

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/57983> holds various files of this Leiden University dissertation.

Author: Cai, F.

Title: Fuzzy systems and unsupervised computing: exploration of applications in biology

Issue Date: 2017/12/12

Chapter 5

A Systematic Study on One Dimensional Gel Electrophoresis Image Analysis

This chapter is based on the following publication:

Cai, F., S. Liu, P. ten. Dijke, and F. J. Verbeek. "Image Analysis and Pattern Extraction of Proteins Classes from One-Dimensional Gels Electrophoresis." J. Bioscience, Biochemistry and Bioinformatics11: 1106-1113. vol. 7, no. 4, pp. 201-212, 2017.

Chapter summary

In this chapter, we focus on estimating the practical performance of fuzzy systems on the data analysis within the scope of protein/DNA phenotypic study. In detail, we are going to address the following research questions.

1. Can the 1-dimension gel electrophoresis data be quantitatively and accurately assessed using newly developed fuzzy-logical based algorithm and fuzzy systems?
2. Can we identify the essentials of protein/DNA, and validate the results of gel electrophoresis from published reports?

Following the workflow of data analysis (cf. Figure 1-2), this chapter is divided into two major sections. First, the design of the fuzzy systems and its solutions are demonstrated. Each fuzzy-logic and unsupervised computing processing step is illustrated and their motivations behind this design are explained. In the context of the applied fuzzy systems and, heterogeneous methodologies are integrated into a global picture thereof. Second, the data from electrophoresis experiments are qualitatively and quantitatively evaluated. The variations in the bands/lanes are derived from numerical measurements. These results are then compared and discussed with other experiments as described in literature.

5.1. Introduction

Mixtures of proteins can be separated and visualized by Sodium-dodecyl sulphate (SDS)-polyacrylamide gel electrophoresis (PAGE); this is a classical tool for protein analysis [23]. Combining this analysis with Western blotting and probing, the filter with specific antibodies, or the extraction of protein from gel and mass spectrometric (MS) analysis, make it a very powerful tool to determine relative quantities and identification of proteins. In addition, prior to SDS-PAGE, proteins can be fluorescently labeled and the resulting images can be captured by a flatbed scanner equipped for fluorescence. During protein sample preparation, protease inhibitors should be taken into consideration to prevent degradation of proteins; on the gel, these proteases appear as faster running protein fragments.

A popular separation technique, capable of fast and easy analyzing less complex samples, is the high-resolution 1-dimensional (D) gel electrophoresis. Proteins, as obtained from cell lysates, are usually dissolved in a SDS containing buffer and are boiled before loading onto the polyacrylamide gel. Subsequently, on the authority of the molecular weight of proteins, they are charged by force to migrate through the gel under the influence of an electric field. Using this method employing SDS in sample buffer, there is, for most proteins a good correlation between polypeptide length and charge. The latter is running the samples under so-called denaturing conditions. On the contrast, proteins can also be separated under non-denaturing conditions (proteins are then still in their folded state); but then they are not only separated by molecular weight but also by their shape. On a gel, multiple samples are loaded along with molecular standards. The gel, referred to a matrix instance used to contain and separate target molecules, is stained (for instance, by coomassie brilliant blue or silver) and then visualized by a lightbox; alternatively, the fluorescently labeled proteins can also be visualized by a laser scanner. Afterward, the resulting gel images consists of several vertical lanes equal to the number of wells in which the protein samples were loaded; and a number of horizontal bands corresponding to proteins or fragments thereof, reflecting the amounts and characteristics of individual proteinaceous components.

The banding patterns and the relative differential intensities of the bands can be converted into graphical, numerical and tenable formats through image processing and analysis techniques. In this manner computing with intelligent techniques prevents subjective and tedious image interpretation; this otherwise may lead to reproducibility issues. With respect to the analysis of gel electrophoresis profiles, the image processing requires three main steps [27] (BBS): (1) Background correction; (2) Bands detection, matching and quantification; (3) Similarity clustering analysis.

Several software systems have been developed for the automated analysis of profile images acquired from gel electrophoresis techniques [1-4]. Some of these platforms are semi-automatic and locate 1D mean profiles on peak/minima valley as either bands or noise for bands selection, and lack quantification on the digital description of bands. Other systems identify and classify lanes and/or bands via employing simple texture features that result in an unambiguous matching and grouping [5]. Nevertheless, these approaches are unable to generally face the challenges of extracting hidden (complexity) patterns within the bands/lanes expression in gel electrophoresis images via just visual examination, or simple BBS processing steps. However, these approaches henceforth demonstrated challenges to recognize objectives in terms of lane distortion, band deformity (including doublet effect which means two or more bands are too close together) and background noise. Therefore, a systematic yet precise approach in the image processing is required. Hence, based on proposed fuzzy-logic based methodologies (cf. Chapter 2, Chapter 3 and Chapter 4), we introduce an elaborate fuzzy systems so as to improve the BBS step as follow (cf. Figure 5-1):

- Background correction using fuzzy DCBC method [6].
- Normalization and fuzzy feature extraction for lanes/bands.
- Rough fuzzy c-means and particle swarm optimization (RFCM-PSO) hybridized clustering analysis.
- Functional analysis approaches.

In brief, the work with a comprehensive way presented in this chapter contributes to extract qualitative and quantitative information from 1D gels, consisting of background noise subtraction, topographical normalization of bands and lanes, phenotypical description of bands, revealing hidden patterns recognition by dealing with clustering of overlapping and indiscernible information.

The remainder of this chapter is organized as follows: in Section 5.2 we introduce the methodology including image acquisition and processing; i.e., several new innovative algorithms and analysis procedures. After image enhancement, phenotype measurements are obtained on each individual band of all different lanes. Next, the categorization of phenotypic stages using feature extraction and selection is illustrated. The best combination group of features is applied in a clustering technique to address biological questions of interest. The experimental results are presented in Section 5.3 via a case study example, and Section 5.4 concludes this chapter.

5.2. Methodology

Modern gel electrophoresis techniques allow visualizing protein level structures so that these can be specifically subject to analysis. These techniques revolutionized the field of proteomics and biomarker discovery in detecting the changes in protein expression [7]. However, a significant amount of wet laboratory expertise is still required. Application of these techniques in higher volumes is beyond the capacity of manual processing. Therefore, image processing and machine learning are invoked to help recognizing patterns and to provide an automated analysis solution for gel electrophoresis experiments. In this section, we will introduce the image acquisition protocol followed by approaches for image and data analysis.

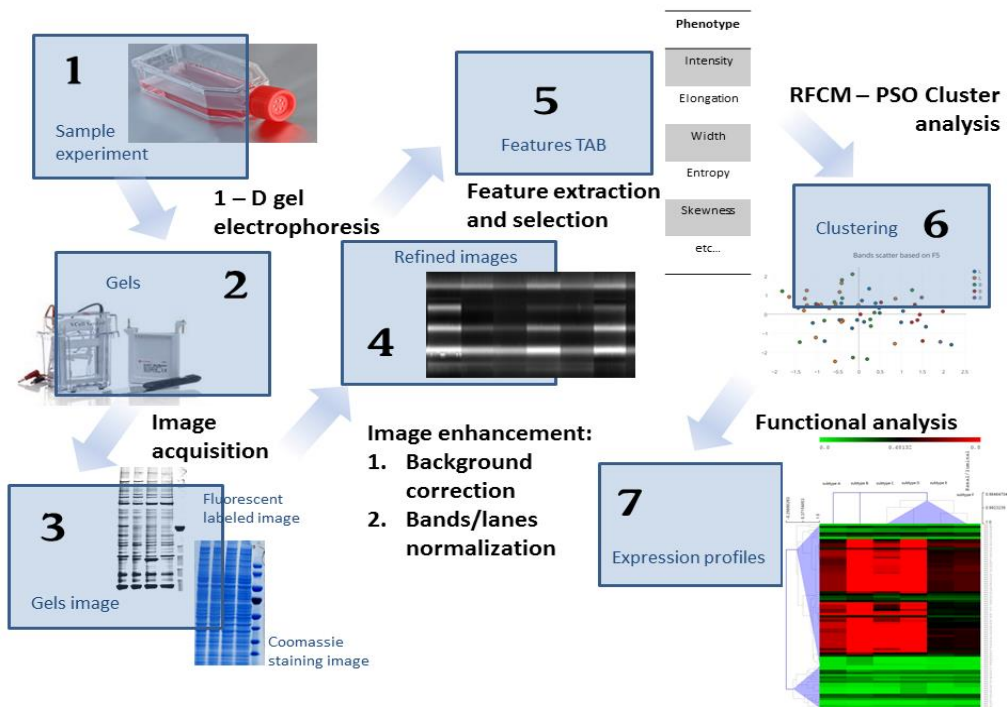


Figure 5-1. Workflow of our 1-D electrophoresis gel image analysis system

A. Image Acquisition

Sample preparation precedes image acquisition (cf. Figure 5-1) and the image of the gel should reflect the differences in composition of the different samples, i.e. each sample represents a particular cell culture. To that end, the samples are labeled with a fluorescently tagged probe. Subsequently, the pre-cast gel (e.g. NuPAGE Bis-Tris) is put into the electrophoresis system (e.g. xCell SureLock Mini-Cell) and in each

