Automatic detection of bands in electrophoresis gel images by means of optimization of a target function

David Soto
Electronics Engineering School
Tecnológico de Costa Rica
E-Mail: dsotto1@hotmail.com

Pablo Alvarado
Electronics Engineering School
Tecnológico de Costa Rica
E-Mail: palvarado@itcr.ac.cr

Abstract—Biocontrols are one of the mandatory factors to be considered in the search for sustainable agricultural activities. The amount of samples to be analyzed while searching for such agents makes unavoidable the use of technological tools that accelerate the processes involved and improve the accuracy of the results. These results are obtained using different techniques, among them the gel electrophoresis, which is widely used to analyze the molecular separation of substances like proteins and nucleic acids. The outcomes given by this technique are based on the final location of bands formed by the molecular segments in the analyzed sample once the process is finished. This work has as main objective to automatically find the location of each of these bands by means using as input gel electrophoresis images. The proposed solution is based in the optimization of a target function using the Downhill Simplex method and genetic algorithms to guide the generation of the multidimensional points to be optimized. The target function is the mean square error between the intensity distribution of a lane in the image and the sum of gaussian functions that model the distribution of DNA/RNA or proteins along the lane. Hence, the intensity level and the central band locations are obtained by finding the set of parameters that best fit the sum of gaussian functions to the lane image.

Index Terms—gel electrophoresis, image analysis, band detection, optimization, genetic algorithms

I. INTRODUCTION

Sustainable development is achieved when a balance among social, economical and environmental factors ensures the needs of present and future generations [1]. Many problems encountered while pursuing that balance require novel technological solutions for cleaner and more effective production methods, and ways to measure and compensate the effects of such methods on the environment and on the human population. In this regard, molecular biology offers tools to precisely characterize organisms, what supports the optimization of quality, quantity and impact of agricultural, forestry, and farming products, and makes possible the control of environmental health by means of biomonitors.

Even though the biomolecular tools and methods make it in principle possible to find novel solutions to present problems, the sophistication and associated costs of the required methodologies and equipment limit their spread use and slow down possible advances toward a sustainable development. Therefore, additional tools have to be created to support and reduce the time required by experts on the analysis of experimental results.

The present paper collaborates with the development of a tool to support the analysis of gel electrophoresis images, which is a technique used to analyze the presence of particular molecular chains in proteins or nucleic acids.

The analysis of gel electrophoresis images is used as a previous phase of DNA sequencing, as it permits to select particular DNA segments that are of interest to the application. These images are also generated as a validation instrument, for instance, for DNA extraction. They can also be used to compare DNA samples at a “coarser” level than sequencing, since they are precise enough for a wide range of applications.

Gel electrophoresis is based on the mobility differences of each molecule according to its size and electric charge while moving through the gel. As a first step, the substances to be characterized, such as DNA or RNA, are combined with a mixture of restriction enzymes in order to cut the molecules at specific recognition nucleotide sites. Subsequently, the sample is injected into an agarose or polyacrylamide gel matrix that is under the influence of an electric field. The results obtained once the electrophoresis process has end are based on the final location of “bands” or molecule fragments within “lanes” formed by the displacement of the samples through the gel. The band locations are used to perform a comparison between two or more lanes.

Figure 1 shows a block diagram of the electrophoresis gel image analysis tool being developed. The present work concentrates on the band detection block.

The manual location of these bands in the electrophoresis gel images is prone to error since gel images usually exhibit noise, distortions and low contrast problems. Another issue that makes the manual location of bands harder is the nonexistence of one unique criterion of what should be interpreted as band, since even though some regions within the lanes have lower intensity levels than the rest of the lane, these can not be considered as bands unless they follow a characteristic intensity spatial distribution, or band profile. Furthermore, two closely located bands can be misinterpreted as a single band.
by the limitations of the human visual system. Therefore, an automatic band detection method based on optimization techniques within a lane’s profile is explored.

This work is structured as follows: on the next section related works on band detection will be introduced, followed by the optimization approach proposed on this work. Section IV presents the results obtained with the hybrid optimization method. The conclusions are summarized in section V.

