Automatic Segmentation and Band Detection of Protein Images Based on the Standard Deviation Profile and its Derivative

Yassin Labyed, Naima Kaabouch, Member, IEEE, Richard R. Schultz, Member, IEEE, Brij B. Singh
Department of Electrical Engineering, University of North Dakota, Grand Forks, ND 58202-7165
1Department of Biochemistry & Molecular Biology, School of Medicine and Health Sciences, University of North Dakota, Grand Forks, ND 58202-7165

Abstract—Gel electrophoresis has significantly influenced the progress achieved in genetic studies over the last decade. Image processing techniques that are commonly used to analyze gel electrophoresis images require mainly three steps: band detection, band matching, and quantification and comparison. Although several techniques have been proposed to fully automate all steps, errors in band detection and, hence, in quantification are still important issues to address. In order to detect bands, many techniques were used, including image segmentation. In this paper, we present two novel, fully-automated techniques based on the standard deviation and its derivative to perform segmentation and to detect protein bands. Results show that even for poor quality images with faint bands, segmentation and detection are highly accurate.

Index Terms—Gel electrophoresis image, protein, band detection, segmentation

1. INTRODUCTION

The development of gel electrophoresis has significantly influenced the progress achieved in genetic studies in the last decade. Electrophoresis is an electrochemical separation process in which protein or RNA/DNA fragments are forced to migrate through a specific substrate, such as a polyacrylamide gel, under the influence of an electric field. The location of these fragments is then visualized on a UV trans-illuminator and captured by a camera. The resulting gel image contains the original location (wells) and the bands corresponding to the fragments that have to be analyzed to extract all the information needed.

Band detection is the ultimate goal in protein gel image analysis. Quantitative information, such as the amount of substance in each band and the molecular weight of each band, is estimated by calculating the area of the band, and by considering the position relative to a predefined reference band, respectively [2]. Several software systems have been developed to automatically analyze and process electrophoresis gel images [1-7]. Some of these systems are semiautomatic and perform band detection by segmenting the image into lanes and locating the peaks of the one-dimensional profile of each lane. However, these methods have major disadvantages because they require the user to manually select the area of interest and adjust different parameters [1, 3].

Other software systems identify bands by analyzing both the horizontal and vertical profiles of the gel image [2]. These systems use edge- and region-based segmentation methods to segment the image into objects and calculate the area of those objects to quantify the area of the bands. Nevertheless, these methods cannot generally locate faint bands, and they sometimes detect false bands due to noise.

Because the existing approaches cannot achieve accurate and automatic results, the analysis can be time consuming and not reproducible. In order to allow researchers to speed up their analyses and obtain more repeatable results, we propose two novel techniques that are fully automatic to perform lane segmentation and band detection.

2. QUALITY OF PROTEIN IMAGES

Protein images obtained through gel electrophoresis can suffer from various types of distortion and degradation. The following are the most frequent of them:

- Geometric distortion of the position of the entire image
- Horizontal lane deformation (smiles)
- Salt and pepper noise with strong non-uniform background
- Low contrast bands
- Smears on the gel image due to many proteins not moving as a global mass during the electrophoresis process
- Physical overlapping of different proteins within the same area of the gel.

Correcting defects of distortion does not represent an essential problem because effective techniques for these tasks are available in many software systems. However, degradation and local artifacts, often present in gel images, are still issues and cause essential problems for band detection.

Fig. 1 and Fig. 2 show two electrophoresis gel images containing protein lanes with nonuniform backgrounds and noisy stains. Some of the bands are very bright with long-tailed shapes, while others are very faint.
3. METHODOLOGY

3.1 Lane segmentation

Our algorithm is based on the computation of the standard deviation (STD) of column pixel intensities for every column of the image. Standard deviation is defined as follows:

$$\sigma(i) = \left(\frac{1}{N-1} \sum_{j=1}^{M} (I_{ij} - \bar{I})^2\right)^{1/2}$$  (1)

Where: $N$ and $M$ are the image numbers of columns and rows, respectively.
$i$ and $j$ are the indices of the image columns and rows, respectively.
$I_{ij}$ is the intensity of the pixel located at column $i$ and row $j$.
$\bar{I}$ is the mean of the intensity values of column $i$ and is defined follows:

$$\bar{I} = \frac{1}{M} \sum_{j=1}^{M} I_{ij}$$  (2)

Fig. 3 and Fig. 4 represent the standard deviation profiles corresponding to Fig. 1 and Fig. 2, respectively. It is apparent from the STD profile that the columns of the image where a lane exists have a higher STD value compared to the columns of the image where there is a space between lanes. Visually assessing the STD profile, we can classify the lanes according to their positions in the image. However, in computation, it is rather difficult to isolate the lanes because of the noisy peaks of the STD profile. On the other hand, the derivative of the STD profile allows us to locate the columns of the image at which the intensity changes rapidly and, consequently, locate the left and right edges of the lanes.

