2010 International Conference on Pattern Recognition

Efficient Quantitative Information Extraction from PCR-RFLP Gel Electrophoresis Images

Christos Maramis and Anastasios Delopoulos
Department of Electrical and Computer Engineering
Aristotle University of Thessaloniki
Thessaloniki, Greece
chmaramis@mug.ee.auth.gr adelo@eng.auth.gr

***Abstract*—For the purpose of PCR-RFLP analysis, as in the case of human papillomavirus (HPV) typing, quantitative information needs to be extracted from images resulting from one-dimensional gel electrophoresis by associating the image intensity with the concentration of biological material at the corresponding position on a gel matrix. However, the background intensity of the image stands in the way of quantifying this association. We propose a novel, efficient methodology for modeling the image background with a polynomial function and prove that this can benefit the extraction of accurate information from the lane intensity profile when modeled by a superposition of properly shaped parametric functions.**

***Keywords*-background component subtraction; polynomial model; PCR-RFLP; gel electrophoresis**

I. Introduction

Gel electrophoresis is a very common technique for separating biomolecules (usually proteins or DNA molecules) on the basis of their size. Digitized images of gel electrophoresis experiments are widely used in many molecular biology applications (e.g., [1]–[3]) to extract valuable information about the biological material on the electrophorized gel matrix.

Although, at first, the extracted information was mainly of qualitative nature [1], modern applications are more and more based on the extraction of quantitative information regarding the size and concentration of the material on the gel matrix [2]. However, in most cases, it is impossible to obtain accurate quantitative information from such images before analyzing and processing them by methods that are able to reveal the underlying biological information. To this direction, we propose a novel methodology for efficient quantitative information extraction from PCR-RFLP gel electrophoresis experiments.

The rest of the paper is structured as follows: Section II describes the information extraction problem we are treating. Sections III and IV present the proposed methodology for dealing with the above problem. Section V includes the experiments that verify the efficiency of the proposed methodology. Finally, Section VI draws the conclusions of this work.

II. Problem Statement

Although the proposed methodology constitutes a generic approach to efficient information extraction from PCR-RFLP gel electrophoresis experiments, we have chosen to state the problem with the help of a specific application, namely the human papillomavirus (HPV) typing.

Molecular biologists attempt to identify the HPV types that have infected a subject by combining the established molecular biology technique of PCR-RFLP with one-dimensional gel electrophoresis [3]. First, a sample from the cervix of the subject is being collected and the HPV DNA that is contained in it is amplified with the use of the PCR technique. Next, the RFLP analysis technique is employed to segment the viral DNA into a set of fragments of predefined length in base pairs. Then, a solution of the resulting material is injected into a gel matrix and is forced by an electrophoretic force to migrate in a direction parallel to the electric field. Larger DNA fragments have lower mobilities thus covering smaller distances, while smaller fragments are more agile and cover greater distances.

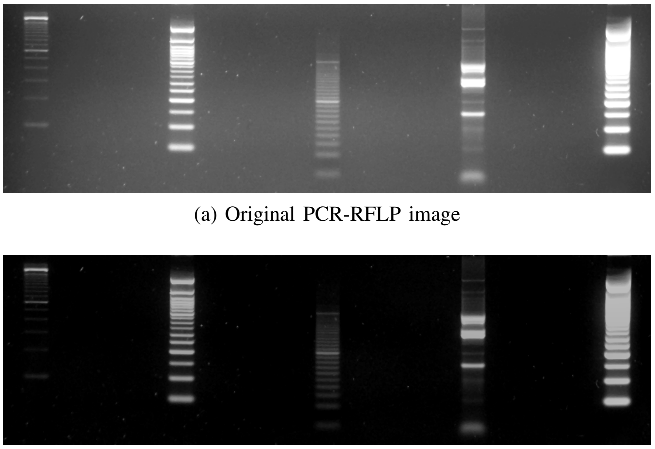
After the end of the electrophoresis, a digitized image of the gel matrix is acquired looking like the one in Fig 1a. Such images consist of isolated vertical stripes (five in the aforementioned image) called lanes which bear the HPV DNA that exists on the gel. On each lane, the DNA fragments of the same length tend to be grouped into blobs of horizontal orientation called bands.

At this point, the molecular biologists analyze - usually with the help of appropriate software - the image in order to discover the HPV types that have infected the subject. The procedure for each lane is summarized in the following steps: First, the positions of the bands on the vertical axis are located. Then, these band positions are associated with the corresponding lengths of the DNA fragments that form the bands. Finally, the set of discovered fragment lengths is compared to the expected pattern of fragment lengths for each virus type and a decision is made regarding the presence or not of each HPV type in the sample.

