting the opened image from the original image. This operation is sometimes referred to as the top hat transform [51]. The features that stand out from the uniform background in the subtracted image, are identified as protein spots. The morphology operations are sensitive to noise, but unlike the second derivative approach, noisy features which are part of larger structures will not influence the output. This means that only noise in areas between the larger structures show up in the final protein spot image, and it is possible to separate the larger structures of interest (protein spots) from the smaller ones (noise). This is done by successive opening and closing of small disks, where the largest size of these disks determine the smallest allowable size of a feature [49]. Larger features are not significantly affected by this operation, while the smaller ones are removed. The results on a small image during the different morphology operations are shown in figure 1.13.

The major problem with the morphology approach is resolving the individual spots in overlapping spot clusters. In order to identify all areas in the image consisting of proteins it is necessary to use a disk that is too large to identify the smaller variations within the larger areas.

1.5.6 Watersheds

Watersheds have also been used in the segmentation of 2-DE images. When using the watershed approach [2, 38, 52] the image is again viewed as a landscape, but this time each peak is a depression on the surface rather then an elevation. The segmentation is based on first identifying all local minima in the landscape, and find the catchment basins associated with each local minima. The concept of a catchment basin can be described by visualising rain falling over the landscape. As the rain is allowed to flow downhill in the landscape, pools will emerge around the local minima. The water in each pool will be collected from a specific area surrounding the local minima, and this area is called a catchment basin. The boundary between several catchment basins are called watersheds. For 2D-gels, each separated catchment basin is said to represent an isolated protein spot. The main difference of this method to the other three approaches, is that all pixels in the image are assigned to a catchment basin, meaning that no pixels are initially said to represent non-protein regions. This results in over-segmentation of the image (figure 1.14(b)), and an additional filtering of the segments is necessary to identify the image seg-

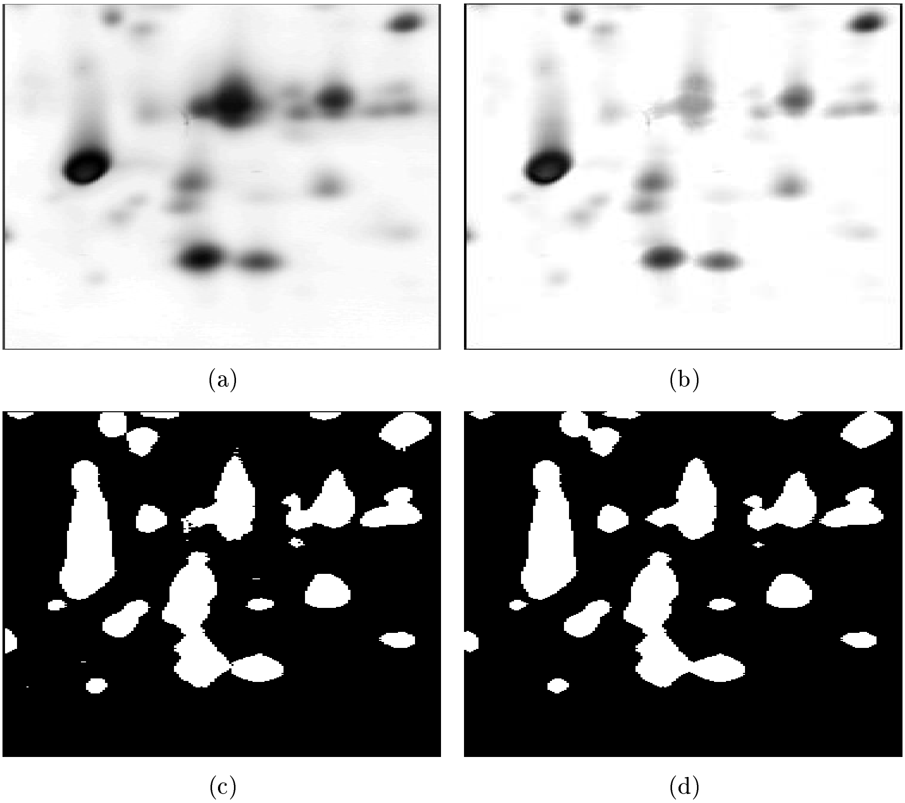


Figure 1.13: Results from applying morphological operations for spot identification. (a): Original image.(b): Image after morphological opening with a line and a disk as structural elements. (c): Binary image after thresholding of image from (b). (d): Binary image after noise-removal with successive openings and closings.

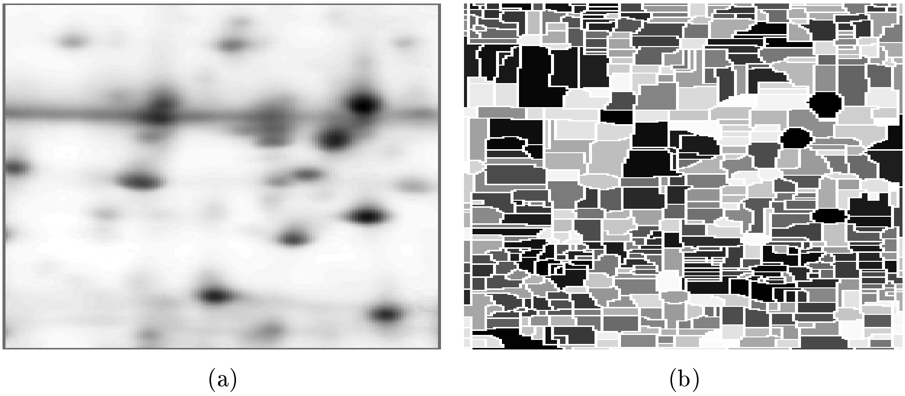


Figure 1.14: Image segmentation by watershed. (a): Original image. (b): Over-segmentation common in watersheds. Some additional thresholding an merging of areas is required to produce the output used for further analysis. Courtesy of Beata Walczak.

ments consisting of true protein spots. Usually this filtering is based on size and intensity of each segment. The method will also have problems detecting shoulders in overlapping spots, since no catchment basins will be recognised if the spot does not have a distinct depression.

1.5.7 Other Methods and combinations

Two other segmentation methods are also mentioned here for completeness. Garrels [6] identified peaks for each line during the scanning procedure of the gel, and then combined the peak-areas on each line to form protein spots in two dimensions. The peaks on each line were identified by a second derivative approach. The presented method is one of the earliest published in this field, and one would not normally mix the image scanning procedure with spot identification today. However, analysing images line-wise may have some advantages, especially when it comes to detection of shoulders and overlap resolution. Prehm et al. [4] used specially customised masks representing curvature to identify protein spots. The masks used in these publication were specially designed for the specific gel-images used, and usually an approach with more general applicability is sought in the present segmentation of 2D-gels.

Several publications have combined the before mentioned approaches in different ways to achieve a better output of the segmentation procedure. Conradsen et al. [29] combine morphology and second derivatives, Olson et al. [47]

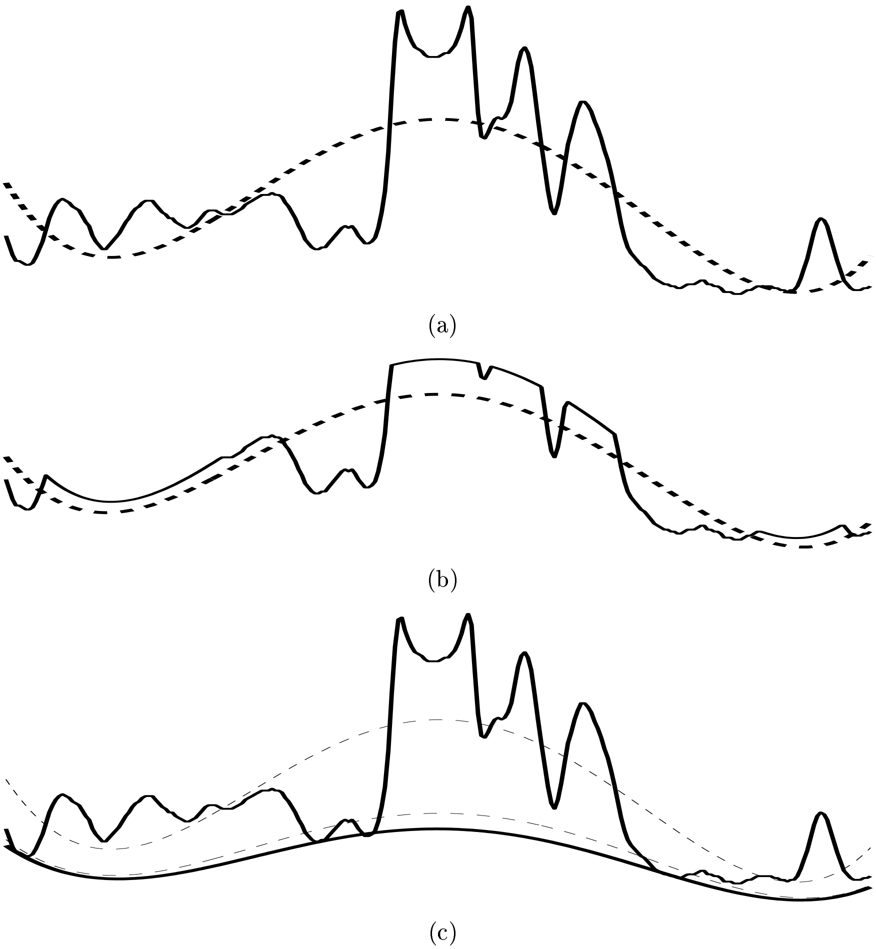
combine stepwise threshold and second derivatives, Takahashi et al. [45] use a morphology based approach with ring-operators combined with region growing and pixel collection, Kim et al. [52] combine watersheds and stepwise threshold, and recently Mannar et al. [50] has introduced and approach they claim uses principles from morphology, watershed and pixel-collection. Combining approaches can generally be useful, since each of the presented methods has different strengths and weaknesses. This is also the approach chosen to build a segmentation procedure used throughout this book.

1.5.8 Background intensity correction

Three methods for background intensity correction will be described in this section: a histogram approach, propagation of local minima and polynomial fit. The histogram approach takes the histogram of all pixel-intensities, and defines the background intensity based on the peaks in this histogram. The estimated background can be applied globally to the whole image [7, 45], or locally to a sub-region of the image [10]. In the latter case an histogram has to be constructed based on the pixels in each sub-image. Tyson et al. [36] used the approach of propagating local minima to estimate the intensity background. All local minima are identified in the image, and each pixel in the image are associated with its closest minima. Thus the minima are propagated throughout the whole image, creating local minima regions. After the propagation, there will be discontinuities on the boundary between the local minima regions, which are smeared out using a smoothing filter. The smoothed local minima image is said to constitute the background intensity. Another popular way of estimating background intensity is the use of polynomial functions [5, 27, 47]. The method consist of fitting pixels constituting the background to a polynomial of some degree, usually of order three or four. In this way the background intensity of every pixel in the image is estimated using the coefficients from the polynomial fit. The disadvantage of this approach is that it needs some pre-estimate of what is considered as background pixels, which the polynomial coefficients are calculated from. Fitting the polynomial to all pixels in the image usually gives a too high estimate of the background. Because of this, background subtraction by a polynomial function is often performed after the segmentation procedure. However, Lieber et al. [53] introduces a modified version of the polynomial fit, which does not need any pre-assumptions on background or foreground, and can be used directly on the original image. The method has not previously been used on 2D-gels, and was designed to remove dominating fluorescence from one-dimensional Raman spectra. The problems encountered in the Raman spectra have many similarities with the background intensity problems in 2D-gels, so this method is adopted here for background intensity corrections. The method is easy to extend to two-dimensional signals, however, the one dimensional approach is here kept, and the intensity correction is performed line-by-line in both vertical and horizontal direction. The reason for this is that the one-dimensional approach will also correct intensities in streaks, which is not possible in the two-dimensional case. The basic idea of the modified polynomial fit is as follows: A first polynomial approximation is calculated based on all pixels in an image. This polynomial will, as already mentioned, have too high values to represent the true background. The intensities above the polynomial fit is therefore subtracted from the original image, creating a second image. In this image the highest peaks are cut off, as shown in figure 1.15(b). A new polynomial approximation is calculated based on the second image. This polynomial will have lower values than the first fit, since the approximation is based on data where the peaks have been removed. It will also be closer to the true background. A third image is then constructed by subtraction in the same manner as in the previous step. A third polynomial can again be calculated based on this image. The process of fitting and subtracting goes on in an iterative fashion, until the approximation becomes stable, or a certain number of iterations has been reached. The polynomial fit is shown for 1,10 and 50 iterations in figure 1.15(c). In this thesis a polynomial of fourth degree and 50 iterations were found sufficient to produce the desired background. It should at last be noted that one should be careful when performing background intensity correction, since this may introduce unwanted variations into the data as reported by Wheelock et al. [21].

1.5.9 Comparing segmentation procedures

It would be a natural next step to compare the outputs of the different segmentation approaches, and thus decide on a best solution. This has, however, proved difficult for several reasons. First there is a great variability when it comes to details and parameters in the different methods. For example, the second derivative approach may work very well if a proper denoising has been performed on the images, while other methods may perform satisfactory without denoising. How should one then decide which one performs best? If a lot of effort and parametrisation has to be put into the denoising method, it may not be worth the effort, even if it produced slightly better results. The number of parameters is an important issue. Methods with only a few intuitive parameters would be preferred, but at the same time methods with more parameters might perform better if its parameters are correctly set. Then there is the issue of the variability in the experimental material. Images from 2-DE can vary



(a)

(b)

(c)

Figure 1.15: Iterative background correction from Lieber et al. [53]. (a): The original signal and the first polynomial fit (dotted line). (b): The signal with peaks removed, and the second polynomial fit. (c): The original signal, and the polynomial after 1, 10 and 50 iterations. The solid line at the bottom is the final estimate of the background.

substantially between experiments and different laboratories. A method that seems to perform well on one type of gel, may fail on other gels because of the variability spot appearance, background and sources of noise. It is also difficult to evaluate the final performance, because the true resolution of all protein spots on a gel is rarely known exactly. Because of the large variability in methods an approaches used to segment 2D-gels, both reported in the literature and in commercial software, it would be a great advantage if one could decide upon a more standardised procedure for segmentation and data-analysis. To accomplish this, one needs to compare the methods on an even basis, on data-sets which comprise the variability in 2D-gels commonly encountered, a so-called "ground truth" dataset. Unfortunately such ground truth data do not exist for 2D-gels, and this, together with the other issues, is probably the reason why a thorough study comparing the different methodologies has not yet been published.

1.5.10 Selected segmentation pipeline

Despite of the difficulties in selecting the best image segmentation in 2-DE, the following procedures were found sufficient to produce satisfactory segmentation results in the following articles. Images illustrating the different steps in the segmentation procedure are shown in figure 1.16.

To identify general areas in the image representing protein spots, image morphology was found to be the most effective method, as presented by Skolnick [49]. Streaks in the image were also removed using the same morphology approach, and the final morphological removal of small noisy features by successive opening and closing was also applied. The only parameter needed for this segmentation is the size of the disk used as structural element, and no pre-processing was generally needed. Noise was generally removed by a median filter or polynomial smoothing prior to the spot identification, or by the morphological post-processing step. Images after the morphology operation are shown in figure 1.16(a). As can be seen from this image, morphology has successfully managed to identify areas where protein spots are present. However, a considerable number of these areas contains overlapping and unresolved spots, so an additional method is needed to separate the unresolved areas into individual spots. One can generally say that morphology is effective in separating out regions in the image where protein spots are present, but fail to resolve these regions into individual protein spots. Images created by morphology are ideal starting points for the pixel-based method presented in chapter 4 and 5.

To resolve the individual proteins in the areas identified by morphology, the stepwise threshold of Vo et al. [9] was adopted. Each region is subjected to the stepwise threshold starting at the lowest intensity in each region and working up. The intensity is usually increased one unit at a time, but longer steps can also be used to make the segmentation faster. As mentioned earlier, this method is sensitive to noisy features. Especially if the intensity values of the original image are used in the thresholding, several small, spurious and ill-shaped spots are often produced, as displayed in figure 1.16(b). The smallest spots can be removed by simple size requirement of an accepted region split, but also another correction was designed to improve the appearance of the resolved protein spots.

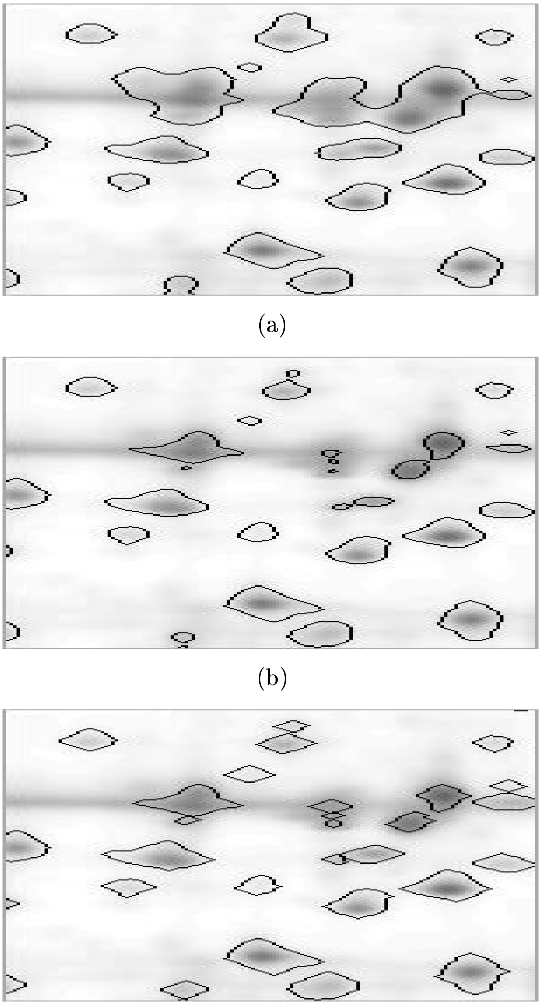
For each identified protein spot by the thresholding procedure, a window is constructed around the image segment constituting the spot. The minimum pixel intensity on this window is located, and all pixels inside the window higher than this minimum value are assigned to the protein spot. One restriction here is that pixels are not assigned if they belong to another protein spot. This window approach may be similar to a method for spot-identification used by the commercial software ImageMaster (GE Healthcare), as mentioned by Mannar et al. [50]. The final image segments representing individual protein spots are shown in figure 1.16(c).

1.6 Multivariate Analysis

As already mentioned, multivariate analysis is the method of choice in this thesis for analysing the output from a segmentation procedure. In the following common multivariate approaches are described shortly.

1.6.1 Principal Component Analysis - PCA

Principal Component Analysis (PCA) [54] is a way of decomposing a data matrix $\mathbf{X}$ of samples and variables into principal components. This decomposition is an effective way of representing data when the number of interesting phenomena in the data are much smaller than the number of variables in $\mathbf{X}$. In such data most of the variables are highly correlated, and analysing the data one variable at a time is not very efficient. Instead it is convenient to search for directions in the data displaying high variability, and identify the original variables which contribute to this variability. In this fashion all variables are considered at the same time, which is the main point of doing multivariate analysis. The directions of main variability constitute the principal components, and the decomposition of $\mathbf{X}$ is mathematically described as:



(a)

(b)

(c)

Figure 1.16: Results from different steps in the selected segmentation procedure. (a): Segments after morphological operations. (b): Segments after stepwise threshold of the larger segments in (a). (c): Final image segments.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \tag{1.4}$$

where **T** and **P** are called scores and loadings, and together they represent the principal components. **E** is the residual matrix, and consist of the variability in **X** which is not explained by the principal components. Principal components are orthogonal, and are subtracted from **X** one at a time, until the residual matrix **E** only consist of normally distributed noise, indicating that there is no interesting phenomena left to model in **E**. As this stage the number of calculated components is usually much smaller than the original number of variables. The loadings are calculated weights for each variable in each principal component, and represents the contribution or importance of a variable is given in a component. The scores are the values each sample is assigned in the new coordinate system resulting from the decomposition. The principal components have only mathematical, and no physical interpretation in themselves, but can still often be interpreted in relation to physical phenomena by looking at plots of scores and loadings. Using score and loading-plots, relationship between samples and variables in the principal components are easily visualised. PCA is unsupervised, meaning that only the maximum variation in **X**, and no outside phenomena, guides the decomposition of **X**.

1.6.2 Partial Least Squares Regression - PLSR

Partial Least Squares Regression (PLSR) [54, 55] is closely related to PCA, and also decomposed **X** into components, only this time they are called PLSR-components. PLSR, however, is a supervised method, meaning the one or several outside phenomena or variables, often called response factors or target variables, guide the decomposition of **X**. Instead of looking for the maximum variation in **X** alone, PLSR looks for the maximum variation in the covariance matrix $\mathbf{X}^{\mathbf{T}}\mathbf{Y}$. In other words PLSR looks for variations in **X** which is important for explaining the variation in **Y**. In addition to the decomposition of **X**, the decomposition of **Y** is given as

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^{\mathbf{T}} + \mathbf{F} \tag{1.5}$$

where **Q** and **F** are the loadings and residual matrix of **Y** respectively. PLSR builds a linear regression model between **X** and **Y** based on the principal components. Since each variable has a weight in each principal component, it is also possible to associate a regression coefficient for each variable towards the response **Y**. This regression model can also be used for prediction purposes. Once a PLSR model has been built based on a data set **X** of samples and variables, and one or several response factors **Y** for each sample, responses for new samples can be predicted based on the calculated regression coefficients. The only technical requirement is that the new samples have the same measured variables in **X** as the data used to create the PLSR-model.

A special version of PLSR is called Discriminant PLSR (DPLSR) [54, 56], where the response factor **Y** is of a specific discrete type, rather than the continuous values used in normal PLSR. The response factor **Y** in DPLSR are limited to ones and zeros, representing class membership. If a sample belongs to a certain predefined class it is assigned a value one in **Y**, and the value zero is assigned if the sample does not belong to this class. Thus the number of responses in **Y** is equal to the number of considered classes, except for the case of two classes, where only one **Y**-variable is necessary. Though the input in a DPLSR regression model are discrete binary values, predicted values from the model are continuous, and limits has to be decided to assign the prediction outputs to the correct class. Thus the prediction part of DPLSR rather takes the form of a classification.

1.6.3 Cross Validation

The predictive ability of regression model created by PLSR needs to be properly validated. The PLSR model itself is not able to assess whether the model is useful for prediction or classification of future samples. The optimal predictive ability of the model will often be subject to overfit, that is, information in **X** which is only randomly correlated to **Y**, and not representing general tendencies valid also outside the model-data, is used to assist in the prediction of **Y**. Especially this is true for data with a large number of variables compared to the number of samples. Some other method is needed to validate a models predictive ability, and to select the optimal number of principal components to use. The most obvious solution is to predict independent test samples not used for modelling, but similar to the samples used in calibration model. The predictive ability the model have on these test-samples gives an indication of the global validity of the model. However, it is argued that using an independent data-set only for testing is a waste of information and resources, and rather than leaving this data out of the analysis, all data should be used for modelling to create a better and more robust model. But how do we then validate the model? To solve this the concept of cross-validation [54, 57] was introduced. The principle of cross-validation is to use all available samples for both modelling and as independent test samples. To achieve this, a number of sub-models is created. In sub-model is calculated based on a sub-set of all the data, with a certain number of samples left out. The left out samples are then predicted using the corresponding sub-model. In this way the left-out samples are not used for modelling, and work as independent test-samples for the model. This procedure is repeated, and new sub-models are created until all samples have been left out at least once. The quality of the predictions of the left-out samples is then interpreted as an estimate of the overall predictive ability of the model. A special version of this validation approach is leave-one-out cross validation. Here only one sample is left out at a time, and the sub-model is calculated based on all other samples. This is repeated for all samples in the data-set.

The multivariate methods described above are commonly used on data from chemical spectroscopy, where the have wide applications. Multivariate approaches have not to the same extent been used on images from 2-DE, though a few publications exist [26, 35, 43, 44]. The challenges inherent in data from 2-DE have many similarities with chemical spectra. Signals appear as peaks standing out a not always uniform surface, and the height of the peaks is related to amount of material analysed. The main difference is that images from 2-DE are two-dimensional, while only one dimension is common in basic spectroscopy. There, however, ways to circumvent this problem, for example by unfolding the data which is described in chapter 4. The basic multivariate approaches should thus be well suited to analyse gel-images in 2-DE.

1.6.4 Outline of scientific papers

The last four chapters is dedicated to the scientific papers produced during this thesis. These articles summarise the research that has been done in the fields described previously, and constitute the main body of this book. The outline is as follows: As mentioned previously the presence of noise and artifacts not related to protein spots is an important concern when doing image segmentation of 2D-gels. In chapter 2 a multivariate classification model is developed, separating image segments consisting of protein spots fro image segments resulting from other sources. The classification method is DPLSR. The use of common spot boundaries has lately been introduced as a solution to the serious problem of matching spots on several gels in 2-DE. However, little has been said on how these common boundaries should be defined. In chapter 3 this issue is addressed, and a method for assigning common spot boundaries in multiple gels is suggested. Chapters 4 and 5 are both concerned with image segmentation. In chapter 4 an alternative way of analysing multiple gels without using image segmentation is introduced. The method consist of multivariate analysis on the pixel-level, and identification of significant areas for protein variations in the gel. The last article in chapter 5 is much founded on the work in chapter 4, where image segmentation is used in combination with the pixel-based analysis to improve the visualisation and output from a 2-DE analysis.

Chapter 2

A multivariate spot filtering model for two dimensional gel electrophoresis

Morten B. Rye and Bjørn K. Alsberg

Department of Chemistry
Norwegian University of Science and Technology
N-7491 Trondheim, Norway

This chapter has been recommended for publication in Electrophoresis.

Abstract

Image segmentation plays an important role in the automatic analysis of protein spots in two-dimensional gel electrophoresis (2-DE). Using image segments representing protein spots, the amount of protein in each segment can be quantified, and corresponding segments can be matched and compared for multiple gels. However, the common presence of image segments caused by noise and unwanted artifacts highly disturb the analysis and comparison of the gels. Common sources of such artifacts are cracks in the gel surface, fingerprints, dust and other pollutions. It would be advantageous to remove these unwanted artifacts during or after the segmentation procedure. To achieve this task a multivariate spot filtering model is developed using image segments from a gel segmentation. Parameters in the model are based on texture, shape and intensity measurements of the image segments. The model successfully managed to separate segments caused by noise, artifacts and cracks from image segments representing true protein spots. The classification method used is Discriminant Partial Least Squares Regression (DPLSR).

2.1 Introduction

Ever since its introduction, two-dimensional gel electrophoresis (2-DE) [1] has been the method of choice for separating, identifying and quantifying a large number of proteins from a cell sample. Parallel to the development of gel methodology, the search for a reliable automatic procedure to analyse the resulting gels has been sought. The alternative to automatic procedures is manual analysis by a human expert, which is both expensive and time consuming.

Several commercial packages for automatic or semi-automatic analysis of 2-DE exist today [20]. Software packages usually include methods for segmentation, alignment, quantification and matching of 2-DE images. Image segmentation in 2-DE is separating the pixels in an image relating to protein spots from pixels resulting from background, noise and unwanted artifacts, and is an important step in the analysis procedure. The final result of a segmentation is usually a binary image which separates the areas of interest (proteins), from areas not interesting for the following quantification and matching. The advantage of doing segmentation prior to spot matching, is to compare objects of interest (protein spots) rather than the individual pixels in an image. However, for such a comparison to be successful, a reliable segmentation has to be performed, meaning that objects compared are reduced to single, isolated protein spots.

There are numerous ways to perform image segmentation. A good review of methods used for 2-DE segmentation can be found in [13]. Despite the differences, it is usually agreed that image segmentation should include methods

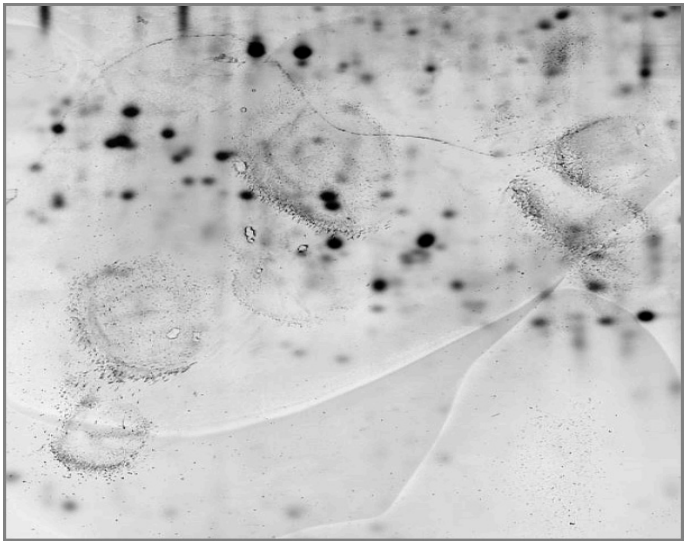


Figure 2.1: Example of a noisy area in a gel including cracks and other artifacts.

for noise removal, background-correction and streak-removal to perform satisfactorily. However, in several cases this is not sufficient to produce a reliable segmentation. It is not uncommon for gels to be exposed to various degrees of pollutions and distortions during their formation process. Common sources of such distortions are dust, fingerprints, and cracks in the gel surface. The segmentation procedure itself cannot distinguish between protein spots and such distortions. Thus unwanted artifacts are interpreted as protein segments after the segmentation procedure is completed, and information that distinguish these areas from protein spots are lost. A typical example of a noisy region causing problems in the analysis is shown in figure 2.1.

One way to handle image segments not resulting from protein spots, is the use of spot filtering methods between the segmentation step and the final analysis. Spot filtering assigns each image segment a score, which describes how similar a segment is to an ideal protein spot. A threshold is then applied to the scores, and segments that deviate too much from the ideal spot can be removed. The threshold value is often selected by a user, who has to decide whether the goal is to remove spurious image segments, or to keep as many of the true protein spot segments as possible. The drawback of selecting a low threshold, is that spurious segments are still included among the proteins, while raising the threshold might remove several of the true protein spots along with the spurious ones. It follows that the success of a spot filtering model depends on how well it manages to separate the image segments in question. The authors were not able to find any reports in the literature handling spot filtering explicitly, however one article describes a filtering model as a part of a 2D-gel segmentation and analysis procedure in general. Cutler et al. [46] use the method of pixel value collection to collect contiguous pixel groups. These groups are then tested against simple shape criteria, to verify if this pixel collection constitutes a protein spot. The exact details of these criteria and how they are calculated and evaluated are not mentioned in the article, apart from that they are based on size, aspect ratio, compactness and spread. Spot filtering methods are also used by some commercial software packages. Especially the Saliency parameter introduced by ImageMaster (GE Healthcare) is of interest, which is a measure based on the spot curvature as explained in their online user manual. The Saliency score indicates how far an image segment “stands out” with respect to its environment. Real spots generally have large saliencies while artifacts and noise have low saliencies. Unfortunately, since ImageMaster is a commercial software, there is no information on how this parameter is calculated, and comparing it to other parameters thus becomes difficult. For that reason it is not included here for comparison. PDQuest (Bio-Rad Laboratories) uses the degree of Gaussian fit as a spot filtering measure. The idea of using the Gaussian fit for spot filtering is based on the common assumption in 2-DE that all protein spots have a Gaussian shape, and deviations from this shape indicates that the image segment in question do not result from a protein spot. The degree of Gaussian fit is included among the parameters used in this study. An evaluation of whether this fit is a good classifier is left to the discussion part of this paper.

Common for the few existing spot filtering methods is that only one parameter is used, or at best, one at a time to assign a score to each image segment. In this study a multivariate approach is presented using several parameters simultaneously based on texture, intensity and shape of the image segments. The common important parameter variations with respect to the classifier are collected in principal components, and a score is assigned to each segment using Discriminant Partial Least Squares Regression [54, 56] (DPLSR). The results from this study indicate that such methods can be very useful for spot-filtering in 2-DE, and the model presented successfully separates image segments resulting from several unwanted artifacts from segments consisting of true protein spots.

The following steps are performed in this study: First two independent, noisy

2D-gel images are subjected to a common image segmentation procedure. The resulting image segments are then classified manually by a human expert to belong to one of several pre-defined classes, comprising different types of proteins and artifacts. A set of parameters are then calculated for each image segment, constituting a descriptor for this particular segment. To create a calibration model, image segments from three of the classes are used to build a calibration model. Then all image segments from the calibration gel and the independent test gel are assigned a score based on this model. Finally the parameters are evaluated and discussed.

2.2 Materials and methods

2.2.1 Gels

Two gels from independent 2-DE experiments are used in this study. Both gels were kindly provided by the Norwegian Food Research Institute (MATFORSK). The gels were silver-stained and scanned using an office scanner with 8-bit colour depth and a resolution of 240 dpi. A more detailed description of the data are given in [58, 59]. An image of each gel is shown in figure 2.2. It should be noted that the presented gels represent a degree of noise and pollution that is probably not acceptable for a protein identification in 2-DE in general. However, these contaminated gels are selected by purpose to represent several sources of noise and artifacts often seen to a lesser extent in 2D-gels, and thus provide sufficient data material for the spot filtering model.

All program code for segmentation and analysis are written in Matlab version 7.2.0 (Mathworks).