well of the gel a sample is loaded. Under an electric current (164-5050 PowerPac) samples migrate over the gel in a linear trajectory and proteins with different molecular weights produce separate bands. Once the current is stopped and the gel is fixed, the result is captured with an imager for the fluorescent signals; i.e. the Typhoon 9410 Gel and Blot Imager as used in this study. Alternatively, after coomassie staining, the gel can be scanned using a lightbox and photo-scanner, i.e. the Microtek ArtixScan F2 scanner. In our experiments described in this chapter, we used images containing fluorescent signals with a spatial resolution of 10 line-pairs/mm, and with a pixel size of 10 μm that produced an image size of 4096x1024 pixels with a 16 bit dynamic range. An example of acquiring images, accompanied with labels and markers is shown in Figure 5-2 and Figure 5-3 (after selection of region of interest) respectively.

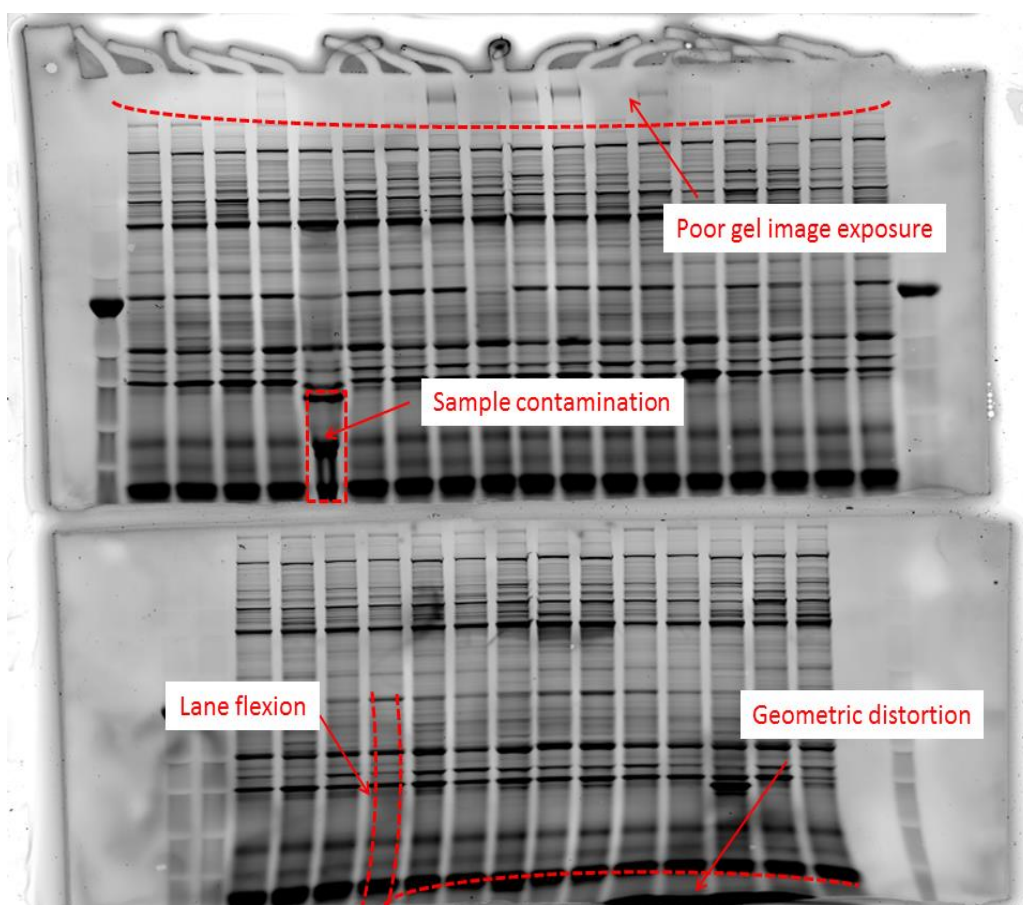


Figure 5-2. A sample of acquiring raw gel electrophoresis images that reveals common challenges for imaging processing, including geometric distortion, cell line flexion, low contrast and noisy illumination.

[illegible]

85

B. Image Enhancement

A problem that arises in gel electrophoresis imaging is the introduction of information that is not part of the original signal. This part of the information should be considered as noise and outlier. Images acquired from the optical detection system may inevitably suffer from the various sources of systematic and experimental variation, through which the “true” information is masked. Hence, a cropping of regions of interest, background correction and data normalization are required.

Region cropping. Before loading images into preprocessing track, bands of interest are firstly resolved in a region that this part of image can be isolated for further analysis by manually cropping and adhering (cf. Figure 5-3 from Figure 5-2, this processing is typically operated by biologists). Cropping process should also be performed to remove regions of gel that show sample contamination and extreme distortion of cell line which could interfere with bands detection.

Background correction. The adjustment or removal of the background signal should be performed to accurately quantify the fragments present in the gel image; i.e. the true signal. Approaches that have been suggested for such background correction, include global minimum subtraction from time domain, signal-pass filtering in wavelet domain or frequency domain and processing using mathematical morphology. A more comprehensive discussion on an amount of approaches can be found in [8] [9].

In our system, an estimate from the background illumination is produced by a morphological subtraction of the gel image. To that end, the fuzzy and rough concepts are employed [6] to an improved dam-based rolling ball method. In this manner the mutual information shared by foreground and background is balanced for an ideal image $f(x, y)$, the background correction procedure can be described as Equation 5-2 and can also be recalled from Chapter 2.

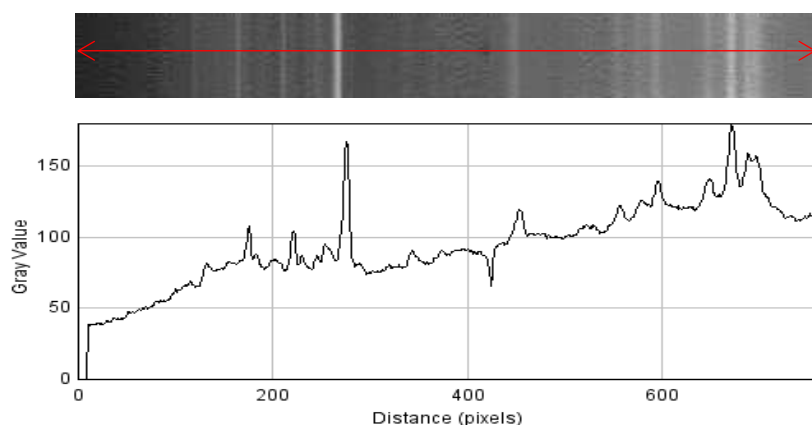
Raw image estimation:

$$\tilde{f}(x, y) = f(x, y) + \text{Background}(x, y) \quad \text{Equation 5-1}$$

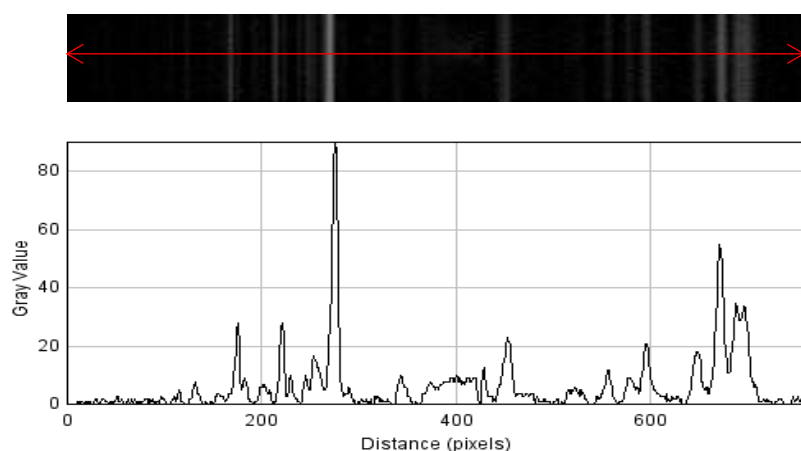
Background subtraction:

$$f(x, y) \approx \tilde{f}(x, y) - \text{Background}(x, y) + I(\tilde{f}(x, y), \text{Background}(x, y)) \quad \text{Equation 5-2}$$

where I represents the mutual information as defined in [10]. Figure 5-4 shows the comparison of selected raw gel image and enhanced gel image.



