Image Analysis and Pattern Extraction of Proteins Classes from One-Dimensional Gels Electrophoresis

F. Cai¹, S. Liu², P. Ten Dijke², F. J. Verbeek¹*

¹ Section Imaging and Bioinformatics, LIACS, Leiden University, Leiden, Netherlands.
² Department of Molecular Cell Biology, LUMC, Leiden University, Leiden, Netherlands.

* Corresponding author. Tel.: +31-(0)624236531; email: f.j.verbeek@liacs.leidenuniv.nl
Manuscript submitted February 12, 2017; accepted May 23, 2017.
doi: 10.17706/ijbbb.2017.7.4.201-212

Abstract: Biochemists and cell biologists use 1D gel electrophoresis to analyze protein compositions in samples. PAGE, a.k.a. sodium-dodecyl sulphate (SDS)-polyacrylamide gel electrophoresis, uses electric current to separate proteins. The proteins are visualized by staining or immunofluorescent techniques. The result of a gel is captured in an image which is usually analyzed by visual inspection. As the banding patterns get more complex and the amount of differential intensity of bands on a gel increases, an elaborate methodology is required for the analysis. For systematic processing and analyzing complex gels, we have developed new techniques for recognition, assessment and quantitative interpretation in bands on individual lanes. In this manner, we compare similarities and differences between lanes of a gel. Specific attention is given on the features that are characterizing the banding pattern. It is important to make sure all lanes of gels in an experiment can be unambiguously assessed. Therefore, lane normalization is applied after background correction. With this regard, groups of gels within an experiment can be quantitative assessed. Automation will further speed up to the process of interpretation that can now be accomplished by objective means.

Key words: 1D electrophoresis, PAGE gel, pattern extraction, feature extraction, image analysis.

1. Introduction

Mixtures of proteins can be separated and visualized by Sodium-dodecyl sulphate (SDS)-polyacrylamide gel electrophoresis (PAGE); it is a classical tool for protein analysis [1]. 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 for determining 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 fluorescent scanner. During protein sample preparation, care should be taken to include protease inhibitors to prevent degradation of proteins, which appear as faster running protein fragments on a gel.

A popular separation technique, capable of fast and easy analyzing less complex samples, is 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 them onto the polyacrylamide gel. Subsequently, according to the molecular weight of proteins, we charged by forcing them 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 can consist of several vertical lanes (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 pre-vents subjective and tedious image interpretation; this otherwise may lead to reproducibility issues. With respect to the analysis of gel electrophoresis profiles, image processing techniques require three main steps (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 [2]-[5]. Some of these plat-forms 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 [6]. Nevertheless, these approaches cannot 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. Therefore a systematic approach in the image processing is required. Consequently, we focus in this paper on proposing more elaborate methodologies so as to improve the BBS steps as follows (cf. Fig. 1):

- Background correction using geometrically constraint rolling ball method [7].
- Normalization and feature extraction for lanes/bands.
- Rough fuzzy c-means and particle swarm optimization (RFCM-PSO) hybridized clustering analysis.
- Function analysis approaches.

In short, the work presented in this paper contributes with a comprehensive way to extract qualitative and quantitative information from 1D gels; this consists 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 paper is organized as follows: in section 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 3 via a case study example, and section 4 concludes the paper.

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 of protein expression [8]. However, still a significant amount of wet laboratory expertise is 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.

![Workflow of our 1-D electrophoresis gel image analysis system.](image)

**2.1. Image Acquisition**

Sample preparation precedes image acquisition (cf. Fig. 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; e.g. the Typhoon 9410 Gel and Blot Imager, for the fluorescent signals as used in this study. Alternatively, after coomassie staining, the gel can be scanned using a lightbox and photo-scanner, e.g. the Microtek ArtixScan F2 scanner. In our experiments described in this paper, we used images containing fluorescent signals with a spatial resolution of 10 line-pairs/mm, and with a pixel size of 10 μm that resulting in an image size of 4096x1024 pixels. The images have a 16 bit dynamic range.

**2.2. 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 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 background correction and data normalization are required.

**Background correction.** The adjustment or removal of the background signal should be performed in
order to be able 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 extent discussion on an amount of approaches can be found in [9], [10].

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 [7] to come 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:

- **Raw image estimation:**
  \[
  \tilde{f}(x,y) = f(x,y) + \text{Background}(x,y)
  \]  

- **Background subtraction:**
  \[
  f(x,y) \approx \tilde{f}(x,y) - \text{Background}(x,y) + I(\tilde{f}(x,y), \text{Background}(x,y))
  \]

where \( I \) represents the mutual information as defined in [11]. Figure 2 shows the comparison of selected raw gel image and enhanced gel image.

Fig. 2. 1D profiles (the intensity in y-axis w.r.t the location in x-axis) scanned along with red line. **Left**: raw image; **Right**: enhanced image after background correction.

