31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

An automatic Method to identify and extract information of DNA bands in Gel Electrophoresis Images

C.M.R. Caridade, A.R.S. Marçal, T. Mendonça, A.M. Pessoa, S. Pereira

Abstract— **This paper presents a system for the automatic processing of Digital Images obtained from Gel Electrophoresis. The system identifies automatically the number and the location of lanes in the digital image, as well as the location of bands on each lane, without any intervention from the user. A reference lane with a know substance is used to compute the molecular weight of the observed (unknown) bands. The system performance was tested using 12 images, obtained from 4 gels with 3 different exposures. A total of 5443 bands were tested in 12 images, 672 reference / observed lane pairs. The average error in the estimation of molecular weight of 9.2%.**

I. INTRODUCTION

Electrophoresis through agarose gels is used to separate, identify and purify DNA fragments. Agarose gels proved convenient for sizing DNA, and the use of ethidium bromide to stain the DNA permits DNA bands to be visualized after the run (for a review see [1]). Ethidium-agarose electrophoresis of DNA was adopted in molecular biology laboratories over the world particularly after Sharp et al [2] used the method to fractionate DNA fragments generated by restriction enzyme digestion.

Fluorescent dyes are used in molecular biology laboratories for the detection and sizing of DNA and RNA in agarose gels. Ethidium bromide, usually abbreviated as EtBr [2], remains the most common dye to visualize DNA or RNA bands in agarose gel electrophoresis. It fluoresces under UV light when intercalated into DNA. Typically, DNA bands containing more than ∼10ng DNA become visible in an EtBr-treated gel viewed under UV light and the fluorescent images can be recorded as photographs or digital images.

Determining the size of DNA fragments is an integral part of other molecular biology procedures, such as physical mapping, subcloning, sequencing or separation of PCR products of defined sizes. Under a constant field strength, a linear duplex DNA molecule migrates through the gel matrix at a rate inversely proportional to the $\log_{10}$ of their molecular weight (or molecular size expressed in number of base pairs) and proportional to the applied voltage [3] [4]. However, with higher voltages (5-10 V/cm) the migration of large DNA molecules increases at a faster rate than small DNA molecules [3]. Linear double-stranded DNA molecules are sized by their relative movement through a gel compared to a molecular weight standard, so mobility measurements are critical to size determinations. To compute the size of unknown DNA fragments separated on gels, a standard curve must be created using fragments of known size from the standard molecular weight markers that are run in parallel with the unknown samples during gel electrophoresis.

Electrophoresis in agarose gels also provides a rapid and convenient way to measure the quantity of DNA. Because the amount of fluorescence is proportional to the total mass of DNA, the amount of nucleic acid can be estimated from the intensity of fluorescence emitted by ethidium bromide. The quantity of DNA in the sample can be estimated by comparing the fluorescent yield of the sample with that of a series of standards [4]. The colours on the Gel Electrophoresis Image (GEI) vary with the dye/stain used, but generally the GEI can be converted to an intensity (or greyscale) image without any loss of information. An example of a greyscale GEI is presented in figure 1 (left). A GEI might contain one or more gels, each with a number of lanes. In the example of figure 1 the image has a single gel with 8 lanes. Each lane has various bands, corresponding to the presence of DNA molecules with a given molecular weight. The intensity of a band depends on the mass (amount, quantity) of DNA.

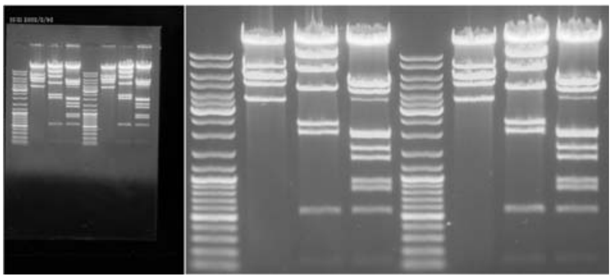

Fig. 1. Example of a Gel Electrophoresis Image (*G1b*) - original image in greyscale (left) and sub-image with the interest area extracted automatically (right).

