Improved Modeling of Lane Intensity Profiles on Gel Electrophoresis Images

C.F. Maramis[1] and A.N. Delopoulos[1]

[1] Dept. Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece

Abstract— **The quantitative information extraction from PCR-RFLP gel electrophoresis images requires the efficient modeling of the lane intensity profiles. To improve the acquired modeling accuracy, we introduce two novel ideas that can be incorporated in the modeling process. The first one proposes the use of the simplified integrated Weibull function as the basis function of the employed superposition model and the second proposes switching the domain of the intensity profile to-be-modeled to the unexploited fragment length domain.**

Keywords— **intensity profile modeling, integrated Weibull, fragment length domain, gel electrophoresis, PCR-RFLP**

I. Introduction

Gel electrophoresis is a very common technique for separating macromolecules (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. DNA footprinting [1], HPV Typing [2]) to extract valuable information about the molecular material that exists on a matrix covered with gel (gel matrix).

Although, at first, the extracted information was mainly of qualitative nature [3], modern applications are more and more based on the extraction of quantitative information regarding the size and the concentration of molecular material on the gel matrix [1, 4]. In most cases, the problem of extracting the above information ends up to be a task of modeling one-dimensional curves by an appropriate model function (see next section). This modeling procedure, however, is often not performed efficiently enough and this has an impact of the accuracy of the extracted information. To this direction, we introduce two novel ideas that can be applied on digitized images of PCR-RFLP one-dimensional gel electrophoresis experiments to help improve the aforementioned modeling task.

The paper is structured as follows: Section II. describes the modeling problem we are treating. Section III. presents the proposed methodologies for improving the modeling results. Section IV. describes the experiments that investigate the efficiency of the proposed methodologies. Section V. comments on the experimental results. Finally, Section VI. draws the conclusion of this work.

II. Problem Statement

Molecular biologists often attempt to identify the DNA macromolecules that exist on a subject's molecular sample by combining the established molecular biology technique of PCR-RFLP with one-dimensional gel electrophoresis (e.g. [2]). First, the sample of interest is being collected and the DNA that is contained in it is amplified with the use of the PCR technique. Next, the RFLP technique is employed to segment the 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. 1. Such images consist of vertical stripes (five in the aforementioned image) called lanes which bear the 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.

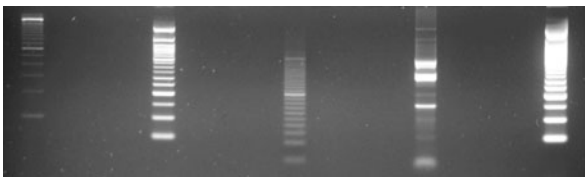

Fig. 1: A sample PCR-RFLP gel electrophoresis image with five lanes.

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 material (material load) at the corresponding position of the gel matrix. Molecular biologists employ this idea in order to identify the DNA molecules that exist on each lane. This identification task involves locating the positions of the bands on the vertical axis and, then, associating these band positions with the corresponding lengths of the DNA fragments that form the bands. The set of discovered fragment lengths provides the information required for identify-

P.D. Bamidis and N. Pallikarakis (Eds.): MEDICON 2010, IFMBE Proceedings 29, pp. 671–674, 2010.
www.springerlink.com

ing the existing DNA molecules.

So far, it may seem to the reader that the band position information alone is sufficient. However, there are applications (for instance HPV typing in the case of multiple infections) for which quantitative information about the volume of the material that forms each band has also to be inferred. In other words, not only the position but also the area of each band has to be computed. 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 material load as the sum of the intensities of the band's pixels. However, these approaches have proved to be inaccurate.

Thus, we have passed to the next generation of methods which are currently in use. These methods involve the extraction of the one-dimensional intensity profile of the lane along the vertical axis (lane's intensity profile), i.e., the mean of the lane's intensity image along the horizontal direction. These approaches assume that the contribution of each band to the intensity profile can be modeled by a parametric function of appropriate shape (e.g. Gaussian [4]). With the appropriate band shape determined, a superposition model of the corresponding basis function is employed to fit the extracted intensity profile and the resulting parameters of the model are used to estimate the position and area of the bands. Unfortunately, this modeling effort rarely results in the desired fitting efficiency and accuracy. Thus, in the following section, we describe two novel approaches for improving the attempted modeling.