**Data normalization.** To make quantitative comparisons between profiles of lanes, and/or position of bands, it is necessary to normalize the distortion. Several approaches have been described [12], [13]. We employ the Sparse Dynamic Time Warping (SDTW) method [14] to yield optimal conjunct alignments, as it is very efficient and maintains the ability of searching for a more optimal solution.
Fig. 3. Normalization procedure: (a) Original profiles in a gel, e.g. fluorescently labeled gel, with density on y-axis, w.r.t position of x-axis. (b) Warping function in SDTW. (c) Mean of warped profiles, as well as a mean with its positive and negative standard deviation. (d) Density profiles of the gels after SDTW quantile normalization. Each profile is normalized to have a similar distribution.

Segmentation of the bands. The aligned profiles of each of the lanes are put up for analysis by separating all of its bands. 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. The mask is subsequently obtained from these threshold values for each of the bands in each of the lanes.

2.3. Feature Extraction and Selection

Feature extraction aims to reduce the data dimensionality and complexity, therefore its application delivers an efficient way to allow for more feasible statistical analysis. Multifactorial classes of algorithms can be applied on 2-D gel images [15], for instance, boundary-based techniques, region-based methods, and hybrid methods that combine boundary and region criteria. However, none of general or optimal procedures for extraction and quantification on 1-D gel images is reported.

Phenotypic features are considered as the composite of observable characteristics or traits for an organism [16], 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>
<thead>
<tr>
<th>Feature name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>size</td>
<td>The surface area of object</td>
</tr>
<tr>
<td>Intensity</td>
<td>Amount of intensity belong to object</td>
</tr>
<tr>
<td>Perimeter</td>
<td>The perimeter of object</td>
</tr>
<tr>
<td>Circularity</td>
<td>Area-to-perimeter ratio</td>
</tr>
<tr>
<td>Extension</td>
<td>Derived from 2&lt;sup&gt;nd&lt;/sup&gt; order invariants of object [17], [18]</td>
</tr>
<tr>
<td>Dispersion</td>
<td>Derived from 2&lt;sup&gt;nd&lt;/sup&gt; order invariants of object [17], [18]</td>
</tr>
<tr>
<td>Elongation</td>
<td>Derived from 2&lt;sup&gt;nd&lt;/sup&gt; order invariants of object [17], [18]</td>
</tr>
<tr>
<td>Orientation</td>
<td>Derived from 2&lt;sup&gt;nd&lt;/sup&gt; order invariants of object [17], [18]</td>
</tr>
</tbody>
</table>

Direct and indirect quantifications include determination of the selected phenotypic measurements (cf. Table 1 and Table 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 a significant step for identification of the phenotype.

Table 2. Texture Measurement of a Phenotype (x Represents the Intensity Value of One Pixel, While H(x) is the Histogram of the Intensities)

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

In order to guarantee that all selected features are independent and equal of variance, the Mahalanobis distance [19] is chosen as the probabilistic distance criterion. Subsequently, we employ the sequential backward feature selection [20] procedure, which selects a subset of features from feature-pool that can best predict and describe the data. The selected features based on the best performance in backward selection approach, are band width, band intensity standard deviation and lane skewness. The original 1-D gel electrophoresis images usually have different widths of bands (cf. Fig. 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. Fig. 3 (c) and (d)) as fragments migrate downstream.

2.4. Information Clustering Analysis

The measurement information is summarized into a matrix of statistics that represents information of the patterns. To date, various pattern finding procedures are concluded, however, information clustering is of great importance in research and implementation. In order to address biological questions accordingly, two of the issues arise: how to partition sets of samples that contain various features into groups among a large number of bands or lanes on electrophoresis gels; and how to figure out difference patterns amidst samples with indistinguishable information and features.

To tackle these issues, we present [21] 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 [22] 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 takes place, the PSO procedure searches for the global optimum by updating the candidate centroid positions of partitioning solutions. A schematic flow of RFCM-PSO can be described in pseudo-code as:

By taking the advantage of both 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. The result of clustering example is shown in Fig. 4.
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

Fig. 4. Clustering result of genetically distinct cancer cell lines (type A in red dot, type B in green dot) as accomplished by RFCM-PSO. 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 2.3.

3. Experimental Results

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

Statistics. With our application, we aim to find 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.
The experiment consists of 60 cell-lines that are loaded on the PAGE gels in 60 separate lanes. After running the gel this results in 60 lanes with 85 bands that we need to compare. For each band the features are computed (cf. Section 2.3). The preprocessing and normalization of the gel lanes make it possible to directly compare the bands for position and intensity. Fig. 5 (a) and (b) show all 60 lanes with their 85 bands in one visualization. The similarity of the band with respect to its corresponding cell lines can be categorized utilizing the F-test. Additionally, for a particular cancer subtype (e.g. type A), the entire set of cell-lines are recognized as two groups in terms of type A positive and type A negative. In this manner, the significance between two groups of type A can be examined by applying multi-variate T-test schema on one band in different lines. In our approach, we propose a comprehensive way employing multi-variates (cf. features from Section 2.3) in order to obtain a measure of significance in a p-value. We therefore use the multi-variate Hoteling T-square test [23].

