An Image Processing Tool for Automatic germline TP53 R337H Mutation Detection through PCR-RFLP

J. J. Dos Santos[a 1], E. Falcon [b] J. L. Vazquez[a] and H. Legal[a,1]
[a]National University of Asuncion, San Lorenzo, Paraguay
[b]Little Prince College, Curitiba, Brazil

Abstract

This article proposes an algorithm for the analysis of electrophoresis gel images, as an auxiliary tool for the automatic identification of the bands that contain these images by the specific PCR-RFLP assay. The correct identification of the bands within each lane leads to the correct identification of positive and negative results of the TP53 R337H germline mutation for each sample. The proposed methodology is composed of the following steps: (1) Separation of the red-color channel because it results in the best image quality. (2) gray level pre-processing of the image using mathematical morphology. (3) automatic binarization using Fuzzy C-Means thresholding technique. (4) filtering of the image using morphological binary recostruction, and (5) Detection of the bands and lanes by mask estimation.

Keywords: DNA gel image processing, thresholding, mathematical morphology, DNA mutation, PCR-RFLP assay.

1. Introduction

The tumor suppressor gene TP53 is the most frequently mutated gene in human cancer, and the germline TP53 R337H mutation is the most common mutation reported to date [1]. Efforts to identify this mutation in Brazil and at the area around the Brazilian-Paraguayan border were motivated by the highest adrenocortical carcinoma (ACC) incidence worldwide in children from the Parana-State [1]. Presence of the germline mutation is performed using the DNA PCR-restriction fragment length polymorphism (RFLP). The final step of PCR-RFLP assay leads to the interpretation of an image captured from an electrophoresis gel illuminated by an ultraviolet light source.

The detection of bands on the DNA electrophoresis gel images is fundamental for the diagnosis of a possible mutation of TP53 gene, but the correct

[1]E-mail Corresponding Author: jds.infosystem@gmail.com

identification of such bands in some cases becomes difficult for the human visual perception, due to artifacts in images such as noise, low contrast or notable variations in the intensity of image gray level without corresponding them to the bands. Such factors could lead to erroneous or unwanted conclusions.

Ismail's proposal [5] for segmentation and detection of the DNA gel images contains two stages, the first one being pre-processing and the seconc one is the detection of the lanes of bands.

Jiann [6] proposes an automatic procedure for analyzing the DNA gel images and exclusion of the unwanted background.

Skutkova et al [7], propose an algorithm for the segmentation and enhancment of the lanes detection from the bands of DNA gel images.

Sara and Daniel [8] propose a technique for the segmentation of images and the automatic detection of bands on images of electrophoresis gel, based on differents methods such as fuzzy C-means and Particle Swarm Optimization

Troy and Steven [9] proposes a tracking algorithm of lane of bands on the images of electrophoresis gel using intensity levels of gray in local maxima.

Akbari and Albregtsen [10] propose an algorithm for the automatic segmentation of the bands of DNA gel, based on the variance, maximun restricted verisimilitude and the equivalent widht.

Leal and Leal [11] propose an algorithm for the reading automation of electrophoresis gel images based on artificial vision with the use of neural networks.

This article proposes a method for the segmentation and automatic detection of the bands on images electrophoresis gel. The scheme presented is detailed as follows. In section 2, the pre-processing of the image and its segmentation are presented. In section 3, the proposed method that is presented is that it consists on the detection of the positions of the bands. In Section 4, the results of experimental tests are detailed with the percentages of errors and successes. Finally, in section 5, the conclusion is presented

2. Pre-processing

The first step of the proposed approach consists in the separation of the three RGB channels and the selection of the red channel. This is done because the objects in the gel electrophoresis image databes used are identified more clearly and with less noise than a standard RGB to grayscale image conversion. Images in the red channel are already considered grayscale, making it possible to process them directly.

2.1. Separation of objects and background

Notable variations in intensity of gray levels as a result of high-frequency noise, are one of the problems in the images of electrophoresis gel. Firstly, segmentation is processed with a mathematical morphology operation on image gray levels, specifically an erosion, where the image is eroded repeatedly with a 6×6 mask. After a substraction operation between the image of the red- color channel and the result of erosion of the same image, the objects that present high gray level intensity are obtained and then the image background is removed.