III. Proposed Methodologies

A. Simplified Integrated Weibull Function

The most important step towards modeling the lane's 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 many functions proposed to serve as basis functions [5] . The Gaussian and the Lorentzian functions are currently the prevailing candidates [4, 5, 6] and, indeed, the great majority of profile peaks can be satisfactorily modeled by one of the above functions: Gaussian for more "sharp" bands and Lorentzian for bands with more prominent tails.

However, there are also cases where the actual shape of some bands lies somewhere in the middle. In such cases, a more "agile" parametric basis function has to employed, a function that has the freedom to take a wide range of shapes (including that of a Gaussian and a Lorentzian). This function, inspired from the probability density function of the integrated Weibull distribution [7], is called *simplified integrated Weibull* and is given by the following equation:

$$W(x;\beta,\gamma,x_0) = \exp(-\frac{1}{\gamma}|\frac{x-x_0}{\beta}|^{\gamma}) \quad (1)$$

where x is the position along the vertical axis, and $W(x)$ is the corresponding mean intensity along the horizontal axis. As one can observe from the Eq. (1), the simplified integrated Weibull function outnumbers both the Gaussian and the Lorentzian function in terms of independent parameters, and thus it is the perfect candidate for expressing a wide variety of band shapes.

If we assume a lane that includes N bands, then the lane's intensity profile can be modeled by a superposition of N simplified integrated Weibull functions. The mathematical expression of the superposition model is given by the following equation:

$$I(x) = \sum_{i=1}^{N} A_i W_i(x;\beta_i,\gamma_i,x_{0i}) \quad (2)$$

B. Switching to the Fragment Length Domain

In all the existing scientific efforts on parametric modeling of intensity profiles, the domain of the intensity profile function is the one of pixel positions [3, 1, 4, 5, 6]. This is also evident in our above methodological proposition by observing Eq. (1) and (2).

However, no matter how common the use of the pixel position domain is, it is not the straightforward approach. Let us elaborate more on this. As we have already mentioned in Sect. II., the ultimate goal of this modeling procedure is the determination of the DNA fragment lengths that are present on a lane and the subsequent identification of the DNA macromolecules that they compose. This means that we do not actually care about the bands positions in pixels, we just employ them to estimate the associated lengths of the DNA fragments that form the bands. If this association between pixel positions and fragment lengths is known, then the utilization of an intensity profile function in the *domain of fragment lengths* makes much more sense.

For this reason, we propose switching the intensity profile function from the pixel position domain to the fragment length domain and attempting to model the switched intensity profile by a superposition of appropriately shaped functions. This domain switching is performed by employing the association between positions along the vertical axis of the lane and DNA fragment lengths, which is provided by special lanes (called ladders) containing DNA of predefined length. The modeling of a intensity profile function which is defined on

the fragment length domain is a much more direct approach, since the modeling result will provide directly the lengths of the DNA fragments that are present in the lane as the centers of the model's basis functions.

Assuming that $x = f(y)$ is the function that associates the DNA fragment lengths (y) to pixel positions (x), the intensity profile function on the fragment length domain which will be employed for the estimation of the parameters of the superposition model is given by the following equation:

$$I_{FL}(y) = I(m(y)) \tag{3}$$

IV. Experimental Results

In order to check whether our proposed methodologies improve the intensity profile modeling procedure, we have designed and executed a series of experiments.

The first experiment investigates whether the simplified integrated Weibull function can be used for modeling single bands and it provides a quick check on whether the proposed function could serve as the basis function of the superposition model. We have located a set of isolated bands, i.e., bands that practically do not overlap with other bands in the lane and we attempted to fit the resulting data points into a Gaussian, a Lorentzian (i.e. the state of the art approaches) and a simplified integrated Weibull function. The results where very satisfactory as in almost all case the simplified integrated Weibull model outperformed the other two models. The outcome of the fitting procedure for four randomly selected isolated bands is given in Fig. 2 and the corresponding Root Mean Squared Error (RMSE) results in Table 1.