The derivative profile is given by the following equation:

$$\sigma'(i) = \frac{\partial}{\partial i} \left(\frac{1}{N-1} \sum_{j=1}^{M} (I_{ij} - \bar{I})^2\right)^{1/2}$$  (3)

$$= \frac{\partial}{\partial i} \left(\frac{1}{N-1} \left(\sum_{j=1}^{M} I_{ij}^2 + \sum_{j=1}^{M} \bar{I}_j^2 - 2 \sum_{j=1}^{M} I_{ij} \bar{I}_j\right)\right)^{1/2}$$  (4)

The computation of the derivative of this complex function of three terms can easily be performed if $I_{ij}(i)$ and $\bar{I}_j(i)$ are known. However, its numerical computation will require a great deal of time. Because the objectives are to locate the important changes in the standard deviation and to speed up the computation time, we used a modified derivative equation. Equation (4) is then replaced by

$$\sigma_n'(i) = \sigma(i+1) - \sigma(i)$$  (5)

$$= \left(\frac{1}{N-1} \sum_{j=1}^{M} (I_{ij+1} - \bar{I}_{ij+1})^2\right)^{1/2}$$

Fig. 3. STD profile corresponding to Fig. 1.

Fig. 4. STD profile corresponding to Fig. 2.
\[-\frac{1}{N-1} \sum_{j=1}^{M} (I_{ij} - \bar{I})^2 \]  \hspace{1cm} (6)

In this last equation, if the mean of column \(i\) is equal to the mean of column \(i+1\), i.e., \(I_{i+1} = \bar{I}\), then

\[ \sigma^2(i) = \begin{cases} > 0 & \text{if} \quad I_{i+1} > I_i \\ \leq 0 & \text{otherwise} \end{cases} \]

For the beginning of any lane, the grayscale intensity changes from a low value to a high value which results in a positive peak for the derivative profile, while the end of the lane presents an opposite behavior and, hence, results in a negative peak for the derivative.

As shown in Fig. 5 and Fig. 6, a large positive value of the derivative corresponds to the leftmost column of the lane, and similarly a large negative value of the derivative corresponds to the rightmost column of the lane.

Based on the principles outlined above, we introduce a four-step algorithm to perform lane segmentation of the gel image:

1. Find the column at which the derivative profile is at a maximum (\(\text{max}\)).
2. Find the column of the first local minimum (\(\text{min}\)) subsequent to the maximum found in step 1.
3. Set the values of the derivative profile corresponding to the detected columns between \(\text{max}\) and \(\text{min}\) to zero.
4. Repeat steps 1 through 3 until the maximum of the derivative profile falls below an experimentally chosen value.

Steps 1 and 2 enable us to find the leftmost and rightmost columns of the lane with the highest intensity, while step 3 extracts that lane and allows us to locate the next high intensity lane of the image. The process of lane segmentation ends when the value of the local maximum found in step 1 is below an experimentally chosen value, hence indicating that all the lanes have been identified.

As shown in Fig. 7 and Fig. 8, although the gel images contain noisy spots, our system accurately separated the 12 lanes from Fig. 1. and the 3 lanes from Fig. 2.

3.2 Band detection

Following the separation of the gel image lanes, we implemented and compared the results of two band detection techniques.

The first technique is based on the standard deviation (STD) profile which is obtained by computing the standard
deviation of the row pixel intensities of the lane. The peaks of the profile correspond to the location of the bands. These bands are located by searching for the local maxima of the profile.

The second band detection approach is similar to the one we proposed for lane segmentation. First, we computed the standard deviation of the row pixel intensities for every row of the lane. We then computed the derivative of the standard deviation according to equation 4 in order to obtain the derivative profile. The left and the right edges of the band are detected by finding the local maxima and minima of the derivative profile, respectively.

4. RESULTS

Ten protein gel images of different quality have been used to test the accuracy and efficiency of the system. Figure 9 represents the standard deviation of row pixel intensities of lane 1 shown in Fig. 8. Fig. 10, Fig. 11 and Fig. 12 show the detected bands, identified by arrows, using the first band detection technique.

Due to the noise in the gel image, the STD profile contains small peaks which can be detected as false bands. Faint bands result in very small peaks of the STD profile and are sometimes undetected. Moreover, closely located bands may appear as only one peak in the STD profile and can result in undetected bands.

Figure 13 shows the derivative profile obtained by computing the derivative of the STD profile of Figure 9. For each band, the maxima of the profile correspond to the left edge of the band, while the minima correspond to the left edge of the band. Fig. 14, Fig. 15 and Fig. 16 show the results of the second band detection technique.

As observed, the band detection technique based on the derivative of the STD profile is more effective in detecting the bands present in the lane than the technique based on the STD profile only.

Table 1 and Table 2 give the results of applying both detection techniques to the good quality image in Fig. 2 and the bad quality image in Fig. 1, respectively. Band detection based on the derivative profile gives an accuracy of 93 % for the good quality image and 74 % for the poor quality image, while the detection based on the STD profile gives an accuracy of 71 % for the good quality image and 60 % for the poor quality image.
In order to decrease false band detections, we filtered the STD and derivative profiles using a Gaussian low pass filter. Fig. 17 and Fig. 18 represent the filtered standard deviation and derivative profiles of Fig. 10, respectively.

As observed from these tables, the low pass filter has increased the accuracy of band detection for a good quality image to 96 % using the STD profile and to 100 % using the derivative profile. The accuracy of band detection for a poor quality image has increased to 85 % using the STD profile and to 91 % using the derivative profile.

5. CONCLUSION

We have developed a technique for segmentation and band detection of protein bands in proteomics. The proposed method is fully automatic and free of user interaction. This technique can be used and applied to any gel electrophoresis image containing protein or DNA. Future work includes automatic thresholding, and quantification of the molecular weight and the amount of protein present in the bands.

ACKNOWLEDGMENT

This work was supported by the ND EPSCoR project through National Science Foundation grant # UND0012168.

REFERENCES