2.2. Binarization of the image with Fuzzy C-Means thresholding method

The problem of the thresholding for the electrophoresis DNA gel images is based on the identification of an optimal threshold for the separation of the object and the image background. After testing 20 thresholding methods, the method based on Fuzzy C-Means Clustering provided the best result for image binarization. The Fuzzy C-means method is an algorithm of partitional clustering, based on objective functions, that depending on the fuzzy partition defines a clustering criterion as an objective function. The method finds C clusters, in which, an element may belong to more than one cluster with a membership certain value. The membership function of an element to the cluster is given between the range $[0, 1]$. In other words, an element may belong to all classes, to one or none [3].

In the case of thresholding of digital images through object and background visualization of fuzzy sets, objects O and B, with each pixel that shows a partial membership to each region depending on its gray level, μ being the membership function, the value for each pixel $x_{i,j}$ may belong to $\mu_{O(x(i,j))} \in [0, 1]$, $\mu_{B(x(i,j))}. \in [0, 1]$ [3].

Where $\mu = (\mu_1, \mu_2, \cdots, \mu_c)$ is the membership function

Fuzzy C-Means algorithm minimizes the following objective function:

$$S_F = \sum_{i,j \in n} \sum_{c=1}^{k} \mu_{c(i,j)}^{q} ||x(i,j) - v_c||^2 \tag{1}$$

that is the Sum Square Error. Where

$$\mu_{c(i,j)} = \left[\sum_{c=1}^{k} \left(\frac{||x(i,j) - v_c||}{||x(i,j) - v_n||} \right)^{2/(q-1)} \right]^{-1} \tag{2}$$

and the centroid in the c-th cluster.

$$v_c = \frac{\sum_{i,j \in n} u^q_{c(i,j)} x(i,j)}{\sum_{i,j \in n} u^q_{c(i,j)}} \tag{3}$$

Where $\mu_{c(i,j)}$ corresponds to the membership function, q controls the amount of diffusion, $x(i,j) \in \{1, 2, \cdots, L-1\}$ belongs to value of the pixel. The Fuzzy C-Means algorithm is detailed in **algorithm 1** [3].

Algorithm 1 Fuzzy C-Means thresholding method

1: Initializes description thresholded μ_O y μ_B satisfying than $\mu_{O(x(i,j))} \in [0,1]$, $\mu_{B(x(i,j))}. \in [0,1]$
2: Compute the mean gray value of both regions using the **equation 3**
3: Assign the membership value with equation
$\mu_O(x(i,j)) = \frac{1}{1+[d(x(i,j),v_O)/d(x(i,j),v_B)]^{2/(q-1)}}$
4: Repeat steps 2,3 and 4 while there are significant changes.

2.3. Image filtering

Due to images containing variations in gray level intensities, which generate a lot of noise, the image is filtered in order to reduce the impact of high frequency noise and also to improve the separation of the bands. The method of morphological filtering by binary erosion [4] is given by:

$$\text{ero}^b(X) = X \ominus B = x \in \epsilon : B_x \subset X \tag{4}$$

Where $ero^b(X)$ is the image result of the erosion, X is the original image to be eroded, $\ominus$ is the morphological operation by a structural element B where $B_x \subset X$. The structuring element of the erosion for the DNA gel images is a horizontal mask of 5 pixels. After eroding repeatly, the reconstruction process is performed with the method of conditional dilation with the same structing element [JAC, 1996]

$$\text{dil}^B_{cX}(\text{ero}^b(X)) = \text{dil}^B(\text{ero}^b(X)) \cap X \tag{5}$$

Where $\text{dil}^B_{cX}(\text{ero}^b(X))$ is the result of conditional dilation, b is the structurin element, $\text{ero}^b(X)$ is the image erosion and X is the original image.