(a) Raw gel image

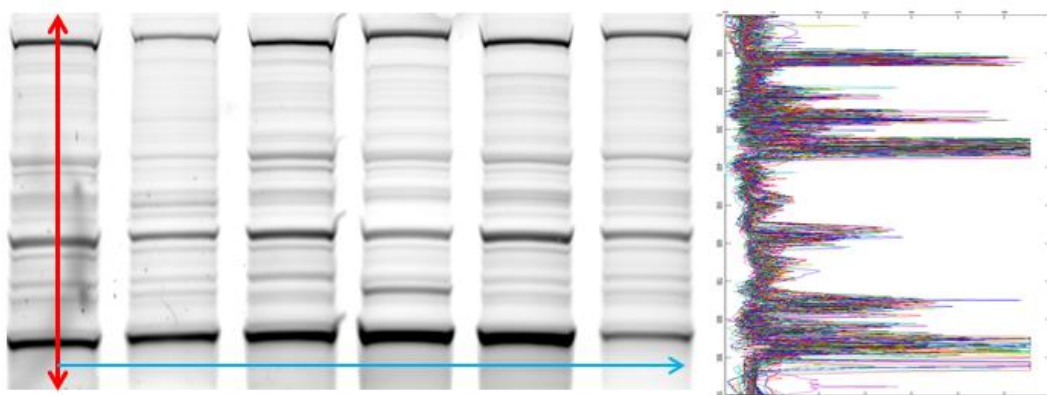


(b) Enhanced image after background correction

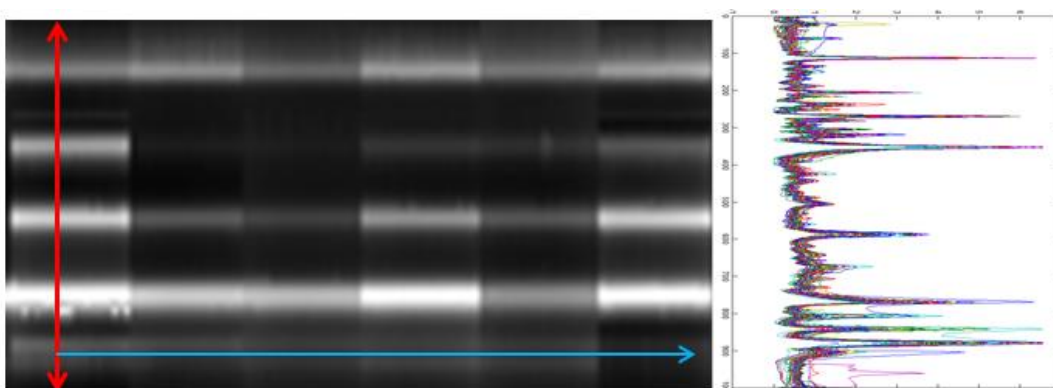
Figure 5-4. 1D profiles (the intensity in y-axis w.r.t the location in x-axis) scanned along with red line.

Data normalization. To make quantitative comparisons between profiles of lanes, and/or position of bands, it is necessary to normalize the distortion. There are two major steps in correcting the distortion of gel electrophoresis images. One is the straightening of vertical flexion of cell line, namely intra-lane alignment of bands. This procedure helps to recognize and relocate cell line. Various methodologies, i.e. interpolation based [28] and grid based [29], are reported to efficiently deform image shape. In this case study, the flexion in cell lines takes place out of our interested region, and thus can be neglected or slightly aligned from horizontal normalization.

Another normalization procedure in our study is the horizontal bands assignment, where several approaches have been described [20] [21]. We employ the Sparse Dynamic Time Warping (SDTW) method [11] to yield optimal conjunct alignments, as it is very efficient and can maintain the ability of searching for a more optimal solution. By this mean, all the separate bands are relocated into a parallel line where the corresponding bands have the same positions as in the different cell lines.



(a)



(b)

Figure 5-5. Bands normalization procedure. The second column is the profile plots on red arrow that along blue direction, and profiles are normalized to have a similar distribution after alignment (drifts are aligned). (a) Original profiles in a gel, e.g. fluorescently labeled gel. (b) Mean of the warped image after band alignment (background between cell line-band is eliminated).

Segmentation of the bands. The aligned profiles of each of the lanes are put up for analysis by separating all of its bands. It performs as an extraction of region of interest mask that image information is ready to be converted to numerical matrix. For each band, a (binary) mask is computed by first finding a (local) peak value from the centerline of the lane, and then established the neighboring (local) valleys for every peak (cf. Figure 5-6). The mask is subsequently obtained from these threshold values for each of the bands in each of the lanes.

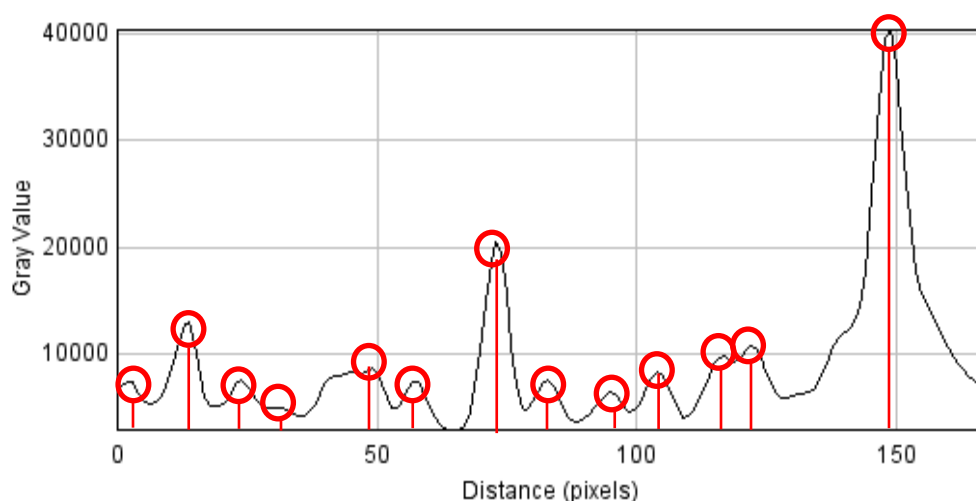


Figure 5-6. Band detection within a cell line. After background subtraction (difference cf. Figure 5-4) and data normalization, the red stems denote peak value on a lane profile, which indicate a possibly location of bands.

C. Feature Extraction and Selection

Feature extraction aims to reduce the data dimensionality and complexity, and therefore its application provides an efficient way to allow more feasible statistical analysis. Multifactorial classes of algorithms can be applied on 2-D gel images [12], for instance, boundary-based techniques, region-based methods, and hybrid methods that combine boundary and region criteria. However, none of the general or optimal procedures for extraction and quantification on 1-D gel images is reported. Targeting on segmented bands of each cell lines, it is crucial to find certain features to represent the characteristics of these bands. This is the key of data analysis of 1D gel electrophoresis images.

Phenotypic features are considered as the composites of observable characteristics or traits for an organism [13], and therefore these are employed in our work. In the

attempt to find prominent phenotypic features to characterize the proteins or fragments, two aspects should be considered: (1) features should be representative and relevant; and (2) features should be robust with respect to the small variations in bands intensities.