2.2.2 Image Segmentation

Both images are pre-processed by polynomial smoothing [9] with a mask size of 13 pixels. Segmentation of images is then performed using image morphology as described by [49, 51]. Streaks are removed using a line structural element of 31 pixels length, and image background estimated by using a circular disk with a diameter of 25 pixels as structural element. After streaks and background are subtracted from the original images, a single threshold is sufficient to identify the features used for further analysis. Some of the identified features consist of larger clusters of proteins and artifacts, which are resolved using a procedure described by Vo et al [9] for splitting overlapping protein spots. This collection of relatively simple steps is found sufficient to produce the image segments

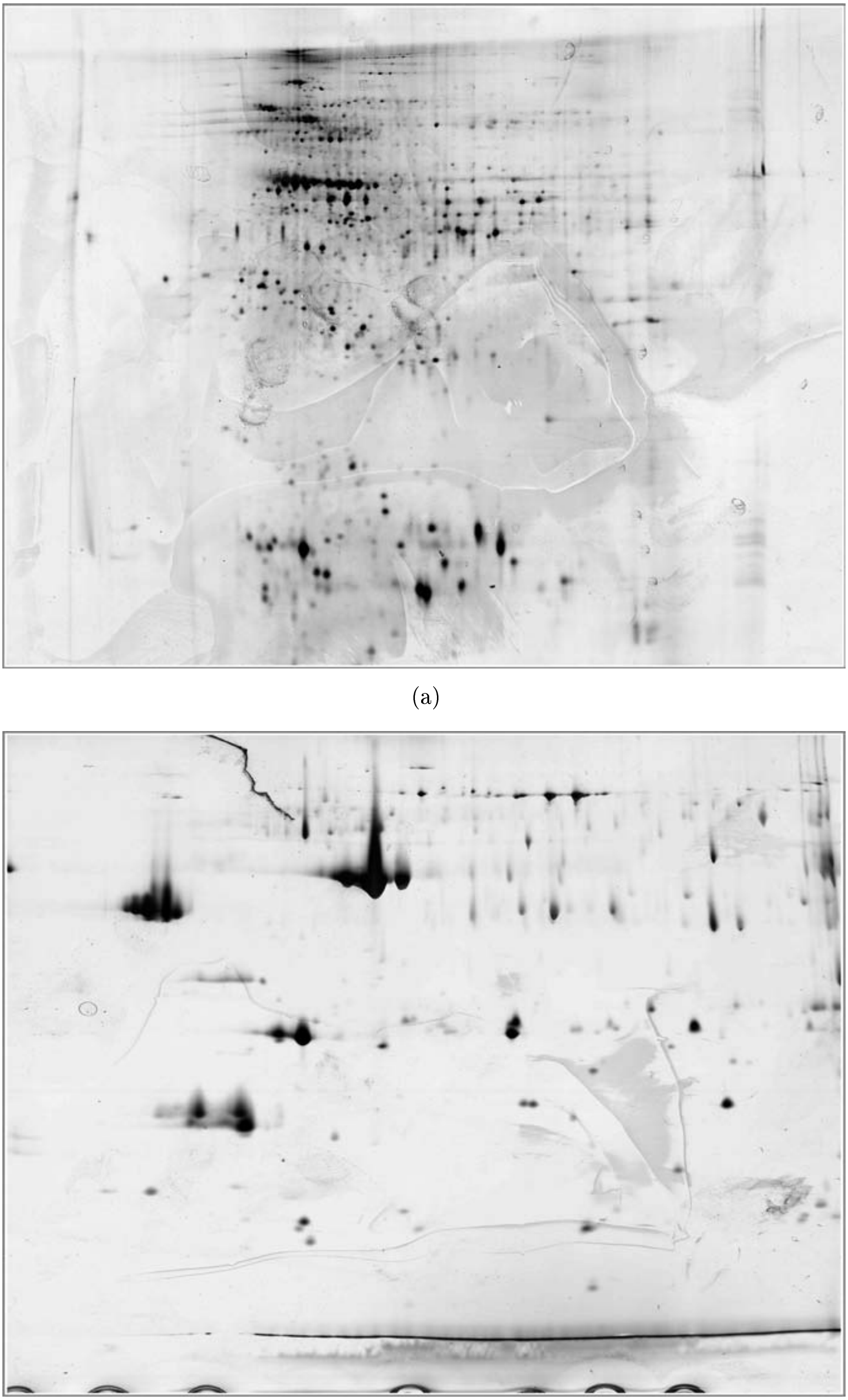


Figure 2.2: 2D-gels used in this study. (a): Calibration gel. (b): Test gel.

necessary for this study. The resulting image segments are manually assigned class membership, and used to build and validate the spot-filtering model.

2.2.3 Spot classes

The image segments from a 2D-gel segmentation procedure consist of protein spots as well as artifacts and other unwanted effects. To account for the variability among the segments it is necessary to classify the segments according to some predefined classes. Classes selected for this study are listed below. All segments were manually assigned to one of the classes prior to the automatic spot filtering. Examples of segments characteristic for each class are shown in figure 2.3.

Noisy features

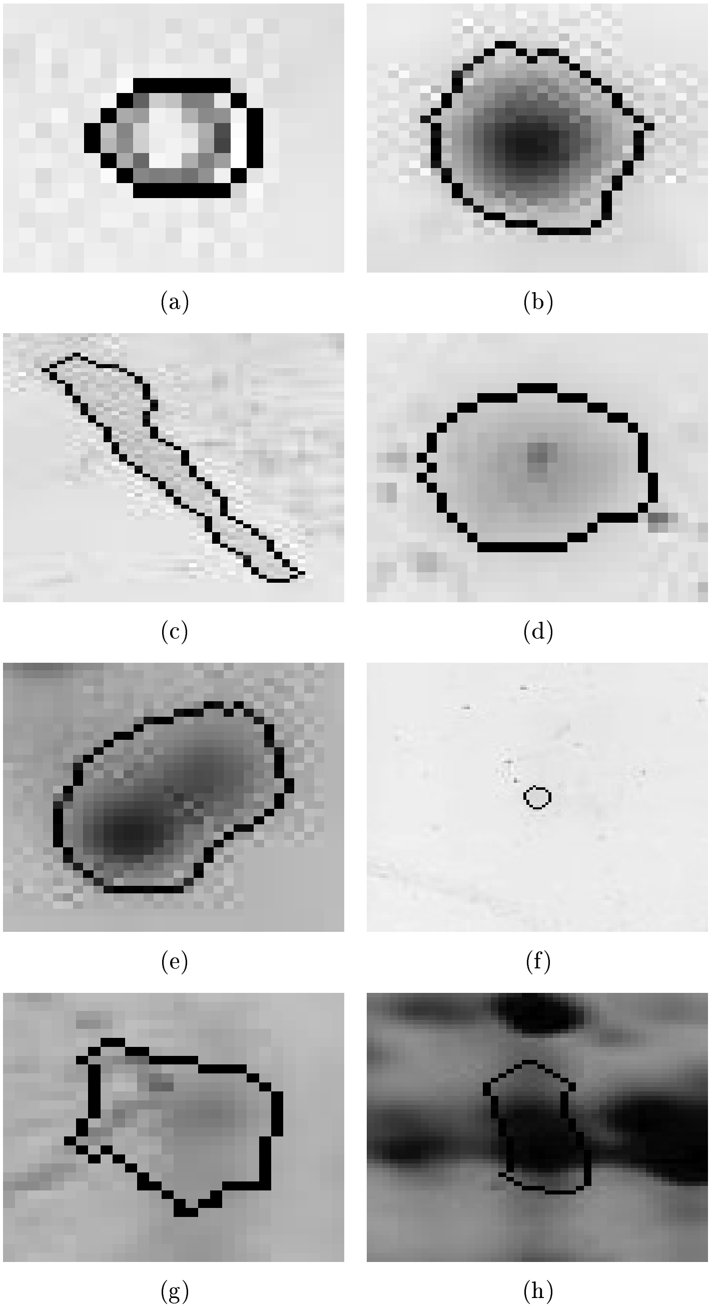
Typical image segments of this class result from fingerprints, dust and other pollutions with an irregular and noisy surface texture. An example is shown in figure 2.3(a).

Single protein spots

These are the ideal protein spot segments, consisting of a single isolated protein spot (figure 2.3(b)).

Cracks

A typical segment caused by cracks in the gel surface is shown in figure 2.3(c). Cracks in a gel can occur in any direction, however cracks oriented along a vertical or horizontal direction are usually removed along with the streaks during the segmentation procedure. Thus the cracks under consideration are limited to the ones oriented diagonally in a gel image. Cracks do not necessarily display a texture that distinguishes them from proteins, but their orientation and deviation from the ideal circular shape should facilitate their separation from protein spot segments.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Figure 2.3: Examples of manually classified image segments. (a): Noisy feature or artifact. (b): Isolated protein spot. (c): Crack in the gel surface. (d): Contaminated protein spot. (e): Overlapping protein spots. (f): Indistinguishable artifact with surroundings. (g): Overlap between protein and artifact. (h): Saturated protein spot.

Contaminated protein spots

Image segments that clearly represent a protein spot, but at the same time contain some contaminating artifact or irregular noisy surface. A typical example is shown in figure 2.3(d).

Overlapping protein spots

Segments consisting of two or more overlapping protein spots (figure 2.3(e)).

Indistinguishable artifacts

There are image segments that could not be clearly distinguished from protein spots based on the appearance of the image segment alone. The classification of these segments is based on their position in the gel and the characteristics of their local environment. An example is shown in figure 2.3(f). In this case the decision is based on the fact that the image segment is located far away from other proteins, and at a position in the gel where proteins are less likely to occur.

Unclassified segments

An image segment where the human observer is unable to decide whether it is caused by an artifact or a protein spot. Image segments from this class were re-classified in a second round, where the human observer is forced to decide whether the segment represents an artifact or a protein. Typical examples are low intensity image segments located in regions where both protein spots and artifacts are present.

Overlap between protein and artifact

Occasionally it is observed protein spots that overlap with artifacts. An example is shown in figure 2.3(g).

Saturated protein spots

This is a class of protein spot segments which has a significant number of pixels above the saturation limit. These spots deviate from the ideal Gaussian shape by displaying a flat surface instead of a peak. An example is shown in figure 2.3(h).

2.2.4 Descriptors

For a human observer it is quite easy to distinguish between image segments consisting of protein spots, and image segments caused by artifacts or cracks in the gel-surface. To automatically reproduce the variability experienced by a human observer, descriptors that represent these variations are necessary. We have focused on descriptors related to the noisy surface texture of unwanted artifacts, and descriptors that represent diagonal and elliptic shapes commonly seen in segments caused by cracks in the gel-surface.

Texture parameters

For texture description several early spatial texture models from the literature are considered [60–63]. The Grey Level Difference Method (GLDM) described in [61, 63] sufficiently describe the expected texture variations in this study. A modified version of the GLDM is also presented to account for intensity combinations of three pixels at a time in addition to the two-pixel interaction described in the original GLDM. This modification significantly improves the spot filtering model, as will be reported later in this study.

The original GLDM approach is based on calculating absolute intensity differences between pairs of pixels for every pixel in the image segment under consideration. If the two pixels are neighbouring pixels, the GLDM is said to be of first order. Generally the GLDM is stated as follows:

Let $I(x, y)$ be the image intensity at coordinates (x, y) in an image, where $I(x, y)$ is the digital image function. For a displacement given by $\delta = (\Delta x, \Delta y)$, the intensity difference between two arbitrary pair of pixels can be written as

$$I_\delta(x, y) = \mid I(x, y) - I(x + \Delta x, y + \Delta y) \mid \tag{2.1}$$

In this study intensity differences between neighbouring pixels based on a 4-connected neighbourhood are considered, meaning that the value of $(\Delta x, \Delta y)$ only can take the values $(0, 1)$ and $(1, 0)$. This leads to the following equation for intensity differences in the horizontal direction:

$$I_\delta(x, y) = \mid I(x, y) - I(x + 1, y) \mid \tag{2.2}$$

For the vertical direction the following equation applies:

$$I_\delta(x, y) = \mid I(x, y) - I(x, y + 1) \mid \tag{2.3}$$

Based on the calculated intensity differences for all pixels in an image segment by equation 2.2 and 2.3, a probability density function associated with each segment can be constructed:

$$f(i \mid \delta) = P(I_\delta(x, y) = i) \tag{2.4}$$

Here P is the probability that an intensity difference in an image segment takes the value i. The probability density function is a vector with the same length as all possible intensity differences (255 for 8-bit images). Each time an intensity difference is calculated, a value of 1 is added to the corresponding position in the vector, representing intensity difference i. After all differences have been added, the vector is normalised by dividing each element in the vector by the total number of calculated intensity differences. This normalised vector is referred to as the probability density function. To account for directional variations, the original GLDM calculated individual density functions for each direction. In many texture applications this makes sense, however, in this study there is no expected variations in texture according to direction. Thus intensity differences for both horizontal and vertical direction are collected in a single density function for each image segment.

After the density functions are constructed, a number of parameters can be computed based on these functions. In this study the following five parameters are calculated and used in the spot-filtering model. Here N is the number of intensity levels in the images.

Contrast:

$$CON = \sum_{i=0}^{N-1} i^2 f(i \mid \delta) \tag{2.5}$$

Angular Second Moment:

$$ASM = \sum_{i=0}^{N-1} f(i \mid \delta)^2 \tag{2.6}$$

Entropy:

$$ENT = -\sum_{i=0}^{N-1} f(i \mid \delta) \log f(i \mid \delta) \tag{2.7}$$

Mean:

$$MEAN = \sum_{i=0}^{N-1} i f(i \mid \delta) \tag{2.8}$$

Inverse Difference Moment:

$$IDM = \sum_{i=0}^{N-1} \frac{f(i \mid \delta)}{i^2 + 1} \tag{2.9}$$

To improve the description of texture differences between image segments, a model for capturing intensity differences for three pixels at a time is also developed. The motivation for developing such a model is the observation that certain constellations of three-pixel interactions are more favourable for protein segments than for segments caused by noise and artifacts, as displayed in figure 2.4.

No references describing such texture parameters have been found in the literature, so the extended GLDM is formulated in a similar notation as the

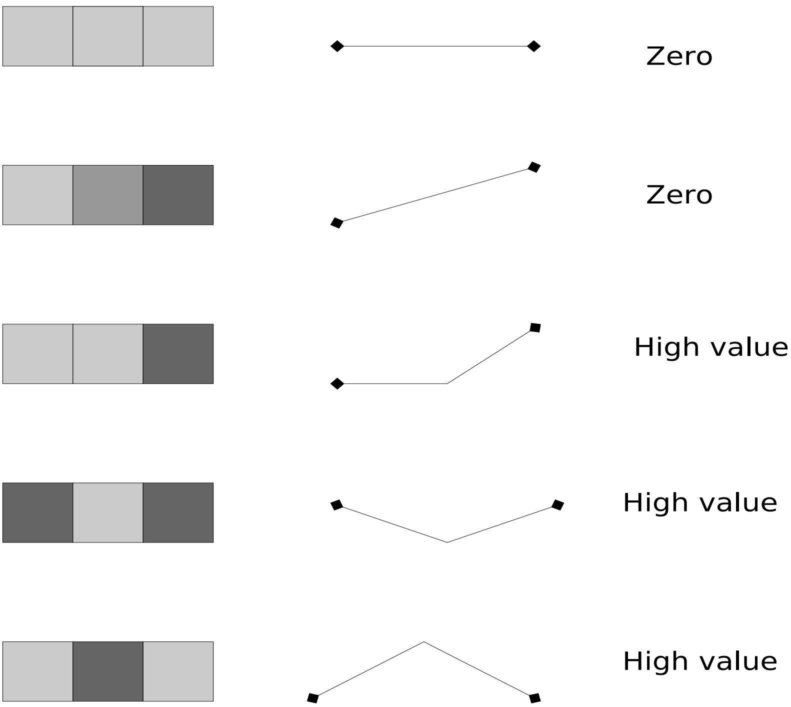


Figure 2.4: Possible combinations of three pixels at a time, and the resulting zero or high input in the modified GLDM model. Pixel greyvalues are shown to the left, pixel intensity displayed as a graph in the middle, and the intensity difference i in $f(i \mid \delta)$ at the centre pixel are indicated to the right.

original GLDM. For three arbitrary pixels with distances $\delta_1 = (\Delta x_1, \Delta y_1)$ and $\delta_2 = (\Delta x_2, \Delta y_2)$ from (x, y) the difference function is written as:

$$I_{\delta_1,\delta_2}(x,y) = \mid [I(x+\Delta x_1, y+\Delta y_1) - I(x,y)] - [I(x,y) - I(x+\Delta x_2, y+\Delta y_2)] \mid \tag{2.10}$$

Again only adjacent pixels in the 4-connected neighbourhood are considered. In this case $\delta_1 = -\delta_2$ and the equation reduces to

$$I_{\delta_1,\delta_2}(x,y) = \mid [I(x-1,y) - I(x,y)] - [I(x,y) - I(x+1,y)] \mid \tag{2.11}$$

in the horizontal direction, and

$$I_{\delta_1,\delta_2}(x,y) = \mid [I(x,y-1) - I(x,y)] - [I(x,y) - I(x,y+1)] \mid \tag{2.12}$$

in the vertical direction.

The probability density function and the five resulting texture parameters were calculated in the same manner as for the original GLDM, the only difference being that the summations are carried out over the $2N-2$ possible intensity differences rather than $N-1$. This gives a total of 10 texture parameters used in the model, 5 from the original GLDM, and 5 for the extended version.

Shape parameters

Several artifacts not related to proteins can be distinguished based on the shape of an image segment. This is especially true for cracks in the gel surface, which often deviates from the circular shape inherited by protein spots, and also by their diagonal orientation in the gel image. Image segments consisting of several proteins may have similar shapes to such cracks, especially when the proteins are located along streaks in the gel image. However, these streaks are always oriented in the protein migration directions (horizontally and vertically), distinguishing them from most cracks. It is concluded from these observations that parameters taking both shape and orientation into account will be useful classifiers.

To account for deviations from circular shape, the ratio between the longest cross-section and the corresponding orthogonal cross-section is calculated. The longest cross-section is here defined as the longest 8-connected path between any two boundary-pixels in a segment. A boundary pixel is a pixel situated on the segment boundary, that is, it will have pixel neighbours not belonging to its own image segment. Another requirement of the longest cross-section is that the segments centre-pixel is included in the cross-section. The centre-pixel is here defined as the image segments centre-of-gravity based on the pixel intensities. After the longest cross-section is identified, its corresponding orthogonal 8-connected cross-section is calculated. The orthogonal cross-section is identified as the 8-connected path between two boundary pixels having angles (in radians) closest to $\frac{\pi}{2}$ and $-\frac{\pi}{2}$ respectively. The angle in question is the angle between the boundary pixel, the centre-pixel and direction of the longest cross-section. The orthogonal cross-section is also required to pass through the centre-pixel. The ratio of the orthogonal (shortest) length to the longest cross-section is calculated and used as a parameter. For a perfect circular image segments the distances of these two cross-sections will be equal, and the ratio will be 1, while long thin image segments characteristic for cracks will have values closer to 0, depending on the distortion from circularity. It should be noted that, in theory, it is possible to have objects that deviate from circularity, but still have a ratio of 1. However, such objects are rarely seen in 2D-gels, and the described ratio (sometimes referred to as the Feret Ratio) is found a suitable classifier for 2D-gel image segments.

The angle the longest cross-section makes with respect to the horizontal or vertical direction in the gel-image is also a useful parameter, considering that artifacts (especially cracks) have a diagonal orientation in the image. Thus the smallest angle the cross-section makes with the horizontal or vertical image axes is also used as a parameter. Angles are first calculated with respect to the horizontal and vertical axes. The smallest of these two angles are then identified and used as a parameter. It should be noted that this single parameter does not distinguish which of the axes the calculation is based upon. Only the smallest angle is selected and compared later in the modelling stage.

Another parameter is calculated to account for variations in orientation. This is the ratio of the cross section through the centre-pixel in the vertical and horizontal direction to the overall longest cross-section calculated previously. The choice of this parameter is motivated by the fact that diagonal image segments will necessarily have a direction of its longest cross-section that deviates from the horizontal or vertical direction. If the image segment in question has a long, thin shape, but with an orientation parallel to the horizontal or vertical direction, its longest cross-section will be similar to the cross-section in the vertical or horizontal direction, and one of the calculated ratios will be close to 1. If the longest cross-section has a diagonal orientation, but the image segment in question has a circular shape, the length of all possible cross-sections will be equal, and both ratios will equal 1. However, if the object is not circular, and at the same time has a diagonal cross-section (which is typical for cracks), none of the ratios will be close to 1 (see figure 2.5(c)). Using the maximum of the two described ratios as a parameter will thus give an indication of both shape and orientation. Again this parameter does not distinguish whether the ratio is calculated with respect to the vertical or horizontal cross-section.

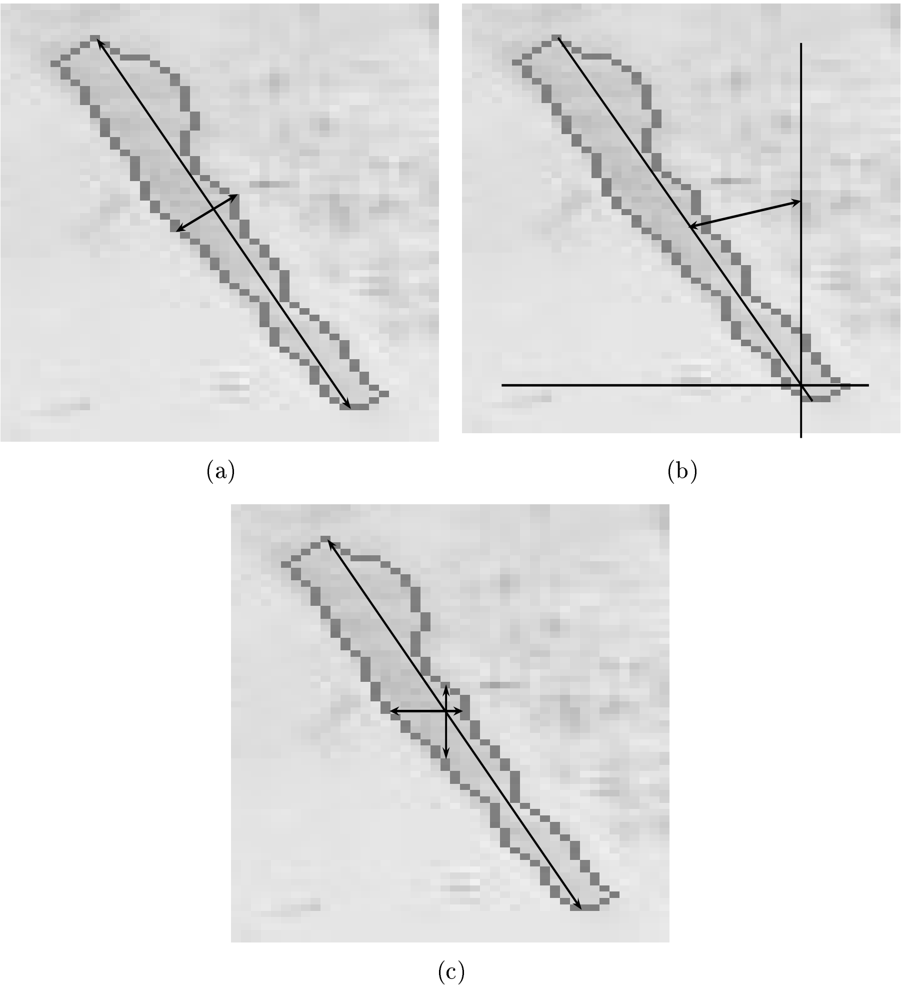
Finally it is convenient to combine the angle and orthogonal-to-maximum cross-section ratio, which gives an indication of both shape and orientation. Since diagonal, thin image segments have a low ratio value and a high angle value, the product of the angle and the inverse ratio is also used as a parameter. Parameters for shape and orientation used in this study are illustrated in figure 2.5.

Gaussian fit

Gaussian approximations to protein spots are commonly used in 2-DE [8, 27, 36]. The idea of using the Gaussian fit as a parameter is based on the assumption that all single, isolated protein spots should ideally display the shape of a perfect Gaussian peak. Thus the degree of Gaussian fit would give an indication of whether an image segment consists of a protein or an artifact. The Gaussian approximation is performed by fitting the following function to all pixels constituting an image segment:

$$G(x,y) = B(x,y) + I(x,y)\exp(-\frac{(x-x_c)^2}{2\sigma_x^2})\exp(-\frac{(y-y_c)^2}{2\sigma_y^2}) \tag{2.13}$$

Here $G(x,y)$ is the Gaussian approximation at image coordinate (x,y), $B(x,y)$ is the background intensity at (x,y), (x_c, y_c) is the coordinates of the centre-pixel, and σ_x and σ_y control the spread of the Gaussian function independently in horizontal and vertical direction. For deciding the optimal Gaussian fit, sum of squares differences between the original pixel intensities (with background subtracted) and the Gaussian approximation were minimised using the function *lsqnonlin.m* in the Matlab Optimisation toolbox (Mathworks). *Maximum Function Evaluations* were set to 50000, and *Maximum Iterations* to 10000. To



(a)

(b)

(c)

Figure 2.5: Illustration of the shape parameters. (a): The longest cross-section between two boundary pixels of an image segment and its orthogonal cross-section. Both is required to go through the centre-pixel of the image segment. (b): The smallest angle of the longest cross-section with respect to the horizontal or vertical axes. In this case the shortest angle is the one to the vertical axis, and is indicated by the arrows. (c): The horizontal and vertical cross-section through the spot-centre.

Table 2.1: Model parameters and abbreviations. “1” and “2” at the end of the texture parameters indicate the original and modified GLDM model respectively.

Abbreviation	Description
SIZE	Size in pixels
AVINT	Average intensity
CON1, CON2	Contrast
ASM1, ASM2	Angular Second Moment
ENT1, ENT2	Entropy
MEAN1, MEAN2	Mean (texture)
IDM1, IDM2	Inverse Difference Moment
RATIO	Ratio of orthogonal to the longest cross-section
ANGLE	Minimum deviation of angle from horizontal or vertical
GAUSS	Gaussian fit
RATAN	Inverse ratio times angle
RATXY	Maximum ratio to horizontal or vertical cross-section

make the Gaussian fit variable directly comparable between image segments of different size, the total deviation from Gaussian fit is divided by the number of pixels in the image segment. All parameters used for the spot-filtering model, together with their abbreviations are given in table 2.1

2.2.5 Multivariate data analysis

Partial Least Squares Regression (PLSR) [54, 55] is a common multivariate method used to build a regression model between a data matrix $\mathbf{X}$ with samples and variables, and a response or variable of interest contained in matrix $\mathbf{Y}$. PLSR makes use of the same basic principle as the well known Principal Component Analysis (PCA) [54], where the original data matrix is decomposed into a set of latent variables (called principal components) and noise. Principal Components can be understood in terms of scores and loadings. Loadings consist of the weights each original variable is given in each principal component, while the scores are the coordinates the samples are assigned in the new coordinate system defined by the principal components. The first principal component points in the direction of the maximum variation in the original $\mathbf{X}$ matrix. The information contained in the first component is then subtracted from $\mathbf{X}$, and the second component maximises the variation in the new $\mathbf{X}$. More components are calculated in the same manner, until there is no structure left in $\mathbf{X}$, and the components start to model noise. Such a decomposition of $\mathbf{X}$ has several advantages. First the number of latent variables is usually much smaller than the original number of variables. The latent variables are also independent (orthogonal) which is rarely the case for the original variables. Finally correlations among the original variables are easily visualised by plotting the resulting scores and loadings, making it easy to interpret relationships within and between samples or variables. PLSR decomposes the $\mathbf{X}$ matrix in a similar fashion, but in this case the decomposition is guided by one or several variables of interest contained in $\mathbf{Y}$. In PLSR it is the variation in the covariance matrix $\mathbf{X}^{\mathbf{T}}\mathbf{Y}$ which is maximised for each PLSR-component. The PLSR algorithm thus looks for variations in $\mathbf{X}$ that are relevant for the prediction of response $\mathbf{Y}$. A full description of PLSR and its algorithm is given in the references [54, 55].

In this analysis a special version of PLSR, called Discriminant PLSR (DPLSR) [54, 56] is used. What separates DPLSR from regular PLSR is that the response consist of logical ones and zeros (a sample either belongs to a group, or not), while continuous response variables are common in regular PLSR. It must also be noted that although the input values in DPLSR are discrete, its output, or predicted values, are continuous.

To avoid overfit and measure the predictive ability, PLSR models are often validated by a method called cross validation [54, 57]. The model is checked by leaving out samples from the calibration set, using them as temporarily test samples. A PLSR model is calculated using the rest of the samples, and the test samples are predicted using this sub-model. The procedure is repeated for all samples in the model. In this way all samples work as independent test-sets for the corresponding PLSR model. Based on the prediction accuracy of the test samples, the optimal number of principal components can be decided, reducing the risk of overfit. In this study a special version is used where only one sample is left out at a time. This method is often referred to as leave-one-out cross validation. In addition to cross validation the model can also be validated with a totally independent test data-set. Both forms of validation are used in this study.

2.2.6 Data set and Models

All image segments from the segmentation procedure were manually assigned class membership according to the classes listed in section 2.2.3. A total of 1310 segments were assigned in the calibration gel, and 527 in the test gel.

Only image segments which belong to noisy features, cracks and isolated protein spots are used to build the calibration model. This is because differences between protein spots and artifacts are best described by these three classes. A total of 709 image segments are selected for modelling, giving a calibration matrix $\mathbf{X}$ of 709 samples and 17 descriptor parameters (or variables) for each sample.

In this study there is only one, binary response variable $\mathbf{y}$. A value of one is assigned to segments belonging to noisy features and cracks, and zero is assigned to isolated protein spot segments. This produces a $\mathbf{y}$-vector of same length as the number of samples (709). It should be noted that during the modelling stage, segments caused by cracks and noisy features are treated as belonging to the same class, because they are both assigned the same value in $\mathbf{y}$. Thus the model does not try to predict which of the twelve classes listed in section 2.2.3 a segment belongs to, but merely whether it results from proteins or is caused by some unwanted artifact.

2.3 Results

After using leave-one-out cross validation, five PLSR-components were found to explain the significant variation in $\mathbf{X}$ with respect to the response factor $\mathbf{y}$. These components also explain 85% of the total variation in $\mathbf{y}$ and 82% of the total variation in $\mathbf{X}$, meaning that the original 17 variables have been reduced to five PLSR-components, which explains the co-variance between $\mathbf{X}$ and $\mathbf{y}$. These five components were used to build the DPLSR model, and all image segment (1310 from the calibration gel and 523 from test gel), are given a predicted score based on this model. A score close to 1 indicates that the image segment most likely results from noise or artifacts, whereas true proteins will have scores close to 0. The performance of the DPLSR spot filtering model is given in the next chapters for the different classes of image segments and some other interesting subsets.

2.3.1 Classes used in the calibration model

The results of the classification are shown in table 2.2 for the calibration data and in table 2.3 for the test data. There is an observed separation between image segments caused by unwanted artifacts (including cracks) and image segments consisting of isolated protein spots. This observation is valid both for the calibration and the test data. Only 6-9% of the image segments have a score in the uncertain interval 0.4-0.6, which also shows that the two groups are well separated by the DPLSR model. Selecting a threshold of 0.6 remove over 90% of the unwanted artifacts for these classes, and only 1% of the true spots are lost. To avoid loosing true spots altogether, 70% of the artifacts can still be removed by selecting a threshold of 0.8. The results are not significantly different for the calibration data and independent test data. It is therefore concluded that spot filtering using DPLSR successfully separates the selected classes with the most distinct features.

Table 2.2: Model performance for calibration data. The numbers are percentages of image segments (samples) assigned a model-score in the interval indicated to the left.

Class	Noisy Features and Cracks	Protein spots
Samples	306	399
Score: < 0.2	0.0	78.7
0.2 - 0.4	1.0	19.8
0.4 - 0.6	4.2	1.5
0.6 - 0.8	22.5	0.0
> 0.8	72.2	0.0

Table 2.3: Test set performance for same classes as in table 2.2.

Class	Noisy Features and Cracks	Protein spots
Samples	219	135
Score: < 0.2	0.0	74.1
0.2 - 0.4	1.8	20.7
0.4 - 0.6	5.5	3.7
0.6 - 0.8	18.3	1.5
> 0.8	74.4	0.0

2.3.2 Other classes

The DPLSR spot filtering model was also used to assign scores to the image segments belonging to the other classes listed in section 2.2.3 None of these image segments were used to build the model, neither in the calibration nor in the test gel. Calculated scores for these image segments are shown in table 2.4 and table 2.5 for the calibration and test gel respectively. It can be seen from the tables that the model assigned low scores to most of the contaminated, overlapping and saturated spots. The results are in accordance with the intention of the model, because these image segments contain protein information. They can thus be kept for further analysis. A few of the saturated spots were assigned higher scores, thus increasing the risk of removal after a threshold is

Table 2.4: Model performance for other classes. Calibration gel

Class	contam- minated spots	Over- lapping spots	indistin- guishable artifacts	Unclassified segments	Overlap protein/ artifact	Saturated spots
Samples	130	73	142	186	21	82
Score: < 0.2	33.1	79.5	6.3	32.8	38.1	76.8
0.2 - 0.4	45.4	19.2	12.0	32.8	23.8	17.1
0.4 - 0.6	20.0	1.4	40.8	28.0	28.6	3.7
0.6 - 0.8	1.5	0.0	31.7	5.9	0.0	1.2
> 0.8	0.0	0.0	9.2	0.5	9.5	1.2

Table 2.5: Model performance for other classes. Test gel

Class	contam- minated spots	Over- lapping spots	indistin- guishable artifacts	Unclassified segments	Overlap protein/ artifact	Saturated spots
Samples	21	22	69	29	15	35
Score: < 0.2	38.1	50.0	29.0	55.2	33.3	65.7
0.2 - 0.4	38.1	22.7	21.7	13.8	13.3	25.7
0.4 - 0.6	19.0	22.7	26.1	17.2	33.3	5.7
0.6 - 0.8	0.0	4.5	18.8	13.8	6.7	0.0
> 0.8	4.8	0.0	4.3	0.0	13.3	2.9

Table 2.6: Model performance for Unclassified segments. Calibration gel.

Class	Artifacts	Proteins
Samples	76	102
Score: < 0.2	9.2	52.9
0.2 - 0.4	32.9	32.4
0.4 - 0.6	46.1	12.7
0.6 - 0.8	10.5	2.0
> 0.8	1.3	0.0

Table 2.7: Model performance for Unclassified segments. Test gel.