II. RELATED WORKS

In this section the state of the art about automatic location of bands in electrophoresis gel images will be presented. Mainly three different approaches were found to perform the task. In [2] the analysis of vertically oriented lanes is made using a smile-shaped template since this is the theoretical distortion model expected on the images. The template is placed over each row of the lane and subsequently the mean value of the intensity of the pixels covered by the template is computed, creating with all the obtained values a 1-D lane. As criterion, the central locations of the bands contained in the lane are the local minima of the 1-D lane. In order to simplify the detection of local minima, an intensity histogram equalization is first made, followed by a morphological segmentation and finally the 1-D lane is low-pass filtered to remove the remaining noise. The algorithm ends with the location of the local minima of the 1-D lane. However, this strategy is based on the premise that all the bands contained in the lane have distortions and are smile-shaped which is not always the case. Besides, the algorithm depends on the template used to perform the analysis and it is not capable to effectively detect bands in sectors with agglomeration of bands since it identify them as a single band.

Bajla et al. [3] propose a technique based on two stages that consider the information of the two dimensions of the lane and the interaction with the user. In the first stage the gel image is regularized in its intensity using a filter based on the principle of geometry-driven diffusion (GDD), subsequently with the lanes vertically oriented a linear detector of bands boundaries is applied which is an accumulator of intensity differences between rows, defined as:

\[ D_i = \sum_j |I_{i+1,j} - I_{i,j}| \]  

(1)

where \( I_{i,j} \) is the intensity level of the pixel located in the row \( i \) and column \( j \). Later the algorithm performs a search of the local maxima in \( D_i \). The rectangle created between each couple of local maxima is considered a band. This completes the first stage and the user can add or delete bands from the current set of bands found. Finally, the second stage uses the gradient of the image in the neighborhood of pixels that were detected as band edges to improve the location of the edges.

In order to achieve this the algorithm creates a rectangle that encloses the boundary or edge given by the gradient, part of the region considered as background and part of region considered as band, obtaining the final indicator of the band boundary edge as the absolute difference between the mean value of the background and the mean of the region considered as band. Due to the intensity regularization performed on the first stage of the algorithm this strategy does not allow to find bands in regions with low contrast, since bands with low levels of intensity will be diffused in the background even though they describe the gaussian intensity profile of a band.

Finally in [4] and [5] two strategies are proposed based in the deconvolution by means of a maximum likelihood estimation of the parameters. The first of them, for a vertical oriented lane, creates a lane with the intensity average of each row of the original lane and subsequently tries to fit it to one of the genotypes previously stored in a data base. Nevertheless, not always a data base with all the genotypes to be tested is available, this limits the method’s usefulness. The second strategy tries to fit the one-dimensional lane representation \( y(s) \) of the DNA sequence with a sum of Dirac’s delta functions

\[ x(s) = A_0 + \sum_{j=1}^{p} A_j \delta(s - \tau_j) + n \]  

(2)

under consideration that the signal \( y(s) \) can be obtained by convolution with

\[ y(s) = x(s) * w(s) \]  

(3)

where \( w(s) \) is the impulse response. The strategy focuses on finding the center band positions \( \tau_j \) and its respective amplitude \( A_j \) [6]. However, it is assumed that there are no bands at both ends of the lane and that there is white noise \( n \) normally-distributed with a low variance across the lane, which is not true on all gel images, specially due to the presence of background noise.

III. BAND ESTIMATION AS AN OPTIMIZATION PROBLEM

The intensity profile of one electrophoresis band can be mathematically modeled by means of a gaussian function [7]:

\[ B(x) = A e^{\frac{x - \mu}{\sigma}} \]  

(4)

where \( \mu \) represents the central band location, \( A \) the intensity level or amplitude on \( \mu \), and \( \sigma \) defines the width of the \( i \)-th band. The intensity distribution of a horizontally oriented lane is modeled as a sum of gaussian functions whose parameters describe every band present along the lane. The solution proposed in this work removes the lane background before processing the lane [8]. It is assumed that the maximal level in a band is represented with the intensity value zero (white)
and the minimal level with 255 intensity units (black), and as design criterion all bands in a lane are supposed to have the same variance \( \sigma_i^2 = \sigma^2 \) which is an input parameter of the algorithm.

Thus, mathematically the band detection problem can be stated as finding the parameters \( A_i, \mu_i \) and \( \sigma_i \) of the intensity profile approximation function

\[
f(x; \sigma_i, A_i, \mu_i) = \sum_{i=1}^{N} A_i e^{\frac{1}{2} \left( \frac{x-\mu_i}{\sigma_i} \right)^2}
\]

that make \( f(x; \sigma_i, A_i, \mu_i) \) most similar to the captured image profile \( \tilde{f}(x) \). This profile \( \tilde{f}(x) \) corresponds to the intensity average distribution of the five central rows of the image lane.