So far, it may seem to the reader that the band position information alone is sufficient for completing the typing

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.627

IEEE computer society



(a) Original PCR-RFLP image

(b) Background-corrected PCR-RFLP image

Figure 1. (a) A sample PCR-RFLP gel electrophoresis image with five lanes. (b) The result of removing the background component from (a) with the proposed methodology.

process. However, this is not true because the fragment length patterns of two different types may be partially overlapping. Thus, in the case of multiple infections, there may be more than one combinations of types that result in the observed set of band positions. In order to deal with such inconclusive cases, quantitative information about the concentration of the material (viral load) that forms each band has to be inferred. In other words, not only the position but also the volume of each band has to be computed. When the viral load of each band is also considered and with the assumption that each virus type contributes to each of its own bands with the same viral load, a more specific decision about the combination of types that have infected the subject can be reached.

The main idea behind the analysis of gel electrophoresis images for quantitative information extraction is the fact that the intensity of the image at some position can be related to the amount of biological material (viral load in our case) at the corresponding position of the gel matrix. However, the intensity at each image position is decomposed into two components: the intensity that is caused by the presence of viral material at this position and the background intensity, i.e., its intensity at the hypothetical case where no viral material was present at this position of the gel matrix. Obviously, when viral load information needs to be extracted only the former intensity component has to be considered. At the informative parts of the image (i.e., the lanes) the presence of HPV DNA hinders the direct computation of the background intensity. The next section deals with this issue; it proposes a methodology for subtracting the unknown background intensity from the observed intensity on the lane areas of the image.

Following background subtraction, the position and the volume of the existing bands have to be estimated. The early approaches to this problem included the binary detection (using some intensity threshold) of the bands on the two-dimensional lane image and the approximation of the viral load of each band as the sum of the intensities of the band's pixels. However, these approaches have proved to be inaccurate. Thus, the next generation of methods involves the extraction of the one-dimensional intensity profile of the lane along the vertical axis. These methods assume that the contribution of each band to the intensity profile can be modeled by a parametric function of appropriate shape (usually Gaussian or Lorentzian [2], [4], [5]). To this direction, a Gaussian or Lorentzian superposition model is employed to fit the extracted intensity profile. The resulting parameters of the model are used to estimate the position and volume of the bands. Section IV deals with the issue of intensity profile modeling.

III. Background Component Subtraction

A. Related Work

The problem of background intensity subtraction on digitized images of molecular biology experiments has received considerable attention within the framework of two-dimensional gel electrophoresis and also DNA microarray applications, giving birth to many background subtraction approaches which are apparently applicable to one-dimensional gel electrophoresis images as well. Nevertheless, these approaches are not the optimal solution in our case, since they do not take into account the special structure of PCR-RFLP images. Among them, the closest to our approach is the work in [6], which also employs a polynomial function of the spatial coordinates to model the background component.

Focusing on the related methods of interest, i.e., the background component subtraction methods which have been devised specifically for one-dimensional gel electrophoresis applications, one can discern two classes. The first class includes methods like the subtraction of a constant intensity value and the subtraction of a locally median filtered version of the image, which are very simplistic and thus perform poorly in the task of eliminating the background intensity contribution.

The approaches of the second class are more sophisticated and apply various mathematical morphology transformations (e.g., the "opening" operator in [4], the "closing" operator in [1], and the "rolling disk" transformation in [7]) on a lane's intensity profile to estimate its background intensity. Such approaches are more efficient in removing the background component from the intensity profile. However, they are sensitive to the order of the applied operators/transformations and their performance deteriorates in the – common in practice – case of overlapping bands.

B. Lane Boundary Detection

The digitized images that capture the result of PCR-RFLP gel electrophoresis experiments consist of rectangle lane areas which are separated from each other by also rectangle virus-free areas (background areas from now on) where, evidently, the observed intensity includes only the background intensity component. Our approach proposes the detection of these background areas by locating the lane boundaries and, subsequently, the utilization of the available background information to reconstruct the background intensity of the entire image by some parametric model.

The algorithm for boundary detection is based on the fact that, since the lane areas are covered with material, they will generally appear lighter than the empty background areas between the lanes. Therefore, we expect strong intensity transitions between lanes and background when moving horizontally. This effect will be magnified if we consider the entire length of a lane. Thus, the algorithm calculates the discrete intensity derivative in the horizontal direction and sums its value across the vertical direction. The resulting one-dimensional curve has local extrema at the boundaries of the lanes with negative sign at transitions from background to lane area (when moving from the left to the right of the image) and with positive sign at the inverse transitions. Each lane is bounded in the horizontal direction by a negative extremum to the left and a positive extremum to the right. The pairing of the local extrema is straightforward and is based on the similarity of their absolute values.