Class	Artifacts	Proteins
Samples	7	15
Score: < 0.2	28.6	53.3
0.2 - 0.4	0.0	20.0
0.4 - 0.6	28.6	20.0
0.6 - 0.8	42.9	6.7
> 0.8	0.0	0.0

Table 2.8: Cracks and size variations for noisy features. Calibration data.

Size (pixels)	Less than 60	Between 60 and 250	Over 250	Cracks
Samples	158	96	33	19
Score: < 0.2	0.0	0.0	0.0	0.0
0.2 - 0.4	0.6	2.1	0.0	0.0
0.4 - 0.6	4.4	3.1	3.0	10.5
0.6 - 0.8	24.1	27.1	3.0	21.1
> 0.8	70.9	67.7	93.9	68.4

Table 2.9: Cracks and size variations for noisy features. Test data.

Size (pixels)	Less than 60	Between 60 and 250	Over 250	Cracks
Samples	103	77	16	23
Score: < 0.2	0.0	0.0	0.0	0.0
0.2 - 0.4	1.0	1.3	12.5	0.0
0.4 - 0.6	4.9	6.5	6.3	4.3
0.6 - 0.8	16.5	19.5	12.5	26.1
> 0.8	77.7	72.7	68.8	70.0

Table 2.10: Overall performance. Calibration gel.

Class	Artifacts	Proteins
Samples	524	786
Score: < 0.2	3.1	67.7
0.2 - 0.4	8.6	25.3
0.4 - 0.6	20.2	6.2
0.6 - 0.8	23.3	0.6
> 0.8	44.8	0.1

Table 2.11: Overall performance. Test gel.

Class	Artifacts	Proteins
Samples	295	228
Score: < 0.2	7.5	65.8
0.2 - 0.4	6.4	23.2
0.4 - 0.6	10.8	8.3
0.6 - 0.8	19.0	1.8
> 0.8	56.2	0.9

applied. Saturated spots can ,on the other hand, be easily identified earlier in the process by their high percentage of pixel intensities above the saturation limit, and do not necessarily need to be subjected to the spot filtering procedure.

No obvious separation is observed for the three remaining classes in this study, which is as expected. The classes of overlapping proteins and artifacts could naturally not be distinguished, which also goes for the unclassified segments, where the human observer was not able to decide whether the image segment contained a protein spot or an artifact. One would have preferred the class *indistinguishable image segments* to have higher scores similar to artifacts. But considering the criteria for this class described in section 2.2.3, the results make sense. The image segments of this class contain artifacts where the observer used other criteria than the appearance of the segments themselves to classify them. The classification was rather based on the surrounding environment and position in the gel. Because such criteria are not used in the descriptor, one does not expect the spot filtering model to handle these segments properly. Creating parameters based on these criteria is not straightforward, since such parameters are themselves based on results from the spot filtering model. One way to deal with such problems is to calculate parameters based on segments with high and low score assignments (above 0.8 and below 0.2 for instance). Thus new parameters, for example the number of spots/artifacts in the neighbourhood, can be used in a second re-classification of all segments, creating new scores. This procedure is continued iteratively until stable results are achieved.

The gels (especially the calibration gel) consist of a significant number of unclassified image segments. These segments were reclassified in such a way that the human observer, though uncertain, was forced to decide whether they contained a protein or an artifact. The results of this second classification are shown in table 2.6 and 2.7. (It should be noted that some of the unclassified segments were found to contain overlap between artifact and possible protein. These segments were kept out of second classification, and is the reason why the numbers in table 2.6 and 2.7 don't add up the numbers in column four in table 2.4 and 2.5.) As can be seen from the tables, there is no conclusive pattern for the two classes. Most segments tend to have lower values, indicating that they are mostly similar to protein spots, but the scores are far from clear compared with results for the real protein spots in table 2.2 and 2.3. It can thus be concluded that on image segments where the human observer was uncertain, the spot filtering model also produced uncertain results.

2.3.3 Size Variations and Cracks

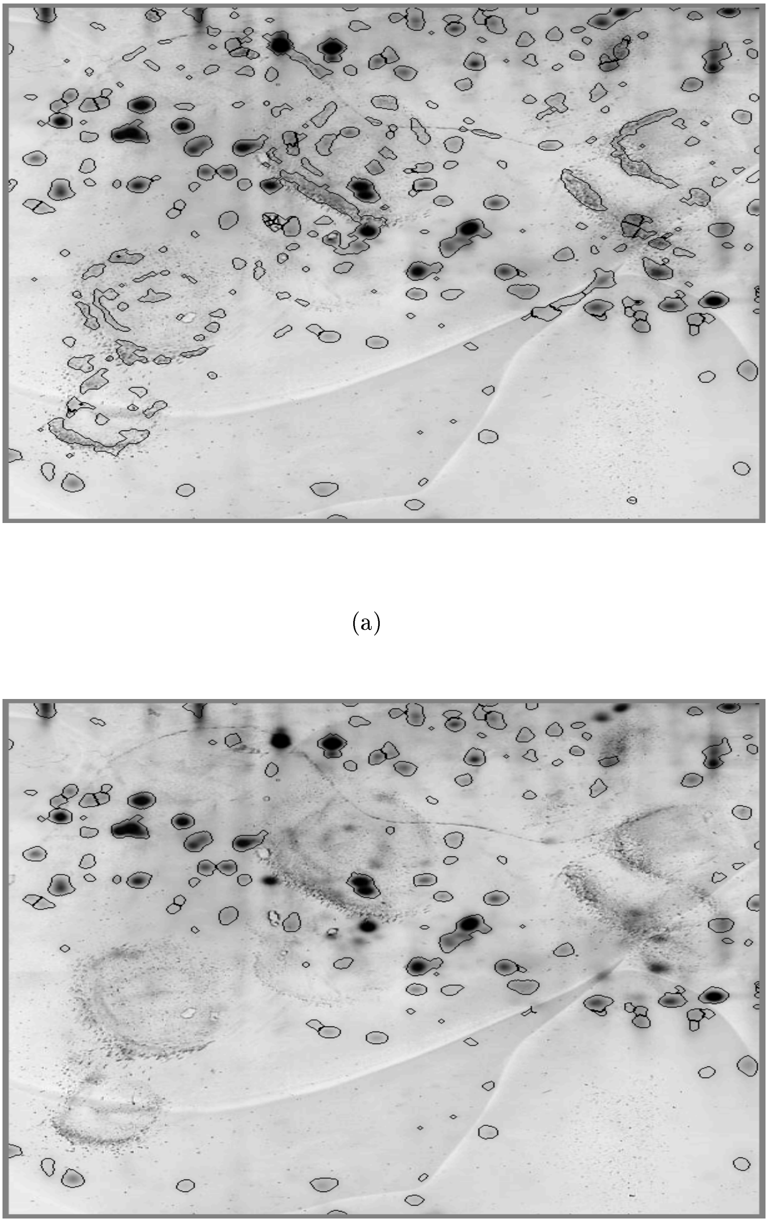
The image segments containing noisy artifacts and cracks were divided into sub-groups, investigating if the spot filtering model assigned similar values to all types of artifacts. The image segments containing noisy features were divided into three sub-groups depending on their size (in pixels). The group limits were set at 60 and 250 pixels. The results are given in table 2.9 and 2.10. Considering the results from both the calibration and the test gel there seems to be little variation between the sub-groups. That is, cracks are discriminated just as well as the noisy features. There is a deviation of the correspondence between the percentages of the larger noisy features. However, it should be noted that the number of samples for this sub-group is rather small for the test-gel, and the score of a single segment will thus have a large influence on the calculated percentages.

2.3.4 Overall performance

The overall performance of the spot filtering model is given in tables 2.10 and 2.11 for the calibration and test gel respectively. Here all classes of image segments are assigned to either protein spots or artifacts. It is concluded that the spot filtering model performs satisfactorily in removing most of the noise and artifacts, and at the same time avoiding removal of true proteins. Segmentation results for some critical areas in the gel before and after the spot filtering model is applied is shown in figure 2.6. There is some variation between the results from the calibration and test gel, which is expected because there will always be some variation between different gels. However, the general conclusions drawn from the calibration gel is also valid for the test gel.

2.3.5 Evaluation of parameters

Plots of scores and loadings for PLSR- component 1 versus PLSR-component 2 are shown in figure 2.7, and PLSR-component 2 versus PLSR-component 3 are shown in figure 2.8. Score and loading plots are a common way to visualise results in multivariate analysis, and should be interpreted simultaneously to investigate which variables in the loading plot contributes to the separation of samples in the score plot. Variables with high absolute loadings are the important variables, and have coordinates in the loading plot far away from the origin, and in the same direction as the sample-separation in the score plot. Less important variables are situated close to the origin, or in directions orthogonal to the separation. As can be seen from the plots of the first two components, there is a clear separation of two classes used in the calibration



(a)

(b)

Figure 2.6: Results of image segmentation. (a): Before spot filtering. (b): After spot filtering. The selected model threshold for removing image segments was set to 0.4 in this case.

model, especially along the first component. As can be seen from the corresponding loading plot, the texture parameters dominate the first component, while the other variables contribute less. It is thus concluded that the texture variables are most important in distinguishing noisy features from protein spots, which is as expected. Plots of the second and third component reveal another relationship. In these components parameters used to distinguish the cracks dominate, while the texture parameters are clustered around the origin. The artifacts resulting from cracks are highlighted with circles in figure 2.8(a), confirming that this separation is dominant in the second and third component.

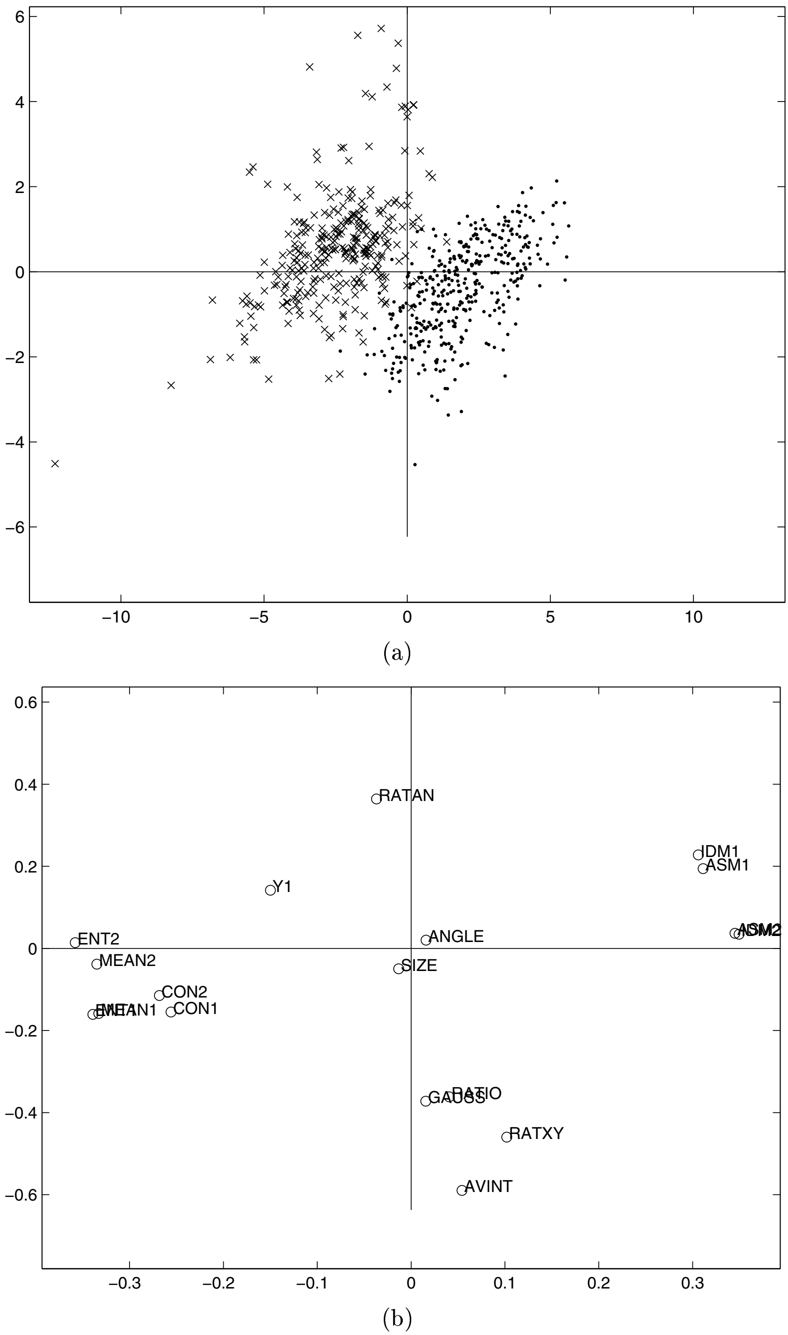
The Gaussian parameter contributes to the separation, especially in the second and third component. It should thus be included in the model. However, a separation based on the Gaussian fit alone will not be sufficient to separate the image segments in this study. Deviations from Gaussian shape is quite common for protein spots [2, 3, 5], so this result is not surprising. It is thus concluded that spot filtering models are greatly improved by evaluating more parameters than the Gaussian fit alone.

The loading plots also justify the inclusion of the modified GLDM texture parameters and the angle times inverse ratio parameter. Most of the former and also the latter contribute significantly to separation without being too correlated to other parameters.

2.4 Discussion

Some suggested improvement is already mentioned in the previous chapter. Here we discuss some aspects of the general validity of the multivariate model presented in this paper.

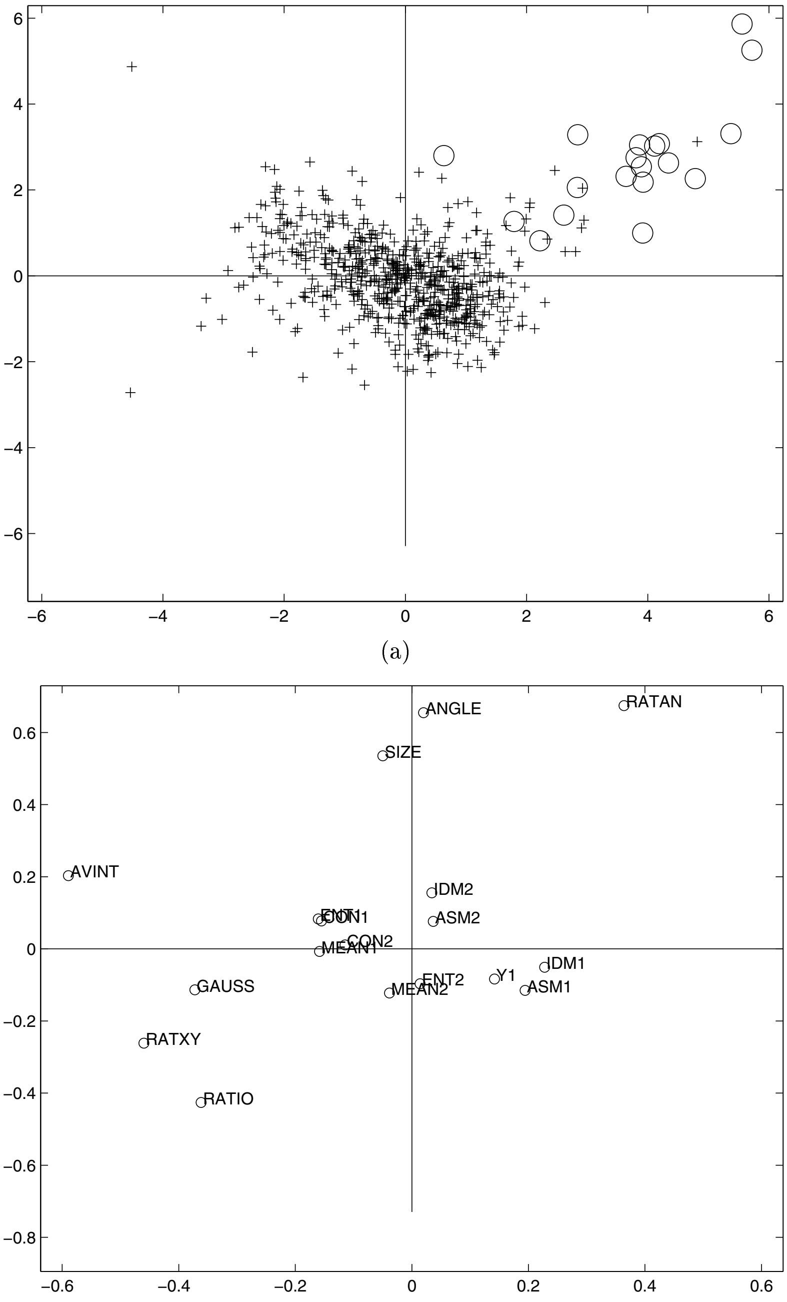
The model managed to produce good results for both the calibration gel and the independent test gel used in this study. Though the two gels are from different biological samples, they also display many similarities. They were both silver-stained, and have approximately the same resolution in size and depth. Gels scanned at higher resolutions, and stained by other methods may perform less favourable when submitted to this particular spot filtering model. The ideal spot filtering model would account for all possible noise and artifact variations for all possible gels at multiple resolutions and differing staining methods. To achieve this, a large database of so called “ground truth” 2D-gels will be necessary, comprising example gels of the different variations. Unfortunately, such a database does not exist, and the performance of the model may thus be poor when samples from other gels are used . For instance image segments



(a)

(b)

Figure 2.7: Plots of scores and loadings. (a): Scores of the first (horizontal) and second (vertical) principal component. Artifacts and protein spots are marked by crosses and dots respectively. It is a clear separation between the two classes. (b): Loading plot of the first two components. Parameter abbreviations are taken from table 2.1.



(a)

(b)

Figure 2.8: (a): Score plot of the second (horizontal) and third (vertical) principal component. Image segments resulting from cracks are marked with circles, and are mostly separated from the other samples. All other image segments, both protein and artifacts, are marked by crosses (b): Loading plot of the second and third component. Parameter abbreviations are taken from table 2.1.

produced from a higher resolution gels may have different texture properties, and thus the presented model will not be valid for these data. However, it is always possible to build local models, representing gels created with a specific procedure at a particular laboratory. For biologist using only gels produced by this equipment, a local spot filtering model can be a useful additional tool to clean the gels from spurious spots and noise, improving the output from the final data analysis.

2.5 Concluding remarks

The multivariate spot filtering model introduced performed successfully in separating image segments resulting from noisy artifacts and cracks, from image segments consisting of protein spots. The concept of evaluating image segments after the segmentation procedure is in general a useful method for reducing the number of unwanted artifacts in 2D-gels.

2.6 Acknowledgements

We would like to thank the people at Norwegian Food Research Institute (MATFORSK) for providing the gel material used for the analysis in this study. We would also like to thank The National Programme for Research in Functional Genomics in Norway (FUGE) for funding.

Chapter 3

A new method for assigning common spot boundaries for multiple gels in two dimensional gel electrophoresis

Morten B. Rye, Ellen M. Færgestad and Bjørn K. Alsberg

Department of Chemistry
Norwegian University of Science and Technology
N-7491 Trondheim, Norway

This chapter has been recommended for publication in Electrophoresis.

Abstract

The benefits of defining common spot boundaries when several gels from 2-DE are compared and analysed have lately been stressed by both commercial software producers and users of this software. Though the importance of common spot boundaries is clearly stated, few reports exist that target this issue explicitly. In this study a method for defining common spots boundaries is developed, called the spot density method. The method consists of the following steps: Segmentation and spot identification on each individual gel, transferring the spot centre coordinates for all gels onto a single new gel, collecting spot-centres clustered together in the new gel and finally assigning pixels and new spot boundaries based on the spots in each cluster. The method is compared to a synthetic gel approach, and validated by visual inspection of three representative areas in the gels. The gel images need to be aligned prior to segmentation and spot identification, but the method can be used regardless of the choice of segmentation procedure. This makes the method an easy extension to existing methods for spot identification and matching. Conclusions based on the visual inspection are that the spot density method identifies both partly overlapping spots and low intensity spots better than the synthetic gel approach.

3.1 Introduction

Ever since the first attempts to create a fully automated analysis in 2-DE, the task of finding corresponding protein spots between different gels has been a major challenge. The usual approach has been to identify protein spots on each gel individually, followed by an algorithm or method to find corresponding spots in all individual gels [9–13, 27, 37, 47]. These methods, commonly referred to as spot matching methods, often use a master or reference gel, which all other gels are compared to. After all spots are matched, the volume and intensity of the corresponding spots can be compared, and proteins differentially expressed can be highlighted.

Spot matching procedures usually make use of a spots position on a gel, together with its local neighbourhood to determine the most likely match. However, several problems arise when this approach is used. Global and local perturbations in the analysed gels complicate the matching procedure significantly. This is especially true in regions where a large number of protein spots are present, resulting in partly merged protein clusters. Two highly overlapping spots might appear as a single spot in one gel, but are detected as separated spots in another gel. An example of such a situation is shown in figure 3.1, where images and spot boundaries are produced using the commercial software ImageMaster 2D Platinum Version 6.01 (GE Healthcare).

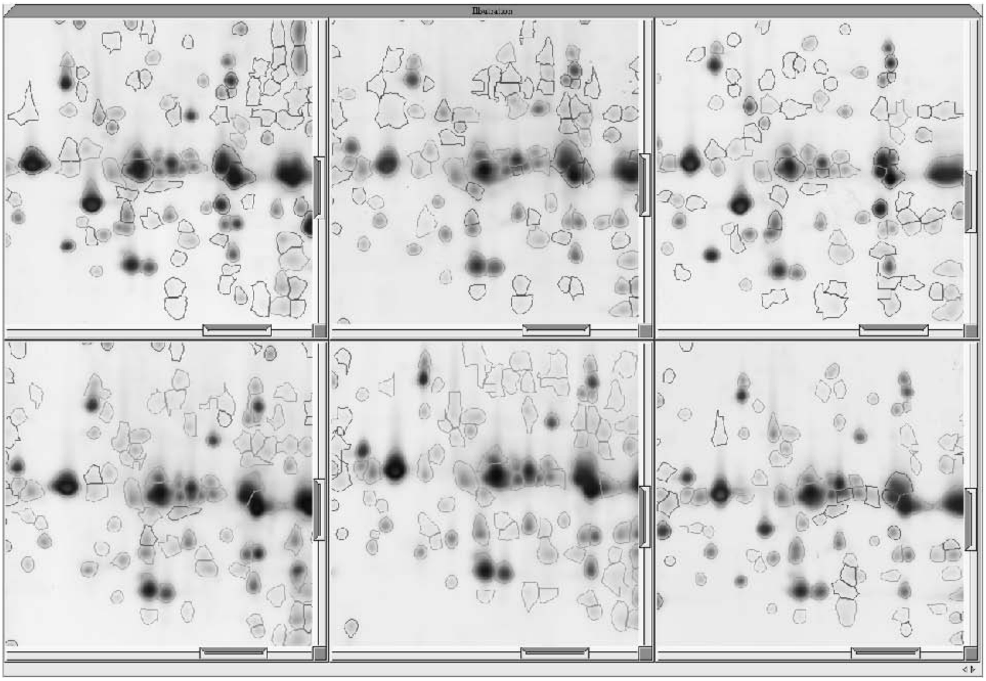


Figure 3.1: Comparable areas from six of analysed gels. Spot boundaries for each gel are defined using the software ImageMaster. Finding complete spot matches across all gels are extremely challenging.

Individual spot boundaries for a corresponding area in six different gels are displayed for a highly clustered region. As can be seen from the figure, it is impossible to find a satisfactory match between all six sub-images. Situations similar to this usually result in the insertion of missing values in the final spot volume data table, leading to serious errors in the following data analysis. The pitfalls of erroneously estimated spot volumes and inserted missing values are lately well described by Færgestad et al. [64]. The number of spots containing a missing value in a match set can be quite large, as shown in a recent study by Grove et al. [40]. It was found that as much as 80% of the spots collected in the match table for analysis are subject to missing values. The number of missing values also escalates with the number of gels compared and analysed. Using a high number of replicate gels does not necessarily produce better results, because of the poor reproducibility of spot boundaries in each individual gel.

One natural way to address this problem is to define a common set of spot boundaries for all gels. The advantages of this approach has lately been stressed by several commercial software packages, especially Progenesis SameSpots (Nonlinear Dynamics) and Delta-2D (DECODON). The view that common spot boundaries are a better approach to the matching problem is also shared by the authors of this paper. When spots are compared between gels, a protein not present on a particular gel will have a volume or intensity close to zero inside the common boundary, corresponding to a true missing value in the spot matching approach. The immediate advantage with this approach is, of course, that matches are always accurate, because all boundaries are identical and represent the same area in all gels. However, for this assumption to be useful, all gels need to be properly aligned before the boundaries are identified and spots compared. Fortunately there exist alignment procedures, both in commercial software and published in the literature [13, 14, 29, 31, 32, 35] that satisfy this assumption, meaning it is possible to warp and transform all images subject to analysis such that the spots compared occupy the same areas in all gels. The next issue will then be how to define the common spot boundaries. Luhn et al. [41] use the concept of a synthetic gel, which is a gel image created by combining pixel intensities from all gels used in the analysis. This is the approach used by the software Delta-2D. A pixel intensity in the synthetic gel is calculated by a weighted average over the corresponding pixel intensities in the individual gels. The synthetic gel is then subject to a spot identification (or segmentation) procedure, and the boundaries identified in the synthetic gel are used for all gels.

To use a synthetic gel image to calculate common spot boundaries is easy and straightforward, but also has some disadvantages. First the boundaries for all gels are based on the segmentation of only one (synthetic) gel. The segmentation of a single gel is almost always subject to errors, and these errors are thus propagated to all gels. Secondly, including spots present on all gels result in an increased number of spots in the synthetic gel, introducing more clusters and merged spots. Thus two spots clearly separated and isolated in individual gels, might be identified as a single spot in the synthetic gel. At last noise and other artifacts not related to proteins are usually present in at least some of the analysed gels. Since the synthetic image comprise information from all gels, these artifacts will also be present in the synthetic image, and are thus also propagated to all gels. It should be mentioned that the impact of the last problem can be reduced by adjusting the pixel weights in each gel during the synthetic gel creation.

In this study we present an alternative approach to define common spot boundaries in a set of 2D-gels subjected to analysis. Instead of using a single synthetic gel for defining the spot boundaries, information from spot boundaries in all gels are used. Spot boundaries are first defined individually for each gel as for the traditional spot matching approach. However, instead of trying to match the produced spot segments, the spot boundaries from each gel are combined to produce a common set of spot boundaries for all gels. The method presented is compared to the synthetic image approach, and validated here by visual inspection of three representative spot clusters from a set of silver stained 2D-gels. Visual results from the full gels are not shown in this article because of resolution capabilities, but can be downloaded as supplementary material.

3.2 Materials and methods

3.2.1 2D-gels

The gels used in this study include 7 Norwegian Red dual-purpose bulls from a performance test station (GENO-Breeding and AI Association) slaughtered at approximately 13 months of age/450 kg live weight in 2004. Muscle samples from the Longissimus dorsi were collected one, two, three, six and ten hours after slaughter, giving a total of 35 gels used in the experiment. The samples were immediately frozen in liquid nitrogen, and proteins were extracted in TES buffer (10 mM Tris (pH 7.6), 1 mM EDTA and 0.25 M sucrose) and analysed by 2-DE.The analytical gels were stained by silver staining. The 2-DE gels were then scanned using an office scanner (Epson Expression 1680 Pro, Epson) with 8-bit colour depth and a resolution of 240-dpi. To remove spikes and noise, all gel images were filtered using a median filter of size 3 pixels in both directions. Gel alignment was performed using the commercial software TT900 S2S (Nonlinear Dynamics Ltd.; www.nonlinear.com). Unless stated otherwise, all program code are written in Matlab version 7.2.0 (Mathworks).

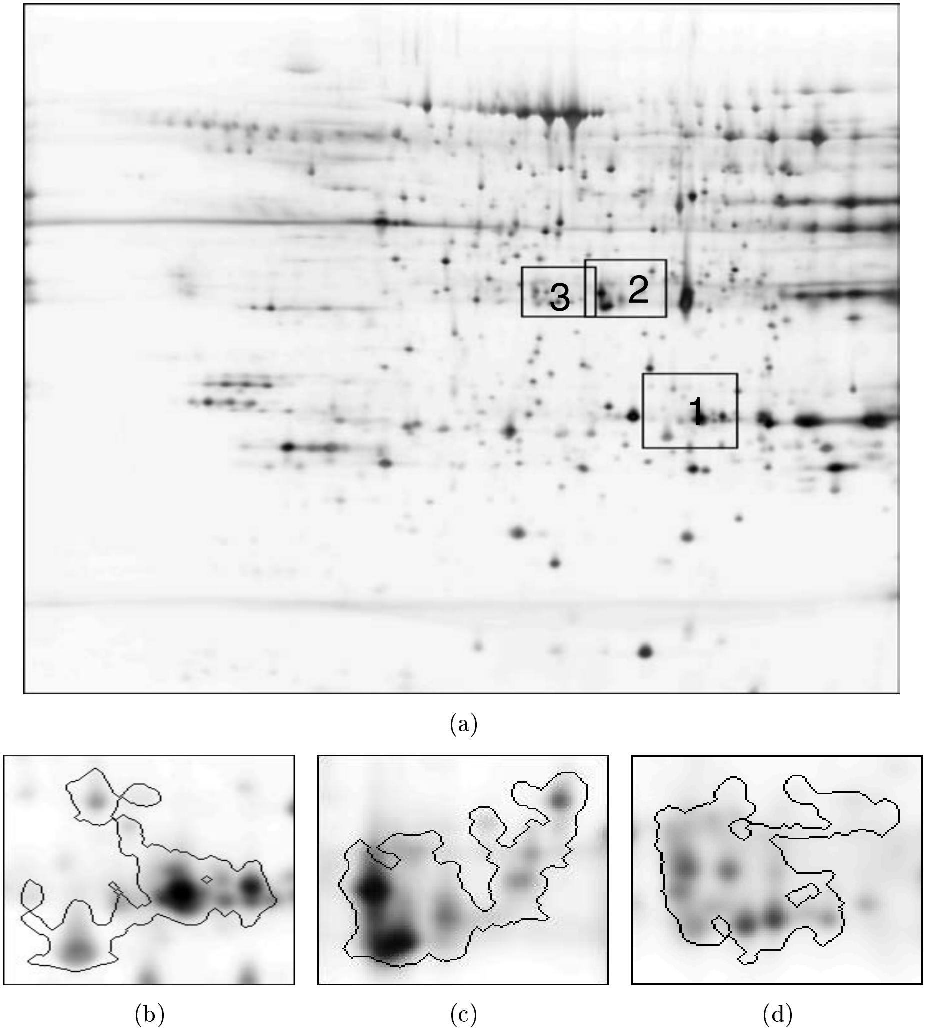
The selected image regions used for visual validation are displayed in figure 3.2, and the spot clusters within the boundaries in figure 3.2(b) to 3.2(d) are the focus for visual inspection. The spot cluster shown in figure 3.2(b) is also visible in figure 3.1.

3.2.2 Spot Identification

In the following description it is assumed that all images are inverted, that is, the image background is dark, and the spots appear as light peaks rising from the background.

To define common spot boundaries by our approach, a method is first needed to identify protein spots on each individual gel. This task is performed using image segmentation procedures. The goal of image segmentation in 2-DE is to separate the background and noisy areas in the gel-images from areas resulting from protein content, and to divide the latter areas into as many individual protein spots as possible. Because of the many subtle variations within 2D-gels, this is not a trivial task, and several methods and approaches are reported in the literature. (See [13] for a good overview of these approaches). Here a three-step procedure is adopted, all based on previously reported ideas. First all areas in the image resulting from protein spots are identified. Streaks are also removed at this stage. Though streaks are caused by the presence of proteins, they are not useful for protein identification and quantification, and should be removed. Each of the identified protein spot areas will normally include several spots, especially in clustered spot-regions, so these areas are further resolved into individual protein spots. Finally boundaries of small spots are redefined. A more detailed description of each step is given below. To visualise each step in the procedure, sub-images are shown for the different stages in figure 3.2 and 3.3.

The first step (streak removal and identification of protein spot areas) is performed using image morphology. Image morphology is an effective segmentation method, and is performed by successively dilating and eroding images with structural elements similar to the features in the image one wants to keep or remove. The use of morphology in 2-DE and for images in general is well described by Skolnick [49] and Sternberg [48]. The selected structural element for streak removal is a line 61 pixels in length, and for identification of protein



(a)

(b)

(c)

(d)

Figure 3.2: Areas in the gel used for visual validation of the presented method. Figure (b), (c) and (d) are the highlighted areas marked 1, 2 and 3 in the larger image. Figure (a), (c) and (d) also display the boundary for clusters selected for visual validation.