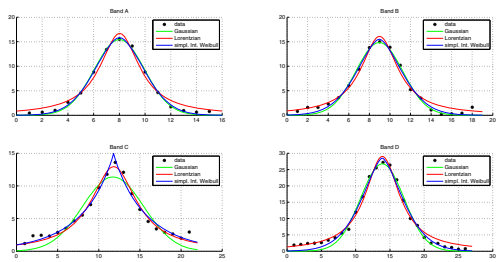

Fig. 2: The fitting result of the three models to four isolated bands.

The first experiment pleads for the appropriateness of the proposed function as a model of single bands. The next step is to check whether the simplified integrated Weibull function can also be used as the basis function for the modeling of entire intensity profiles. This role is obviously more demanding since an intensity profile usually consists of many bands that are often overlapping. For this reason, in the second experiment we have extracted the intensity profile of a number of lanes with the methodology described in [8] and attempted to fit the resulting data points into three parametric models: A superposition of Gaussian functions, a superposition of Lorentzian functions, and a superposition of simplified integrated Weibull functions. For each profile, the number of components for the three superposition models has been set equal to the number of bands on the corresponding lanes by visually inspecting the lane's image. This experiment has revealed that the proposed function generally provides better results in fitting the extracted intensity profiles when used as the basis function of the superposition model in comparison with the Gaussian and the Lorentzian function. The results of the fitting procedure for the leftmost lane of Fig. 1 is given in Fig. 3.

Table 1: RMSE of fitting for four isolated bands.

	Gaussian	Lorentzian	int. Weibull
Band A	0.4179	0.8652	0.3704
Band B	0.7353	0.8325	0.6693
Band C	1.2647	0.5434	0.5466
Band D	1.2919	1.1984	1.0669

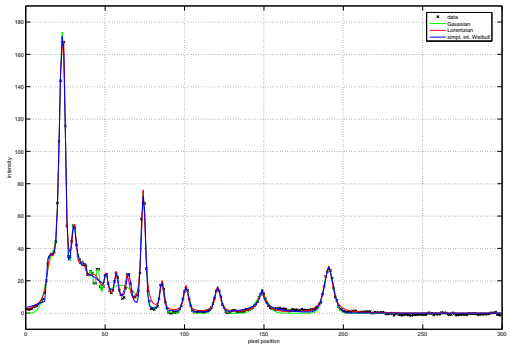

Fig. 3: The fitting result of the three models to the lane's intensity profile on the pixel position domain.

The last experiment deals with the second proposed methodology, namely the switching of the intensity profile to the fragment length domain and aims at investigating the effect of this approach on the modeling process. In other words, in this experiment we change the domain of the intensity profile that were extracted for the previous experiment and fit our new datasets into the three superposition models described above (Gaussian, Lorentzian, and integrated Weibull) in order to examine whether this domain switching improves or deteriorates the fitting results when compared to the previous

experiment. This experiment reveals that, contrary to what we expected, in the domain of fragment lengths the simplified integrated Weibull is not the prevailing basis function. However, the domain switching approach results in a noticeable improvement of the fitting results for the Gaussian basis function. The results of the fitting procedure for the leftmost lane of Fig. 1 is given in Fig. 4. The RMSE of fitting for the same lane in the pixel position domain (experiment 2) and in the fragment length domain (experiment 3) are presented in Table 2.