Table 5-1. Basic measurements for a phenotype

Feature name	Description
size	The surface area of object
Intensity	Amount of intensity belong to object
Perimeter	The perimeter of object
Circularity	Area-to-perimeter ratio
Extension	Derived from 2 nd –order invariants of object [14][15]
Dispersion	Derived from 2 nd –order invariants of object [14][15]
Elongation	Derived from 2 nd –order invariants of object [14][15]
Orientation	Derived from 2 nd –order invariants of object [14][15]

Table 5-2. Texture measurement of a phenotype (x represents the intensity value of one pixel, while $H(x)$ is the histogram of the intensities)

Feature name	Expression	Description
Avg	$f_1 = \mu$	Average intensity in a region of object.
Std	$f_2 = \sqrt{\sum_x (x - f_1)^2 H(x)}$	Standard deviation of intensity in a region of object.
Smoothness	$f_3 = 1 - \frac{1}{(1 + f_1^2)}$	Relative smoothness of intensity in a region.
Skewness	$f_4 = \sum_x (x - f_1)^3 H(x)$	Deviation from symmetry of mean intensity
Uniformity	$f_5 = \sum_x H^2(x)$	Sum of squared elements in histogram
Entropy	$f_6 = -\log_2 H(x) \sum_x H(x)$	Statistical measure of uncertainty

Direct and indirect quantifications include determination of the selected phenotypic measurements (cf. Table 5-1 and Table 5-2), in which each result is calculated from the pixels that define the shape of lanes/bands. This procedure quantifies the information pattern of 1-D gel electrophoresis images into distinct measurements, which requires further selection of features. The manner in which prominent features

are chosen to represent the dynamics of fragments-migrating process becomes an important step for identification of the phenotype.

In order to guarantee that all selected features are independent and equal of variance, the Mahalanobis distance [16] is chosen as the probabilistic distance criterion. Subsequently, we employ the methodology, the Fuzzy Criteria in Multi Objective Feature Selection Algorithm [17] (cf. chapter 3), which selects a subset of features from feature-pool that can best predict and describe the data. The resulting solutions from this method lead to a set of candidate feature combinations. This will facilitate the bio-scientist in selecting the proper features to predict a biological phenomenon and provide a guideline for new experimental design. In this case study, the selected features based on the best performance in the approach from multi-objective optimization, are band width, band intensity standard deviation, and lane skewness. The original 1-D gel electrophoresis images usually have different widths of bands (cf. Figure 5-1) at different positions reflecting the molecular weights of fragments. The “band intensity standard deviation” is a global index for a detected band; while the “lane skewness” is a vertical descriptor to understand the deviation from symmetry (cf. Figure 5-5 (a) and (b)) as fragments migrate downstream.

D. Information Clustering Analysis

The measurement information is summarized into a matrix of statistics that represents the patterns information. To date, various pattern finding procedures have been settled. However, for research implementation the information clustering is particularly important. In order to address biological questions accordingly, two issues come up: how to partition sets of samples that contain various features into groups among a large number of bands or lanes from electrophoresis gels; and how to figure out different patterns amidst samples with indistinguishable information and features.

To tackle these issues, we present [18] (cf. Chapter 4) an innovative and efficient approach that is capable of clustering information from overlapping and otherwise indiscernible partitions. This method, a.k.a. Rough Fuzzy C-means and Particle Swarm Optimization hybridization (RFCM-PSO), combines the RFCM clustering algorithm [19] with an optimization technique. In RFCM, the rough approximation sets are employed to constrain the *fuzzifier* membership index. Subsequently, the iterative procedure of partitioning is then minimizing the RFCM objective function. Whereas the optimization of the clustering results occurs, the PSO procedure searches for the global optimum by updating the candidate centroid positions of partitioning solutions. The pseudo-code of RFCM-PSO is shown:

Algorithm *Rough fuzzy c-means and PSO hybrid method:*

Input: fuzzifier, weighting factor, cluster number, and controlling parameters in RFCM**Given:** integral population and generation in PSO**Initializing:** stochastic centroids, membership matrix, position and velocity at first generation**for** each generation **do****1. training RFCM parameter:**

Compute similarity distance for each object to its belonging cluster centroids.

if Rough approximation condition **then**

Reset membership matrix from FCM.

end if

Update centroids.

Update membership matrix.

2. Optimization procedure:

Computing the personal best and global best positions.

Update position and velocity for each particle.

Convergence check: break main loop**end for**

By taking the advantage of both the RFCM clustering method and the intrinsic characteristics of PSO, this combined-model can now deal with overlapping partitions, uncertainty and vagueness of information. At the same time, the optimization procedure in PSO demonstrates the ability of searching optimal solutions. In this case study, the information clustering takes place at two aspects: 1) clustering of bands and 2) clustering of cell-lines. In the first aspect, a band of all cell lines is selected. For instance, band number 65 in all 60 cell lines is chosen, then 1x60 samples are obtained and will contain several selected features (e.g. band width, skewness, local entropy). Thereafter via a clustering of bands, we can notify the intrinsic property of a certain band that affect expressions (captured by 1D gel electrophoresis, and represented as different intensity on image data) in a variety of cell lines. Another implementation of the clustering technique is on the cell line. With little labelled cell type (cf. Figure 5-3, some types of cell lines are considered as null or unverified), we utilize results from the decision tree (cf. Figure 5-7), and cluster the cell line based upon their selected feature performance. The result of clustering example is shown in Figure 5-8.

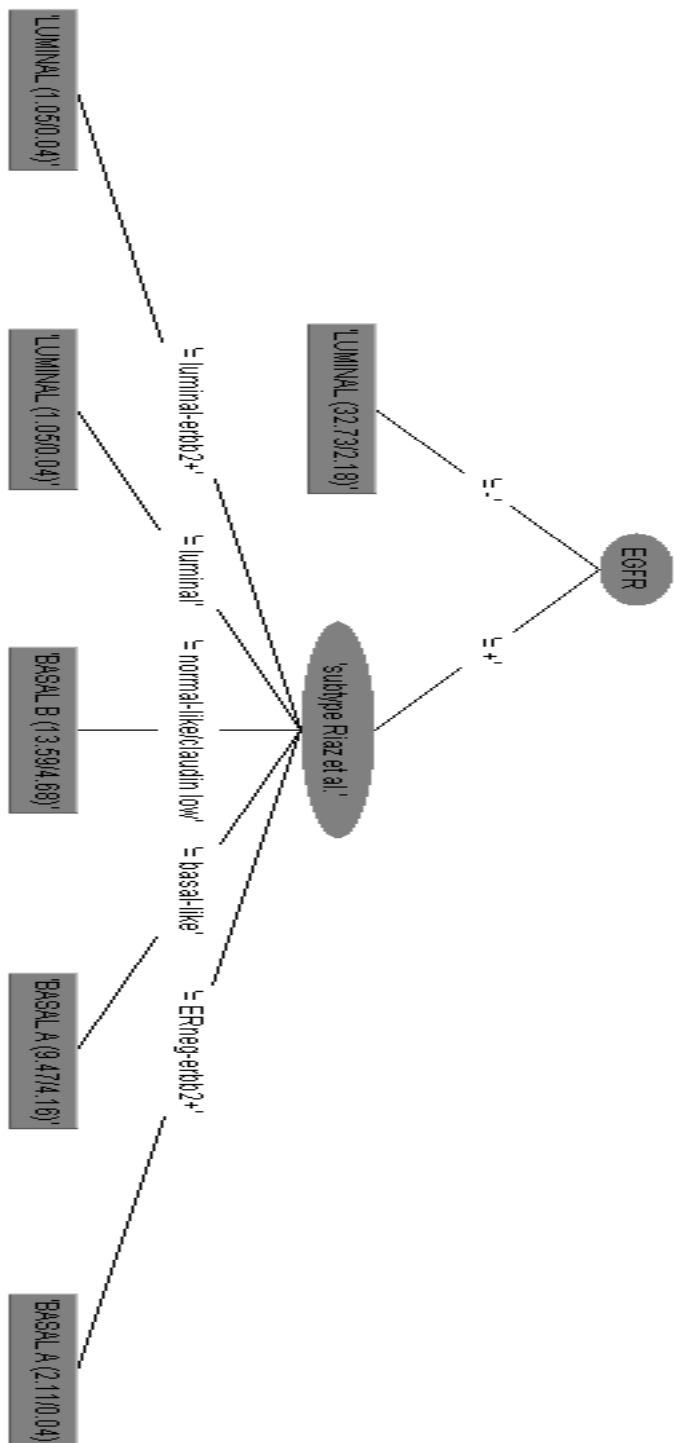


Figure 5-7. Decision tree view of chances that consequent clustering outcome may be significantly affected by cell line types w.r.t their subtype and reactor.