3. Identification of lanes and bands

After the morphological filtering, the number of lanes and their positions could be determined through the detected bands. However, due to the fact that some bands may not be identified in the electrophoresis gel images, this could lead to undesirable conclusions with the diagnoses, because of the misidentification of positions of bands in the lanes. Thus, the proposed algorithm aims to correct the identification of each lane and positions of bands through the projection of an estimated mask. The image is scanned for storage of the coordinates of the pixels marked $[M(i,j), M(x,y)]$, intervals corresponding for each horinzontal lane, where M is the matrix of image segmented. Within the coordinates of horizontal lanes obtained, the image scan is performed in the given ranges, where the size average of the bands and the distances between them, are calculated vertically and horizontally. Identifying the coordinates of the first band for each horizontal lane is carried out for estimating the mask. The estimation mask is given by calculating positions from the first lane, using the average size for bands and space between them. The mask is projected onto the segmented image, where the search is performed to pixels marked at intervals $[M(i,j) + k(i,j), M(x,y) + k(x,y)]$ corresponding to lanes and bands. If at least one pixel is marked in these intervals, the band is detected in its respective correct position. For the diagnosis of positive cases, the process algorithm must check whether three horizontal lanes, are located in an average calculated interval. Detection of one or more positive cases are identified in the specified columns. The algorithm for the detection of bands is detailed in **algorithm 2** and in **figure 1**.

Algorithm 2 identification of Lanes and bands

1: The image is scanned searching for marked pixels corresponding to the positions of the bands
2: Storing, $M(i,j)$ the first pixel and the last $M(x,y)]$, coordinates for the positions of each horizontal lane
3: The image is scanned vertically in the horizontal lanes interval
4: Store the coordinates $M(i,j)$ y $M(x,y)]$ for each band
5: Calculate the average size of the bands and space between them
6: Calculation of the mask estimation K by the average obtained
7: **if** $[M(i,j)+k(i,j), M(x,y)+k(x,y)] = 1$ **then**
8: band exists
9: **end if**
10: **if** bands_horizontal $= 3$ in $[M(i,j), M(x,y)]$ **then**
11: positive cases exist
12: **end if**

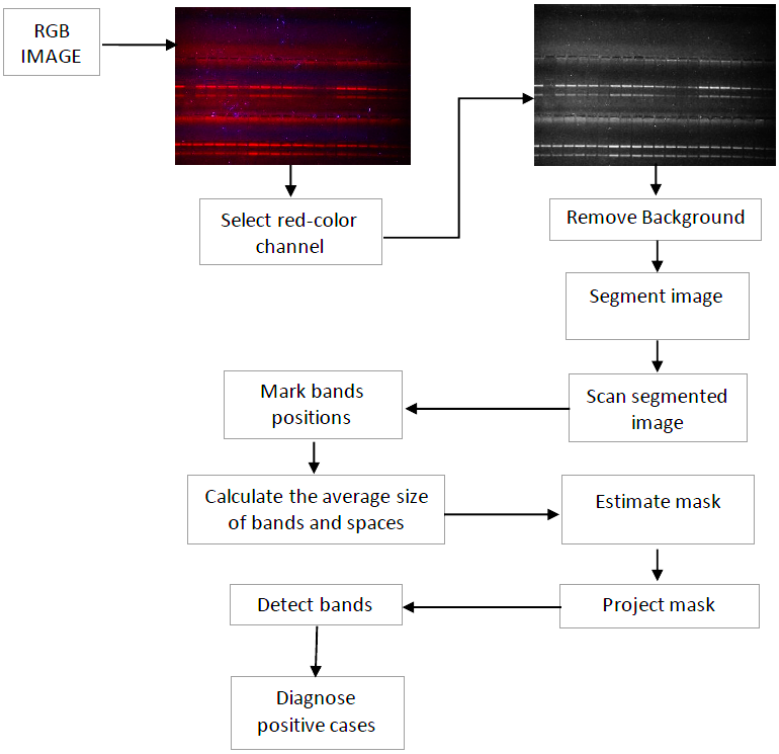

Figure 1 - Diagram of the proposed algorithm.

4. Results

Experimental tests were performed on gel electrophoresis images, obtained from a data base with 10,000 samples [2] (5 TP53 R337H germline positives and 9,995 negatives). 216 samples of the electrophoresis gel images processed **(showed Figure 2 as an example)** were analyzed using the proposed methodology. Results were validated by a medical expert in the specific DNA process. 202 samples from the 216 were correctly identified leading to a success rate of 93.52% and an error rate of 6.48%. (See **table 1**).

Taking into account the positive cases diagnosed in the germline mutation TP53 R337H, another test to validate the efficiency of correct positive germline mutation was performed searching all the bands of each specific lane. In this case, results lead to a success rate of 100%.