In order to estimate the multidimensional vector containing the set of parameters \( A_i \) and \( \mu_i \), for \( i=1 \ldots N \) that describe all bands along the lane, the Downhill Simplex optimization algorithm [9] is used in combination with a modified PESA genetic approach that improves finding global over local minima [10].

The lane under analysis is partitioned in overlapping windows to reduce the dimensionality of the optimization problem. The target function used in the optimization process is defined as the mean square error \( MSE \) between the intensity average distribution \( f(x) \) taken from the real image, and the profile model \( f(x; \sigma_i, A_i, \mu_i) \):

\[
MSE = \frac{1}{w} \sum_{x=0}^{w-1} (f(x + W) - \tilde{f}(x + W; \sigma_i, A_i, \mu_i))^2
\]

where \( w \) is the window width and \( W \) the starting position of the window in the lane. Knowledge about the valid range of possible values for the parameters \( \sigma_i, A_i \), and \( \mu_i \) allow to modify the target function such that the error function MSE excludes invalid values. This is achieved by replacing all the parameters by a corresponding sigmoid function:

\[
A_i \leftarrow A(\Psi_i) = \frac{255}{1 + e^{-\Psi_i}}, \quad \mu_i \leftarrow \mu(\Omega_i) = \frac{w - 1}{1 + e^{-\Omega_i}}
\]

where the new set of parameters are \( \Psi_i \) for the amplitude and \( \Omega_i \) for the central band positions [10].

The MSE function is an error surface in \( \mathbb{R}^N \) parameter space with \( N = 2B_{pw} \) and \( B_{pw} \) the maximum number of bands in a window. Since this surface contains local minima and saddle points, it is not possible to ensure that an optimization algorithm searching in a path on the error surface converges to the global minimum, when the algorithm starts from any random initial point.

Due the presence of saddle points it is not possible to use optimization algorithms than consider gradient information (e.g. conjugate gradients or steepest descent), since such algorithms stand still on the saddle points where, exactly as the minima, the gradient has zero magnitude [10]. For this reason the present solution uses Downhill Simplex. Although global convergence is not ensured, this algorithm is insensitive to saddle points as it only evaluates of the target function as such while searching for the minimum.

The optimization strategy proposed starts several Downhill Simplex path from different points spread over the error surface. This improves the probability that one of them hits the global minimum. The selection of the initial \( N \)-dimensional points is performed by the genetic algorithm which tends to place them in promising sites while it avoids improbable parameter sets.

By convention, a point \( P \) in the \( \mathbb{R}^N \) parameter space is given by

\[
P = [\Psi_0, \Omega_0, \Psi_1, \Omega_1 \ldots \Psi_{B_{pw}}, \Omega_{B_{pw}}]
\]

Each point \( P \) is considered as an individual, where (9) is its phenotype representation that later is converted to a binary representation called its genotype. A diagram of the optimization process is shown in Fig. 2.

![Figure 2. Optimization strategy modular diagram](image)

For each iteration the genetic algorithm PESA [11] generates a new internal population \( \rho_i \) with \( n_i \) new individuals, subsequently these are converted into their respective phenotypes, resulting in the phenotype population to optimize \( F_{\rho_i} \). Each phenotype of this population is used as point \( P_0 \) or vertex to create a new simplex to optimize.

The \( n_i \) simplices created are used as starting points by the Downhill Simplex algorithm to optimize the function MSE using \( f(x) \) as intensity reference. For each one of these simplices the optimizer returns the minimum on which it converged and the error \( \epsilon \) on that minimum, giving as result the population \( S_{\rho_i} \), of \( n_i \) locally optimized individuals. Each individual in the population \( S_{\rho_i} \) has a fitness factor

\[
\Lambda = \frac{1}{1 + \epsilon}
\]

which indicates a degree of match between the parameterized model \( \tilde{f}(x) \) and the image profile \( f(x) \). Finally, the population \( S_{\rho_i} \) is converted to its genotype representation \( C_{S_{\rho_i}} \) and then it is sent to the genetic algorithm PESA with its fitness factors, where the Pareto front is rebuilt on each new iteration according to the returned \( S_{\rho_i} \) and the current external population \( \rho_E \). Hence, the first internal population used during the first iteration by the genetic algorithm is randomly generated and the remaining internal populations are
derived from the mutation and crossover of individuals that are locally optimized and are in the current Pareto front.