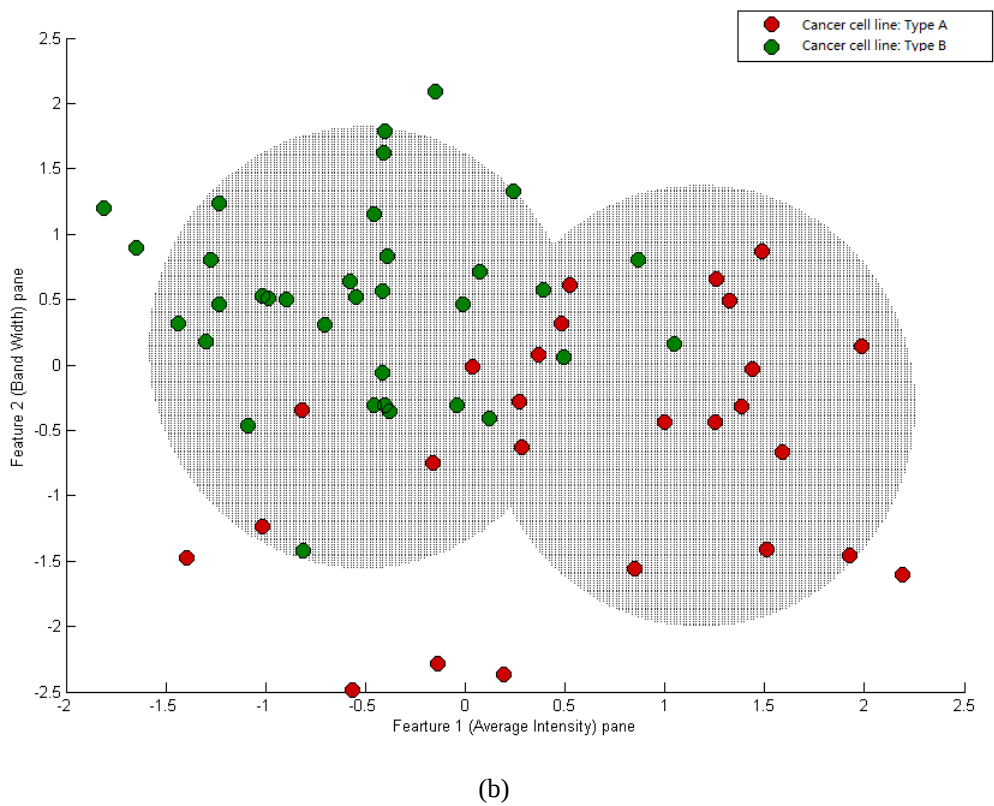
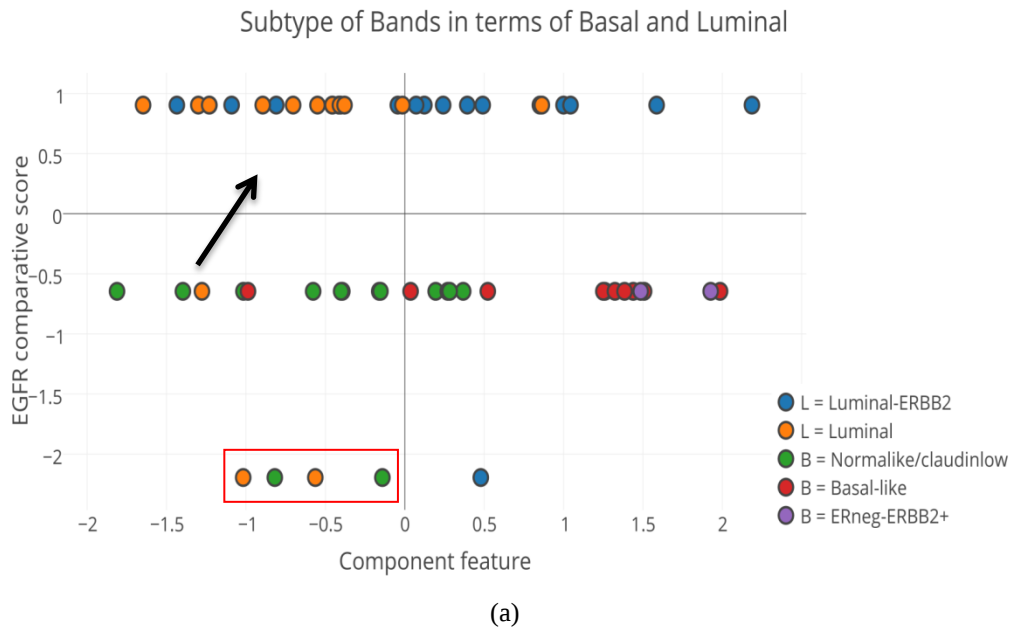


Figure 5-8. Clustering result of genetically distinct cancer cell lines. (a) All 60 cell lines should be clustered into two types, either one band should belong to “Basal” or “Luminal”. In Figure 3, cluster L contains all the subtypes that belong to “Luminal”, and, the rest of subtypes are clustered to B cloud, namely the “Basal” cloud. However, the Line T47D is occurred in the B cloud when it has quite similar feature properties as its clustered neighbors. Moreover, lines laid in red region are indiscernible, though they can also have a similar feature property. When checking the line information table (cf. figure 5-2), the abovementioned lines are number 57 to 60 respectively, and with a void score for their corresponding EGFR value. Here, 27 of basal bands are compared with 33 different luminal bands based on the difference in their intrinsic feature. (b) We combine the selected features together, and perform a clustering analysis against the information table in terms of cell line type Basal and Luminal (type A in red dot means basal, type B in green dot means Luminal). Every green and red dot represents an integration of features from detected bands which belong to a list of cell lines (LCLs). Features are normalized, and cell lines are categorized unsupervised. For this particular clustering result, the F-score is 0.92 using the feature pool as reported in section 5-2.D.

5.3. Measurements and Results

In this section, we apply our approaches to illustrate the fuzzy system research strategy on a dataset of cancer cell-lines; the cell-lines have different genotypes and from gene expression profiling, it becomes clear that they should be distinguished in different subgroups.

To estimate and investigate how quantitatively and accurately the proposed fuzzy system can be employed on interpreting biological questions, dedicated solutions are conducted and analyzed on this dataset. The experiment consists of 60 cell-lines that are loaded on the PAGE gels in 60 separate lanes. After running the gel, they will result in a sequence of images that contains 60 lanes with 85 detected bands. For each band the features are measured (cf. Section 5-2.C). The preprocessing and normalization of the gel lanes make it possible to directly compare the bands for position and intensity.

We aim at finding a pattern in a series of gel images to characterize groups of cell-lines. To this end, we formulate a null hypothesis (H_0) and test if the statistical inference of the underlying distribution can be considered significant. In other words, attributes (features) of bands, within the scope of different cancer cell-lines, will be carefully measured by applying T-test schema. It is different from what typically used in student T-test, where samples contain only one or two attributes. In this case study, we propose a more comprehensive Hotelling T-square test [22] employing multi-variates (cf. features from Section 5-2.C) to obtain an accurate measurement

of the significance. Ultimately, the resulting p-value is used to weight the strength of the evidence, demonstrating the significance of protein/DNA expressions in different cell type groups. We have found several bands that perform significantly in either cell-type, subtype, or the reaction of regulator (cf. Table 5-3 and Table 5-4). Each band of all cell lines are compared separately, and their statistical performance are shown in ranges of either $p \leq 0.05$ or $p \leq 0.01$ (where bands have p value ≤ 0.01 are involved in those bands with p value ≤ 0.05). The latter value of p indicates a more significance in expression (effect) of the group (population). Table 5-3 gives an example of one series gel images, where the significances of detected bands are explored.

In Table 5-4, accordingly, a significance test of Basal and Luminal cell line is conducted using top 5 best descriptors, e.g. intensity standard deviation, intensity, entropy, skewness, and width. For either Luminal or Basal cell lines, the detailed and numerical attributes of tested bands are averaged (Total) and compared with Luminal group average (Luminal) and Basal group average (Basal). When the total expression of a band (all features are counted) is close to the expression of Luminal group (same band), we consider this band is Luminal-significant; and vice versa for the band of Basal-significant. Alternatively, the value of the features in Table 5-4 with N/A means a specific band has no expression in such groups. For example, the number 77 band (cf. Table 5-4), as it is recognized as UCHL1 band [26] that has higher significance in Basal group but lower significance in Luminal group. A bar plot of UCHL1 expression in two separate experiment is shown in Figure 5-9. Gels containing both 20 cell lines and 60 cell lines (partially overlap for validation) are conducted on the experiments, where both T-test results depict a high significance, $p \leq 0.05$. The visualization of employed feature, particularly in intensity, denotes that band UCHL1 in some cell lines has higher significant expression in basal group, and its average intensity in luminal group is lower compared with the basal one. The combination of the selected features provides a guide-line for clustering and statistical analysis of bands with respect to corresponding cell lines. These results have the same comparisons and validations in either mass spectrometry and/or reports [23] [26].