Table 1: Results of the lanes identification

Samples Analized	**Samples Detected correctly**	**Undetected samples**	**Error rate**	**Acerts rate**
216	202	14	6.48%	93.52%

Table 2: Results of the cases positive identification

Samples Analized	**positives cases**	**Negative cases**	**Acerts Rate**	**Error rate**
216	5	211	100%	0%

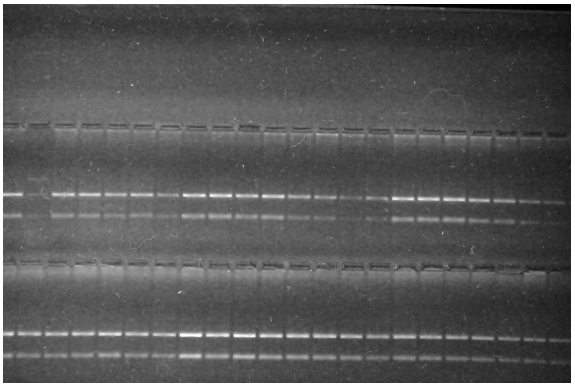

Figure 2 - Elecrophoresis gel image with 44 samples

5. Conclusion

This article presented a scheme for automatic identification of bands in the electrophoresis gel images processed by the morphological filtering and a binarization method based on fuzzy C-Means algorithm. This algorithm was proposed based on the projection of a mask estimated for the correct identification of the bands and diagnostic of possible cases. Proposed approach resulted in a success rate of 93.52% and an error rate of 6.48% of samples identification. Efficiency of diagnosis of correcto positive germiline mutation lead to a success rate of 100%, but it may require futher validation, since the prevalence of the mutation is only 5/10,000 in the Paraguayan studied population.

References

[1] Ribeiro, R. C., Sandrini, F., Figueiredo, B., Zambetti, G. P., Michalkiewicz, E., Lafferty, A. R., ... & Cat, I. (2001). An inherited p53 mutation that contributes in a tissue-specific manner to pediatric adrenal cortical carcinoma. Proceedings of the National Academy of Sciences, 98(16), 9330-9335.

[2] Falcon-de Legal, E., Ascurra, M., Custódio, G., Ayala, H. L., Monteiro, M., Vega, C., ... & Ribeiro, E. M. (2015). Prevalence of an inherited cancer predisposition syndrome associated with the germ line TP53 R337H mutation in Paraguay. Cancer epidemiology, 39(2), 166-169. doi: 10.1016/j.canep.2015.01.005.

[3] Jawahar, C. V., Biswas, P. K., & Ray, A. K. (1997). Investigations on fuzzy thresholding based on fuzzy clustering. Pattern Recognition, 30(10), 1605-1613.

[4] Facon, J. (1996). Morfología Matemática. Teoría y ejemplos. Curitiba Brasil, CITS.

[5] Ismail, I., Eltaweel, G. S., & Nassar, H. (2014). Bands detection and Lanes segmentation in DNA Fingerprint images. J Inf Comput Sci, 9(4), 243-51.

[6] Lee, J. D., Huang, C. H., Wang, N. W., & Lu, C. S. (2011). Automatic DNA sequencing for electrophoresis gels using image processing algorithms. Journal of Biomedical Science and Engineering, 4(08), 523.

[7] Skutkova, H., Vitek, M., Krizkova, S., Kizek, R., & Provaznik, I. (2013). Preprocessing and classification of electrophoresis gel images using dynamic time warping. International Journal of Electrochemical Science, 8, 1609-1622.

[8] Ibrahim I.S., Allah Makhlouf, M. A., El-Tawel Ghada.S. and Wahed M.E., "Swarm Optimization Techniques for Segmenting Gel Electrophoresis Images", Department of Information System, Suez Canal University, Ismailia, Egypt - 18-06-2016, DOI: 10.3844/ajbsp.2016.18.33

[9] Zerr, T., & Henikoff, S. (2005). Automated band mapping in electrophoretic gel images using background information. Nucleic acids research, 33(9), 2806-2812.

[10] Akbari, A., & Albregtsen, F. (2004, September). Automatic segmentation of DNA bands in one dimensional gel images produced by hybridizing techniques. In Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE (Vol. 2, pp. 2852-2855). IEEE.

[11] Leal, E., & Leal, N. (2010). Automatización de la Prueba HLA mediante análisis de imágenes de gel de Electroforesis empleando Visión Artificial.