After the iterative process, the element lying in front of the 1-D Pareto front corresponds to the global minimum of the MSE and therefore it contains the set of parameters, intensity amplitude and central location for the bands contained in the analyzed window. Starting from this point a last Downhill Simplex optimization is performed with a lower tolerance as stop criterion to improve the location of the global minimum.

As a design criterion in the proposed solution every individual generated by the genetic algorithm with a parameter greater than the possible ranges will have a fitness factor equal to zero in order to avoid its use as a parent for the next generation. The band estimation algorithm concludes with an stage responsible of removing all those bands that were duplicated, using techniques as thresholding and fusion of bands according with the distance between bands [10], where all those bands separated by no more than \( \Delta \mu \) pixels will be merged, giving as output a final multidimensional vector that describes all the bands located in the lane.

IV. RESULTS

The proposed method has been tested with real lanes extracted from electrophoresis gel images as well as with synthetic lanes, i.e., lanes created by algorithms with an number of bands manually configured and randomly distributed along the lane and with random intensity levels. For the latter kind of lanes, the set of parameters that describes the intensity distribution along the lane is known, allowing the evaluation of aspects like the difference between the number of real and estimated bands and the average deviation between the real and estimated central band locations.

Three tests were made. The first one uses for the optimization the exact number of bands present in the synthetic image shown in Fig. 3.

![Figure 3. Synthetic lane of size 100×20 pixels used to evaluate the optimization of the target function. It contains six bands of width \( \sigma = 1 \).](image)

Table I

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Real Value</th>
<th>Estimated values</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda_0 )</td>
<td>234</td>
<td>253,82</td>
</tr>
<tr>
<td>( \lambda_1 )</td>
<td>204</td>
<td>208,57</td>
</tr>
<tr>
<td>( \lambda_2 )</td>
<td>185</td>
<td>0,003</td>
</tr>
<tr>
<td>( \lambda_3 )</td>
<td>241</td>
<td>69,62</td>
</tr>
<tr>
<td>( \lambda_4 )</td>
<td>255</td>
<td>79,43</td>
</tr>
<tr>
<td>( \lambda_5 )</td>
<td>255</td>
<td>255</td>
</tr>
<tr>
<td>( \mu_0 )</td>
<td>16</td>
<td>16,44</td>
</tr>
<tr>
<td>( \mu_1 )</td>
<td>43</td>
<td>42,79</td>
</tr>
<tr>
<td>( \mu_2 )</td>
<td>62</td>
<td>62,2</td>
</tr>
<tr>
<td>( \mu_3 )</td>
<td>70</td>
<td>69,62</td>
</tr>
<tr>
<td>( \mu_4 )</td>
<td>79</td>
<td>79,43</td>
</tr>
<tr>
<td>( \mu_5 )</td>
<td>86</td>
<td>86,25</td>
</tr>
</tbody>
</table>

Table II

| Band | \( \lambda \) (pix) | \( \mu \) (pix) | \( |S_l - E_l| \) |
|------|------------------|----------------|------------------|
| 1    | 67,6473          | 67,6478585228 | 0,0005           |
| 2    | 125,646          | 125,6625537665 | 0,0165           |
| 3    | 182,739          | 182,7499728758 | 0,0002           |
| 4    | 187,274          | 187,2702830142 | 0,00037          |
| 5    | 205,532          | 205,536700029 | 0,0015           |
| 6    | 215,728          | 215,7142432794 | 0,0137           |
| 7    | 248,608          | 248,606449616 | 0,0012           |
| 8    | 256,459          | 256,4493157084 | 0,0096           |
| 9    | 271,955          | 271,963417659 | 0,0085           |
| 10   | 278,473          | 278,470623450 | 0,0039           |
| 11   | 420,798          | 420,7861601729 | 0,0118           |
| 12   | 479,628          | 479,819239763 | 0,1912           |
| 13   | 482,44 | 482,700695630 | 0,3001           |
| 14   | 492,197          | 492,1977915196 | 0,0008           |
| 15   | 501,954          | 501,9635734031 | 0,0096           |
| 16   | 542,639          | 542,639002006 | 0,0060           |
| 17   | 557,859          | 557,852046994 | 0,0038           |
| 18   | 376,86 | 376,8402920150 | 0,0197           |
| 19   | 583,909          | 583,937232930 | 0,0283           |
| 20   | 620,664          | 620,6300742855 | 0,0339           |
| 21   | 627,593          | 627,594325958 | 0,019           |
| 22   | 644,647          | 644,636241135 | 0,0104           |
| 23   | 663,442          | 663,4429345640 | 0,0009           |
| 24   | 671,996          | 672,000113064 | 0,004           |
| 25   | 689,243          | 689,2471432225 | 0,0041           |