The calculation of the molecular weights and mass for an observed substance is done using a reference in one of the lanes. The reference is a standard substance, with the molecular profile of the various bands known. There are a number of software tools available for GEI analysis, such as Gel Doc XR (BioRad GmbH, Germany) and Kodak 1D software (KodaK, USA). However, in all of these softwares there are several steps that require a considerable interaction from the operator, including the identification of the exact

C.M.R. Caridade is with Instituto Superior de Engenharia de Coimbra, R. Pedro Nunes, Qt.Nora, 3030-199 Coimbra, Portugal and Faculdade de Ciências, Univ. Porto, DMA, R. Campo Alegre, 687, 4169-007 Porto, Portugal caridade@isec.pt

A.R.S. Marçal and T. Mendonça are with Faculdade de Ciências, Univ. Porto, DMA, R. Campo Alegre, 687, 4169-007 Porto, Portugal andre.marcal@fc.up.pt tmendo@fc.up.pt

A.M. Pessoa and S. Pereira are with Faculdade de Ciências, Univ. Porto, Dep. Botânica, R. Campo Alegre, 46, 4169-007 Porto, Portugal pessoa.am@gmail.com mspereir@fc.up.pt

978-1-4244-3296-7/09/$25.00 ©2009 IEEE

location of lanes and layers.

The purpose of this work is to present a methodology to process GEI fully automatically. This includes the automatic identification of interest area, lanes and bands, as well as the calculation of the molecular weight profile for each lane observed, given a reference lane.

II. METHODS

A. Pre-processing

Initially the original GEI is subjected to a number of pre-processing tasks. This includes the conversion from color to greyscale, which is done simply by averaging the R G and B components. The greyscale image is then converted to a binary image using global thresholding, with the threshold value computed by the Otsu method [5]. The noise in the binary image is removed using the morphological operation open, with a 5 pixel radius circular structuring element [6].

Cumulative line and column histograms are used to determine the number of gels present in the image (up to 3). The process, which devides the original image in up to 3 sub-images, is described in detail in [7]. The section of the image containing the interest area (only the lanes of a single gel) is then obtained. The first non-void column in the binary image (starting from left and from right) define the left and right limits of the interest area [7]. The same approach is done for image lines. An example of the resulting image with only the interest area is presented in figure 1 (right).

B. Lane Detection

Once the interest area of the GEI is established, the next step is to identify the number of lanes and their location. The number of pixels ON for each column in the binary image is used to produce a histogram function *f*. The function *f* for test image *G1b* is presented in figure 2 (top). The interest area width is divided by an integer *n*, between pre-established minimum and maximum values (e.g. 3 to 25). For each value of *n*, the lane width (W_n) is estimated, as well as the location of the central column and edge location for all lanes. The assumption is that if the value of *n* is correct, there will be high values of *f* in and around the central columns for most lanes, and low values around the edges between lanes. Two functions are used to compute the sum of *f* values on the predicted lane edges, $F_e(n)$ using (1), and on the predicted lane centres, $F_c(n)$ using (2). A function (ϕ) based on the difference between F_c and F_e is used to evaluate the most plausible value for *n*. A plot of function ϕ for test image *G1b* is presented in figure 2 (bottom). In this case the maximum of ϕ is 8, which is the correct estimate of the number of bands. More detailed, about this algorithm can be found in [7].

$$F_e(n) = \frac{\sum_{i=1}^{n} \sum_{j=4}^{W_n-3} f(W_n \times (i-1) + j)}{(W_n - 6) \times n} \qquad (1)$$

$$F_c(n) = \frac{\sum_{i=1}^{n-1} f(W_n \times i - 1) + f(W_n \times i) + f(W_n \times i + 1)}{3 \times (n-1)} \qquad (2)$$