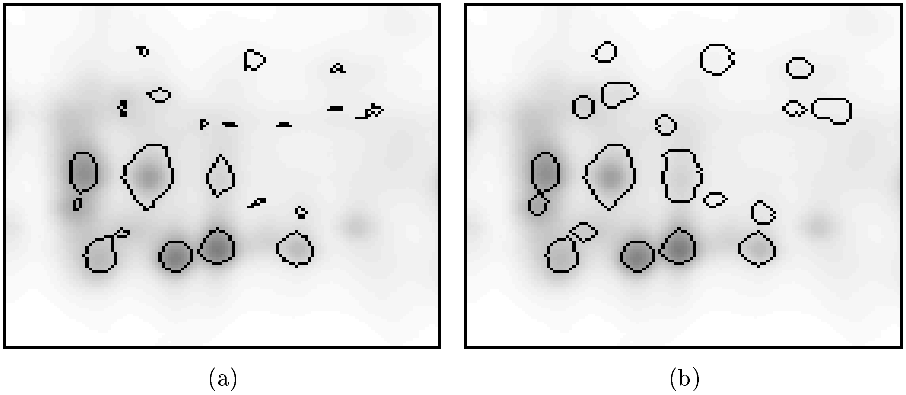


Figure 3.3: (a): Boundaries after resolving the partly overlapping spot cluster in figure 3.2(d). (b): Redefined spot boundaries.

spot areas a circular disk with a radius of 20 pixels is chosen. After dilating and eroding the gel-image with these structural elements, a single threshold is sufficient to highlight the spot areas. The final threshold value is set to 0.025 for image intensity values between 0 and 1. The identified spot areas after morphology are shown for the sub-images in figure 3.2. As can be seen from this figure, most of the identified areas consist of several protein spots, especially in the clustered regions. These regions also include overlapping spots, so some method is needed to resolve these areas into individual spots, which is the second step of the spot identification procedure. Except for a proposed method to identify individual spots in highly saturated areas [65], no studies have been found addressing the challenge of overlapping spot clusters explicitly. However, several methods for resolving overlapping spots have been reported together with general segmentation procedures [7, 9, 45, 47]. Here we have adopted an approach based on a method used by Cutler et al. [46] to identify the individual spots in overlapping spot clusters. The method is only dependent on the image grey values, and uses no derivatives. Identification of isolated spots is achieved in a method called pixel value collection, and is performed by examining the image at each of its intensity planes, starting from the highest intensity values and working down. At each level the contour of the image is analysed, and adjacent pixels are collected as the protein spots grow from the highest to the lowest intensity value. A spot stops growing when its shape and size no longer satisfy some predefined protein spot criteria, and the spot boundary is defined at this level. In our study this method is used to identify individual spots in the clusters defined by the morphology approach. The method is also modified to work the other way around, that is, starting at low levels and working towards higher levels. Merged areas are split, until no more splits are possible, and the boundary is defined as the level where the last split happened. No criteria are defined for accepting spots after a split, other than that it should consist of at least 5 pixels. If one of the spots after a split consists of less than 5 pixels, the split is rejected, and the boundary is defined at the level where the last split appeared. The individual spot boundaries after resolution are shown in figure 3.3(a).

The chosen method for resolving spot-clusters is quite noise-sensitive, meaning that irregular or noisy spot surfaces results in spurious spots with shape and spot-boundaries that differ from the ideal shape. This is especially true for small spots with size less than 30-40 pixels. Apart from disregarding noisy segments with size less than 5 pixels during the resolution procedure, accepted image segments with size less than 100 pixels were subject to an additional procedure to redefine the spot boundaries. To refinement method is based on constructing a window around the spot in question, and defines a threshold used inside the window based on the intensity values situated on this window. This is an approach similar to the one presumably used by the commercial software ImageMaster as mentioned by [50], and works as follows. The lowest intensity value inside the original spot boundaries is first identified. Secondly the mean of all intensity values situated on the window boundary below this lowest value is used as a new threshold. The spot is then redefined as consisting of all pixels inside the window higher than the new threshold. Spots in the sub-image after running the refining method is shown in figure 3.3(b). In this study a window-size of 2 pixels outside the original spot pixel-coordinates is used. Finally, to correct for irregular boundaries, all spot shapes are smoothed by morphology using disks with radii of 1 and 2 pixels successively as described by Skolnick [49]. After the individual spot shapes are defined within each cluster, the 4-connected boundary is identified, and the spot-centre coordinate is calculated using the geometric mean.

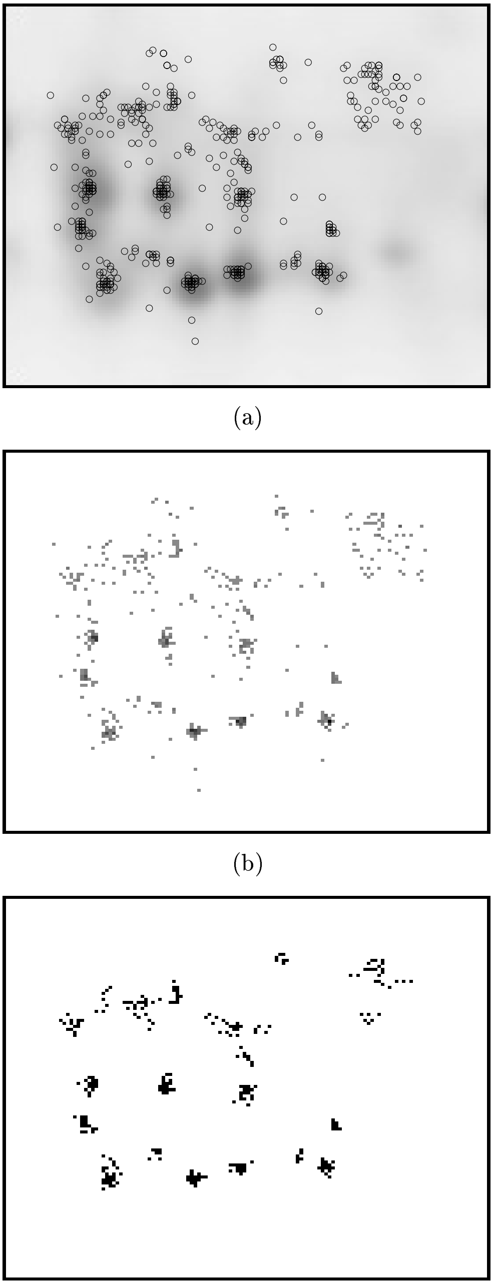
Sometimes more sophisticated methods for spot identification are used by commercial software packages. In many circumstances these software packages outperform the methods mentioned above. However, since the software is commercial, their methods and algorithms are not published, and the results are thus difficult to reproduce. The method for finding common spot-boundaries, which is the main topic in this study is not restricted to the choice of segmentation method. Generally any method for spot identification can be selected, as long as it provides a spot boundary and spot-centre coordinates for each spot.

3.2.3 Defining common spot boundaries

After all individual spot boundaries and the spot centre coordinates have been defined for all gels, common spot boundaries are calculated by the following procedure. First a spot density image is constructed based on the spot centre coordinates of all spots in all images. From the density image, spots are selected based on the clustering of spot centres, and a common boundary is assigned to each cluster of centres based on image pixels from the individual protein spots.

The spot density image is constructed by adding spot-centres from all images used in the analysis onto a new image. The image has the same size as the analysed images, and starts with only zero values. If a spot centre in one of the images is found at coordinates (x, y), an intensity value of one is added at coordinates (x, y) in the spot density image. All spot centres are thus accumulated in the new image by raising the intensity value in the spot density image by one at (x, y) each time a spot centre appears at this set of coordinates. If, for example, three spots from different gels have identical spot centre coordinates, the intensity value will be three in the spot density image at this coordinate. The idea is that spots appearing in several (or all) images will have similar spot centre coordinates. These spot-centres are clustered together as shown for the example image in Figure 3.4(a). The spot density image for this area is shown in Figure 3.4(b), where dark pixels indicate that several spots share this particular spot-centre.

Spot centres clustered together can be collected, and new spot boundaries are constructed by combining the spot boundaries from all spots belonging to a cluster. The collection of spot-centre belonging to comparable spots is guided by two parameters. The first parameter decides how close two spot-centres must be to belong to the same cluster, and depend on the resolution of the image, expected spot size, and precision of the alignment method. Another parameter is the number of spots necessary for the cluster to be included as a final spot. Most gels are subjected to noise and spurious image segments not related to proteins, and these segments should not be included among the final spot boundaries. These segments are, however, not present repeatedly in all the gels, and the centre of these segments will usually turn up as isolated single spot-centres in the spot density image. Protein spots subjected to analysis are, on the contrary, present in many gels with similar spot-centre. Thus a cluster with many spot-centres close together indicates the presence of a significant protein spot, while few isolated spot-centre indicates the presence of noise and



(a)

(b)

(c)

Figure 3.4: (a): Peak centre identified after the segmentation procedure has been performed on each gel individually. (b): Spot density image. (c): Accepted clusters of spot centres.

structures inherited from only one or a small number of the gels. Since one want to accumulate information common for most of the gels, a threshold is useful to decide whether a cluster should be included in the final model or not. After the significant clusters are identified, the spots in each cluster are combined to decide the common spot boundary for each cluster. The common pixels (and boundary) are assigned in the following way:

1. The spot pixels from the original segmentation for each spot in all clusters are collected.
2. The average size (in pixels) over all spots in each cluster is calculated.
3. All clusters are sorted in an ascending order according to their average size.
4. For each cluster pixels are assigned as the union of all pixels belonging to the spots constituting the cluster. The cluster with the lowest average size is assigned first, followed by all clusters in ascending order.
5. For each time new pixels are assigned a check is performed if the pixels are not already assigned to another cluster. Pixels assigned to another cluster are not assigned to the cluster in question.
6. Finally boundary pixels closer to another cluster-centre (that is, the average of the centres constituting a cluster) than the cluster-centre in question are discarded.

After pixels are assigned to all clusters in the spot density image, the spot boundary for each cluster is identified as described previously. This new boundary will then constitute a common spot boundary used for all gels in the analysis.

3.2.4 Selected parameters

In the present study the two parameters mentioned in the last section are determined manually. All spot centres separated by one pixel or less in the spot density image are said to belong to the same spot, and all spot-centres connected in this way constitutes a final spot for which the common spot boundary needs to be decided. This means that all spot-centres connected within the 24-neighbourhood of each spot are said to belong to the same final spot. The spot-centres are collected using the algorithm presented by [10].

The second threshold parameter concerning the minimum acceptable size of a spot-centre cluster is set to five, meaning that a spot must appear in at least five gel-images to be assigned a common boundary. Otherwise the cluster is regarded as noise and omitted from the final set of spot boundaries. The significant spot-clusters for the example image are outlined in figure 3.4(c). The two selected parameters are found to work well for this study, but can easily be changed depending on the circumstances and size of the experiment.

3.2.5 Synthetic Gel

To validate the results of the presented method, the sub-images were compared with the synthetic gel approach described by Luhn et al. [41]. This is also the approach used by the commercial software Delta-2D. The synthetic gel approach creates an image based on weighted pixel intensity averages from all gels. The synthetic image has the same appearance and size as a real gel, and should include all important spots present in all gels. The intensity of each pixel in the synthetic image is calculated by the following formula

$$I_{syn}(x,y) = \frac{\sum_{i=1}^{n}(w_i(I_i(x,y))I_i(x,y))}{\sum_{i=1}^{n}(w_i(I_i(x,y)))} \tag{3.1}$$

where I_{syn} is the synthetic image, I_i is an arbitrary gel image in the analysis and n is the total number of gel images. The weights $w_i(I_i(x,y))$ are defined by the function

$$w_i(a) = g_{max} - a \tag{3.2}$$

where g_{max} is the maximum intensity value in the image and a is an arbitrary intensity value.

The resulting synthetic image is subjected to the spot identification procedure, and the resulting spot boundaries of the synthetic image are used as common boundaries for all gels.

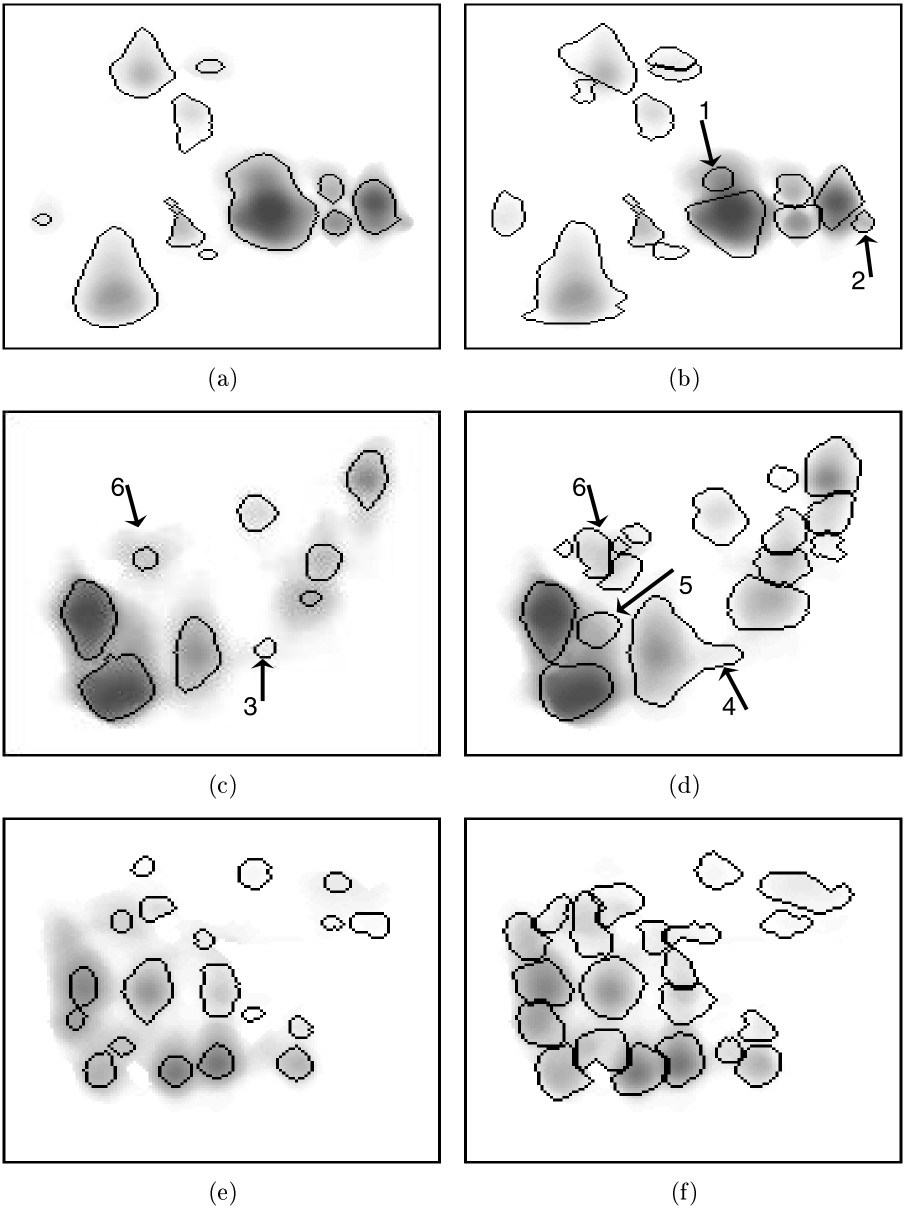
The results using the synthetic gel approach are compared to the spot density approach, and validated by visual inspection of three randomly selected spot-clusters.

3.3 Results and discussion

Three different areas in the image are analysed using both the synthetic gel and the spot density approach. The analysed areas are displayed within the frames in figure 3.2. Results from the full gels can be downloaded as supplementary material. Validation of results in this study is thus qualitative, and visual inspection decides whether the resulting common spot boundaries are sensible and reproduce the impression of spot appearances when gels are analysed manually. The validation areas are all selected where spot identification and definition of spot boundaries are challenging, including highly overlapping spots in complex regions of the gel. Results produced in these regions are thus considered representative for the gel as a whole.

The resulting common boundaries are displayed for the three different areas in figure 3.5 using the spot density approach and the synthetic gel approach. The synthetic gel is the underlying image in all figures. Some observations are immediately evident when inspecting figure 3.5. Though the main boundary features are the same, the spot density approach produces a higher number of segments than the synthetic gel approach, and each segment covers a larger area in the gel. The latter observation is expected, since each spot consist of a union of the pixels from the original spot identification. The increase in the total number of spots is mostly due to the ability for the spot density method to identify weak spots or spots partly overlapping with larger spots or each other, which the segmentation of the synthetic gel fails to register. Spot number 1 and 2 in figure 3.5(b) are good examples of this situation. The presence of spot number 5 in figure 3.5(d) is more uncertain, and it cannot be concluded from the synthetic image whether the inclusion of this spot is justified. However, inspection of the individual gels, as shown in figure 3.6, leaves no doubt about a spot in this region. This is a good example of an advantage of the spot density method to the synthetic gel approach. Introducing all spots in the synthetic gel increases the degree of overlap, which complicates the detection of isolated protein spots. Small spots close to larger spots separated in many of the individual gels tend to merge in the synthetic gel, and weak protein spots present in only a few gels tend to have intensity towards zero, resulting in a failed detection of these spots. Because the spot density approach uses information from the individual gel segmentation, spots only need to be identified in some of the gels (five in this case) to be detected.

In some cases the increased number of spots may be caused by a misalignment of the individual gels. The result of this is that identical proteins are detected at different positions in the various gels, and registered as independent spots



(a) (b) (c) (d) (e) (f)

Figure 3.5: Results from the spot density method, (b), (d) and (f) compared with the results using the synthetic image approach, (a), (c) and (e). The images displayed are the calculated synthetic images. The spots pointed to by arrows are explained in the text.

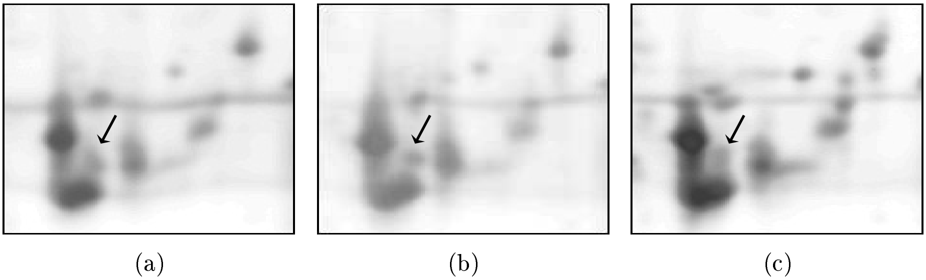


Figure 3.6: Three individual gel images of the area in figure 3.2(c). The images clearly show the presence of a spot at the end of the arrow, confirming that this spot should be included in the final boundary image, as shown in figure 3.5(d) for spot number 5.

in the common boundary image. One possible example of this situation is spot cluster number 6 in figure 3.5(c) and 3.5(d). Theoretically, this problem should cause more spots to be detected using both approaches. However, this particular case seems to be handled better by the synthetic gel, where only one boundary is detected. It should be noted that some of the individual gels have more than one spot detected in this area (results not shown), so whether the possible over-segmentation is the result of misalignment, or is actual due to the presence of several individual weak spots cannot be concluded with certainty.

The strange boundary of spot number 4 in figure 3.5(d) is the result of a streak which was not removed in two gels during the first spot identification procedure. Because this streak is present in at least one of the spots in this cluster, it is also included within the final common boundary. In this case the synthetic image performs better, and the streak is not included, though a weak spot (number 3 in figure 3.5(c)) is identified at the end of the streak. One way to avoid this situation in the spot density approach is to be more selective to pixels included in the final spot. In this study the union of all pixels for all spots in one cluster are included. However, a threshold can be defined, demanding that a pixel is present in more than one of the individual spots to be included within the final spot.

Images of spot boundaries for full gels, both using synthetic gel and the approach presented in this study can be downloaded as supplementary material. Some results regarding the total number of spots identified in these gels are of interest, and are thus mentioned here. The number of spots in each gel identified by the procedure outlined in section 3.2.2, are normally distributed from 826 to 1539 with a mean of 1118 for the 35 gels analysed. The number of spots identified in the synthetic gel using the same approach is only 644, clearly indicating that information has been lost when going from the 35 original gels to a single synthetic gel. The final number of spots identified by the approach presented here is 1316, which is more in accordance the number of spots in each individual gel.

The validation in this study is qualitative rather then quantitative. The areas selected for closer inspection should be representative for the gel as a whole, and the conclusions drawn from these areas valid in the task of defining common spot boundaries in general. Because results in 2-DE are often influenced by subjective visual inspection, and spot boundaries often defined with heavy user interaction, which measures are difficult to quantify, the visual qualitative validation is preferred to the quantitative approach.

At last in this study some of the choices made according to the pixel assignment procedure in section 3.2.3 need to be pointed out and explained. The reason for sorting the clusters in ascending order is to avoid the domination of larger spots. A large spot from one of the individual gels might consist of several, possibly overlapping protein spots. Overlapping protein spots are often regarded as a single protein spot in some gels, but split into several isolated spots in other gels. If the splits are similar in a substantial number of gels (at least five in this study), they should be included in the final boundary image. This is achieved by assigning pixels to the clusters with the lowest average size first. An example of final boundaries using sorted and unsorted clusters are shown in figure 3.7. The final step in Section 3.2.3 is included to avoid spots with discontinuous pixels in the final image. The presence of small spots in the vicinity of larger spots, might split the latter so pixels are no longer connected, or create small spots as holes in the larger spots. An example of a hole is shown in figure 3.7. To avoid this situation, pixels closer to the centre of the smaller spot are removed in the larger spot.

Generally the authors feel the presented spot density approach offers an improved alternative to existing methods for defining common spot boundaries. The spot density approach improved the identification of spots only present in a few of the gels, overlapping spots and small spots in the vicinity of larger spots compared to the synthetic gel approach. The method can also be used with any segmentation procedure, and has only two (or possibly three) parameters that need to be defined by a user. These depend on the resolution of the image, the number of gels in the experiment and the precision of the gel alignment. It is believed that when implemented in a user friendly environ-

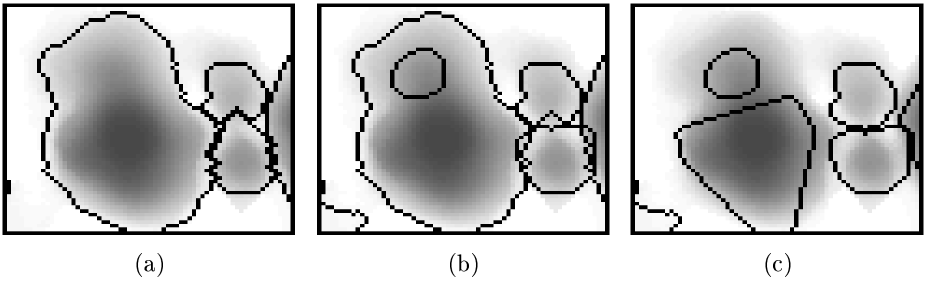


Figure 3.7: (a): Consequence of not sorting the spots in ascending order. The small spot in the upper left corner of figure 3.7(c) is completely dominated by its larger neighbour. (b): Consequence of omitting the last step in the pixel assignment procedure. The small spot creates a hole in the larger spot. (c): Final result when using both sorting and the last pixel assignment step.

ment, the method will simplify and improve the identification and comparison of protein spots in a multiple gel experiment.

3.4 Acknowledgements

We would like to thank Xiaohong Jia at Norwegian Food Research Institute (MATFORSK) for providing the gel material used for the analysis in this study. We would also like to thank The National Programme for Research in Functional Genomics in Norway (FUGE) for funding.

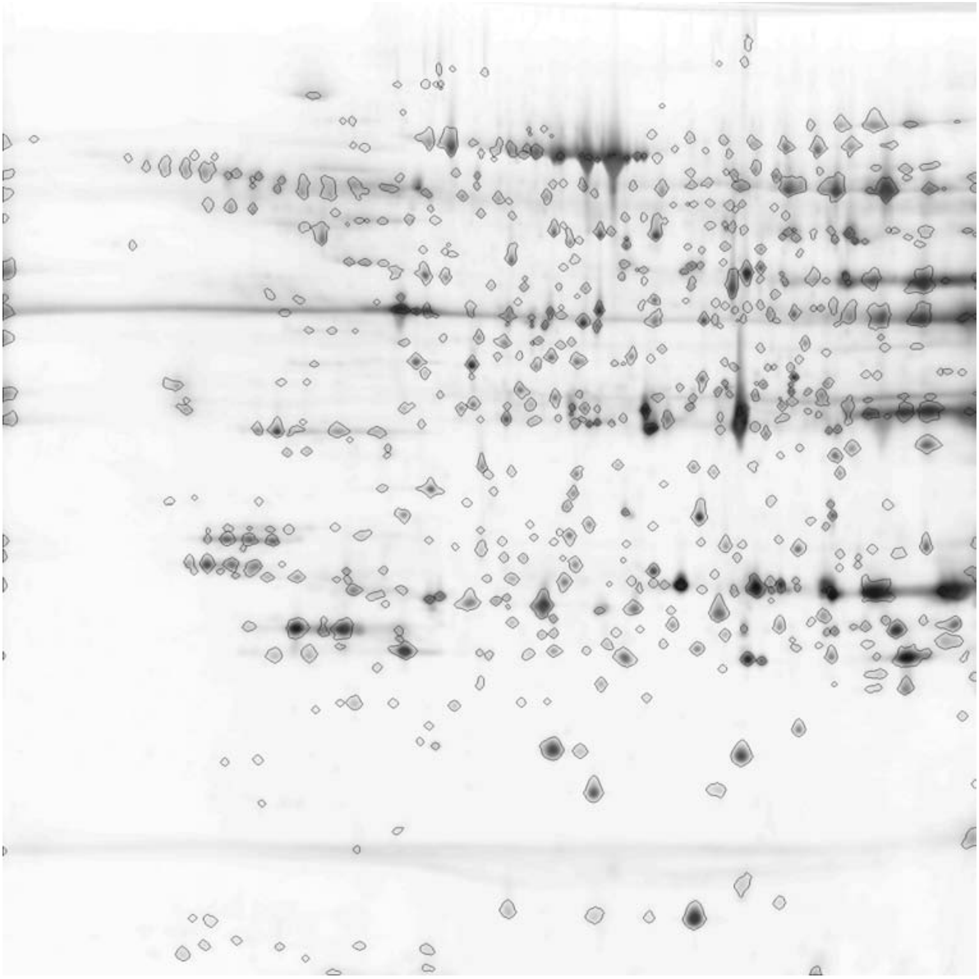


Figure 3.8: Supplementary figure. Spot boundaries identified using the synthetic gel approach.

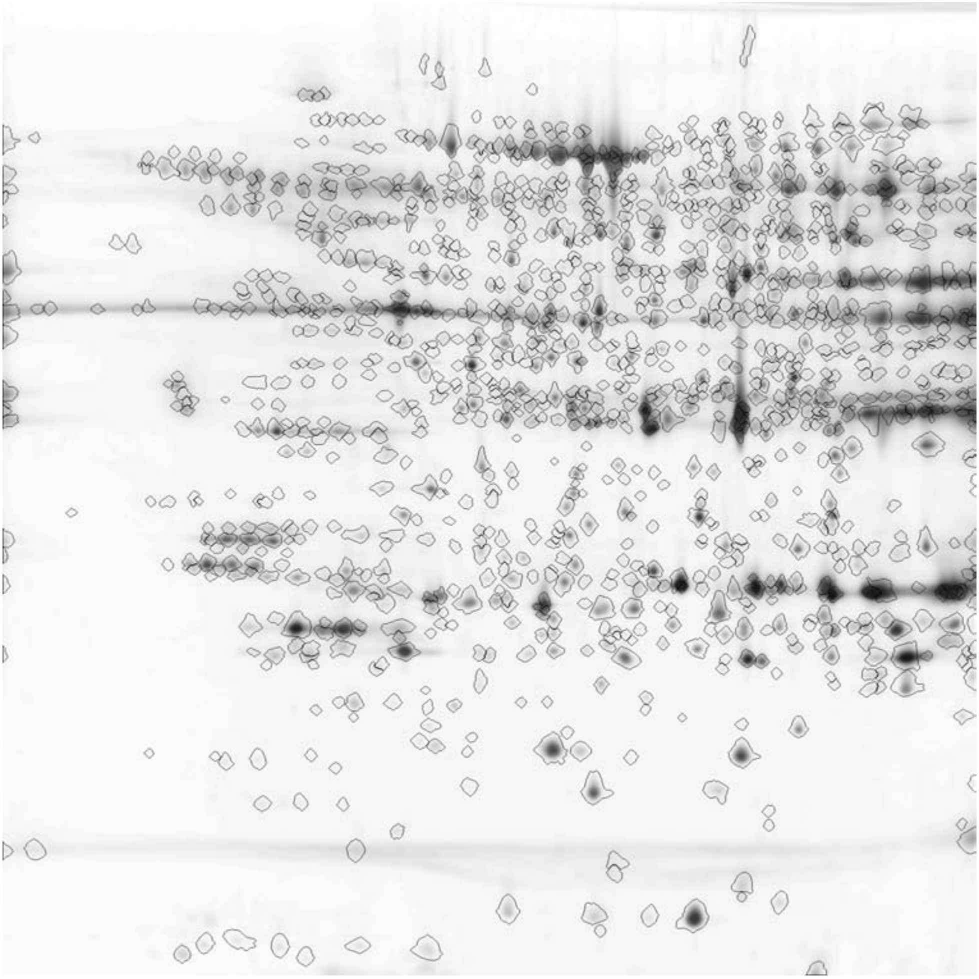


Figure 3.9: Supplementary figure. Spot boundaries identified using the common spot boundary approach.

Chapter 4

Pixel based analysis of multiple images for identification of changes; a novel approach applied to unravel proteome patterns in 2D electrophoresis gel images

Ellen M. Færgestad, Morten B. Rye, Beata Walczak, Lars Gidskehaug, Jens Petter Wold, Harald Grove, Xiaohong Jia, Kristin Hollung, Ulf G. Indahl, Frank Westad, Frans van den Berg and Harald Martens

Department of Chemistry
Norwegian University of Science and Technology
N-7491 Trondheim, Norway

This chapter has been accepted for publication in Proteomics.

Abstract

A novel approach for revealing patterns of proteome variation among series of 2D electrophoresis gel images is presented. The approach utilizes images alignment to ensure that each pixel represents the same information across all gels. Gel images are normalised and background corrected before they are unfolded to 1D pixel vectors which are subjected to multivariate data modelling. Information resulting from this data analysis is then refolded back to the image domain for visualisation and interpretation. The method is rapid and suitable for automatic routines applied after the gel alignment. The approach is compared with spot volume analysis to illustrate how this approach can solve persistent problems like mismatch of protein spots, erroneous missing values and failure to detect variation in overlapping proteins. The method may also detect variations on the contour of saturated proteins. The approach is given the name Pixel based analysis of Multiple images for identification of Changes (PMC). The method can be used for multiple images in general. Effects of pre-treatment of the images on the method is discussed.

4.1 Introduction

Two-dimensional gel electrophoresis (2-DE) is an important technique in proteome research, and a large number of proteins can be separated by this technique. However, there are presently major difficulties in the subsequent analysis of the resulting gel images. Traditionally, data from 2-DE are analysed by identifying spot boundaries for each protein spot in each gel, matching the spots from each gel to a reference gel, and comparing the calculated spot volumes between all gels. Ideally, one would like each spot to represent only one single protein. This could be achieved if 2-DE were performed on samples with a small number of well-separated proteins. However, 2-DE is commonly used to reveal changes in proteome patterns of samples consisting of a large number of unknown proteins. Such complex samples result in images with a high number of overlapping protein spots. Thus, the simple assumption that one spot always represents only one protein is not valid. Campostrini et al. [17] found that with 1mg of total protein applied on 2-DE and at least 1000 spots visualized, the spot singlets were in the minority, rarely exceeding 30% of all the spots analysed. The remaining spots were envelopes of two or more proteins. Similar results were also found in a theoretical study of 2-D images in general [18]. If an envelope of several proteins is identified as one spot, and the various proteins in this envelope respond differently to the experimental design parameters, it may be difficult or impossible to detect the effects by spot volume analysis. Based on the high number of overlapping proteins appearing on

2-DE images, new data analytical methods which can detect variability within envelopes of overlapping proteins are needed.

Tabulating protein spot volumes to a numerical data table for subsequent data analysis also involves the risk of imposing errors in the data. The most critical step is to define the boundaries between overlapping proteins. For overlapping proteins the decision of spot boundaries may be different from one gel to another. If the spot boundaries are set for one gel at a time, as is traditionally done, serious errors may be introduced into the data. For overlapping protein spots the program may set different boundaries between the spots in different samples which may result in missing values of a spot despite its present on the gel. Grove et al. [40] investigated technical replicates of silver stained 2-DE gels. Using state-of-the-art 2-DE software, they found missing values for as much as 40% of the spot volumes in the data table, and as much as 80% of the variables contained one or several missing values. A closer inspection of the results revealed that the appearance of missing values was not consistent with the presence or absence of proteins for any of these variables. When viewing only a small subsection of the gel containing just a few samples, one is visually able to distinguish between missing values arising from mismatches of proteins and truly absent proteins. However, for the whole gel consisting of hundreds of proteins, methods which distinguish between the true absence of a protein and volume detection failure are needed.

The use of difference gel electrophoresis (DIGE) makes the spot detection and matching more reliable than traditional 2-DE systems. By the DIGE approach, three samples are run simultaneously. The matching of protein spots within gels run simultaneously will then be perfect, but the matching between gels is still a challenge. An alternative to defining the spot boundaries for each gel separately is to define common spot boundaries for all samples in the experiment after an alignment of the gel images. This approach avoids the missing value problem, but the challenges in distinguishing effects of overlapping proteins within envelopes of several proteins are not solved. Considering the high number of overlapping proteins expected in 2-DE, this is an important concern.

The approach of aligning the gel pattern followed by multivariate data analysis of the digitalised scan for identification of changes in protein patter was first performed for 1D electrophoresis gels at Ås campus, Norway in 1990 [66]. To correct for migration problems, three spots being present in most tracks on the gel were identified and used for adjusting migration variation, and when absent their position could be set by viewing neighbour samples on the gel plate. Adjustment for background staining, which increased gradually from the top to the bottom of the gels was performed by subtracting a linear function for each track. The data of each sample were viewed as a 1D vector where the amounts of different more or less overlapping proteins are seen as peak height in a line plot. Multivariate data analytical tools could then be used to reveal significantly changes among the proteins. The approach was inspired by Dr. Harald Martens pioneer work on multivariate analysis in chromatography and spectroscopy [54, 67]. The 1D representation of the proteome pattern and spectroscopic data share common challenges with respect to both overlap issues and pre-treatment due to vertical shift problems. The proteome pattern introduces a new dimension of a horizontal shift problem which needs to be solved, and this has later received increasing attention in proteomics as well as in other areas like spectroscopy [56, 68]. For the 1D gels, significant variation in the proteome pattern was detected even for overlapping proteins [66]. These variations were discovered in a gel-region with relatively sparse information. An extension of this approach to 2D image data is now feasible due to the existence of improved tools for aligning 2-DE images to ensure that each pixel represent the same information across all gel images.