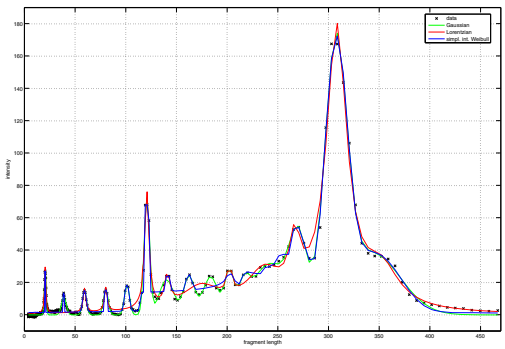

Fig. 4: The fitting result of the three models to the lane's intensity profile on the fragment length domain.

Table 2: RMSE of fitting for the lane's intensity profile on the two domains.

	Gaussian	Lorentzian	int. Weibull
Pixel Position Domain	1.9793	1.8658	1.4241
Fragment Length Domain	1.7749	3.5506	3.3108

V. Discussion

The first two conducted experiments have confirmed our guess: The simplified integrated Weibull function can very efficiently serve as the basis function for modeling the intensity profiles of interest. In fact, it is better than the current "golden standards", i.e., Gaussian and Lorentzian function since it has the ability to take the shape of both the latter functions with the appropriate selection of parameters. This feature is extremely useful in the case where, in the same lane, other bands are better expressed by a more sharp function (currently modeled as Gaussians) and others by functions with more prominent tails (currently modeled as Lorentzians). Based on our experience, such lanes with bands of diverse types do exist and they cannot be treated by the classic approaches. Thus, it seems to us that the simplified integrated Weibull approach can be proved very helpful in intensity profile modeling.

Regarding the third experiment, the results were not the expected ones. It seems that the change of the intensity profile domain does not improve the overall best model which is the simplified integrated Weibull. However, it appears to sensibly improve the Gaussian model. This provides a hint that domain switching could be scientifically/biologically sound, and a motivation to further investigate the proposed idea. Moreover, in the case where all the bands of a lane are sharp enough, i.e. when the Gaussian is the prevailing model in the pixel position domain, the proposed domain switching will indeed improve the overall modeling accuracy.

VI. Conclusion

In this paper, we have presented the issue of modeling the one-dimensional lane intensity profiles from digitized images of PCR-RFLP gel electrophoresis experiments. Seeking ways to improve the modeling accuracy, we have introduced two innovations at the modeling procedure. There are the use of new function to serve as the basis function of the parametric superposition model and the switching of the intensity profile function, which is the data to be modeled, from its commonly used domain to an unexploited domain. Finally, we have presented a series of experiments that investigate the effect of the proposed approaches on the modeling accuracy.

References

1. Das R., Laederach A., Pearlman S.M., Herschlag D., Altman R.B.. SAFA: semi-automated footprinting analysis software for high-throughput quantification of nucleic acid footprinting experiments *Rna.* 2005;11:344.
2. Santiago E., Camacho L., Junquera M.L., Vázquez F.. Full HPV typing by a single restriction enzyme *Journal of clinical virology.* 2006;37:38–46.
3. Horgan G.W., Glasbey C.A.. Uses of digital image analysis in electrophoresis *Electrophoresis.* 1995;16:298–305.
4. Takamoto K., Chance M.R., Brenowitz M.. Semi-automated, single-band peak-fitting analysis of hydroxyl radical nucleic acid footprint autoradiograms for the quantitative analysis of transitions *Nucleic Acids Research.* 2004;32:e119.
5. Vohradskỳ J., Pánek J.. Quantitative analysis of gel electrophoretograms by image analysis and least squares modeling *Electrophoresis.* ;14:601–612.
6. Shadle SE, Allen DF, Guo H., Pogozelski WK, Bashkin JS, Tullius TD. Quantitative analysis of electrophoresis data: novel curve fitting methodology and its application to the determination of a protein-DNA binding constant *Nucleic acids research.* 1997;25:850.
7. Geusebroek J.M., Smeulders A.W.M.. A six-stimulus theory for stochastic texture *International Journal of Computer Vision.* 2005;62:7–16.
8. Maramis C.M., Delopoulos A.N.. Efficient Quantitative Information Extraction from PCR-RFLP Gel Electrophoresis Images in *International Conference on Pattern Recognition*(Istanbul, Turkey) 2010.