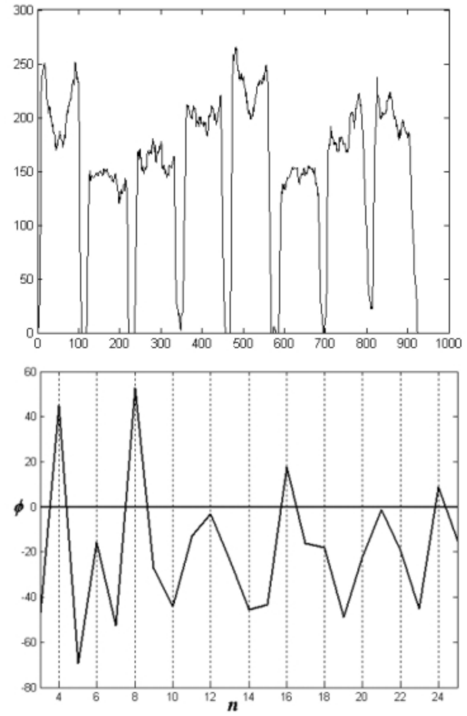

Fig. 2. Cumulative function *f* (top) and function ϕ (bottom), both for test image *G1b*.

C. Band extraction

A band is an area of high density of pixels ON in the binary image, with a roughly rectangular shape. A band is represented as a local maximum in the histogram function obtained for the number of pixels ON per line (for one lane). This function is calculated using only the central $2/3$ of the lane's width. A margin of $1/6$ of the lane width is used at both sides of the lane. The weak local maxima (small height or small width) are eliminated, as they represent false bands. The remaining local maxima are considered the centers of the lane's bands.

D. Reference Calibration

In order to evaluate the performance of the algorithms proposed, 5 standard molecular weight DNA markers were used and compared: MassRuler^{TM} DNA Ladder Mix (A)[1], GeneRuller^{TM} DNA Ladder Mix, ready-to-use (B)[2], Lambda DNA/EcoRI Marker (C), Lambda DNA/HindIII Marker (D), and Lambda DNA/EcoRI+HindIII Marker (E)[3] (Fermentas, Lithuania). Figure 3 shows the standard signature of the 5 DNA markers used.

The characteristics of each reference substance are unique, in terms of the number of strong and weak bands present, and their relative locations along the vertical axis (molecular weight). The matching process between the expected and

[1] MassRuler^{TM} DNA Ladders, LabAid^{TM}, Fermentas, 2006 - http://www.fermentas.com/pdf/labaids/labaid_massruler2006.pdf

[2] GeneRuller^{TM} DNA Ladders, LabAid^{TM}, Fermentas, 2006 - http://www.fermentas.com/pdf/labaids/labaid_generuler2006.pdf

[3] Conventional Lambda DNA Markers, LabAid^{TM}, Fermentas, 2005 - http://www.fermentas.com/pdf/labaids/labaid_lambdamarkers2005.pdf

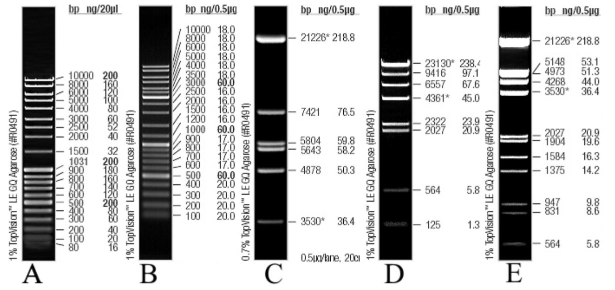

Fig. 3. Reference substances used (see text for details).