The result of the F-test, as depicted in Fig. 5 (a), reveals that the ability for sub-type-clustering of each separate band varies with subtypes in the population of cell-lines (a priori knowledge) as well as with the result from the gel analysis. The intrinsic properties of the bands are directly related to the cell subtype that they originate from. Some bands exhibit very low F-values in all the six subtypes, these bands originate from proteins that are not present or not activated in the cell under the conditions of the experiment. In Fig. 5 (b), it is shown that bands can be reported to have significant differences ($p$-value < 0.01) from a classification into the six subtypes. This means that these bands, i.e. proteins, represent uniqueness expression profiles.

Experiments with gel electrophoresis support the understanding of the relationship between sample groups. A clustering analysis, hereby, targets to find hidden patterns in data. Hierarchical clustering revealed distinct positive (red) and negative (green) expression of proteins (cf. Fig. 6 (b)) in terms of the
normalized features quantified from gel electrophoresis images. In Figure 6 (a), cell-line groups of different subtypes are compared to examine the differences in the patterns.

Fig. 6. Heat map of a functional analysis, i.e. protein occurrence as analyzed from bands on the gel. (a) Heat map of group A vs. group B. Brighter colors represent a higher response in terms of their genotypic expression. The arrows indicate a specific type of cancer cell-line in group A/B with strong responses; (b) Clustering analysis of essentiality of detected bands (vertical) with respect to cancer subtypes (6 types) (horizontal). The colors indicate a degree of correlation between band and subtype.

4. Conclusion

This paper provides an efficient analysis system for images from 1-D PAGE gel electrophoresis. Compared to conventional gel images analysis system, we improved the analysis in a four-stage strategy. New and
innovative algorithms ensure a preprocessing that allows for an accurate comparison between the bands in the gels. Quintessential are the background correction, position normalization and intensity normalization. These elaborated approaches contribute to phenotypic quantification and henceforth classification. The pattern extraction and recognition aims to support genotype analysis. In a case study in the experiment shows that such can be accomplished by investigating and understanding the identity of characteristics of proteins that are distinct/shared between different subgroups of cancer cell-lines. In addition to our case study, in protein characterization, also DNA and RNA fragments can be separated by 1D electrophoresis. The proposed method can, in the same manner, also be applied for a systematic analysis of DNA and RNA patterns.

Acknowledgment

This work is partially supported by the Chinese Scholar Council (CSC).

References

electrophoresis protein patterns from few samples. *Journal of Proteome Research*, 7(3), 1288-1296.


**Fuyu Cai** was born in Canton, China on November 16th, 1988. He graduated with a master’s degree from the Department of Biomedical Engineering at Northeastern University in China. He is a Ph.D candidate currently studying in the bioinformatics and imaginary group of Institute of Advanced Computer Science, Leiden University, the Netherlands. He focuses on the high through put or/and high content microscopy image processing, including the analysis before, or after Fluorescence and Bright field microscope acquired images, and effect of generic expressing via patterns in image.

**Sijia Liu** was born in Hebei, China in 1988. She graduated with a master’s degree from the Department of Biochemistry at Sichuan University in China. She is a Ph.D candidate currently studying in the Leiden University Medical Center, the Netherlands. She focuses on the TGFβ pathway in breast cancer.

**Peter ten Dijke** received his Ph.D degree in 1991 from Wageningen University, the Netherlands. He did his postdoc studies with Kohei Miyazono and Carl-Henrik Heldin at the Ludwig Institute for Cancer research (LICR), Uppsala, Sweden. In 1994, he became group leader at LICR and in 1999 he moved to Netherlands Cancer Institute (NKI). From 2005 till now, he is currently a professor of molecular cell biology at Leiden University. His laboratory studies the molecular mechanisms by which TGFβ family members elicit their cellular effects via (co)receptors and intracellular SMAD effectors, and how subverted TGF family signaling is involved in cancer, vascular and bone diseases.

**Fons J. Verbeek** received his Ph.D degree in Applied Physics (pattern recognition group) in 1995 from Delft University of Technology, the Netherlands. Currently, he chairs the Imaging and Bioinformatics group at the Leiden Institute of Advanced Computer Science (LIACS), Leiden University, the Netherlands. In his group the research focusses on development of robust methods for “large-scale” image processing and analysis in the biosciences. The image
processing domain is in multi-dimensional microscope images. In addition, methods are being developed to connect image analysis results to biosciences repositories so as to further augment information with available knowledge. Pattern recognition procedures are complemented with Information Visualization to communicate outcomes in the best possible way. In his research the zebrafish model system is used in a range of different application areas.