Here we present a 2D approach of pixel based analysis of multiple images where the resulting patterns from unfolded multivariate data analysis are refolded back to the 2D image domain for visualisation and identification of changes. The method requires alignment and background correction of images before analysis. Analysis of 2-DE images on the level of pixels by unfolding the images to 1D pixel vectors was also presented by Schultz et al. [35] for classification purposes, but the information was not traced back to the image domain for identification of changes in the proteome pattern.

The approach presented here is given the name Pixel based analysis of Multiple images for identification of Changes (PMC). The method was first presented on the conference "From Proteome to Genome" in Sienna, Italy, September 2006. In the present rapport the results of the PMC approach are thoroughly presented and compared with standard methods using spot volume analyses. Proper alignment and background correction are necessary to produce satisfactory results from PMC. Thus a description of the alignment process is included, and influences on the results due to different choices of background correction are also considered. The purpose of the present paper is to present the PMC approach for analysing 2-DE and compare the result of such methods with spot volume analysis.

4.2 Material and methods

4.2.1 Animal samples

The samples used to illustrate the method are taken from meat science where proteins are extracted in a time series after slaughter for seven animals to study changes in the proteome pattern after death. The study includes 7 Norwegian Red bulls (Bos taurus L.) (breed for both meat and milk) from a performance test station (GENO-Breeding and AI Association) slaughtered at approximately 13 months of age/450 kg live weight in 2004. The animals are given the numbers; 4363, 4366, 4368, 4370, 4389, 4391, and 4407. Samples from the longissimus dorsi muscle were collected one, two, three, six and ten hours after slaughter (denoted h1, h2, h3, h6, and h10, respectively). The samples were immediately frozen in liquid nitrogen. Proteins were extracted in TES buffer (10 mM Tris (pH 7.6), 1 mM EDTA and 0.25 M sucrose) and analysed by 2 DE as described previously [59], where gels are stained by silver staining [69]. The whole time series of one animal was run simultaneously in one batch as the focus of the present study was changes in proteome pattern occurring along the time series after death. The different animals were run as different batches. Thus, the effect of animal is confounded by batch variation.

4.2.2 Image analysis

Image alignment

For the PMC approach, the 2 DE images were aligned by the program TT900 S2S (Nonlinear Dynamics) using the gel image from animal 4370 analysed three hours after slaughter (h3) as reference. By this alignment method each gel image is aligned towards a reference. An initial anchor was manually set on a protein that could easily be identified on both images. Based on this, the software automatically suggested a large number of additional, spatially distributed anchors. After a visual inspection of these anchors and removal of unwanted anchors, the resulting anchors were used for automatic alignment. A further fine tuning of the alignment was performed by adding anchors manually. Pre-treatment, unfolding and analysis of the unfolded gel images were performed using Matlab version 7.3 (Mathworks).

Pretreatment of images

Pre-treatment of the gel images was performed by normalisation to adjust the pixel intensity to a constant protein amount and by background correction to remove effects of streaks and background staining. The normalisation was performed by dividing each image by the total intensity of the image, followed by rescaling to a common 0-1 scale for all images. For background correction we compare two approaches based on different principles. A method consisting of repeatedly fitting polynomial curves to a signal was adopted from Lieber et al. [53] originally created to subtract fluorescence background in Raman spectra. Raman spectra are 1D signals. Here we have adapted the method to 2D images by applying the algorithm in a line-by-line fashion in both horizontal and vertical direction. The method basically consists of first fitting a polynomial curve of some degree to a signal values. The signals above the polynomial curve are removed and a new approximation is performed. The method is repeated until convergence or till the number of iterations has reached a certain point resulting in polynomial curves following the lower boundary of the peak intensity. When the function is applied sequentially in vertical and horizontal direction of the images, the highest value of the polynomial curves is used to construct a background image. The degree of polynomial order will affect the flexibility to follow the curvature of the raw data. In the present study we used a polynomial degree of 4. The algorithm is intuitively simple and involves only one parameter.

Another approach of background-correction was performed by the Penalised Asymmetric Least Squares (PALS) approach extended to the data on the 2D grid [70, 71]. By this approach, image background is approximated using different, i.e. asymmetric, weighting of data points, and it is constructed from the tensor product of B-splines [72]. Its smoothness is controlled by the penalty terms, representing differences on neighbouring coefficients of the tensor products.

4.2.3 Spot volume analysis

Spot volume analysis was performed using Image Master 2D Platinum Version 6.01 (GE Healthcare), where the boundaries around the spots were defined for each gel separately. Spot volume analysis was also conducted by the software SameSpot (Nonlinear Dynamics) where common boundaries are set across all images.

4.2.4 Multivariate data analysis

The 1D pixel vectors of the unfolded images were first analysed by the multivariate data analysis method Principal Component Analysis (PCA) [54] after mean centring of the data [56]. PCA is an unsupervised, explorative method which transforms the 1D pixel data into new variables called principal components (PCs). The PCs are constructed to account for, in decreasing order, as much variance across the unfolded images as possible. The first few PCs will cover the majority of the variability in the original data. Each PC is defined as a linear combination of the original variables. The PCs can also be viewed as linear combinations of the samples. Thus, the PCA is bilinear having one vector related to each sample (called scores), and one vector related to each variable (called loadings). Plots of scores for the samples and plot of loadings for the variables give a visual impression into the main variability of the data. The loadings were refolded back to the 2D image domain for visualization. The data were mean centred prior to PCA and PLSR.

To relate the variation in the proteome pattern across the gel images to the experimental design parameter "time after slaughter", we used the multivariate linear regression modelling technique Partial Least Squares Regression (PLSR) [54–56]. PLSR is, like PCA, a bilinear method, where new variables (PLSR components), are generated as linear combinations of both samples and variables. PLSR thereby benefits from data reduction principles similar to PCA. By PLSR we obtain PLSR components which maximize the covariance between two blocks of data. By considering two blocks of data, PLSR is a supervised method in the sense that PLSR components are calculated by relating the vectorised gel images as regressor variables to a set of response variables or to the experimental design factor.

For validation of the regression, the analysis was repeated leaving out one of the samples at a time in a cross-validation routine, with a subsequent prediction of the response variables in the omitted sample from its gel image pixels. This gives a series of estimated regression coefficients. To test for significant changes among different pixels, the variability of the perturbed regression coefficients was used to estimate their significance using the Jack-knife method adapted to bilinear analysis [73]. A t-test is then performed on the different estimates of the regression coefficients from the various cross validation segments. The t-test is performed to test whether or not the regression coefficients are significantly different from zero. Pixels with regression coefficients that vary depending on the individual samples included in the models can thereby be excluded, and only pixels with stable regression coefficients are regarded as significant.

In some situations one may have very small regression coefficients close to zero which happen to be stable over the perturbed regression coefficients and therefore turn out as significant by the Jack-knife method. For the 2-DE images, where a large part of the pixels are in regions without any protein spots, this is a potential problem. When the number of variables is large, such unreasonable significance results can add up to a large number. To avoid erroneous significance of very low regression coefficients we therefore restricted the standard deviation of the perturbed regression coefficients to have a minimum value. This is described generally for t-tests by Allison et al. [74] and adapted to Jack-knife significance tests for bilinear regression methods by Gidskehaug et al. [75]. In the present study, the estimated variance of each pixel across cross-validation segments is weighted with one tenth of the mean variance across pixels.

Finally, we used the spatial location of the significant pixels as an extra visual validation of the selection of significant regression coefficients. When the 2-DE image is unfolded and analysed, all spatial information is excluded from the modelling. The spatial information can therefore be used for validation of the results when refolded back to an image, as regression coefficients reflecting significant variation in protein spots will occur as a large number of concentrated pixels on top of the protein spot.

4.2.5 Data compression and reduction

The gel images analysed consist of approximately 3.3 million pixels, which gives 3.3 million variables per sample when analysed at the level of pixels. This creates challenges with respect to the data capacity, as well as for the significance testing. First, we removed 50 pixels around the gel boundary, as this area was noisy. Furthermore, we reduced the resolution of the images by a factor 0.5 via bicubic interpolation. This resulted in a total of 187 165 variables per sample for the whole gel. For visualisation of changes in proteome pattern we also conducted analysis of a sub region of the gel consisting of 254 × 399 pixels without data reduction.

Large data sets may cause memory problems if analysed directly, but the analysis is simplified by acknowledging that all possible variation between the available samples is kept in the scores from a PCA. There is no need to calculate 187 165 PLS components, because only 34 components hold any information when 35 samples are available. Similarly, the multivariate regression may be performed using the matrix of 34 PCA score-vectors as input instead of the 187 165 variables of the original data. The regression coefficients and the PLS-loadings are subsequently inflated by multiplication with the original PCA-loadings. With this method, the data capacity is only limited by the capacity to run the PCA.

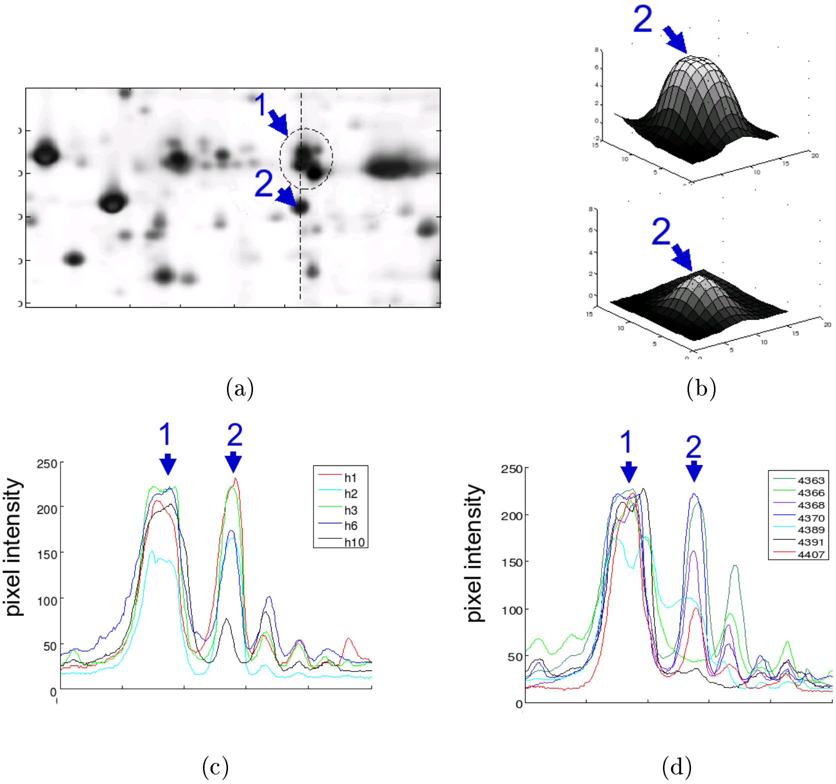
4.3 Results and discussion

4.3.1 Pixel based analysis of multiple images for identification of important features

Image alignment

The alignment of the gel image was controlled by plotting lines across different images as illustrated for a subsection of the gel in figure 4.1(a). Figure 4.1(b) shows changes in expression for a protein spot between two different gels (h3 and h10 for animal 4370). The line drawn vertically in 4.1(a) is plotted for time series h1-h10 for animal 4370 in figure 4.1(c) and for h3 across all animals (or batches) in figure 4.1(d). The images were inverted after unfolding to allow the black pixels on the white background to appear as positive peaks when unfolded. When comparing images within the same batch (figure 4.1(c)) after alignment, all peaks were nicely positioned on top of each other meaning that pixels should be directly comparable across all gels. Figure 4.1(d) also illustrates differences in migration across batches. The protein envelope marked 1 is better resolved in one gel compared with the others. This is a typical batch to batch variation arising in 2-DE due to different migration times, and it is a challenge inherited in 2-DE analyses. The problem of migration differences in 2-DE makes it necessary to apply quality control, and only gels relatively similar in migration are recommended for further data analysis, regardless of the choice of data analytical tools. For the present material the gel images were acceptably homogenous across different batches, and the gel alignment as such performed well both within and between batches.

A gradual intensity decrease with time after slaughter for protein number 2 is evident for both the 3D representation in figure 4.1(b) and the unfolded representation in figure 4.1(c). The most important benefit of unfolding is the 1D representation of the information obtained for each image. For the whole experiment this gives a data table where the rows constitute different gels or samples, and the columns represent pixels. Standard multivariate data analytical tools can then be used to analyse differences between the samples.



(a)

(b)

(c)

(d)

Figure 4.1: (a): A sub-image of the gel 4370-h3. The selected region is the highlighted frame in figure 2(a). (b): 3D representation for spot number 2 in (a), shown for h3 (upper figure) and h10 (lower figure). (c): The vertical line in (a) displayed as a spectrum section for animal 4370 at different times (h1-h10) after slaughter. (d): The same spectrum as (c), but now shown for different animals (batches) at time h3. Migration differences are clearly evident for peak number 1.

A closer inspection of the 1D representation in figure 4.1(c) and 4.1(d) reveals baseline variation among the samples resulting from non-uniform background intensity in the gels. Thus, before unfolding and subsequent data analysis, normalisation and background correction are necessary.

Normalisation and background correction

A number of different approaches can be conducted for normalisation and background correction of 2-DE images. This choice of pre-treatment can affect the output and modify the proteome patterns identified. This is true for the PMC, as it is for any data analytical tool used to analyse the 2 DE images [21]. The normalisation chosen for the present study produces the same total intensity for each image. The rational behind this choice is that proteins are loaded on the gel on an equal protein basis. If also the streaks and background colour on the gel are proteins which has not been focused into spots, this normalisation should adjust the gels to an equal protein basis. A support for this normalisation approach is that background streaks may frequently be seen along the tracks of heavily stained proteins, and general background colours are typically observed in areas of the gels with abundance of proteins. However, if there are major variations in staining intensity across the gels, a number of proteins might not show up on weakly stained gels despite its presence in the samples. This would result in fewer spots in some gels compared with other gels. In such situations adjustment to a constant staining intensity would imply unreasonable elevation of proteins on gels where few proteins are present. This would be a problem for any data analysis conduced. Again the recommendation is to apply quality control of the gels images and only samples with relatively similar staining intensities should be analysed together. For the present material, the staining intensity was relatively similar from one gel to another.

Results from two different background corrections are shown in figure 4.2 for gel 4370-h3, where the raw data image are shown in figure 4.2(a), the background image estimated by 1D polynomial fit is shown in figure 4.2(b), and the background image estimated by the PALS approach adapted to 2D-grids is shown in figure 4.2(d). The background images are subtracted from the raw gel images, giving the resulting background subtracted images shown in figures 4.2(c) and 4.2(e).

The two different principles for background correction produce different results. The background image obtained by the line-wise polynomial fit is able

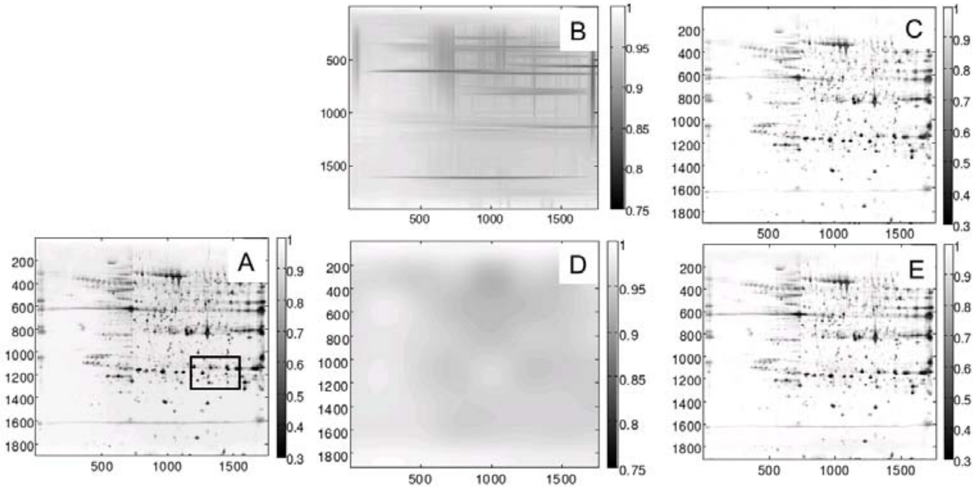


Figure 4.2: (a): Gel image 4370-h3 after normalisation and prior to background correction. (b): Background image estimated by 1D polynomial fit. (c): Resulting background subtracted image using background from (b). (d): Background image estimated by the PALS-approach. (e): Resulting background subtracted image using background from (d).

to identify the streaks running across the images in horizontal and vertical direction, in addition to the more smooth background staining present over larger regions in the gel (figure 4.2(b)). These streaks are not captured by the PALS approach(figure 4.2(d)). The reason for this is that the polynomial fit is applied in a line-by-line fashion ,while the PALS background is estimated simultaneously for the whole image.

When evaluating the model-explained variance of the response using cross-validated PLSR, as shown in figure 4.3, the images corrected by the 1D polynomial fit performed somewhat better than the images corrected by PALS. Both performed considerably better than images analysed without background correction. Obtained results using different background corrections are expected to vary, depending on the nature and level of the background. The presence of streaks can give erroneous impact on proteins of interest, and in these cases an algorithm which effectively corrects for the intensity elevations caused by streaks would be preferred. For the present data streak correction seems to be appropriate, and in the following text results from the images corrected by the 1D polynomial fit is presented. Comments will be given to results on images analysed without background correction and images corrected by the PALS

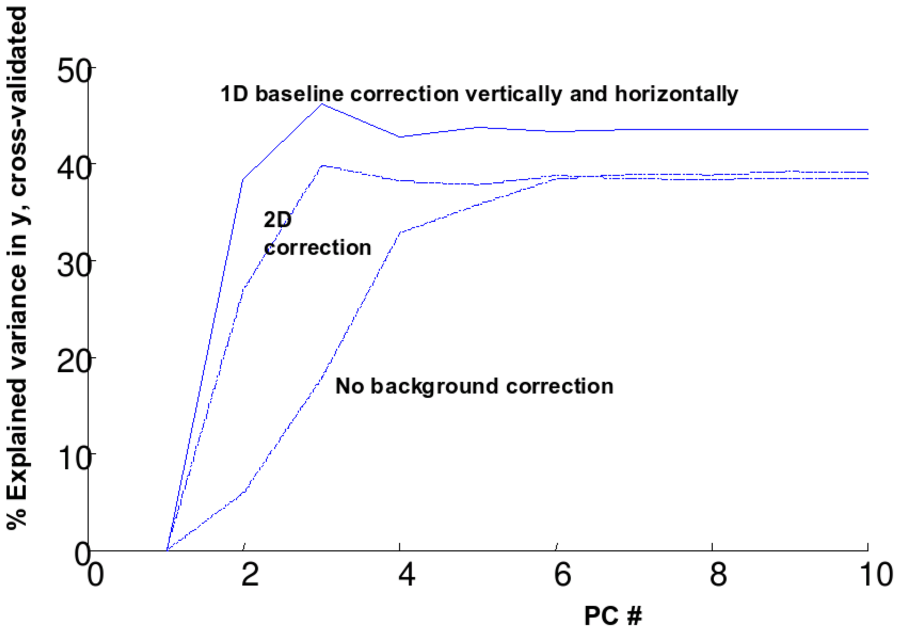


Figure 4.3: Percentage explained variance for the response variable (time after slaughter) using different background corrections. The measure gives an indication of how well the different models perform. (The models are created using a log_{10} transformation of time as response). The reason for this is to improve the linear relationship between the data and the response)

algorithm when necessary.

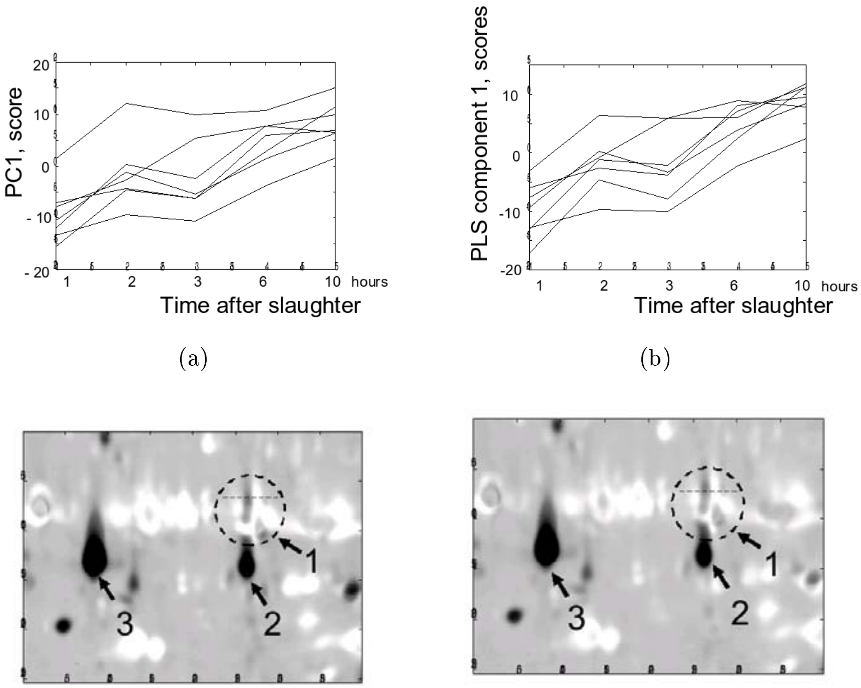
Multivariate analysis

To obtain insight into the main variability of the 2-DE images, unsupervised PCA [54, 56] is performed on the unfolded 1D representation of the gel images. As can be seen from figure 4.4(a), sample scores from the first, most important PC increase gradually with time after slaughter for all animals. The first PC thus reflects variations related to protein changes in the time course after slaughter which is consistent for all animals, even though there are individual differences between the batches. There are differences between the animals, as seen by comparing the profiles in 4.4(a), however this effect can not be distinguished from batch effects as the animals were run in separate batches.

Sample scores are also shown for the first component for supervised PLSR in figure 4.4(b). Here the unfolded image pixels are used as regressor variables and log_{10} of time as response. The reason for the logarithmic transformation is to improve the linear relationship between the regressor variables and the response. The results from PLSR are similar to those from PCA.

Figure 4.4(c) and figure 4.4(d) show loading plots of the pixels from PCA and PLSR respectively, where the first principal component is refolded back to the original 2D image domain. Black pixels on the 2D loading image are negatively related to the scores and reflect pixels corresponding to a decrease in intensity as time after slaughter increases, while white pixels increase in intensity with time. The similarities between the unsupervised and the supervised method confirm that the main variations in the present data do reflect variation related to time after slaughter. For protein envelope number 1 in figure 4.4 the loading images reveal different relations to the time after slaughter of the various protein spots within this envelope. As an example the horizontal lines drawn through these envelopes shifts from black (decrease with time) to white (increase with time). In both loading images the protein spots numbered 2 and 3 are dark, indicating that the pixels are negatively correlated to time after slaughter, meaning this spot will decrease in intensity as time increases. The mean intensity over all animals for protein spot number 3 from figure 4.4 for time h1, h3 and h6 are shown in figure 4.5(a). The cross section indicated in figure 4.5(a) is plotted for all times (h1-h10) in figure 4.5(b). Again each line is calculated as the mean of all animals, and the decrease with time is clearly visible. The flat curvatures on the top of the peak in two of the lines also illustrate the challenge of saturated protein spots in silver stained gels. Due to the saturation challenge, the largest differences among the samples are not necessary seen in the centre of the protein spot, but closer to the boundary as illustrated by a 1D loading plot of this cross-section from PCA in figure 4.5(c).

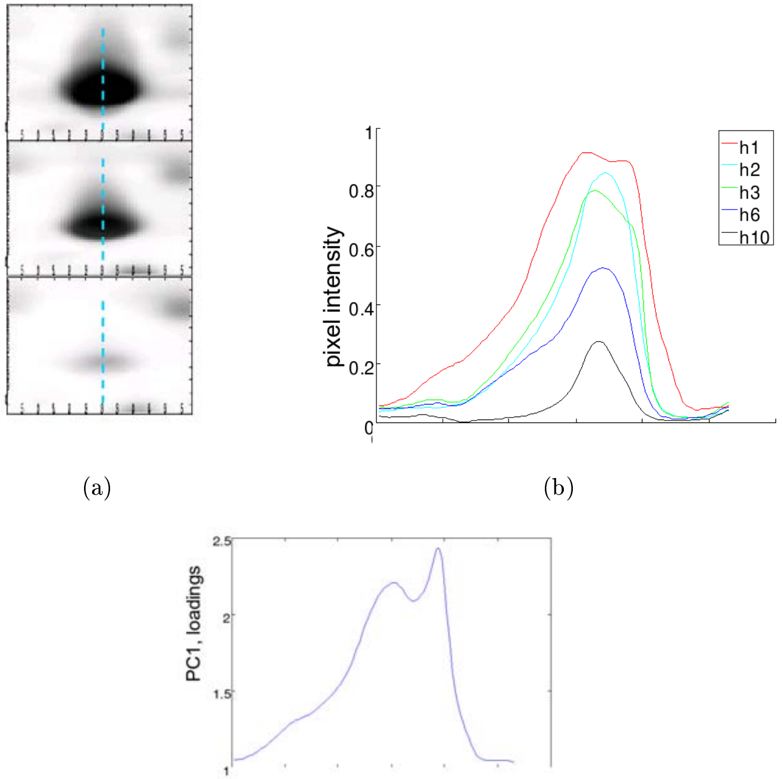
The decreases and increases observed in the time course after slaughter are interpreted relatively as both the application of proteins on the gel, and the subsequent analysis of the gel images are performed on an equal protein basis. Thus, the results show that protein spots number 2 and 3 constitute a successively smaller proportion of the total amount of protein in the sample as time after slaughter increases. Similarly, the proteins showing an increase constitute a successively larger proportion of the total amount of the proteins in the extract.



(a) (b)

(c) (d)

Figure 4.4: (a): PCA-scores from first principal component for each animal (batch) plotted as a function of time. (b): PLSR-scores from first PLSR-component for each animal. It can be seen that the PLSR-scores are similar to the PCA-scores. (c): Loadings from PCA refolded back to the image domain. Dark regions decrease in intensity with time, while white regions increase in intensity with time. (d): Loadings from PLSR refolded back to the image domain.



(a)

(b)

(c)

Figure 4.5: (a): Mean images for spot number 3 in figure 4.4 for 1, 3 and 6 hours after slaughter (top to bottom). (b): Line plot for the cross-section marked in (a) for mean over each time. (c): Loadings from PCA for the same line-segment as in (b), indicating the saturation effect. The most important variations are detected around the spot-centre, but closer to the boundary.

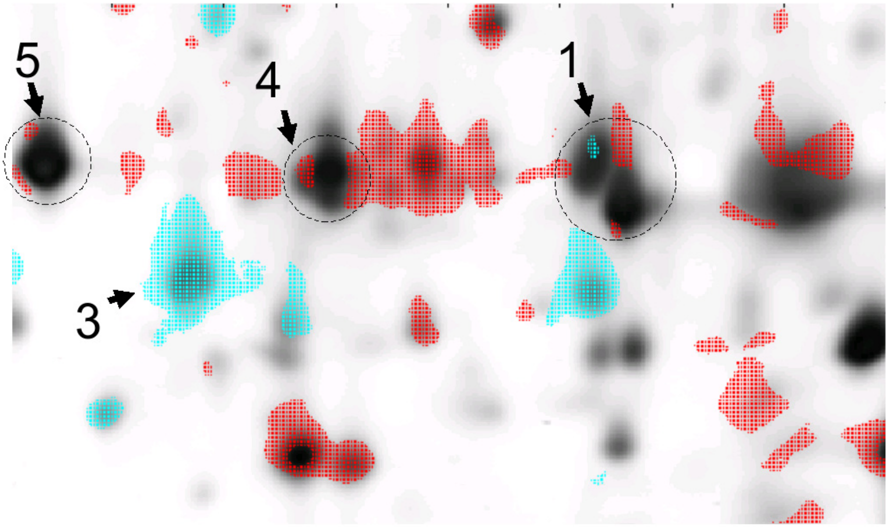
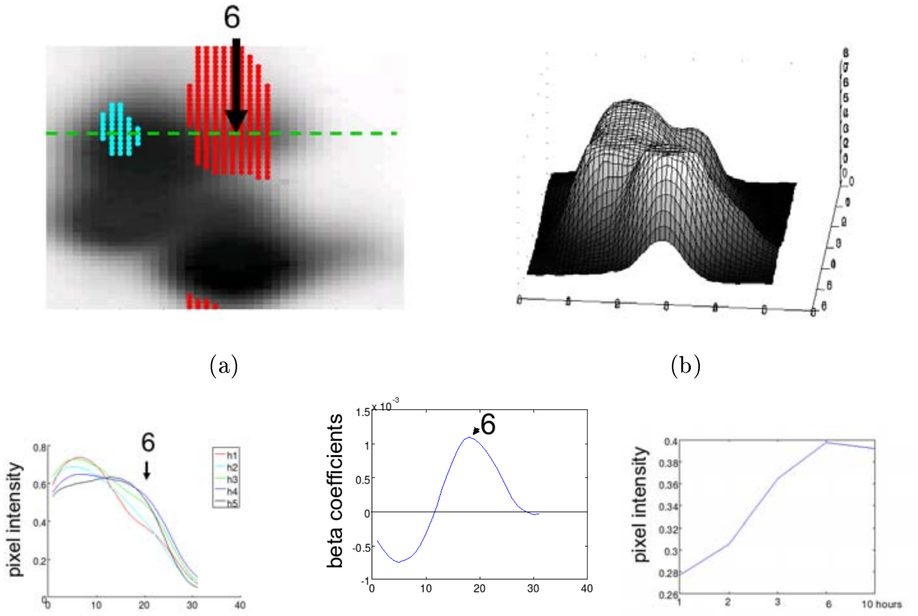


Figure 4.6: Significant pixels displayed on top a synthetic image, created by taking the average over all gels. Pixels negatively correlated with time are coloured blue, and positively correlated pixels are coloured red.

The first component from the PLSR model is found to be most significant with respect to the response variable (time after slaughter). Regression coefficients for each pixel are thus estimated based on this PLSR component, and significant pixels are identified by the Jack-knifing procedure explained previously. To highlight areas in the image containing important variations with respect to time, significant pixels identified by Jack-knifing are plotted on top of a constructed average image of all samples in figure 4.6. Pixels negatively correlated to time are coloured blue, and positively correlated pixels are displayed in red. The pattern of significant protein changes are similar for both images background corrected by PALS and images with no background correction, the most pronounced difference being the occurrence of significant regions outside protein spots, especially for the images where background correction is not applied. A high number of spots are found to change significantly, which is expected, because the protein-balance of meat changes dramatically in the first hours after slaughter.

The significant pixels tend to cluster in regions on top of or close to protein spots. Sometimes these regions cover entire spots, while in other cases only smaller parts of the spots are covered. This may be understood in terms of two phenomena: 1) The highest-density spots usually display the well-understood technical problem of saturation around their maximum; a phenomenon expected to stop pixels near the spot-centre from appearing as significant. 2) More interestingly, significant pixels near the periphery of some spots may indicate the presence of partially unresolved proteins. A closer view of protein number 1, 4 and 5 from figure 4.6 is enhanced in figure 4.7 and figure 4.8. Significant pixels are displayed on top of image 4370-h3 in figure 4.7(a) where protein spot number 6 is seen as a shoulder to another, larger protein spot. By plotting the horizontal cross section in figure 4.7(a) for average images of each time (h1-h5 in figure 4.7(c)) a gradual increase in intensity with time is observed for this shoulder. Similarly, there is a gradual decrease with time for the left neighbouring protein coloured blue in figure 4.7(a). The pixels' regression coefficients are thus negative for the large protein to the left, but switches to positive values for protein number 6 (figure 4.7(d). The centre-pixel for protein number 6 marked by an arrow in figure 4.7(a). is plotted as a function of time in figure 4.7(e) showing a gradual increase in intensity up to 6 hours.

Protein envelope 4 from figure 4.6 is another example where the PMC approach has revealed significant changes in small shoulders of larger proteins, as seen in figure 4.8. PMC are also able to detect significant changes on the boundary of saturated proteins as illustrated for protein spot number 5 from figure 4.6, also displayed in figure 4.8. Silver staining of proteins in 2-DE is sensible, enabling detection of large numbers of proteins, but silver staining also suffer from a low dynamic range, leading to severe saturation problems. The amount of protein in a spot is however, not only manifested as the intensity at the spot-centre, but also by the width of the spot of which protein spot number 3 in figure 4.5 is a typical example. Despite saturation in the central part of a protein spot it is possible, by using the PMC approach, to detect significant variation on shoulders and close to boundaries of larger protein spots. This will of course depend on how severe the saturation problem is, but the PMC approach has the general ability to reveal significant variability in complex areas consisting of overlapping proteins and highlight significant changes closer to the boundary of saturated protein spots, which is not always feasible in other programs for 2-DE analysis.