The optimization of the target function was performed 5 times to produce the variations between real and estimated values of \( \mu_i \) and \( A_i \) shown in Table I. A hit occurs when the algorithm successfully estimates the \( \mu_i \) parameter within a maximum difference of 2 pixels of the real value. Even though not all the estimated values for the intensity are approximated to the real values, the accuracy of the algorithm finding the central position of all bands allows to read the intensity value directly from the image. It is possible to enhance the hit rate increasing the amount of generations evaluated by the genetic algorithm (in this case 100 generations are used).

Moreover, in order to test the average deviation between the estimated and real bands a synthetic lane \( S_l \) of 720×20 pixels with 25 bands and \( \sigma=1 \) is used with the proposed solution to optimize the target function \( MSE \). The analysis is made using windows of 60 pixels considering a maximum of 8 bands per window, with a threshold value of 20\( \mu \) (intensity units) and \( \Delta \mu=2 \) pixels.

To obtain the average deviation between real and estimated bands, each band in \( S_l \) is associated to the closest band in \( E_l \). In this case, since the algorithm is accurate detecting the bands, each band in \( S_l \) has only one corresponding band in \( E_l \).

The distance between the positions of each pair of corre-
sponding bands is computed and added up, to obtain a value of 0.71 pixels as is shown in Table II. This means an average deviation of 0.029 pixels. Figure 4 shows an illustration of this result.

![Figure 4](image4.png)

Figure 4. Lanes involved in the average deviation measuring process. From top to bottom: synthetic lane with background, synthetic lane without background, estimated lane, central position of the estimated bands.

Finally, the target function is optimized using as reference the intensity distribution of two real lanes extracted from an electrophoresis gel image, each one with a different density of bands in order to evaluate its behavior in different situations.

The first lane, with the higher density of bands allows to observe the algorithm’s behavior in situations with agglomeration of bands. The estimation is made using windows of 60 pixels and a threshold value of 20 iu considering 4 bands per window. The lanes obtained are shown in Fig. 5. In this case, the algorithm evaluates the target function 5785905 times in 30 s.

![Figure 5](image5.png)

Figure 5. Real lane used to evaluate the algorithm in situations with agglomeration of bands. From top to bottom: real lane with background, real lane without background, estimated lane, central position of the estimated bands.

In order to evaluate the algorithm’s behavior on lanes where there are regions without bands, the target function is optimized using the lane shown in Fig. 6. In this case, the target function was evaluated 3608133 times by the optimizer in 20 s.

![Figure 6](image6.png)

Figure 6. Real lane used to evaluate the algorithm in situations with more regions free of bands. From top to bottom: real lane with background, real lane without background, estimated lane, central position of the estimated bands.

V. CONCLUSIONS

This work has presented an optimization approach to find the position and amplitude of bands in a lane of a gel electrophoresis image. It makes use of the known facts that, first, all bands exhibit a Gaussian profile, and second, all bands have the same width given through the variance $\sigma^2$. Other methods proposed in the literature for band localization try to detect the band edges. Those methods are incapable of differentiating bands lying close together since in those cases the bands are visually merged.

The proposed method solves the deconvolution problem via a hybrid optimization concept which uses a Downhill Simplex algorithm to find local minima and a genetic algorithm to supervise where the local optimizations are performed.

 Tests performed on synthetic data as well as on real data have shown the effectiveness of the method. To conclude a completely automatic band detection process, a method to automatically estimate the variance and number of bands in a lane has to be devised.

ACKNOWLEDGMENT

This work has been part of the project 5402 1360 2601 supported by the Research Vicerectory of the Tecnológico de Costa Rica.

REFERENCES