Table 5-3. Significant bands description in different type of cell line (85 detected bands in total)

	ER	PgR	ERBB2	EGFR	CK5	Basal/Luminal
Bands description (number)	14	46	Null	Null	29	49
$P \leq 0.01$	69	54				
Bands description (number)	24	24	7	17	8	71
$P \leq 0.05$	46	25	23	46	57	77
	66	26	27		64	78
	71	36	47		74	80
		42	81		75	
		51				
		56				
		82				
		80				

Table 5-4. Significance performance of bands with $p \leq 0.05$ (basal/ luminal cell type)

Band Nr.	Type(ave)	Int Std	Intensity	entropy	skewness	width
# 49	Basal	10.461	315.509	-295.555	2.729	1.259
	Total	10.803	180.136	-159.980	1.564	0.750
	Luminal	N/A	N/A	N/A	N/A	N/A
# 71	Basal	N/A	N/A	N/A	N/A	N/A
	Total	22.274	238.182	-599.722	2.303	2.408
	Luminal	40.498	433.058	-109.040	4.187	4.378
# 77	Basal	198.614	3776.476	-538.034	30.059	17.518
	Total	178.773	4776.549	-846.869	32.663	18.95
	Luminal	161.353	5467.407	-109.478	34.079	19.909
# 78	Basal	115.256	3255.756	-291.145	9.448	9.740
	Total	117.027	4232.853	-403.740	8.965	11.008
	Luminal	117.144	4897.807	-489.061	8.560	11.803
# 80	Basal	27.359	2358.341	-175.230	13.616	5.314
	Total	13.108	1736.411	-86.669	6.697	2.800
	Luminal	1.449	1227.559	-14.210	1.037	0.742

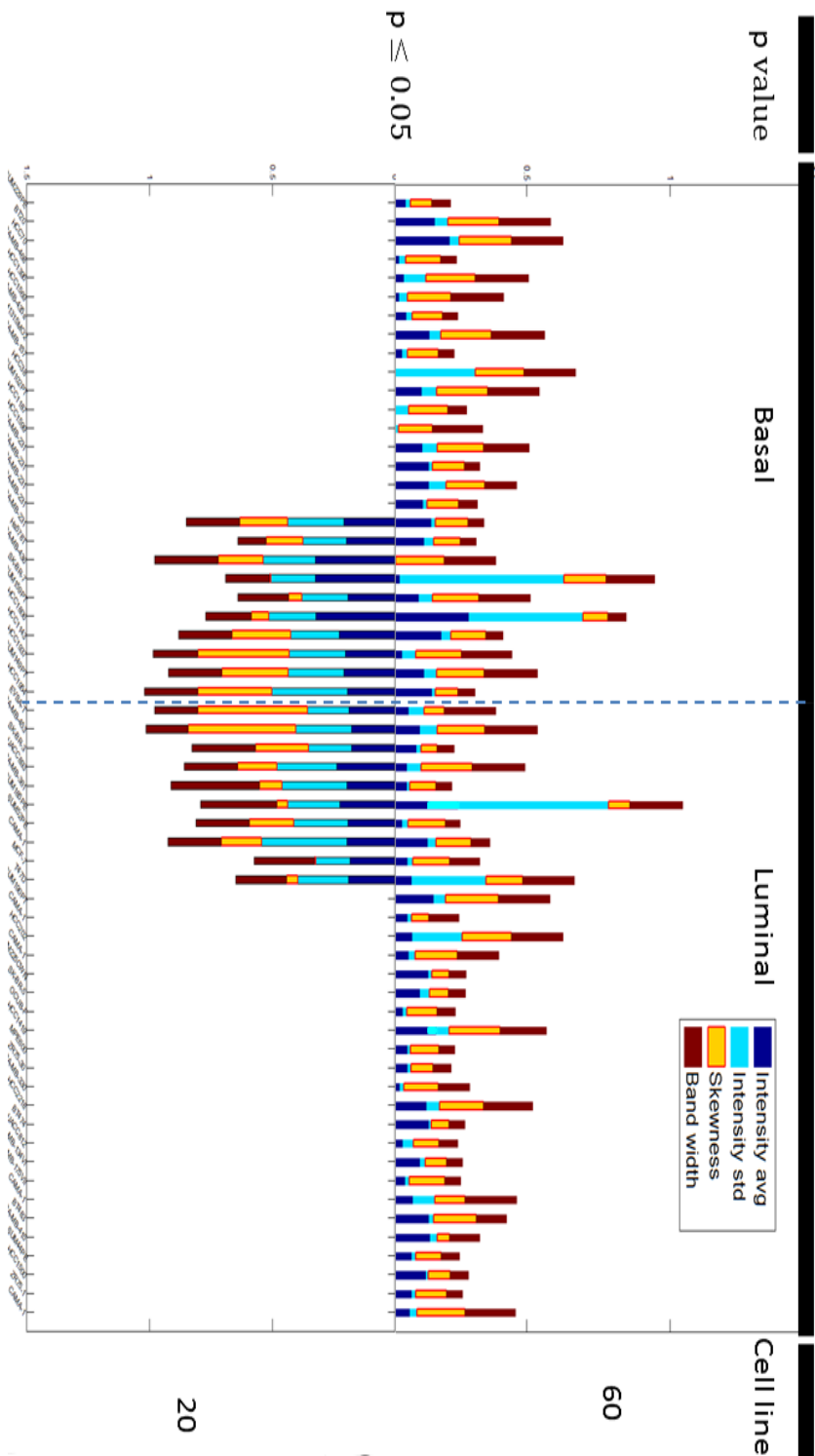
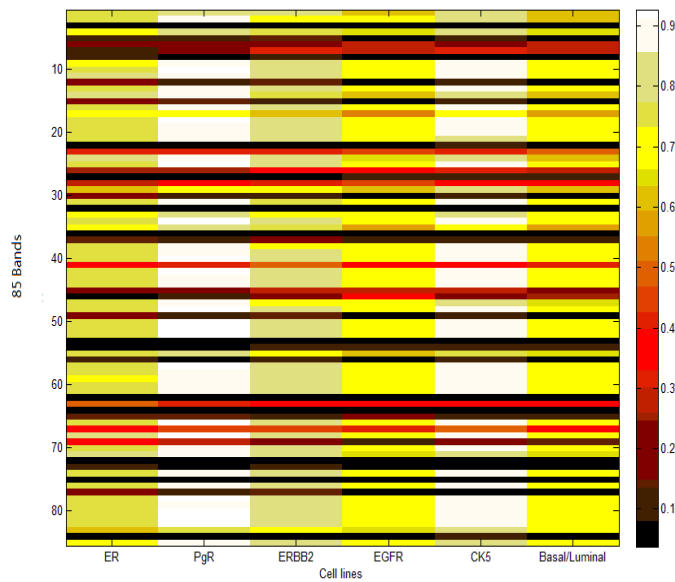
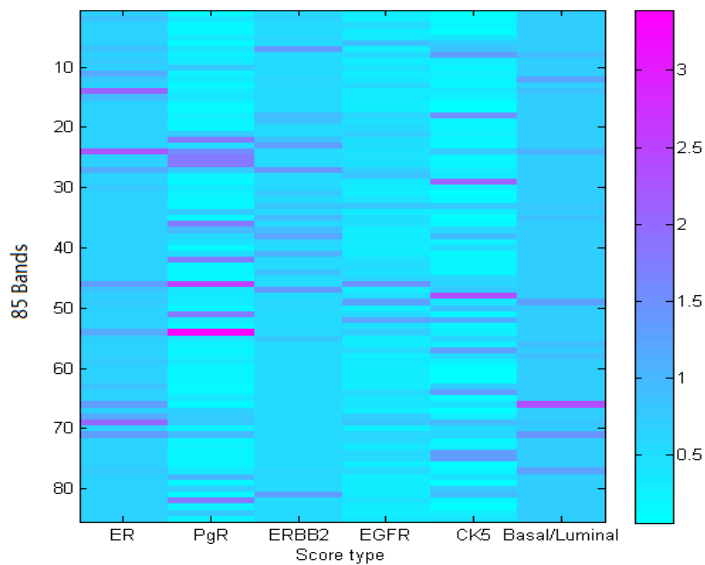


Figure 5-9. Features visualization of band UCHL1 in two gels (20 cell lines and 60 cell lines).