(a)

(b)

(c)

(d)

(e)

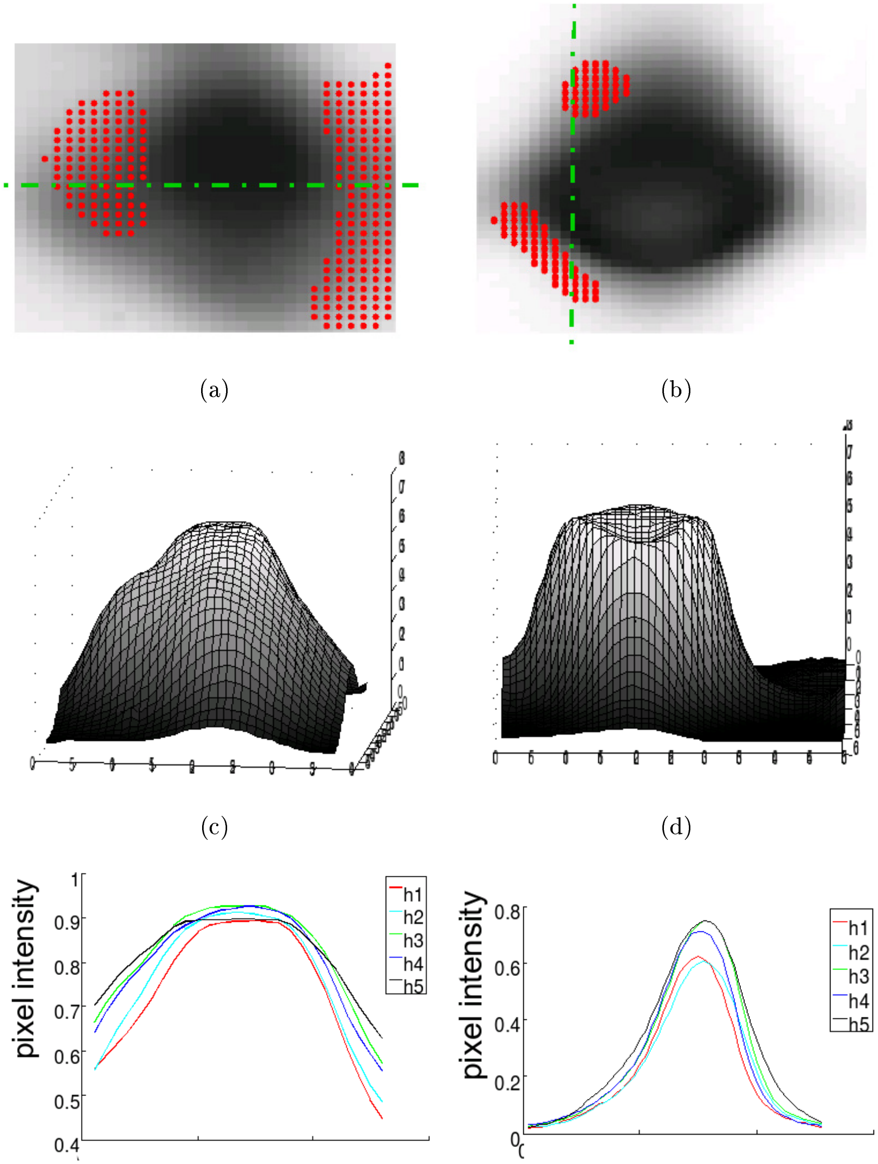
Figure 4.7: (a): Protein complex number 1 in figure 4.6. Significant pixels are displayed in blue and red for a decrease and increase in intensity with time respectively. (b): 3D view of the same protein complex as in (a). (c): Plot of the horizontal cross-section from (a) for each time. Each curve is averaged over each batch. (d): Regression coefficients for the same cross-section as in (c). (e): The intensity of the centre-pixel marked by an arrow in (a) plotted for each time. A gradual increase with time up to 6 hours is observed, which is in accordance with the positive regression coefficient of this pixel.

4.3.2 Analysis by spot volume

For comparison, the same sub-image as highlighted in Figure 2(a) is analysed by two different spot volume approaches; one where spot boundaries are set for each gel individually, and one where spot volume analysis is conducted using common spot boundaries across all samples.

Sub-images of 2-DE gels are shown in figure 4.9 for six of the 35 gels included in the present study. The images are from animal 4363, 4366, and 4370 observed three hours (h3) and ten hours (h10) after slaughter. Gel image 4366-h10 are used as reference gel for spot matching between multiple gel images. Figure 4.9(a) shows results from the analysis where spot boundaries are defined for each gel individually followed by spot matching. The spots with green boundaries indicate protein spots where a match could be found in the reference gel, while the red boundaries indicate spots that could not be matched to any spot on the reference gel. As is seen in the figure, a large number of spots failed to be matched to a protein spot on the reference gel. By comparing protein envelope number 1 for gel image 4366-h3 to the corresponding envelope in the reference image (4366-h10) one can see that this envelope of proteins is split into three spots on image 4366-h3, whereas only two spots are recognised on the reference image. In this situation there is no way to perform correct matching of these spots, and consequently the software failed to match any of the proteins in this envelope resulting in missing values in the spot volume data table. When viewing this protein envelope on gel image 4363-h10, another problem is encountered which is even more difficult to detect. Here two protein spots are identified on image 4363-h10, as was also observed in the reference image, and apparently the matching of these spots is successfully performed. However, the spot boundaries are set differently in these two images, resulting in incorrect estimates of the spot volumes for both proteins identified. When viewing the whole gel consisting of hundreds of proteins, such errors are extremely difficult to detect.

Another important problem is also revealed for protein spot number 2 in figure 4.9. This protein is absent in the reference gel, resulting in failure to detect it on the other gels. If a mixture of the samples is run as a reference gel, one could expect protein spot number 2 to be detected when present, and absent when truly absent. However, from the spot volume table, we could not distinguish the true absence of proteins with the erroneous missing values as observed for image 4366-h3 for protein spot number 1. Thus, the conventional protein spot lists do not give a clear answer to the presence or absence of protein spots, which is a major concern in the analysis of gels from 2-DE.



(a)

(b)

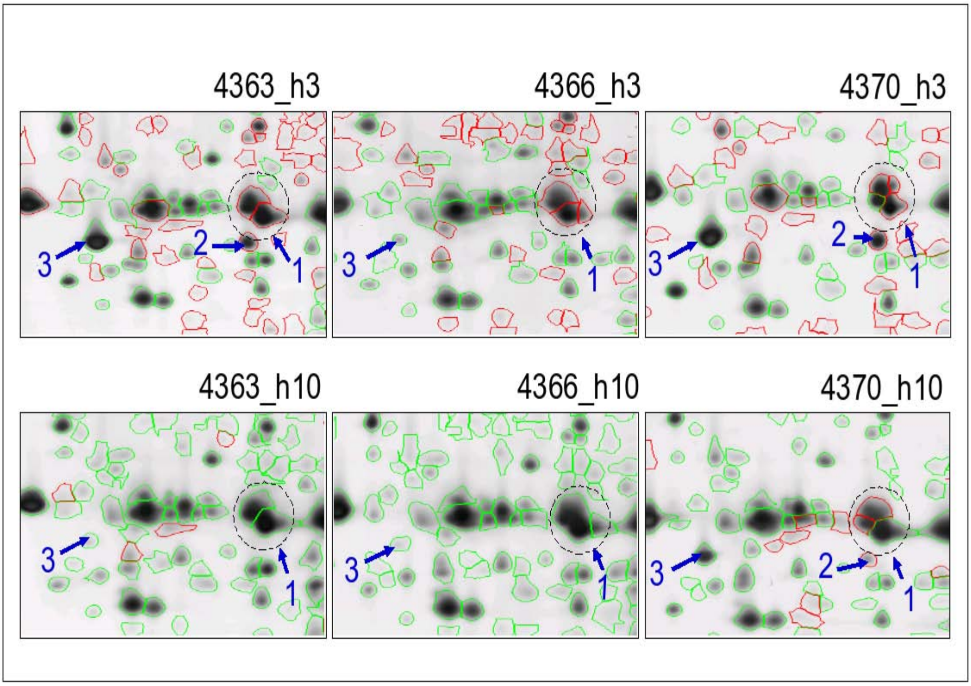
(c)

(d)

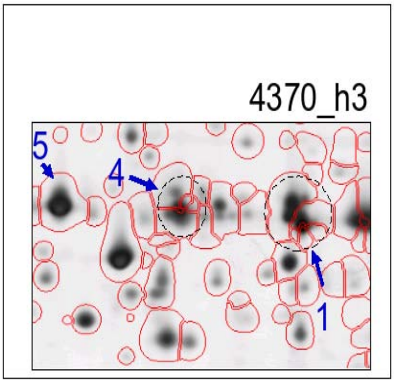
(e)

(f)

Figure 4.8: (a),(c) and (e): Protein envelope number 4 from figure 4.6. (b),(d) and (f): Protein number 5 from figure 4.6.



(a)



(b)

Figure 4.9: (a): Results from analysis where spot boundaries are defined individually for each gel. The software used is ImageMaster (GE Healthcare). (b): Results from analysis using common spot boundaries. The software used is Progenesis SameSpots (Nonlinear Dynamics).

If missing values in a spot volume table are treated in the subsequent data analysis as unknown, the information about presence vs. absence of proteins is lost for all proteins. If the missing values are instead replaced by zeros one would introduce major errors to the data in situations where the missing values result from misdetection rather than from the absence of proteins. As all spot volumes on a 2-DE gel are positive values, erroneous insertion of the value zero, the most extreme value on the scale, gives strong false impacts on the results. Instead of inserting zeros when a protein is not detected it is generally more correct to treat them as unknown by inserting missing values. However, this will lead to a data table with a very high number of missing values, which gives major challenges for the subsequent data analysis [39, 40]. The approach presented in the present paper, where the 2-DE images are aligned, unfolded and analysed at the level of pixels provides a novel approach to deal with this challenge.

When spot volume analysis is performed using common boundaries for all samples, the resulting spot pattern (figure 4.9(b)) results in a data table without missing values. However, for envelopes of several proteins, (1, 4 and 5 in figure 4.9(b)), the spot volume analysis is not able to detect differences. Thus, the PMC approach was also superior to the spot volume method based on defining common boundaries for each spot, as it enables identification of variations within envelopes of overlapping proteins, and detects changes close to the boundary of saturated spots.

4.3.3 Visual aspect

The present study has shown that viewing and analysing the images as 1D vectors of pixels, rather than as lists of spot volume, gives novel insight into the proteome which can not be achieved by spot volume analysis. The results from the multivariate analysis performed at the level of pixels gives images in the same 2D domain as the original gel images, assisting with the interpretation of the results. The reliable co-variation patterns among a long series of individual gel images can thus be represented in terms of a few PC images and score plots. The researcher can thus trace the results at all levels, viewing each individual sample as well as viewing the variability across all samples. One can switch from windows focusing closely on one or a few protein spots of interest to an overall view of the results.

When analysing the 2-DE images at the pixel level, all information on the spatial location of the spots are excluded when performing the multivariate data analysis. The ability to identify significant changes in proteins while strongly overlapping with other proteins was possible as the pixels were analysed independently of each other. Furthermore, as the spatial information was not used during the modelling, it can be used as a visual validation of the selected pixels.

Using pixel loadings and clusters of significant pixels is an easy way for the biologist to visualize areas in the image containing important protein variation. By overlaying the significance image on the original image (figure 4.6) it is immediately evident where to start looking for proteins that change in concentration. When analysing protein spot segments in the traditional way, the analyst usually has to validate and edit numerous protein spot segments, most of which are not significant, and the analysis will be highly dependent on the quality of the segmentation procedure. Spot segments should ideally consist of single, isolated proteins for the spot volume analysis to be performed correctly. However, highly overlapping protein spots are often considered as single segments. If only one of the overlapping proteins varies, or if the variation in content for both proteins differs amongst several gels, an analysis based on them being a single entity might not be able to detect this difference. When performing the analysis at the level of pixels, significant differences in highly overlapping protein clusters and spot "shoulders" are also detected. Several areas marked as significant consist of only parts of protein clusters, and some areas even appear as parts of what seem to be single protein spots. These results are, however, not surprising when one considers the findings of Campostrini et al. [17] concluding that most of the proteins in a gel are at least doublets or triplets. Thus, what appears as partially significant changes in a single protein might actually reflect two different proteins where only one is displaying significant variation. Taking the overlap challenge into consideration, the approach described here is a better way to view the 2-DE images than traditional analysis of protein spots. It offers an immediate and intuitive view of changes of the image, which are not restricted by pre-defined spot boundaries. The highlighted areas, especially those areas where many significant pixels cluster together, should then be subject to further investigation and protein identification.

4.3.4 Further development

A large number of pixels in the gel image reflect background with no protein spot information. Being more selective in which pixels to analyse may both reduce the noise in the data as well as contribute to data reduction. Most of the pixels in a gel image reflect only background, and are not related to protein spot information. Including these regions is likely to introduce noise in the analysis. Rye et al. [76] show that using image segmentation prior to the multivariate analysis to select pixels of interest is a useful way to reduce the amount of data without losing important information. Their segmentation procedure constructs an image mask, which can be used on all gels to highlight areas in the image related to proteins.

4.3.5 Interpretation of the findings

The present study has revealed major changes in the proteome pattern in the time course after slaughter where some proteins decrease in relative intensity and other increase. Further studies of identification of the significantly changes proteins and the interpretation of the findings will be given elsewhere.

4.4 Conclusion

Analysing 2-DE images at the level of individual pixels, by unfolding each gel image and then refolding the results back to 2D images for visualization, provides a strong novel tool for detecting variation in proteomes, even in overlapping, and it may detect significant changes in the border of saturated protein spots. The analysis is based on tools for analysing multivariate data. The method gives a visual insight into the results where all samples can be viewed simultaneously. When implemented in user-friendly software, such a modelling approach will be rapid and easy for the biologist to execute. Novel insights into the results of 2-DE, which can not be obtained by traditional spot volume analysis, are then possible.

4.5 Acknowledgements

We would like to thank Dr. Peter M Chandler, CSIRO, div. plant industry, Canberra A.C.T. 2601, Australia and the reviewers for useful comments to the manuscript.

Chapter 5

An improved pixel-based approach for analysing images in two-dimensional gel electrophoresis.

Morten B. Rye, Ellen M. Færgestad, Harald Martens, Jens Petter Wold and Bjørn K. Alsberg

Department of Chemistry
Norwegian University of Science and Technology
N-7491 Trondheim, Norway

This chapter has been recommended for publication in Electrophoresis.

Abstract

An improved pixel-based approach for analysing 2-DE images is presented. The key feature of the method is to create a mask based on all gels in the experiment using image morphology, followed by multivariate analysis on the pixel level. The method reduces the impact of noise and background by identifying regions in the image where protein spots are present, but make no assumption on individual spot boundaries for isolated spots. This makes it possible to detect significant changes in complex regions, and visualise these changes over multiple gels in an easy way. False missing values and spot volumes caused by imposing erroneous spot boundaries are thus circumvented. The approach presented gives improved pixel-based information from the gels, and is also an alternative to existing methods for data-reduction, significance testing and visualisation of 2-DE data. Results are compared with software using a common spot boundary approach on an experiment consisting of 35 full size gel images. Gel alignment is required before analysis.

5.1 Introduction

The use of two-dimensional gel electrophoresis (2-DE) is an important technique in proteome research. However, the analyses of such gels are both expensive and time-consuming, and due to the complex nature of 2-DE data, there are risks of imposing errors to such data when transferred to a table for subsequent analysis. The multivariate nature of data from 2-DE makes analysis by multivariate approaches a natural choice [20, 26, 35, 43, 44, 77]. Presentation of the results in an easily accessible way is also a challenge when a large number of proteins are found to be relevant. Procedures which are rapid, gives the best possible representation of the information on the gel, perform multivariate data analysis, and present the results to the analyst in a simple and convenient way are therefore needed.

Data from 2-DE is traditionally analysed by identifying spot boundaries for each gel separately. The spots from each gel are then matched to the spots on a reference or master gel. However, serious problems occur when one is unable to perform such a match, and missing values are introduced to the data table. A missing value might be due to the absence of a spot on the gel, or failure to find matching protein spots. Introducing a missing value is a correct choice in the first case, but will lead to serious challenges in the second case [39, 40, 64]. One way to overcome the problem of missing values is to use a common set of spot boundaries for all gels. This has lately been the method of choice for popular commercial gel-analysis packages such as Progenesis SameSpots (Nonlinear

Dynamics) and Delta-2D (DECODON). Defining common spot boundaries are commonly performed by the use of a synthetic gel, calculated as a weighted average of the pixel intensities of all gels analysed [41]. Though this solves the missing value problem by having a complete boundary correspondence between gels, the difficult task of defining boundaries for each individual spot in the gel still has to be performed. Identifying such boundaries is a major challenge in the case of overlapping protein spots. If two overlapping spots are confined within the same spot boundary, significant changes in these spots may not be detected if they respond differently to a response variable or a design factor. The use of image segmentation to define individual spot boundaries always produces image segments consisting of two or more overlapping protein spots. Especially this would be true for a synthetic gel, where spots from several individual gels are introduced to the same gel prior to segmentation. According to a late study by Campostrini et al. [17], the number of multiple spots can be quite large in a 2-DE experiment.

One way to overcome both the missing value problem and the challenge overlapping protein spots, is to use the pixel-values from the image directly in the analysis, as done by Færgestad et al. [64]. When using the pixel values, no assumptions are made on spot-entities, and every pixel in the sample gel will have a corresponding pixel in the reference gel. Both the common spot boundary and the pixel-based approach rely on the images to be properly aligned before analysis.

In the pixel-based approach, unwanted background variation outside the protein spots are not removed prior to the analysis, making it sensitive to noise and artefacts in the gel not related to protein content. When analysing such data, unwanted artefacts or cracks in the gel surface might show up as important sources of variation. An analysis using common spot boundaries do not suffer from this problem, because a proper spot detection routine will remove sources of unwanted variation prior to the data analysis. It should be noted that if the segmentation procedure is not carried out properly, unwanted artefacts can be introduced to the synthetic gel and cause problems to the final segmentation when common spot boundaries are used.

A pixel-based approach offer several advantages as shown by Færgestad et al. [64]. However, when this approach is used on the entire gel, large regions in the gel with no information are included in the analysis. In a general 2-DE experiment most regions in a gel are known a priori not to display any significant changes. Especially this is true for image background, which occupies the larger part of a 2D-gel. Leaving such regions out of the analysis is advantageous with respect to both data reduction and interpretability, while the risk of removing important variations is minimal. Introducing image segmentation in 2-DE, the image background is removed prior to analysis. However, image segmentation in 2-DE is also concerned with identifying spot boundaries for all individual protein spots in a gel. The regions defined by these boundaries are then used for further analysis. This approach makes sense if the goal of the experiment is to identify all existing proteins in a cell sample, it may not be efficient for comparative studies with a large number of gels. Realising that the majority of protein spot do not change significantly in most experiments with multiple gels, resolution and identification of all individual spots should be unnecessary. Considering the findings of Campostrini et al. [17], it is also questionable whether a full resolution of all individual spots is feasible by image analysis alone. While the identification of regions in a gel where protein spots are present can be performed without too much effort (for example by image morphology), the resolution of these regions into isolated spots is tedious and almost always subject to errors. This applies to the common spot boundary approach, as well as for the individual segmentation of each gel in spot matching procedures. In such circumstances alternative approaches, like the pixel-based approach, may utilize the data from multiple gels better. In this study a method is presented which utilises the advantages of the pixel-based approach, while simultaneously constrain the analysis to protein spot regions without the necessity of individual spot boundaries. The problematic spot identification procedure is circumvented, and the identification of significant changes performed directly on the protein spot regions created by image morphology. Significant pixels are identified by multivariate analysis, and collected into significant features where they cluster together. These features may resemble protein spots, clusters of spots and parts and shoulders of overlapping spots. The important aspect is that significant features are identified without imposing the constraint of individual boundaries representing isolated proteins. Highlighted significant features are easily visualised by a user, and can be selected for further investigation. The method is compared with results from common commercial software. While the analysis In Færgestad et al. [64] was restricted to a smaller sub-image, the improved pixel-based method presented here make it possible to analyse full size images with no scale reduction.

5.2 Materials and methods

5.2.1 2D-Gels

Gel images are prepared from samples of muscles of 7 Norwegian Red dual-purpose bulls collected 1, 2, 3, 6 and 10 hours after slaughter. The materials are described in more detail elsewhere [64].

5.2.2 Image segmentation

In the following description it is assumed that images are inverted, that is, the image background is dark, and the spots appear as light peaks rising from the background. This gives a zero baseline and positive protein spots intensity values, which is convenient for this analysis. Unless stated otherwise, all program code in this study is written in Matlab version 7.2.0 (Mathworks).

To make the pixel grey values in each image comparable, all images need to be normalised and aligned properly. Alignment was performed using the commercial software TT900 S2S (Nonlinear Dynamics). Details of the image alignment are given by Færgestad et al. [64]. The images were then normalised by dividing each image with the total intensity of this image. Uniform background was removed by subtracting from each image its minimum intensity value. A typical normalised and aligned gel is shown in figure 5.1. A sub image of this gel is also selected (figure 5.2(a)) to better illustrate the steps in the analysis procedure. A flow chart of all the steps involved in the analysis is shown in figure 5.3.

To identify regions in the image related to proteins spots, segmentation by image morphology is performed on all 35 images. It should be noted at this stage that the pixel-based method introduced can be used with any spot identification procedure, as long as all regions consisting of protein spots are identified. No assumptions on individual spot boundaries are made within these regions. Other procedures or segmentation outputs from commercial 2-DE software can also be used to produce the regions used in this study. Each of the 35 images in the experiment is processed in the following way: Single spikes are removed using a median filter of size 3. The images are then corrected for streaks and non-uniform background by image morphology. Image morphology is performed by successively dilating and eroding images with structural elements similar to the features in the image one wants to keep or remove. The use of morphology in 2-DE and for images in general is well described by Skolnick [49] and Sternberg [48]. The selected structural element for streak identification is

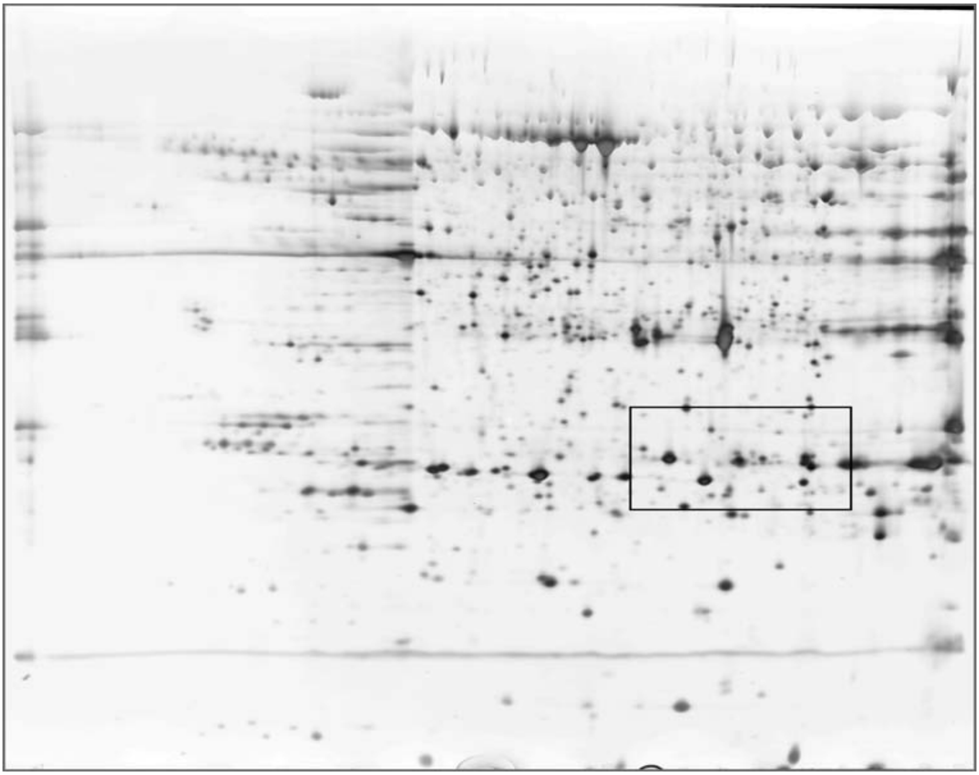
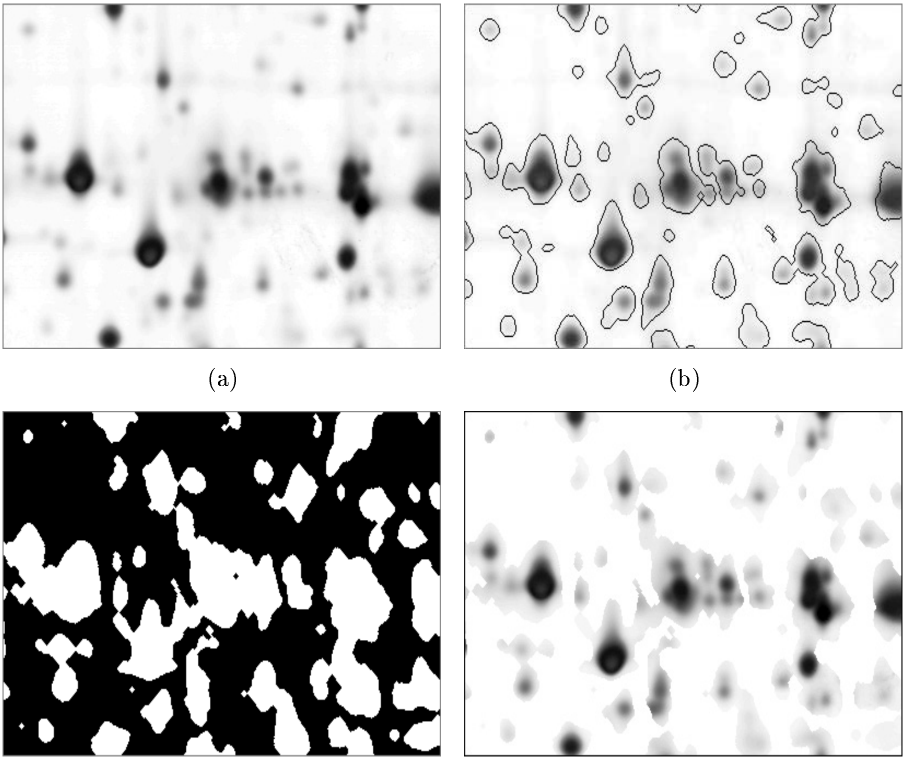


Figure 5.1: One of the gels used in the experiment. The frame indicates the sub-image used to illustrate the steps in the rest of the analysis.



(a)

(b)

(c)

(d)

Figure 5.2: Sub-images illustrating the different steps in the analysis procedure. (a): Normalised image. (b): Boundaries for protein spot areas after segmentation by image morphology. (c): Mask based on all gels. (d): Masked images.

a line 61 pixels in length, and for protein spot regions a circle with a diameter of 40 pixels is chosen. After performing these morphology operations, a single threshold is sufficient to identify the spot regions in the resulting image. The threshold value is set to 0.025 for images with intensity values between 0 and 1. The resulting protein spot regions in the selected sub-image are shown in figure 5.2(b). Note that the goal at this stage is purely to find areas in the image were protein spots are present. Any assumptions on the individual spot boundaries in each area or merged spots are not considered.

The identified spot regions produce a binary image for each gel, where 1 denotes areas with protein spots, and 0 denotes image background and other

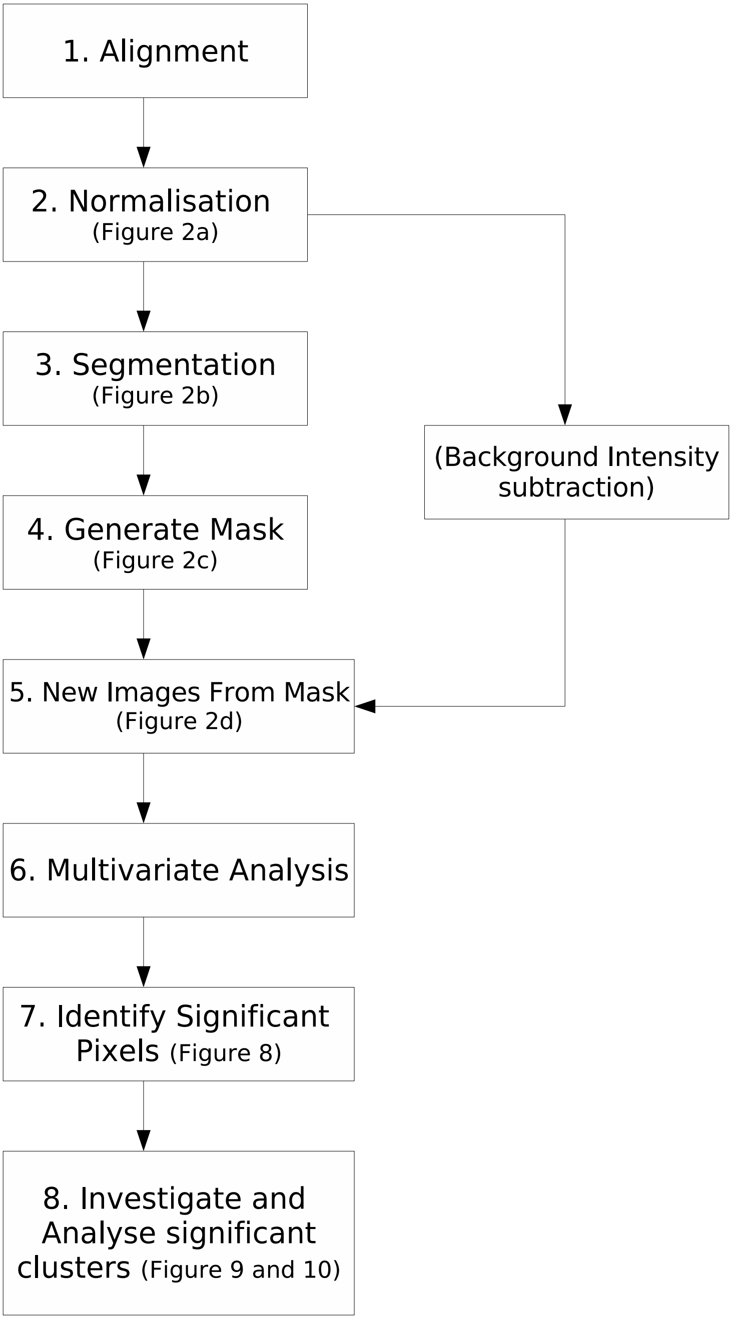

1. Alignment
2. Normalisation (Figure 2a)
3. Segmentation (Figure 2b)
4. Generate Mask (Figure 2c)
5. New Images From Mask (Figure 2d)
6. Multivariate Analysis
7. Identify Significant Pixels (Figure 8)
8. Investigate and Analyse significant clusters (Figure 9 and 10)
(Background Intensity subtraction)

Figure 5.3: Flow chart showing the steps in the analysis procedure.

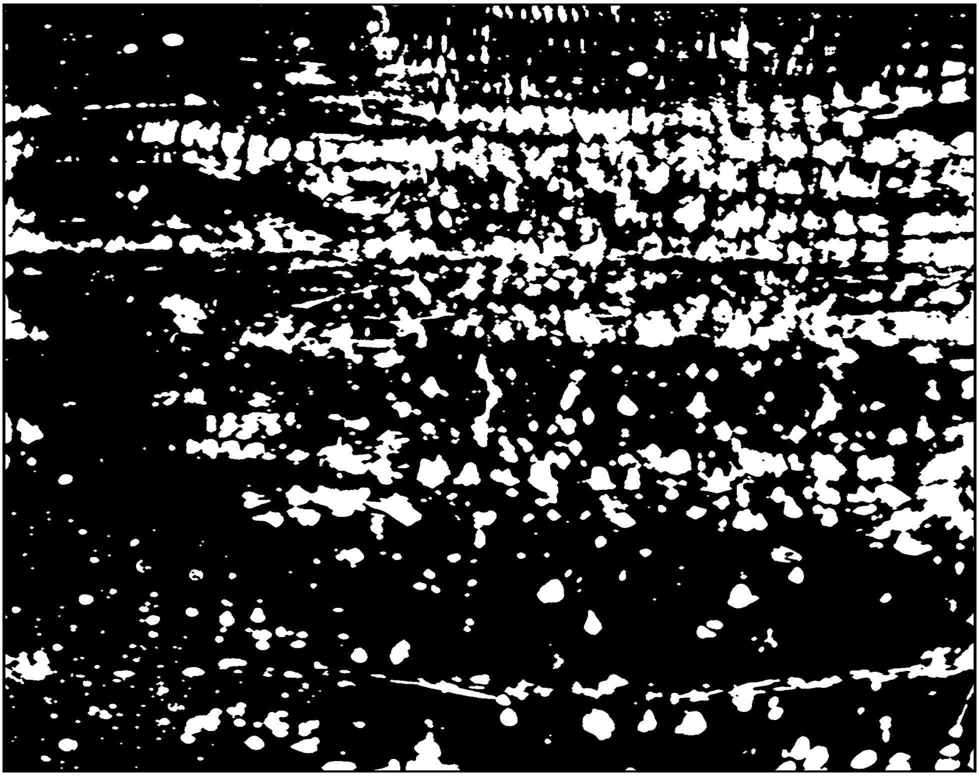


Figure 5.4: Mask used on all gels in full size.

artefacts. To select the protein spot areas used for analysis, it is important that information present in all gels is considered. Based on the union of all pixels representing protein spot areas in each individual gel, a mask is created. This means that if a pixel has value 1 in only one of the binary images, it will be included in the mask. New images are created by applying the resulting mask to each aligned and normalised gel image. The mask for the area occupied by the sub-image is shown in figure 5.2(c) and the resulting new image, with background and noise blanked out, is shown in figure 5.2(d). The mask for the full gel is shown in figure 5.4.

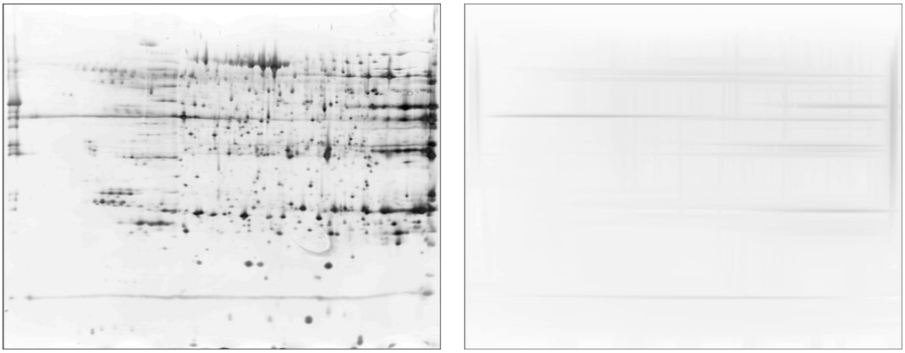
It should be noted at this stage that the median filtering, streak and spot identification in step 3 in figure 5.3 is done merely to assist the segmentation. The new intensity values resulting from the operations performed are not used in the analysis. The analysis can be performed on the intensity values directly as they appear after the aligning and normalisation step, but only using the pixels specified by the mask in step 4.