C. Background Component Modeling

Regarding the modeling of the background, an appropriate function had to be selected. By inspecting a number of horizontal and vertical segments of typical gel electrophoresis images, we have noticed that the intensity variations in both directions are quite smooth and this led us to the assumption that they could be modeled by a polynomial function. After experimenting with these segments regarding the appropriate polynomial degree, we selected as the parametric model of the background intensity the fourth degree polynomial of two variables, which is given by the following equation:

$$I_{back}(x,y;\boldsymbol{\theta}) = \theta_1 x^4 y^4 + \theta_2 x^4 y^3 + \theta_3 x^3 y^4 + \ldots + \theta_{25} \quad (1)$$

If $I(x,y)$ is the intensity of a digital gel image of size $M \times N$ and the set Y_{back} contains the indices of the columns that belong to the background, then our aim is to minimize the sum of squared residuals/errors with respect to $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_{25}]^T$. Thus, we seek the vector:

$$\boldsymbol{\theta}_{opt} = \arg\min_{\boldsymbol{\theta}} \sum_{x=1}^{M} \sum_{y \in Y_{back}} \{I(x,y) - I_{back}(x,y;\boldsymbol{\theta})\}^2 \quad (2)$$

Since the objective function of the optimization problem has quadratic form with respect to $\boldsymbol{\theta}$, it follows that $\boldsymbol{\theta}_{opt}$ is the solution of the corresponding linear "normal equations".

For each image, the parameter vector that best fits the intensities of the background pixels is calculated. Then, the parameters are used to produce an estimation of the background intensity of the image, i.e., the parametric model is used to reconstruct a hypothetical image of the gel matrix where no material has been loaded to the gel. Finally, the estimated background intensities of the lane pixels are subtracted from the corresponding observed intensities.

IV. Intensity Profile Modeling

When the background has been removed, the one-dimensional intensity profile for each lane is extracted. This is accomplished by taking the median value of each line of the background-corrected lane image.

The first step towards modeling the intensity profile by a superposition of parametric functions is to determine the shape that best describes the contribution of each band to the profile. A lot of attention has been drawn to this issue, with the Gaussian and the Lorentzian function being the prevailing candidates [2], [4], [5]. Indeed, our experience shows that almost all profiles can be accurately modeled by one of the above functions. This is why we propose employing both functions and comparing their goodness of fit on each lane's profile in order to determine which model will be adopted for the lane.

The intensity profile modeling process is outlined in the following paragraph. First, the peaks of the profile are detected by the watershed algorithm [8]. Their number serves as an initial estimation of the number of components of the model and also their position, height and width are used to calculate the initial value of the model's parameters. Then, a round of fitting a number of candidate superposition models to the profile by the least squares criterion begins. Supposing that $g(x;\boldsymbol{p})$ is the employed parametric basis function, then the superposition model $P(x)$ consisting of K basis functions can be expressed by the following equation:

$$P(x) = \sum_{i=1}^{K} a_i \cdot g(x;\boldsymbol{p}_i) \quad (3)$$

The candidate models differ only in the shape of their basis function and in the number of their components. The allowed range of values for the latter is as narrow as possible and centered around the aforementioned number of the profile's peaks. Finally, the adopted model is the one that minimizes the mean squared residual/error metric.

V. Experimental Results

In order to investigate the effectiveness of the fourth degree polynomial on modeling the image background intensity, we designed the following experiment: The lane-background boundaries of each image are detected and the

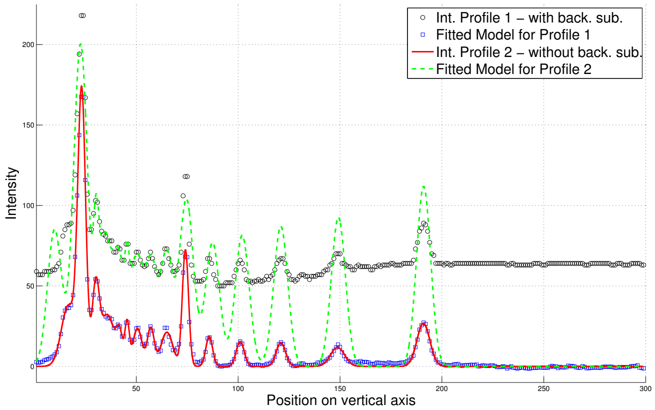


Figure 2. The result of modeling the intensity profile of the first lane of Fig. 1a by a superposition of 15 Gaussian functions with and without the proposed background subtraction.

background areas are used for estimating the parameters of the polynomial background model. Then, the background is reconstructed and the resulting Peak SNR (PSNR) metric[1] is calculated. Next, we select regions of the background areas of the same size with the lanes and treat them as if they were actually lanes; we exclude them from the background area and repeat the background modeling and reconstruction steps. The PSNR of reconstruction for the excluded areas is calculated.