(a)

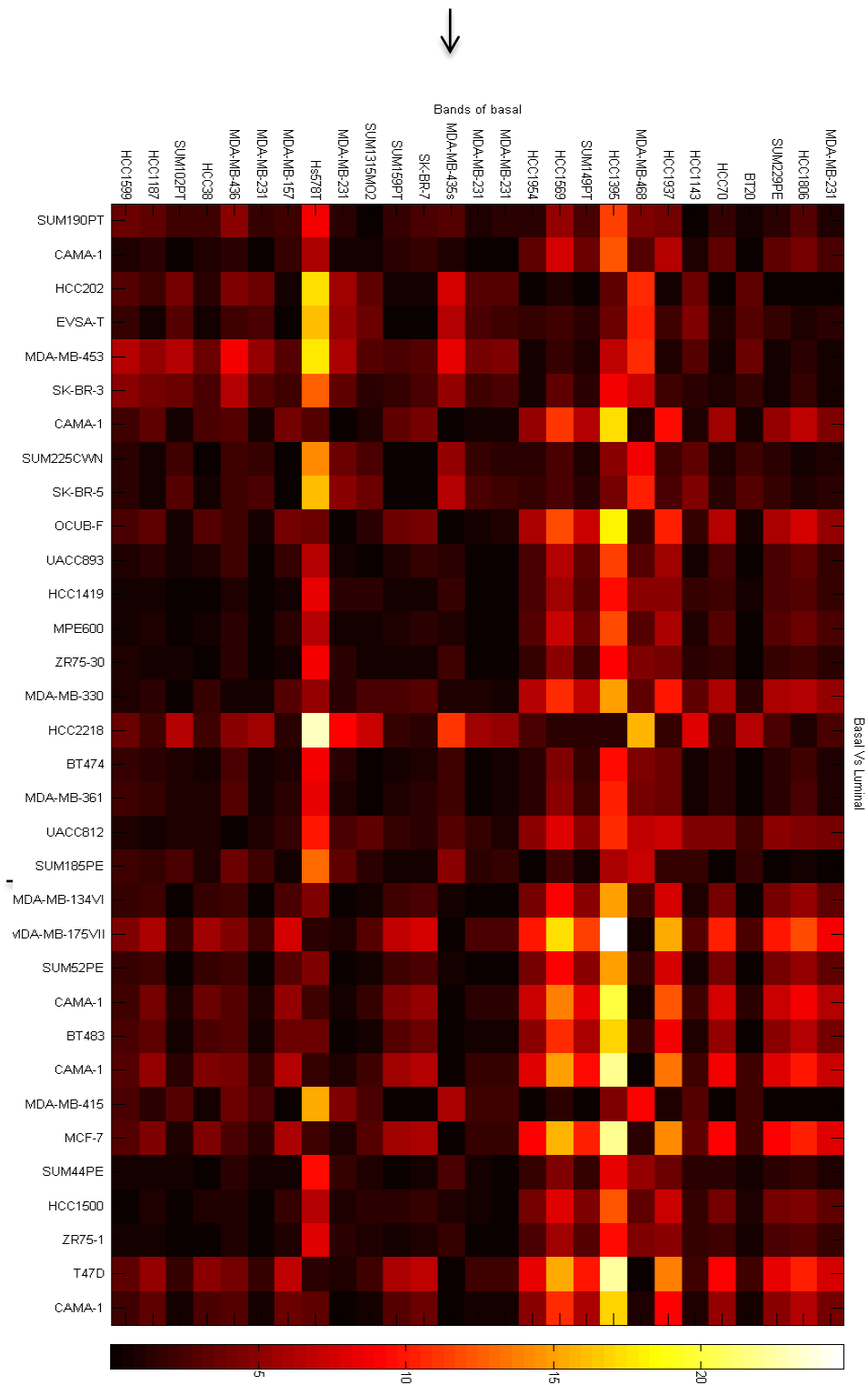


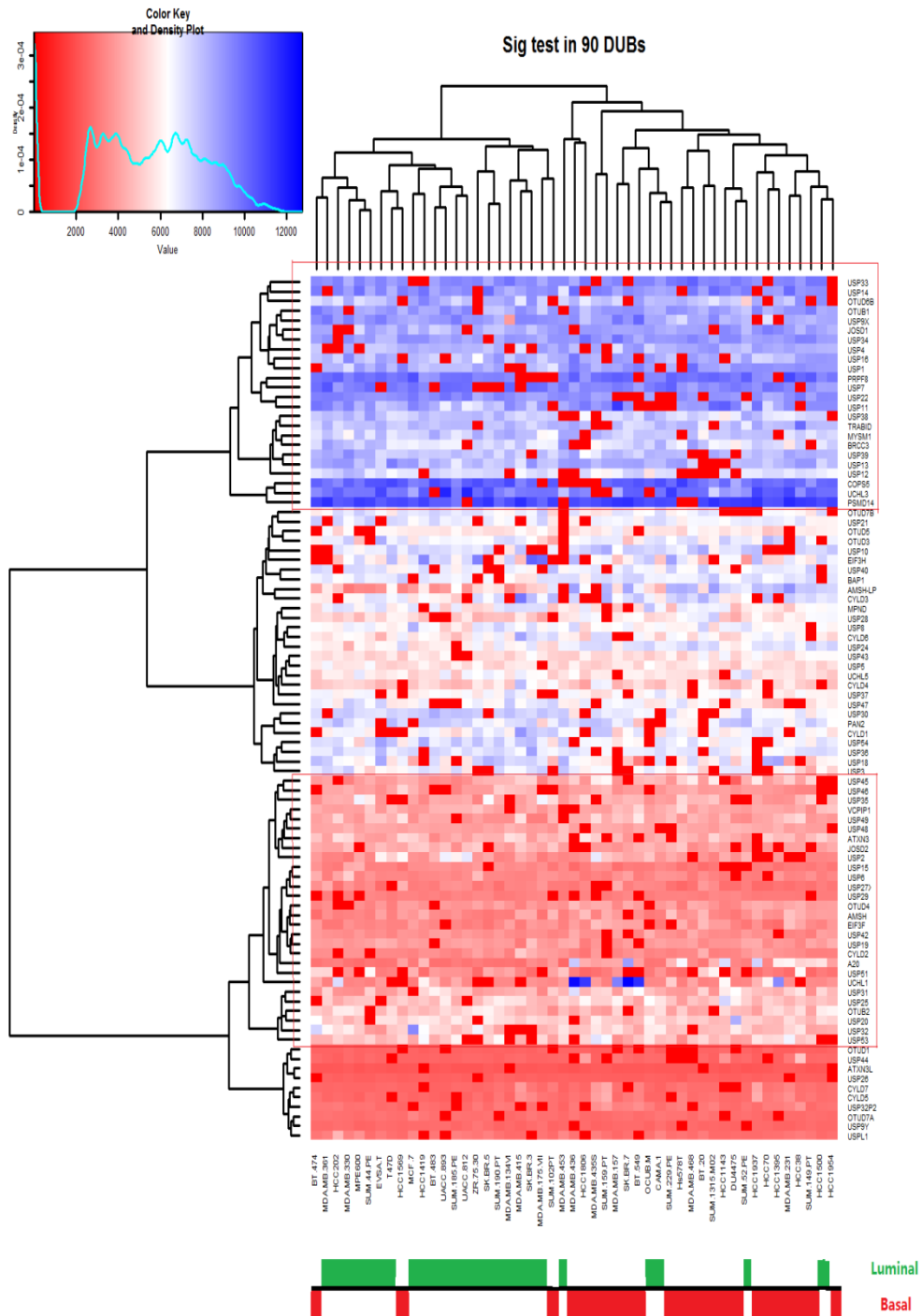
(b)

Figure 5-10. Heat map of statistical tests on six cell-line type respectively. The y-axis represents the number of detected bands and six different types in x-axis are investigated for: (a) F-accuracy test in terms of classification using the feature information of bands; (b) Resulting p-value (log-transformed) from Hotelling T-square significant test. Brighter values indicate a higher accuracy (significance) in the classes.

Figure 5-10 (a) and (b) show the results of F-test and T-test of all 60 lanes with their detected 85 bands in one visualization. The similarity of the band with respect to its corresponding cell lines can be categorized under an unsupervised condition utilizing the proposed fuzzy system pipeline and the accuracies of bands clustering are verified by F-test. Additionally, for a particular cell line subtype, the entire sets of cell-lines are recognized as two groups, e.g. ER positive and ER negative, PgR positive and PgR negative, ERBB2 positive and ERBB2 negative, CK5 positive and CK5 negative, as well as cancer type basal and type luminal.