5.2.3 Background intensity subtraction

The morphology operations described in the previous chapter manages to identify protein spot areas successfully. However, the image intensities in the resulting images are generally altered too much, and one should be precocious in using them for analysis. Instead it is preferred to use the original image intensities. The original images usually still have background intensities not related to proteins, which would be advantageous to subtract before the analysis on the pixel level. For this background intensity subtraction, a method introduced by Lieber et al. [53] is adopted. The method was originally used to remove dominating fluorescence in one-dimensional Raman spectroscopy. The basic idea of this method is first to fit the original signal to a polynomial of some degree. The polynomial fit is then subtracted from the original signal, to create a new signal more similar to the background and with the highest peaks removed. A polynomial of the same degree is then fitted to this new signal, giving a second estimation of the background, which is again subtracted. This procedure goes on in an iterative fashion, until the estimate of the background becomes stable, or a certain number of iterations are reached. For this study a 4th degree polynomial and 50 iterations were found sufficient to produce the desired results. The method was originally introduced for one-dimensional signals, but is easily extendable two 2D-images. However, in this case the one-dimensional approach was preferred, because of the possibility to correct background in streaks. Streaks are not considered as background if the 2D-approach is chosen. The polynomial was thus fitted in the described fashion line-by-line in both vertical and horizontal direction, producing two background images. A final background intensity image was produced by selecting the highest values for each pixel in the two images. A typical background image is shown in figure 5.5(b).

5.2.4 Analysis by software using common spot boundaries

Results from the pixel-based approach are compared with results using the commercial software Progenesis SameSpots (Nonlinear Dynamics). Analysis by SameSpots is based the use of common spot boundaries, that is, spot boundaries are defined commonly for all gels and the spots compared between gels have identical boundaries in all gels. In SameSpots no parameters are necessary to perform spot identification. All 35 gels analysed were grouped according to the time response, and significant spots identified by Analysis of



(a) (b)

Figure 5.5: (a): Original image. (b): Estimated background image.

Variance (ANOVA) at 95% significance level.

5.2.5 Multivariate analysis

Principal component analysis (PCA) [54] is a well known multivariate analysis method where a data matrix **X** is decomposed into a set of latent variables (called principal components) and noise. Principal Components can be understood in terms of scores and loadings. Loadings consist of the weights each original variable is given in each principal component, while the scores are the coordinates the samples are assigned in the new coordinate system consisting of the principal components. In this study the matrix **X** consists of the 35 experimental gels as samples, while the intensities of the 672 948 pixels constituting the mask makes up the variables for each sample, giving an **X**-matrix of size 35 × 672 948. The first principal component maximises the variation in the original **X**-matrix. After the calculation of the first component, the information contained in this component is subtracted from **X**, and the second component maximises the variation in this new **X**. New components are calculated in the same manner, until there is no structure left in **X**, and the components start to model noise. Such a decomposition of **X** has several advantages. First the number of latent variables is usually much smaller than the original number of variables. The latent variables are also independent (orthogonal) which is rarely the case for the original variables. Finally relationships between samples and variables are easily visualised by plotting the resulting scores and loadings.

Partial Least Squares Regression (PLSR) is another common multivariate method closely related to PCA, and is used to build a regression model between a data-matrix $\mathbf{X}$ with samples and variables, and a matrix $\mathbf{Y}$ with responses or design factors. In PLSR the data matrix $\mathbf{X}$ is decomposed in a similar fashion as PCA. However, in this case the decomposition is guided by one or several responses of interest contained in matrix $\mathbf{Y}$. In PLSR it is the variation in the covariance matrix $\mathbf{X^TY}$ which is maximised for each PLSR-component. The PLSR algorithm thus looks for variations in $\mathbf{X}$ that are relevant for the prediction of response $\mathbf{Y}$. A full description of PLSR and its algorithm is given in [54, 55].

Because of the large number of variables with respect to the number of samples in this study, there is a chance of over-fitting the data when using PLSR. To avoid overfit PLSR models are often validated by a method called cross validation [54, 57]. In cross validation the PLSR-model is checked by leaving out a number of samples from the calibration set, and using them as temporarily test samples. The PLSR model is re-calculated using the remaining samples, and the left-out samples are predicted using the new model. This procedure is repeated, leaving out a different sub-set for each new model. A special version of this procedure is the leave-one-out cross validation, where only one sample is left out at a time, and the procedure is continued until all samples have been left out once. In this way all samples work as independent test-sets for the corresponding PLSR sub-model based on the remaining samples. Cross validation is an established way of validating the number of significant components in a PLSR-model, and avoiding the problem of over-fitting by including noisy components. Another advantage of using cross-validation, is that a set of regression coefficients for each variable (in this case the pixels are the variables) are calculated for each sub-model. This introduces the possibility to identify significant variables (pixels) based on the stability of the regression coefficient over the different sub-models using simple statistical tests.

All pixels identified by the mask are compared and analysed using the multivariate methods described above. To apply the multivariate methods, the gels need to be unfolded prior to the analysis. Analysing unfolded gels in this way has many applications and advantages, and is described thoroughly by Færgestad et al. [64]. PCA is used to identify the most important variations in the data in general, and PLSR is used to find the variations in the data best correlated to the response variable which is time after slaughter. PCA is thus an unsupervised method for looking at variations in the data, while PLSR is supervised by the response variable. Significant pixels are identified by performing a t-test on the regression coefficients from a PLSR leave-one-out cross validation. To validate the importance of applying a mask, the same multivariate methods are applied to unmasked images as they appear on step 2 in figure 5.3.

When comparing results from the masked dataset to the unmasked, it is necessary to reduce the size of the images. The unmasked images, consisting of approximately 3 million pixels, are too large to be handled by Matlab, whereas the masked images are reduced to 672 948 pixels. For the comparison both the segmented and the un-segmented images were reduced by one half in each direction using the function "imresize" in Matlab with interpolation option "bicubic".

The data matrix **X** for the reduced unmasked image contains 749 490 pixels (variables) for each of the 35 samples The reduced masked image contains only 168 185 pixels. The response variable **Y** constitutes time after slaughter for the 35 samples, and is thus a vector of the same length as the number of samples.

5.3 Results

5.3.1 Multivariate analysis

All data matrices were mean-centred and standardized by dividing each variable (pixel) with its standard deviation across all samples. When standardising each variable, all pixels are given the same impact on the multivariate model. The scores for the first two principal components from a PCA are plotted for all 35 samples in figure 5.6. As can be seen from the figure, there is considerable variation between the 7 animals with respect to the response variable (hours after slaughter). However, there is a general tendency for the response variable to increase along the first component. This tendency becomes clearer when considering figure 5.7(a) where the average score over all seven animals for each response are plotted for the first two components. From this figure a linear dependence between the response and the first two components can clearly be seen. PCA was also performed on the unmasked data, and the results for average scores are shown in figure 5.7(b). In this case the pattern is more complex, and the nice and simple correspondence between the first two components and the response variable is lost. It can thus be concluded that sources of variation in the data not related to the response variable have been removed in the morphology procedure, which was as intended.

Figure 5.8shows the explained **Y** variance with respect to the number of components used in a PLSR for four different models. In the PLSR-model, the response variable (time after slaughter) is used as a target variable in the

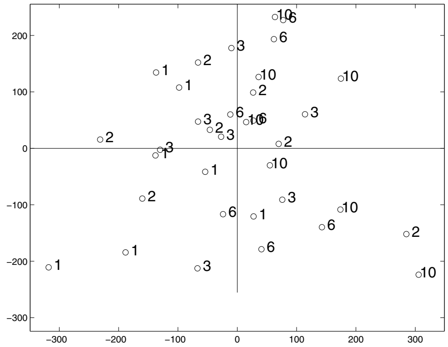
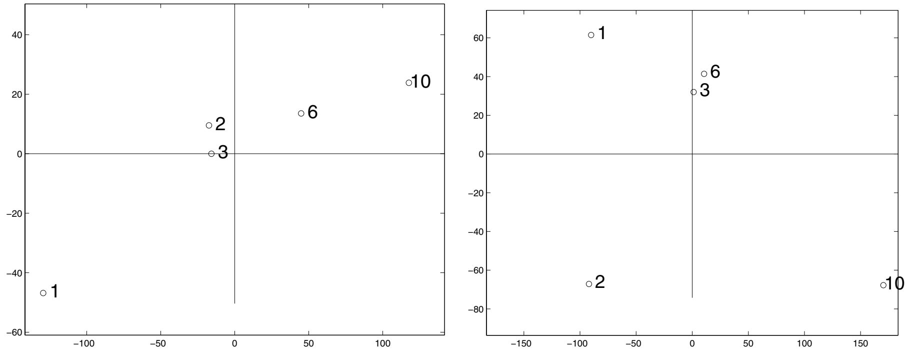


Figure 5.6: Score plot for the first two principal components for all 35 samples. The samples are numbered according to the response variable **Y** (hours after slaughter).



(a)

(b)

Figure 5.7: verage score for all samples at each time in the first two components. (a): Masked data. (b): Unmasked data.

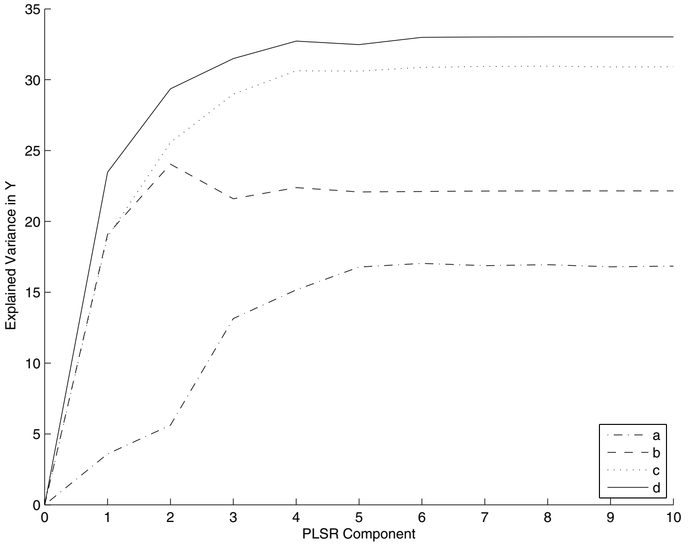


Figure 5.8: Explained variance of response variable Y for four models. (a): Unmasked image with no background subtraction. (b): Unmasked images with background subtracted. (c): Masked images with no background subtraction. (d): Masked images with background subtracted.

decomposition of the data matrix as explained in the previous section. Explained **Y**-variance is an important measure of the performance of a PLSR model. A high percentage of explained **Y** variance indicates that variations in the response **Y** have been captured by the model, while a low percentage indicates that the model has failed to do so. Naturally one seeks to maximise the explained **Y**-variance, without over-fitting the data. To avoid over-fit, all components in the PLSR-models are validated using leave one out cross validation. A summary of results are given in table 5.1.

Several interesting conclusions can be drawn when considering figure 5.8 and table 5.1. It is immediately clear that a model performs poorly when neither background correction nor a mask is applied. The background corrected data performs similarly in the first component for both masked and unmasked data, however the explained **Y** variance increases for the masked data when more components are used, which is not the case for the unmasked data. It is also interesting to observe that the masked data where no background intensity correction is performed also performs better than the unmasked background corrected data if more than one component are considered. The difference in performance between the two masked models (with and without background correction) is marginal, at least in the latter components. The interpretation of this must be that the background intensity correction and the identification of spot regions perform similar tasks, which is to remove unwanted sources of variations not related to protein spots. The morphology procedure does this by removing pixels not related to protein spots, while the background correction changes the image intensity values outside the protein spots to reduce their influence on the model. Performing background correction will, however, also alter the intensity values of the protein spots analysed. Whether this is desired or not, depends on if the proteins are placed on top of the background, or if they are merged with the background. In the first case background correction would be appropriate, while in the latter case background correction would alter protein spot intensities erroneously. For these data, background correction gave a marginally higher explained **Y** variance for the masked data, however the difference is too small to draw a general conclusion. At this point it should be considered whether applying a mask gives any benefits compared to performing an analysis on unmasked images. Figure 5.8 shows that results are poor using unmasked images directly in the analysis, but improve considerably when a mask is applied. Background correction makes analysis on unmasked images possible, but has only a marginal effect on the masked images. It can thus be concluded that applying a mask makes analysis possible without background correction, which is not possible when analysing unmasked images. When

Table 5.1: Summary of results for five models. Masked data are pixels selected by the morphology procedure. The data in the table are from the reduced images (one half in each direction) unless stated otherwise.

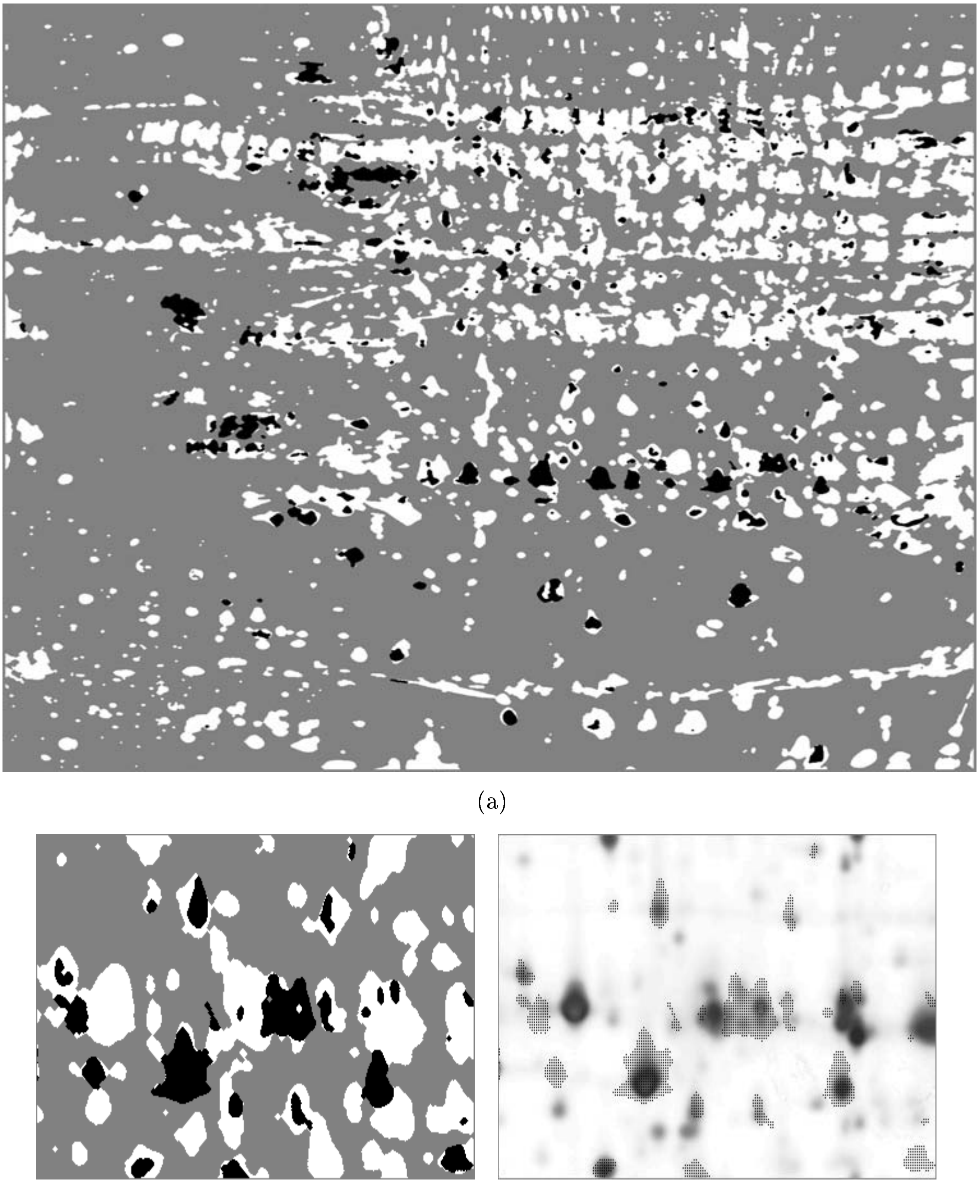
	Pixels analysed	Explained **Y** (first component)	Significant pixels
Masked data	168 185	19%	25 670
Masked data with background correction	168 185	23%	26 654
Masked data with background correction (not reduced)	672 948	23%	106 185
Unmasked data with background correction	749 490	24%	70 920
Unmasked data	749 490	4%	23 368

looking at table 5.1, it is evident that background correction and applying a mask give similar results with respect to the **Y** variance, however the number of significant pixels used in the two models are very different. The masked data need about 1/3 of the significant pixels used for the unmasked model to achieve a similar explained **Y** variance. Færgestad et al. [64] also addressed the issue of too many significant pixels, especially in regions constituting image background where no protein spots are expected to be present. They solved this problem by using a cut-off on the regression coefficients. When performing data-reduction by image-segmentation, no such cut-off is needed, because the background has already been removed. Thus applying a mask offer the same degree of explanation as Færgestad et al. [64], but using far less input variables. Finally it should be noted that the conclusions above where all drawn based on the reduced images. However, as shown in table 5.1, the explained **Y**-variance is identical for the reduced and original images in the case for masked data. This was generally true for all the masked models (results not shown). It is concluded that information drawn from the reduced images is transferable to the images using all pixels. Analysis on full, unmasked images were not possible because of storage problems due to the large number of pixels to be analysed (approximately 3 million).

5.3.2 Significant areas

An important step in the presented analysis is the identification of significant pixels with respect to the response variable (time after slaughter). The data selected for this analysis are the masked and background-corrected images, since this dataset gave the best performance with respect to explained **Y**-variance and for visualisation purposes. Significant pixels where identified using a two-sided t-test on the regression coefficients from the 35 sub-models from the cross validation. The significance cut-off value was set at 0.05. Because the number of significant pixels is quite large, a high number of false positives are expected. This creates a problem if all pixels (variables) are independent, because it is not possible to decide for sure whether a variable is a true or false positive. However, in gel images it is expected that a pixel will generally be correlated with its neighbouring pixels. If one pixel belongs to a certain protein spot, it is expected that this protein will be correlated with other pixels belonging to this spot. This means we can make assumptions on significant areas in the gel by identifying areas where significant pixels cluster together, as shown in figure 5.9(c).

Another observation is that several of the areas marked as significant have



(a)

(b)

(c)

Figure 5.9: Significant pixels highlighted in black. (a): Full size image. (b): Sub-image. (c): Significant areas indicated on top of the sub-image in figure 5.2(a).

pixels with stable, but very small, regression coefficients, meaning that the maximum intensity difference over all images in this area is very small compared to the total intensity range. Such small variations are not visible to the eye, and are of minor interest to the analyst. The number of significant features in the images can thus be reduced by requiring that they have a certain size (in connected pixels) and a minimum intensity span over all images. In this study the minimum size were set to 5 pixels, and the minimum intensity span to 0.07 (on a zero-to-one scale), but these parameters can be changed according to the circumstances. A total of 277 significant features were identified to satisfy the criteria stated above. These features are highlighted (in black) in figure 5.9(a) for the sub-image, and figure 5.9(b) for the full image. Significant areas indicated on top of the original image are shown in figure 5.9(c) for the sub-image. Three selected areas are displayed with marked cross-section in figure 5.10 to figure 5.12, to show that the identified features truly represent protein variations with respect to time after slaughter.

An advantage of not using fixed individual spot boundaries in the analysis is the ability to detect significant variations in spot-shoulders in highly overlapping proteins. These variations are often lost when boundaries are applied. The individual boundaries are often not able to separate multiple spots, and will thus treat multiple spots as single entities. Because weak spot shoulders are often dominated by larger, un-significant variations in the rest of the cluster, the variations due to the weaker spot is lost. Using a pixel-based approach increases the ability to discover such variations. Three examples are shown in figure 5.13 to figure 5.15.

5.3.3 Comparison with software using common spot boundaries

The significant regions displayed in figure 5.10 to 5.15 are also analysed by the commercial software Progenesis SameSpots (Nonlinear Dynamics). Results using SameSpots on these regions are shown in figure 5.16, where a-f correspond to the regions in figure 5.10 to 5.15 respectively. SameSpots performs similar to the pixel-based analysis for the spots in figure 5.10 and 5.12. Both of these spots are identified as significant by SameSpots. SameSpots is not able to identify a spot representing the exact region of figure 5.11, but the spot highlighted in figure 5.16(b) is the closest significant match and has a similar profile. The three spots to the left in figure 5.16(b) are not found significant, however the spot to the lower right is significant. This spot is not detected by the pixel-based approach, where it is interpreted as a streak and thus not included in the mask of analysed pixels. Only pixels included in the mask

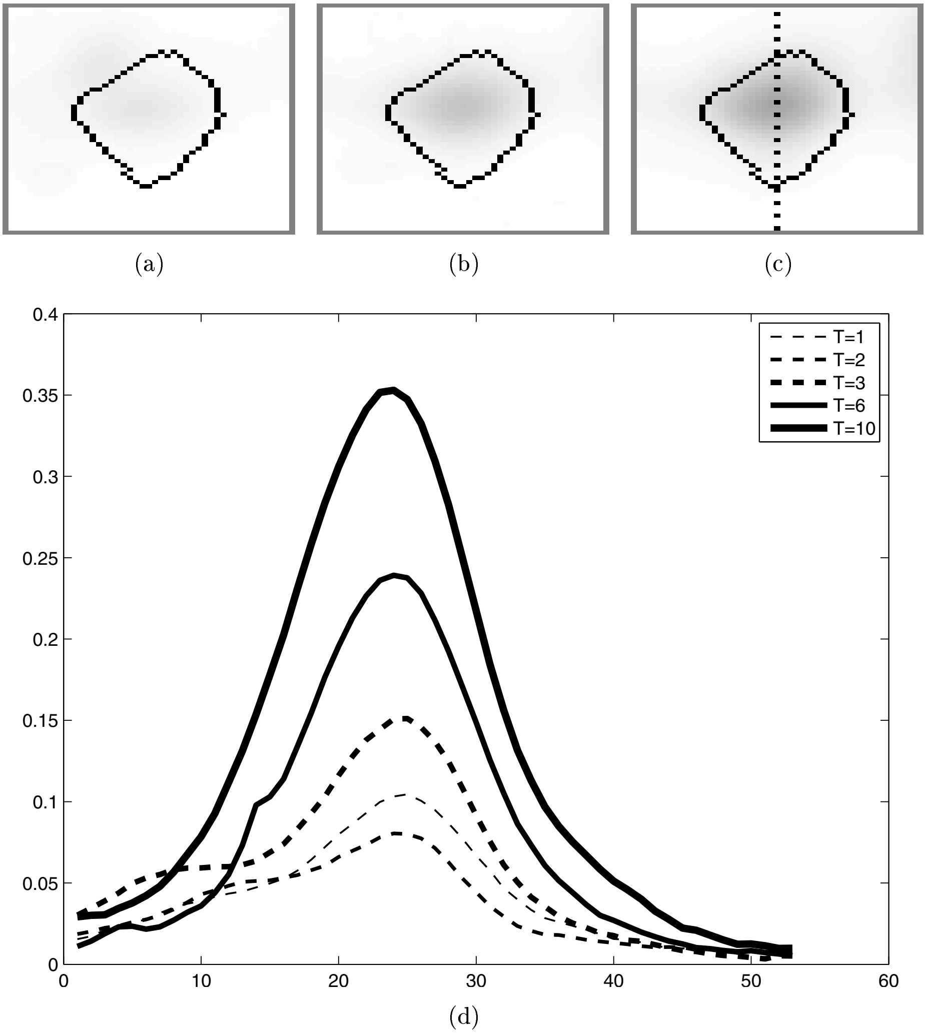


Figure 5.10: Closer inspection of significant pixel cluster. (a),(b) and (c) are average images of the spot taken 1, 6 and 10 hours after slaughter respectively. (d) is the cross-section shown in (c) displayed as a curve for each of the five time averages. This single significant spot clearly increases in intensity with time.

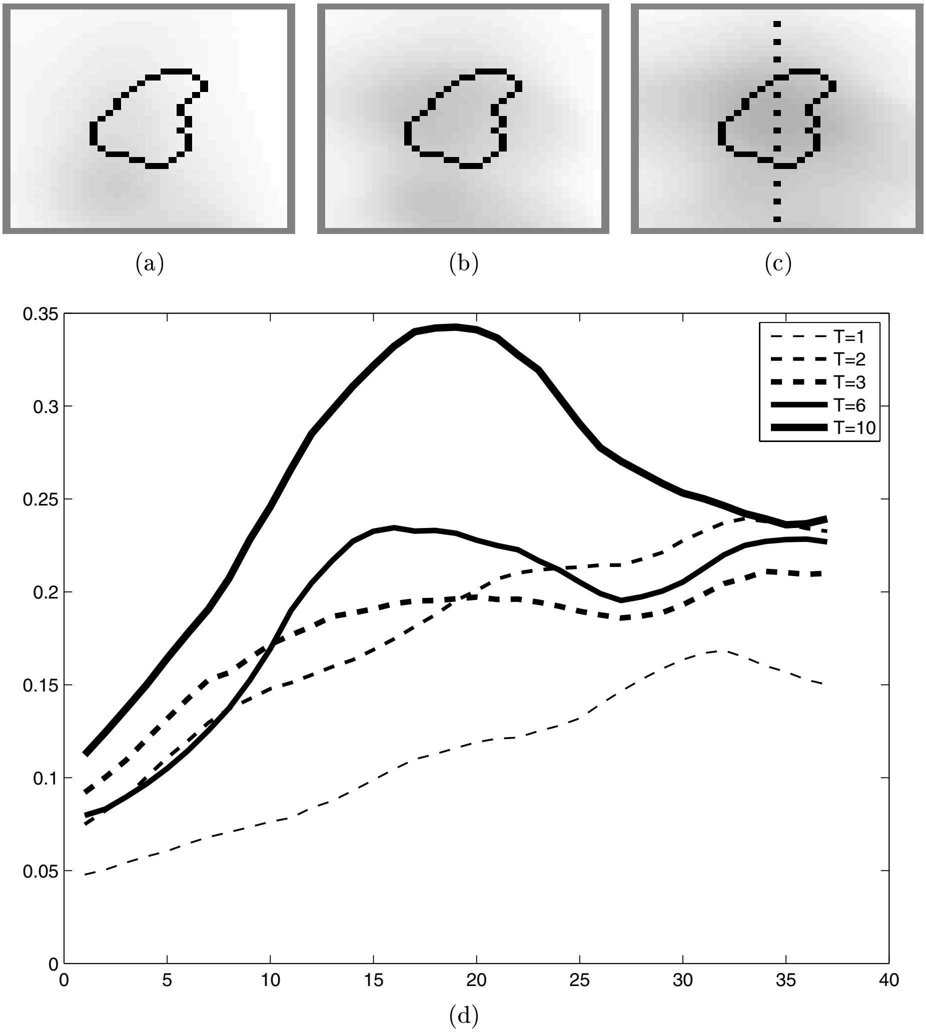
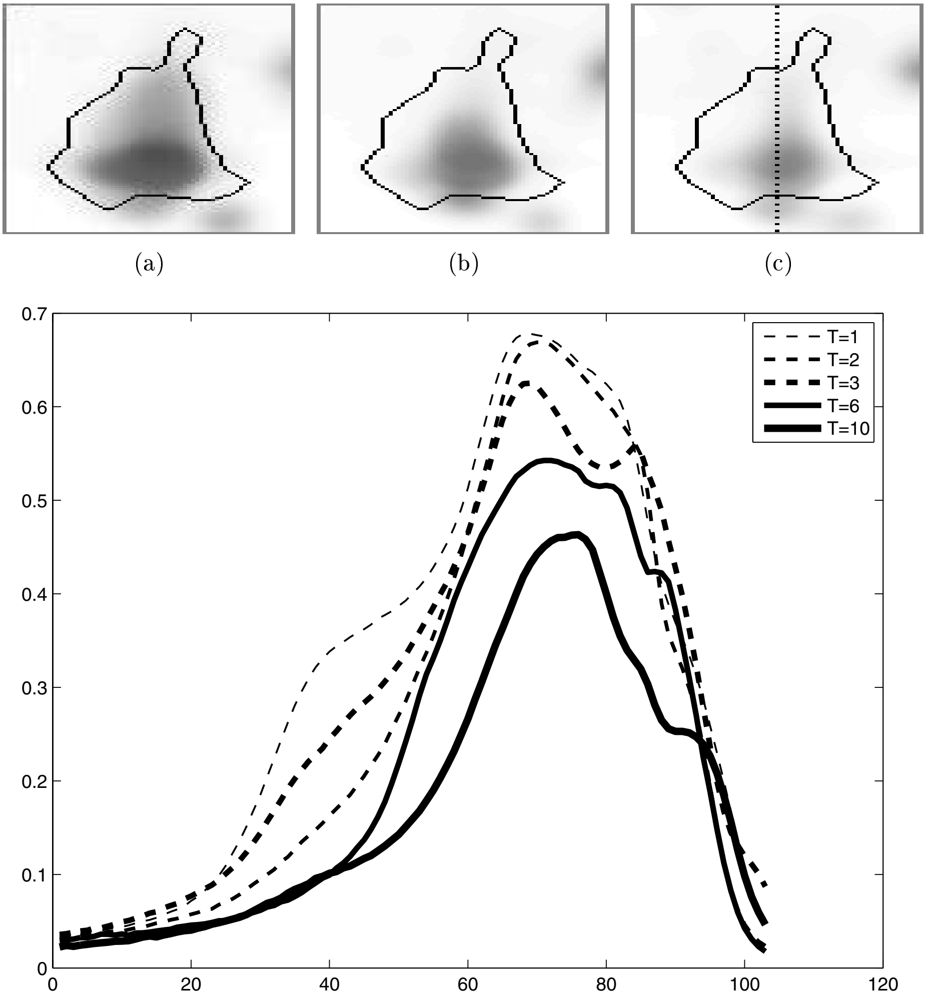


Figure 5.11: Closer inspection of significant pixel cluster. (a),(b) and (c) are average images of the spot taken 1, 6 and 10 hours after slaughter respectively. (d) is the cross-section shown in (c) displayed as a curve for each of the five time averages. The marked region increases with time, while the neighbouring spot remains constant.


(a) (b) (c)

(d)

Figure 5.12: Closer inspection of significant pixel cluster. (a),(b) and (c) are average images of the spot taken 1, 6 and 10 hours after slaughter respectively. (d) is the cross-section shown in (c) displayed as a curve for each of the five time averages. This spot displays a decrease in intensity with time.

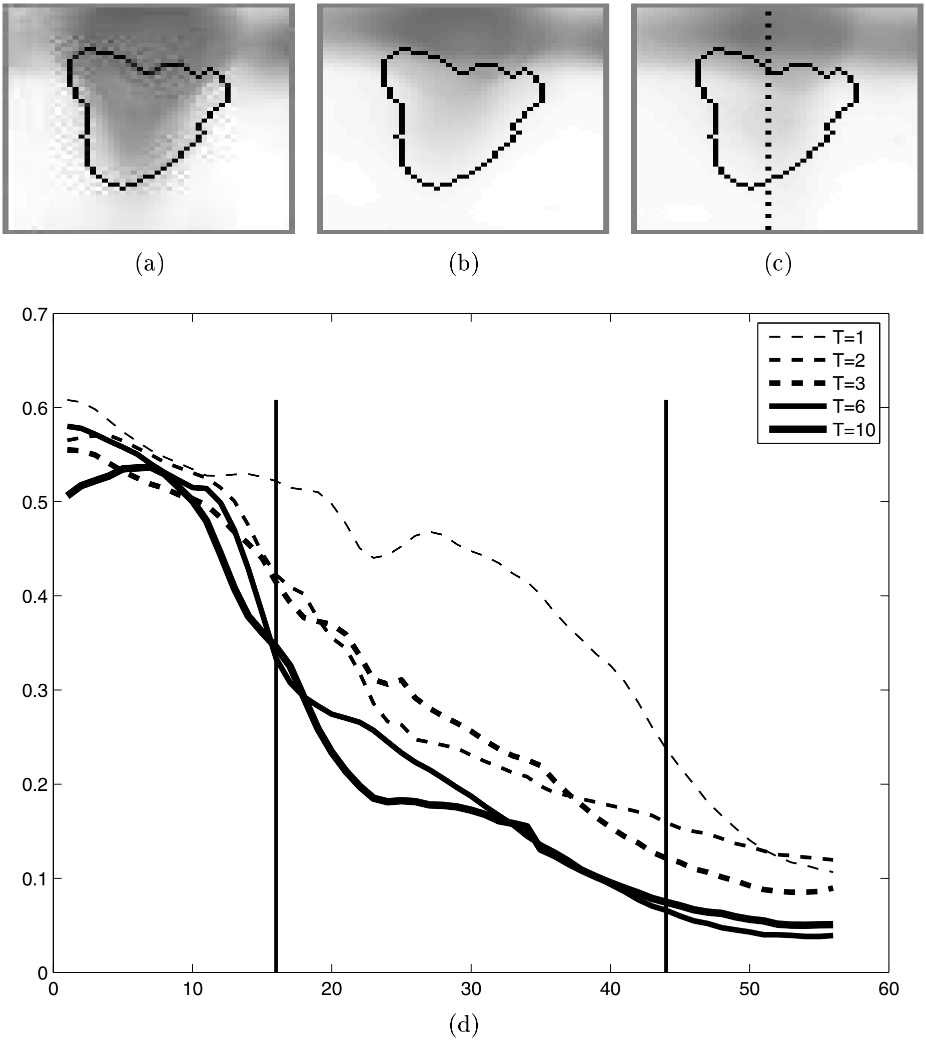
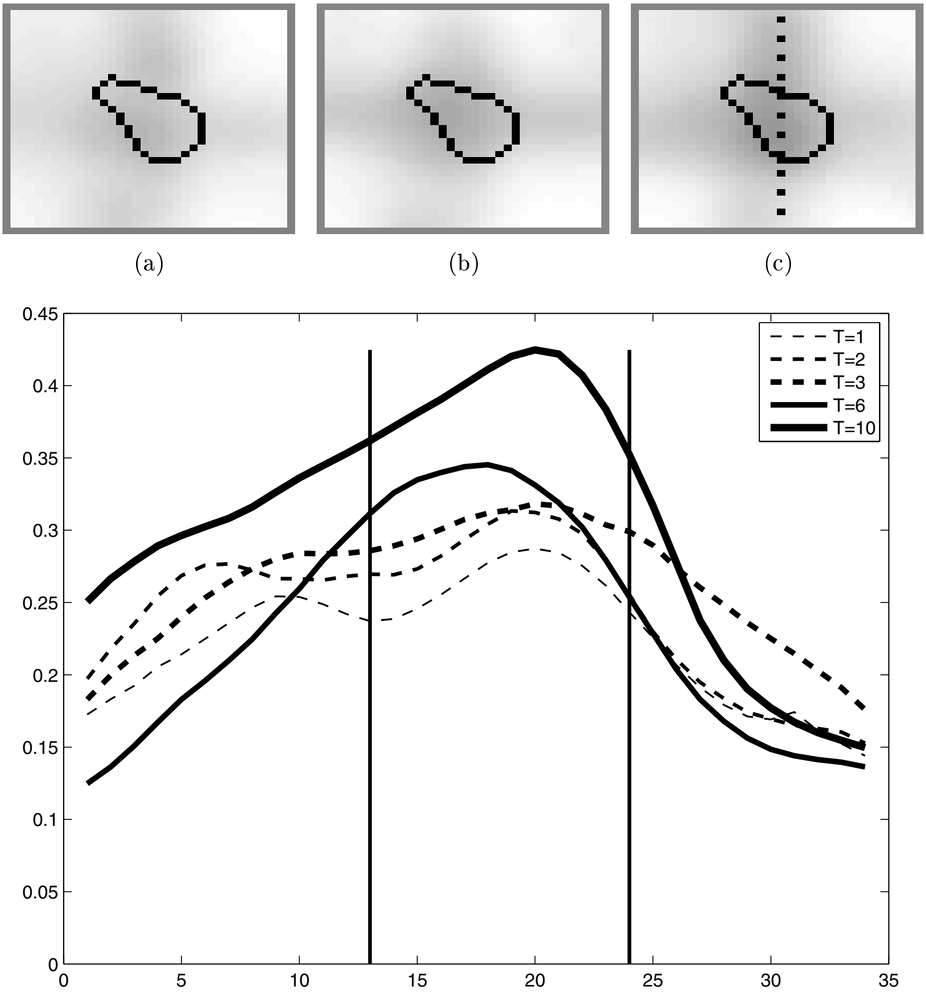


Figure 5.13: Closer inspection of significant pixel cluster. (a),(b) and (c) are average images of the spot taken 1, 6 and 10 hours after slaughter respectively. (d) is the cross-section shown in (c) displayed as a curve for each of the five time averages. The horizontal lines indicate the boundary of the area marked as significant. Typical example of a shoulder decreasing in intensity with time, while its larger neighbour remains constant.