The experiment was conducted on the available set of electrophoresis images and showed that the fourth degree polynomial is capable of modeling the background intensity very accurately. Specifically for the image of Fig. 1a, the PSNR of the entire background reconstruction is 34.243 and the mean PSNR of the excluded background areas reconstruction is 34.638 . The result of background component subtraction for the image of Fig. 1a is given in Fig. 1b.

The next experiment investigated the influence of background subtraction on the efficiency of the intensity profile modeling by comparing the fitting results of the intensity profile with and without the proposed background subtraction approach. See for example the fitting results of the first lane of our sample image in Fig. 2, where the fitted parametric function (dashed line) fails to model the points of the intensity profile without background subtraction (circular data points). The conclusion is that, as expected, the complete lack of background subtraction leads the proposed intensity profile modeling approach to failure.

We next compared the proposed methodology against two common background component subtraction strategies: (i) the subtraction of a constant intensity value, and (ii) the subtraction of a locally median filtered version of the lane area. The results of this experiment indicate that the proposed background subtraction methodology provides intensity profiles that can very accurately be fitted by a superposition of Gaussian or Lorentzian functions, when compared to other background subtraction strategies. The resulting PSNR of modeling for the first three lanes of the image in Fig. 1a are summarized in Table I.

Table I
PSNR OF INTENSITY PROFILE MODELING FOR DIFFERENT BACKGROUND SUBTRACTION METHODS.

	Lane 1	Lane 2	Lane 3
Proposed methodology	38.973	28.253	30.041
Constant value subtraction	23.654	26.376	14.311
Local median filtering	30.616	11.358	16.063

VI. Conclusion

In this paper we have dealt with the problem of efficient quantitative information extraction from PCR-RFLP gel electrophoresis images. We have explained why the removal of the image background intensity and the modeling of the lane's intensity profile are of major importance for our problem and proposed a novel methodology that tackles both issues. The proposed methodology allows for quantitative information to be extracted accurately, and moreover, in a completely automated and robust manner, since – in contrast to the related methods – it does not rely on the empirical determination of any parameters (such as, for instance, the order of the mathematical morphology operators/transformations in [1], [4], [7]). The presented experimental results prove the effectiveness of our methodology.

References

[1] G. Horgan and C. Glasbey, "Uses of digital image analysis in electrophoresis," *Electrophoresis*, vol. 16, no. 3, pp. 298–305, 1995.

[2] K. Takamoto, M. Chance, and M. Brenowitz, "Semi-automated, single-band peak-fitting analysis of hydroxyl radical nucleic acid footprint autoradiograms for the quantitative analysis of transitions," *Nucleic Acids Research*, vol. 32, no. 15, p. e119, 2004.

[3] E. Santiago, L. Camacho, M. Junquera *et al.*, "Full HPV typing by a single restriction enzyme," *Journal of clinical virology*, vol. 37, no. 1, pp. 38–46, 2006.

[4] J. Vohradskỳ and J. Pánek, "Quantitative analysis of gel electrophoretograms by image analysis and least squares modeling," *Electrophoresis*, vol. 14, no. 1, pp. 601–612.

[5] S. Shadle, D. Allen, H. Guo *et al.*, "Quantitative analysis of electrophoresis data: novel curve fitting methodology and its application to the determination of a protein-DNA binding constant," *Nucleic Acids Research*, vol. 25, no. 4, p. 850, 1997.

[6] R. Appel, J. Vargas, P. Palagi *et al.*, "Melanie II–a third-generation software package for analysis of two-dimensional electrophoresis images: II. Algorithms." *Electrophoresis*, vol. 18, no. 15, p. 2735, 1997.

[7] M. Skolnick, "Application of morphological transformations to the analysis of two-dimensional electrophoretic gels of biological materials," *Computer Vision, Graphics, and Image Processing*, vol. 35, no. 3, p. 332, 1986.

[8] F. Meyer, "Topographic distance and watershed lines," *Signal Processing*, vol. 38, no. 1, pp. 113–125, 1994.

[1] For the discrete signal S and its reconstructed version S_R, PSNR is defined as: $\mathrm{PSNR}(S, S_R) = 10 \cdot log_{10}(\max(S^2)/\mathrm{MSE}(S, S_R)))$