The result of the F-test, as depicted in Figure 5-10 (a), reveals that the ability for subtype-clustering of each separate band varies with subtypes in the population of the cell-lines (a priori knowledge) as well as with the gel analysis results. The intrinsic properties of the bands are directly related to the cell subtype that they originate from. Some bands exhibit extremely low F-values in all the six subtypes which originate from proteins indicating an absent or are not activated in the cell under the conditions of the experiment. In Figure 5-10 (b), it is shown that bands can be reported to have significant differences ($p\text{-value} < 0.01$) from a classification into the six subtypes. This indicates that the detected bands of proteins or DNA, can represent a uniqueness expression profiles.





(b)

Figure 5-11. Heat map of the 1D gel electrophoresis phenotypic analysis, where protein/DNA occurrence as analyzed from bands on the gel. (a) Heat map of normalized features of intensity and band width for group basal vs. group luminal. Brighter colors represent a higher response in terms of their phenotypic expression. The arrows indicate a specific type of cancer cell-line in group A/B with strong responses; (b) Hierarchical clustering analysis of essentials of detected bands (vertical) with respect to cancer types (60 cell-lines involved in either basal or luminal group) (horizontal). The colors indicate a degree of correlation between bands and subtypes.

Experiments with gel electrophoresis support the understanding of the relationship between sample groups. A clustering analysis, hereby, targets to find the hidden patterns in data. Hierarchical clustering revealed distinct positive (blue) and negative (red) expression of proteins (cf. Figure 5-11 (b)) in terms of the normalized features quantified from gel electrophoresis images. It helps in validating predicted preference of cell-line type. In Figure 5-11 (a), cell-line groups of different subtypes are compared to examine the differences in the patterns. Two maximum variations of bands in conflict the subtype are pointed with arrows.

5.4. Conclusions

This chapter investigates and illustrates the ability of proposed fuzzy-logic based methodologies and their integrated fuzzy system pipeline for data analysis from 1-D PAGE gel electrophoresis. The adequate processing algorithms and heterogeneous information are thereof composed into a global picture. It is demonstrated via a practical implementation on a series of bio-imaging experiments that this system is reliable and is capable of qualitatively and quantitatively assessing information. Quintessential are the fuzziness background correction, feature extraction and selection of region of mask based upon fuzzy criteria. These elaborated approaches contribute to phenotypic quantification and henceforth unsupervised classification. The pattern extraction and recognition aim to support phenotype analysis. In this chapter, the experiment shows that employing the proposed fuzzy system it can be accomplished by investigating and understanding the identity of proteins characteristics which are distinct/shared between different subgroups of cancer cell-lines. In addition to this case study, in protein characterization, DNA and RNA fragments can also be separated by 1D electrophoresis. In the same manner, the proposed method can also be applied for a systematic analysis of DNA and RNA patterns.

5.5. References

- [1] Machad, Alexei, et al. "An iterative algorithm for segmenting lanes in gel electrophoresis images." *Computer Graphics and Image Processing*, 1997. Proceedings, X Brazilian Symposium on. IEEE, 1997.
- [2] Lin, Yun-Liang Yang, et al. "Automatic method to compare the lanes in gel electrophoresis images." *Information Technology in Biomedicine*, IEEE Transactions on 11.2 (2007): 179-189.
- [3] Adiga, PS Umesh, et al. "Automatic analysis of agarose gel images." *Bioinformatics* 17.11 (2001): 1084-1089.
- [4] Kaabouch, Barry Milavetz, et al. "A Novel Automated Analysis System for DNA Gel Electrophoresis Images." In *Artificial Intelligence and Pattern Recognition*, pp. 36-41. 2007.
- [5] Ye, Xiangyun, et al. "A recent development in image analysis of electrophoresis gels." *Vision Interface'99*, Trois-Rivières 19.21 (1999): 432-438.
- [6] Fuyu Cai, and Fons J. Verbeek. "Dam-based rolling ball with fuzzy-rough constraints, a new background subtraction algorithm for image analysis in microscopy." *Image Processing Theory, Tools and Applications (IPTA)*, 2015 International Conference on. IEEE, 2015.
- [7] Kimberly F. Sellers and Jeffrey C. Miecznikowski. *Statistical Analysis of Gel Electrophoresis Data*, *Gel Electrophoresis - Principles and Basics*, Dr. Sameh Magdeldin (Ed.), InTech, (2012): DOI: 10.5772/36959.
- [8] Kaczmarek, Krzysztof, et al. "Preprocessing of two - dimensional gel electrophoresis images." *Proteomics* 4.8 (2004): 2377-2389.
- [9] Vovk, Botjan Likar, et al. "A review of methods for correction of intensity inhomogeneity in MRI." *IEEE transactions on medical imaging* 26.3 (2007): 405-421.
- [10] Papoulis, S. Unnikrishna Pillai, et al. *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education, 2002.
- [11] Al-Naymat, Javid Taheri, et al. "Sparse dtw: A novel approach to speed up dynamic time warping." *Proceedings of the Eighth Australasian Data Mining Conference-Volume 101*. Australian Computer Society, Inc., 2009.
- [12] Nedenskov Jensen, Bo M. Jørgensen, et al. "Multivariate Data Analysis of Two-Dimensional Gel Electrophoresis Protein Patterns from Few Samples†." *Journal of Proteome Research* 7.3 (2008): 1288-1296.
- [13] Cowin, John Wysolmerski, et al. "Molecular mechanisms guiding embryonic mammary gland development." *Cold Spring Harbor perspectives in biology* 2.6 (2010): a003251.
- [14] Hu, Ming-Kuei. "Visual pattern recognition by moment invariants." *information Theory, IRE Transactions on* 8.2 (1962): 179-187.
- [15] K. Yan, Fons J. Verbeek, et al. (2009), *Cell Tracking and Data Analysis of in vitro Tumor Cells from Time--Lapse Image Sequences*. In: *Proceedings VISAPP 2009*. 281-287.

- [16] Cai, Fuyu, et al. "Fuzzy Criteria in Multi-objective Feature Selection for Unsupervised Learning." *Procedia Computer Science* 102 (2016): 51-58.
- [17] Jain, Jianchang Mao, et al. "Statistical pattern recognition: A review." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22.1 (2000): 4-37.
- [18] F. Cai, F.J. Verbeek. "Rough Fuzzy C-means and Particle Swarm Optimization Hybridized Method for information Clustering Problem". *Journal of Communications*. ISSN: 1796-2021 (in press)
- [19] Ji, Zexuan, et al. "Generalized rough fuzzy c-means algorithm for brain MR image segmentation." *Computer methods and programs in biomedicine* 108.2 (2012): 644-655.
- [20] Lemire, Daniel. "Faster retrieval with a two-pass dynamic-time-warping lower bound." *Pattern recognition* 42.9 (2009): 2169-2180.
- [21] Salvador, Philip Chan, et al. "Toward accurate dynamic time warping in linear time and space." *Intelligent Data Analysis* 11.5 (2007): 561-580.
- [22] Marozzi, M. Multivariate multi-distance tests for high-dimensional low sample size case-control studies. *Statist. Med.*, (2015), 34: 1511–1526. doi: 10.1002/sim.6418.
- [23] Simpson, Erica Golemis, et al. *Basic methods in protein purification and analysis*. Cold Spring Harbor Laboratory Press, 2009.
- [25] Kaur, Amanpreet, and M. D. Singh. "An overview of pso-based approaches in image segmentation." *Int J Eng Technol* 2.8 (2012): 1349-1357.
- [26] Marcotte, Richard, et al. "Functional genomic landscape of human breast cancer drivers, vulnerabilities, and resistance." *Cell* 164.1 (2016): 293-309.
- [27] Intarapanich, Apichart et al. "Automatic DNA Diagnosis for 1D Gel Electrophoresis Images Using Bio-Image Processing Technique." *BMC Genomics* 16.Suppl 12 (2015): S15. PMC. Web. 28 July 2017.
- [28] Schaefer, Scott, Travis McPhail, and Joe Warren. "Image deformation using moving least squares." *ACM transactions on graphics (TOG)*. Vol. 25. No. 3. ACM, 2006.
- [29] Igarashi, Takeo, Tomer Moscovich, and John F. Hughes. "As-rigid-as-possible shape manipulation." *ACM transactions on Graphics (TOG)*. Vol. 24. No. 3. ACM, 2005.