(a)

(b)

(c)

(d)

Figure 5.14: Closer inspection of significant pixel cluster. (a),(b) and (c) are average images of the spot taken 1, 6 and 10 hours after slaughter respectively. (d) is the cross-section shown in (c) displayed as a curve for each of the five time averages. The horizontal lines indicate the boundary of the area marked as significant. Part of a larger spot complex increasing significantly with time.

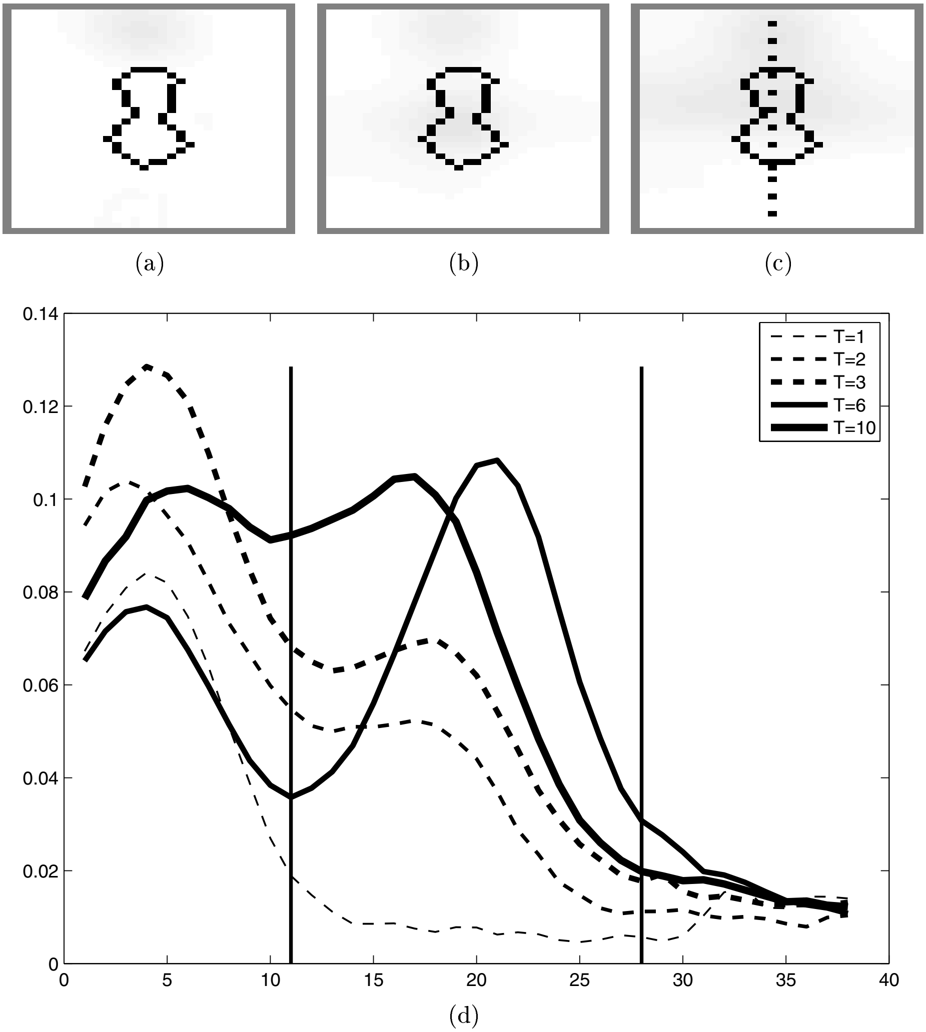
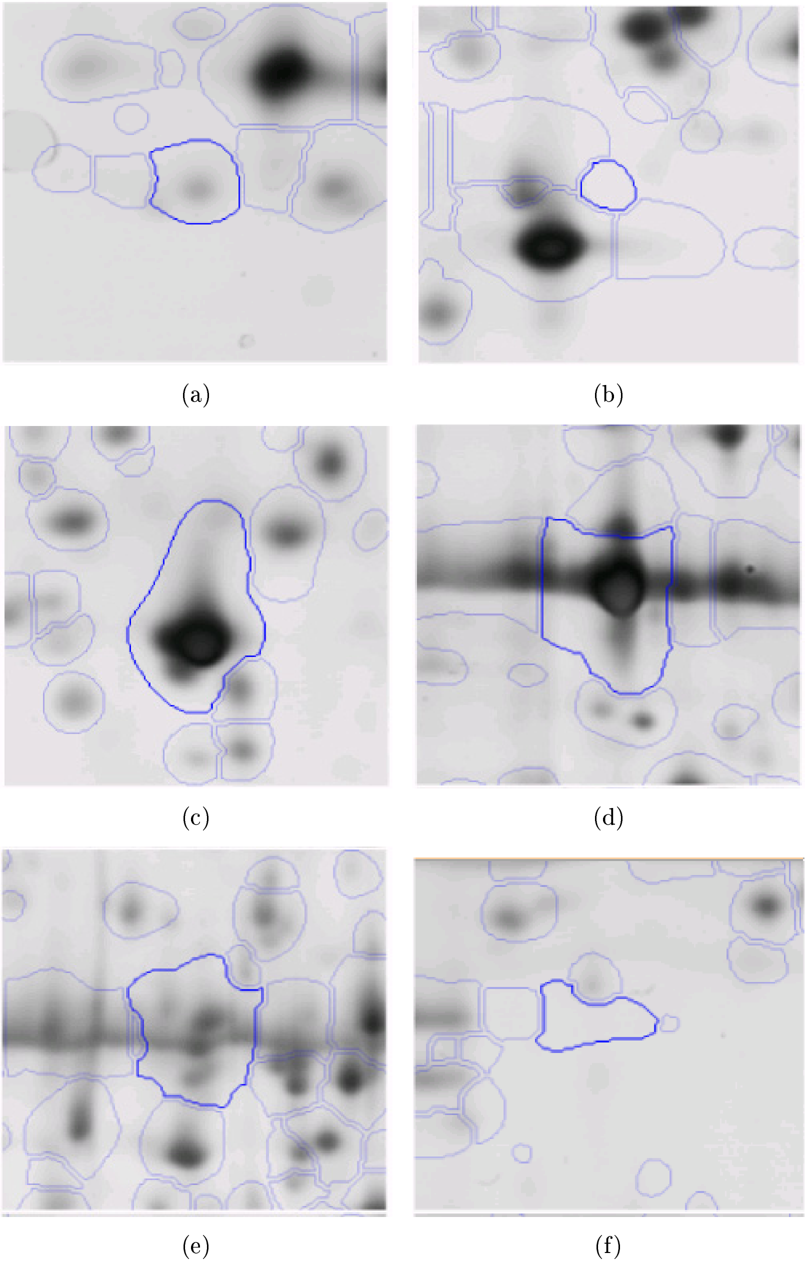


Figure 5.15: Closer inspection of significant pixel cluster. (a),(b) and (c) are average images of the spot taken 1, 6 and 10 hours after slaughter respectively. (d) is the cross-section shown in (c) displayed as a curve for each of the five time averages. The horizontal lines indicate the boundary of the area marked as significant. Weak overlapping region where only one of the spots changes significantly.

are subjected to analyses as described previously. A spot corresponding to the significant shoulder in figure 5.13 is highlighted in figure 5.16(d). The spot shows significant variation, however SameSpots is not able to distinguish the shoulder from the larger neighbouring spot, resulting in wrongly estimated spot volumes. In such situations the pixel based approach is preferred, because of the ability to target significant regions more directly. Another situation where the pixel-based method performs better is the one shown in figure 5.14 and 5.16(e). The complex region is not resolved properly by SameSpots, and several spots are located within the same boundary. The whole region does not display enough significant variation to be detected by SameSpots. The pixel-based approach does not consider such boundaries in complex regions, and thus smaller local variations within these regions are more easily detected. This is also true for the region in figure 5.15 and 5.16(f). SameSpots located a spot corresponding to this region, however this spot is not found to be significant. The spot boundary found by SameSpots is substantially different from the one identified by the pixel-based approach, and is probably the reason for this discrepancy. The result is again a failure to detect a protein displaying significant variation.

Defining boundaries for individual spots has some advantages in regions where several neighbouring spots are significant, as shown in figure 5.17. In such situation more information is achieved if the larger significant region is resolved into components resembling protein spots (figure 5.17(b)). However, a resolution of such regions can always be performed after the significant pixels are identified. It would thus be possible to concentrate the effort of defining isolated protein spot entities to regions which display significant variations, without the need to define spot boundaries in other parts of the gel. It should also be noted that the significant shoulder pointed to by an arrow in figure 5.17(a) is not detected by SameSpots.

A total of 155 spots are identified as significant by SameSpots using ANOVA on each spot with at a 95% significance level. This is fewer than for the pixel-based approach, however, a direct comparison is difficult because the spot boundaries defined by SameSpots are very different to the boundaries created by significant regions in the pixel-based approach. 119 of the 155 significant spots from SameSpots have corresponding significant regions in the pixel-based approach. Some variations between the results are expected, because ANOVA is different test than the multivariate regression analysis. Other inconsistencies are due to removed streaks (described previously) and small, low intensity spots not detected during the morphology identification of spot regions.


(a)

(b)

(c)

(d)

(e)

(f)

Figure 5.16: Significant regions from figure 5.10 to 5.15 analysed by the software Progenesis SameSpots. The regions a-f corresponds to figure 5.10 to 5.15 respectively.

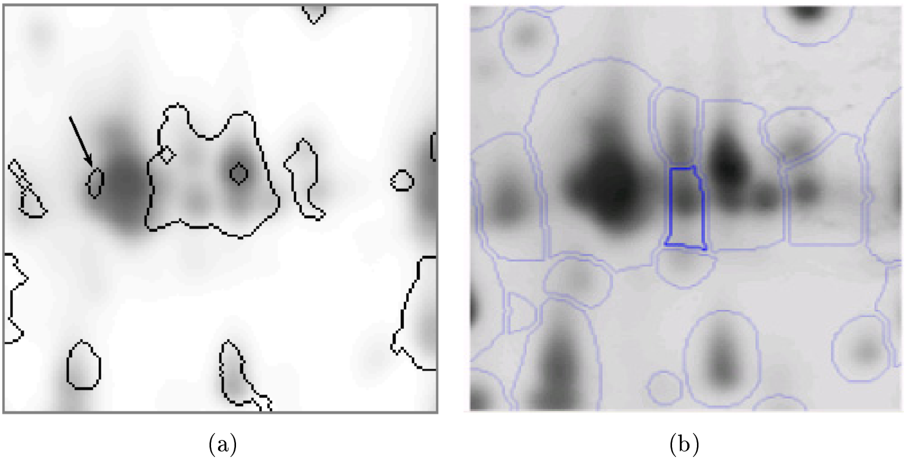


Figure 5.17: (a): A complex region with several neighbouring significant spots analysed by the pixel-based approach. The regions within the boundaries are identified as significant. (b): The same region analysed by Progenesis SameSpots. The significant shoulder pointed to by an arrow in (a) is not detected by SameSpots.

5.4 Discussion

Doing data reduction by image morphology, followed by multivariate analysis on the selected pixels has several advantages. First there is the data reduction potential. Considering the large amount of pixels in gel-images (almost 3 million in this case) some way of reducing or organising the data is usually needed to perform the analysis. Creating spot-lists is one common way of doing this, limiting the comparison analysis to a maximum of a few thousand spots rather than millions of pixels. However, as described previously, spot-lists inherit the pitfall of missing values and wrongly estimated spot volumes caused by the uncertainties in the boundaries of the spot segments. Another way of doing data reduction is to reduce the image size by replacing several pixels with their average or interpolated value. This makes the images smaller in size, but there is the risk of losing information. Noise and unwanted artifacts in the images are also propagated to the reduced images. Data reduction by applying a mask makes it possible to be considerable more selective in what information to keep, and what to throw away. It is a natural assumption that important sources of variation in the image are limited to areas where protein-spots are found. By performing image morphology or other segmentation procedures, background, noise and artefacts disturbing the analysis can be removed, and focus can be put on the areas where proteins are present. Because most of the gel image is unrelated to proteins, the amount of pixels after the segmentation is sufficiently reduced, so no average or interpolation is needed for the remaining pixels. In this study only 672 948 of the original 2 999 620 pixels were selected after morphology, making analysis possible without image reduction, and at the same time improving interpretability by removing undesired variations in the data. There is, of course, also the risk of losing information when performing segmentation. However, most gel segmentation procedures today performs satisfactorily when it comes to identifying areas in the image including proteins. The 2-DE segmentation challenge today is to isolate the individual proteins in these areas. When doing analysis on the pixel level, isolation of individual protein spots is not necessary. By using the union of all pixels as identified by the segmentation of each individual gel, the risk of leaving out important areas is also reduced. Identifying protein spot regions prior to analysis also gives improved multivariate models. Figure 5.7 and 5.8, together with the results in table 5.1 clearly show the importance by applying a mask. Reducing the data amount in this way makes it possible to emphasise the important variations in the data, improving the interpretability and highlight significant areas in the image.

Another advantage of doing analysis on the pixel level is the possibility to easily visualise significant areas, as shown in figure 5.9. Highlighting significant pixels in this way makes it a very intuitive indication from a user point-of-view where to look for important protein variations. By not focusing on specific image segments in spot-lists, it is also possible to highlight important sub-areas in unresolved spots. When using spot-lists, these segments (if unresolved) are wrongly treated as single entities, introducing serious errors when compared between the gels. Merged protein spots are generally a major challenge in automatic 2-DE analysis, and until this problem is properly handled, we consider the visualisation of significant pixels to be a valuable alternative to spot-lists when viewing significant sources of protein spot variation in 2-DE.

5.5 Concluding remarks

The authors find the approach of identifying spot regions by image morphology, followed by multivariate analysis on resulting pixels a useful alternative to existing methods in the areas of data-reduction, interpretability, significance testing and visualisation of 2-DE data.

5.6 Acknowledgements

We would like to thank The National Programme for Research in Functional Genomics (FUGE) in Norway for funding.

Concluding remarks

In the previous chapters several new models and approaches were introduced based on image segmentation in combination with multivariate analysis. Hopefully the reader has the same feeling as the author that the presented methods introduce ways to look at data from 2D-gels which can improve the interpretability and output of the final analysis. In a way this thesis follows the last years course of 2D-gel analysis in general. First the traditional spot-model is considered, and a method for filtering the output of such a model is introduced. The common spot boundary approach has lately gained popularity compared with the previously much used spot-matching procedures, and a new method for assigning such boundaries is presented. Both of these projects has the protein spot model as a basis. Last in this thesis the pixel-based methods are introduced, moving away from the necessity of defining spot boundaries before analysis. This last approach has not been explored much, neither in the literature nor by commercial software. Results from the pixel-based approach are promising, and it should be worthwhile to develop these ideas further. Finally it should be mentioned that it is doubtful whether the total protein content of a cell sample can be determined by using 2D-gel technology alone. Probably it must be used in conjunction with other methods, for example MS-technology, to completely fulfil the task of identifying, quantifying and measuring changes of all proteins in an organism.

Bibliography

[1] P.H. O'Farrell. High resolution two-dimensional electrophoresis of proteins. *Journal of Biological Chemistry*, 250:4007–4021, 1975.

[2] P.E. Bettens, J. Scheunders, D. VanDyck, L. Moens, and P. VanOsta. Computer analysis of two-dimenisonal electrophoresis gels: A new segmentation and modelling algorithm. *Electrophoresis*, 18:792–798, 1997.

[3] M Rogers, J. Graham, and R.P. Tonge. Statistical models of shape for the analysis of protein spots in two-dimensional electrophoresis gel images. *Proteomics*, 3:887–896, 2003.

[4] J. Prehm, P. Jungblut, and Klose J. Analysis of two-dimensional electrophoretic protein patterns using a video camera and a computer - II. adaptation of automatic spot detection to visual evaluation. *Electrophoresis*, 8:562–572, 1987.

[5] W.A. Lutin, C.F. Kyle, and J.A. Freeman. Quantitation of brain proteins by computer-analyzed two dimensional electrophoresis. *Electrophoresis*, 2:93–106, 1978.

[6] J.I. Garrels. Two-dimensional gel electrophoresis and computer analysis of proteins synthesized by clonal cell lines. *The Journal of Biological Chemistry*, 254:7961–7977, 1979.

[7] J. Bossinger, M.J. Miller, K.P. Vo, E.P. Geiduschek, and N.H. Xuong. Quantitative analysis of two-dimesional electropheretograms. *Journal of Biological Cehemistry*, 254:7986–7998, 1979.

[8] N.L. Anderson, J. Taylor, A.E. Scandora, B.P. Coulter, and N.G. Anderson. The TYCHO system for computer analysis of two-dimensional gel electrophoresis patterns. *Clinical Chemistry*, 27:1807–1820, 1981.

[9] K.P. Vo, M.J. Miller, Geiduschek E.P., C. Nielsen, A. Olson, and N.H. Xuong. Computer analysis of two-dimensional gels. *Analytical Biochemistry*, 112:258–271, 1981.

[10] P.F. Lemkin and L.E. Lipkin. GELLAB: a computer system for 2D gel electrophoresis analysis I: segmentation of spots and system preliminaries. *Computers and Biomedical Research*, 14:272–297, 1981.

[11] P.F. Lemkin and L.E. Lipkin. Gellab: A computer system for 2d gel electrophoresis analysis. ii. pairing spots. *Computers and Biomedical Research*, 14:355–380, 1981.

[12] P.F. Lemkin and L.E. Lipkin. GELLAB: a computer system for two-dimensional gel electrophoresis analysis III. multiple two-dimensional analysis. *Computers and Biomedical Research*, 14:407–446, 1981.

[13] A.W. Dowsey, M.J. Dunn, and G.Z. Yang. The role of bioinformatics in two-dimensional gel electrophoresis. *Proteomics*, 3:1567–1596, 2003.

[14] J.S. Gustafsson, A. Blomberg, and M. Rudemo. Warping two-dimensional electrophoresis gel images to correct for geometric distortions of the spot pattern. *Electrophoresis*, 23:1731–1744, 2002.

[15] I. Humphrey-Smith, S.J. Cordwell, and P.B. Walter. Proteome research: Complementarity and limitations with respect to the RNA and DNA worlds. *Electrophoresis*, 18:1217–1242, 1997.

[16] T. Rabilloud. Two-dimensional gel electrophoresis in proteomics: Old, old-fashioned, but still climbs up the mountains. *Proteomics*, 2:3–10, 2002.

[17] N. Campostrini, L.B. Areces, J. Rappsilber, M.C. Pietrogrande, F. Dondi, F. Pastorino, M. Ponzoni, and P.G. Righetti. Spot overlapping in two-dimenisonal maps: A serious problem ignored for much too long. *Proteomics*, 5:2385–2395, 2005.

[18] F.J. Oros and Davis J.M. Comparison of statistical theories of spot overlap in two-dimensional separations and verification of means for estimating the number of zones. *Journal of Chromatography*, 591:1–18, 1992.

[19] J. Barrett, P.M. Brophy, and J.V. Hamilton. Analysing proteomic data. *International Journal of Parasitology*, 35:543–553, 2005.

[20] E. Marengo, E. Robotti, F. Antonucci, D. Cecconi, N. Campostrini, and P.G. Righetti. Numerical approaches for quantitative analysis of two-dimensional maps: A review of commercial software and home-made systems. *Proteomics*, 5:654–666, 2005.

[21] Å.M. Wheelock and Buckpitt A.R. Software-induced variance in two-dimensional gel electrophoresis image analysis. *Electrophoresis*, 26:4508–4520, 2005.

[22] J.C. Nishihara and K.M. Champion. Quantitative evaluation of proteins in one- and two-dimenisonal polyacrylamide gels using a fluorescent stain. *Electrophoresis*, 23:2203–2215, 2002.

[23] B. Raman, A. Cheung, and M.R. Marten. Quantitative comparison and evaluation of two commercially available, two-dimensional electrophoresis image analysis software packages, Z3 and Melanie. *Electrophoresis*, 23:2194–2202, 2002.

[24] H.M. Maurer, R.E. Feldmann, J.O. Bromme, and A. Kalenka. Comparison of statistical approaches for the analysis of proteome expression data of differentiating neural stem cells. *Journal of Proteome Research*, 4:96–100, 2005.

[25] P.S. Arora, H. Yamagiwa, A. Srivastava, M.E. Bolander, and G. Sarkar. Comparative evaluation of two-dimensional gel electrophoresis image analysis software applications using synovial fluids from patients with joint disease. *Orthopaedic Science*, 10:160–166, 2005.

[26] S. Jacobsen, H. Grove, K.N. Jensen, H.A. Sørensen, F. Jessen, K. Hollung, A.K. Uhlen, B.M. Jørgensen, E.M. Færgestad, and I. Søndergaard. Multivariate analysis of 2-DE protein patterns - practical approaches. *Electrophoresis*, 28:1289–1299, 2007.

[27] R.D. Appel, J.R. Vargas, P.M. Palagi, D. Walther, and D.F. Hochstrasser. Melanie II - a third-generation software package for analysis of two-dimensional electrophoresis images: II. Algorithms. *Electrophoresis*, 18:2735–2748, 1997.

[28] K. Kaczmarek, B. Walczak, S. deJong, and B.G.M. Vandeginste. Preprocessing of two-dimensional electrophoresis images. *Proteomics*, 4:2377–2389, 2004.

[29] K. Conradsen and J. Pedersen. Analysis of two-dimensional electrophoretic gels. *Biometrics*, 48:1273–1278, 1992.

[30] Z. Smilansky. Automatic registration for images of two-dimensional gels. *Electrophoresis*, 22:1616–1626, 2001.

[31] S. Veeser, M.J. Dunn, and G.Z. Yang. Multiresolution image registration for two-dimensional gel electrophoresis. *Proteomics*, 1:856–870, 2001.

[32] K. Kaszmarek, B. Walczak, S. DeJong, and B.G.M. Vandeginste. Feature based fuzzy matching of 2D gel electrophoresis images. *J.Chem. Inf. Comput. Sci.*, 42:1431–1442, 2002.

[33] J. Salmi, T. Aittokallio, J. Westerholm, M. Griese, A. rosengren, T.A. Nyman, R. Lahesmaa, and O. Nevalainen. Hierarchical grid transformation for image warping in the analysis of two-dimensional electrophoresis gels. *Proteomics*, 2:1504–1515, 2002.

[34] A.M. Woodward, J.J. Rowland, and D.B. Kell. Fast automatic registration of images using the phase of a complex wavelet transform: application to proteome gels. *The Analyst*, 129:542–552, 2004.

[35] J. Schultz, D.M. Gottlieb, M. Petersen, L. Nesic, S. Jacobsen, and I. Søndergaard. Explorative data analysis of two-dimensional electrophoresis gels. *Electrophoresis*, 25:502–511, 2004.

[36] J.J. Tyson and R.H. Haralick. Computer analysis of two-dimensional gels by a general image processing system. *Electrophoresis*, 7:107–113, 1986.

[37] J. Panek and J. Vohradsky. Point pattern matching in the analysis of two-dimensional gel electropherograms. *Electrophoresis*, 20:3483–3491, 1999.

[38] K.P. Pleissner, F. Hoffmann, K. Kriegel, C. Wenk, S. Wegner, A. Sahlstrom, H. Oswald, H. Alt, and Fleck E. New algorithmic approaches to protein spot detection and pattern matching in two-dimensional electrophoresis databases. *Electrophoresis*, 20:755–765, 1999.

[39] T. Voss and P. Haberl. Observations on the reproducibility and matching efficiency of two-dimensional electrophoresis gels: Consequences for comprehensive data analysis. *Electrophoresis*, 21:3345–3350, 2000.

[40] H. Grove, K. Hollung, A.K. Uhlen, H. Martens, and E.M. Færgestad. Challenges related to analysis of protein spot volumes from two-dimensional gel electrophoresis as revealed by replicate gels. *Journal of Proteome Research*, 5:3399–3410, 2006.

[41] S. Luhn, M. Berth, M. Hecker, and J. Bernhardt. Using standard positions and image fusion to create proteome maps from collections of two-dimwnsional gel electrophoresis images. *Proteomics*, 3:1117–1127, 2003.

[42] M. Baker, H. Busse, and M. Vogt. An automatic registration and segmentation algorithm for multiple electrophoresis images. *Proceedings of SPIE*, 3979:426–436, 2000.

[43] F. Jessen, R. Lametsch, E. Bendixen, I.V.H. Kjærsgård, and B.M. Jørgensen. Extracting information from two-dimenisonal electrophoresis gels by partial least squares regression. *Proteomics*, 2:32–35, 2002.

[44] E. Marengo, E. Robotti, P.G. Righetti, and F. Antonucci. New approach based on fuzzy logic and principal component analysis for the classification of two-dimensional maps in health and disease - application to lymphomas. *Journal of Chrmatography A*, 1004:13–28, 2003.

[45] K. Takahashi, Y. Watanabe, M. Nakazawa, and A. Konagaya. Fully-automated spot recognition and matching algorithms for 2-D gel electrophoretogram of genomic DNA. *Genome Inform. Ser. Workshop Genome Inform.*, 9:161–172, 1998.

[46] P. Cutler, G. Heald, I.R. White, and J. Ruan. A novel approach to spot detection for two-dimensional gel electrophoresis images using pixel value collection. *Proteomics*, 3:392–401, 2003.

[47] A.D. Olson and M.J. Miller. Quantitative computer analysis of sets of two-dimensional gel electropheretograms. *Analytical Biochemistry*, 169:49–70, 1988.

[48] S.R. Sternberg. Grayscale morphology. *Computer Vision, Graphics, and Image Processing*, 35:333–355, 1986.

[49] M.M. Skolnick. Application of morphological transformations to the analysis of two-dimensional electrophoretic gels of biological material. *Computer Vision, Graphics, and Image Processing*, 35:306–332, 1986.

[50] R.K. Mannar, D.J. Smiraglia, C. Plass, and R. Wenger. Contour area filtering of two-dimenisonal electrophoresis images. *Medical Image Analysis*, 10:353–365, 2006.

[51] G.W. Horgan and C.A. Glasbey. Use of digital image analysis in electrophoresis. *Electrophoresis*, 16:298–305, 1995.

[52] Y. Kim, J. Kim, Y. Won, and Y. In. Segmentation pf protein spots in 2D gel electrophoresis images with watershed using hierarchical threshold. *Lecture notes in computer science*, 2869:389–396, 2003.

[53] C.A. Lieber and A. Mahadevan-Jansen. Automated method for subtraction of fluorescence from biological raman spectra. *Applied Spectroscopy*, 57:1363–1367, 2003.

[54] H. Martens and T Næs. *Multivariate Calibration*. Wiley, Chichester, 1989.

[55] S. Wold, H Martens, and H. Wold. The multivariate calibration problem in chemistry solved by the PLS method. In A. Ruhe and B. Kagstrom, editors, *Proc. Conf. Matrix Pencils, Lecture notes in mathematics*, pages 286–293. Springer Verlag, Heidelberg, 1983.

[56] H. Martens and M. Martens. *Multivariate analysis of quality: An Introduction*. Wiley, Chichester, 2001.

[57] M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. B*, 36:111–147, 1974.

[58] N.E.S. Flaete, K. Hollung, L. Ruud, T. Sogn, E.M. Faergestad, H.J. Skarpeid, E.M. Magnus, and A.K. Uhlen. Combined nitrogen and sulphur fertilisation and its effect on wheat quality and protein composition measured by SE-FPLC and proteomics. *Cereal Science*, 41:357–369, 2005.

[59] X. Jia, K.I. Hildrum, F. Westad, E. Kummen, L. Aass, and K. Hollung. Changes in enzymes associated with energy metabolism during the early post mortem period in longissimus thoracis bovine muscle analyzed by proteomics. *Journal of Proteome Research*, 5:1763–1769, 2006.

[60] R.M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE T. Syst. Man. Cyb.*, 3:610–621, 1973.

[61] J.S. Weszka, C.R. Dyer, and A. Rosenfeld. A comparative study of texture measures for terrain classification. *IEEE T. Syst. Man. Cyb.*, 6:269–285, 1976.

[62] R.M. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67:786–804, 1979.

[63] R.W. Conners and Harlow C.A. A theoretical comparison of texture algorithms. *IEEE T. Pattern. Anal.*, 2:204–222, 1980.

[64] E.M. Færgestad, M.B. Rye, B. Walczak, L. Gidskehaug, J.P. Wold, H. Grove, X. Jia, K. Hollung, U.G. Indahl, F. Westad, F. VenDenBerg, and H. Martens. Pixel based analysis of multiple images for identification of changes; a novel approach applied to unravel proteome patterns in 2D electrophoresis gel images. *Proteomics. Accepted for publication.*

[65] A. Efrat, F. Hoffmann, K. Kriegel, C. Schultz, and C. Wenk. Geometric algorithms for the analysis of 2D-electrophoresis gels. *Journal of Computational Biology*, 9:299–315, 2002.

[66] E. Mosleth and A.K. Uhlen. Identification of quality-related gliadins and prediction of bread-making quality of wheat from the electrophoretic patterns of gliadins and high molecular weight subunits of glutenin. *Norwegian Journal of Agricultural Sciences*, 4:27–45, 1990.

[67] H. Martens. Factor-analysis of chemical mixtures - nonnegative factor solutions for spectra of cereal amino-acids. *Analytica Chimica Acta*, 4:423–442, 1979.

[68] F. Westad and H. Martens. Shift and intensity modeling in spectroscopy - general concept and applications. *Chemometrics and Intelligent Laboratory Systems*, 45:361–370, 1999.

[69] H. Blum, H. Beier, and H.J. Gross. Improved silver staining of plant proteins, RNA and DNA in polyacrylamide gels. *Electrophoresis*, 8:93–99, 1987.

[70] P.H.C. Eilers, I.D. Currie, and M. Durban. Fast an compact smoothing on large multidemnsional grids. *Computational Statistics & Data Analysis*, 50:61–76, 2006.

[71] K. Kaszmarek, B. Walczak, S. DeJong, and B.G.M. Vandeginste. Baseline reduction in two dimensional gel electrophoresis images. *Acta Chromatographica*, 15:82–96, 2005.

[72] P.H.C. Eilers and B.D. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11:89–102, 1996.

[73] H. Martens and M. Martens. Modified jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). *Food Quality and Preference*, 11:5–16, 2000.

[74] D.B. Allison, X. Cui, G.P. Page, and M. Sabripour. Microarray data analysis from disarray to consolidation and consensus. *Nature Reviews Genetics*, 7:55–65, 2006.

[75] L. Gidskehaug, E. Anderssen, and B.K. Alsberg. Cross model validated feature selection based on gene clusters. *Chemometrics and Intelligent Laboratory Systems*, 84:172–176, 2006.

[76] M.B. Rye, E.M. Færgestad, H. Martens, J.P. Wold, and B.K. Alsberg. An improved pixel-based approach for analysing images in two-dimenisonal gel electrophoresis. *Electrophoresis. Recommended for publication.*

[77] T.G. Kleno, L.R. Leonardsen, H.Ø. Kjeldal, S.M Laursen, O.N. Jensen, and D. Baunsgaard. Mechanisms of hydrazine toxicity in rat liver investigated by proteomics and multivariate data analysis. *Proteomics*, 4:868–880, 2004.

Sometimes the road less travelled is less travelled for a reason.

- Jerry Seinfeld