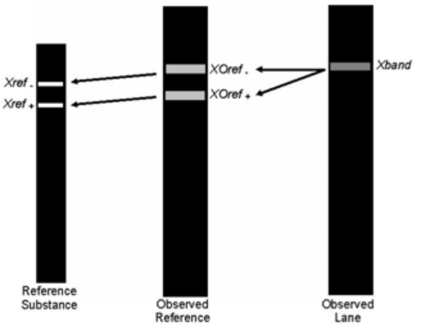

Fig. 4. Example of the interpolation process used to estimate the molecular weight of a band.

observed signatures for the reference substance is done in a two step process. First the strong bands are matched, providing an initial linear relation between the observed and expected location of the reference bands. This relation is used to predict the location of the weak bands. If there is a band within a reasonable distance image of the predicted location (10 image lines), it is paired with the corresponding band in the reference. Otherwise that band is ignored. This approach prevents the misplacement of bands in case one or more bands are not visible in the observed lane.

The characteristics of each type of reference substance are used to perform the matching. For reference C the initial step searches for 5 bands, with the a single band left to the second step (see figure 3). This band is weak and often does not appear in observed lanes. The number of strong and weak bands considered for the remaining references are: 2 strong and 18 weak for A, 3 strong and 18 weak for B, 6 strong and 2 weak for reference D and 8 strong and 3 weak for E.

E. Lane Analysis

The matching between the expected and observed signatures for the reference substance provides a function between the image position within a lane and the molecular weight (base pairs - *bp*).

The estimation of the molecular weight of an observed band is obtained by linear interpolation between the closest values along the vertical axis in the reference lane, using Equations (3) and (4)

$$bp = exp[M * (ln(Xband) - ln(XOref_{-})) + ln(Xref_{-})] \tag{3}$$

$$M = \frac{ln(Xref_{+}) - ln(Xref_{-})}{ln(XOref_{+}) - ln(XOref_{-})} \tag{4}$$

where *Xband* is the *x-axis* position of a band, $XOref_{+}$ and $XOref_{-}$ are the *x-axis* position of the closer observed reference bands and $Xref_{+}$ and $Xref_{-}$ are the x-axis positions of the reference bands matching with $XOref_{+}$ and $XOref_{-}$. The process is illustrated in figure 4.

III. EXPERIMENTAL SETUP

DNA electrophoresis experiments were prepared to evaluate the performance of the algorithms proposed, according to standard molecular biology procedures [4]. The 25 mL gels used contained 1% (w/v) agarose (Molecular Biology Grade - Bioron GmbH, Germany) and 0,2 μg/mL Ethidium Bromide (BioRad GmbH, Germany), dissolved in 1X TAE (40 mM Tris, 20 mM Sodium Acetate, 2 mM EDTA (pH=8,0) - BioRad GmbH, Germany). Each gel was loaded with four of the DNA markers twice (1 μg each); the Lambda DNA Markers were mixed with 0,20 volumes of 6X Orange DNA Loading Dye (Fermentas, Lithuania). The electrophoresis were performed for 50 min, using 80V (5 V/cm) and 1X TAE as the running buffer [4] and conducted using the LifeTechnologies Horizon 58 apparatus (GibcoBRL, UK). The GEI were acquired using the Kodak EDAS 290 imaging system and Kodak 1D software v.3.5.4 (Kodak, USA).

A typical GEI obtained in this experiment is presented in figure 1. It has a total of 8 lanes, corresponding to the following reference substances (from left to right): B, C, D, E, B, C, D, E. Four different gels were prepared (*G1*, *G2*, *G3* and *G4*), with 3 of different exposures (*a*, *b*, and *c*) used for each, resulting in a total of 12 images (all with 8 lanes). This provides 96 different test cases, using each lane as reference at a time ($4 \times 3 \times 8 = 96$) and the remaining 7 lanes as observations (672 in total, 96×7).

IV. RESULTS

The automatic detection of the region of interest was successful in all 12 images tested, as well as the detection of the number and location of lanes. The correct detection of bands depends on several factors, including the band visibility (some are indistinguishable from the background) and the amount of drag and noise.

As the range of molecular weight (*bp*) covered by the different references vary, not all bands in the observed lanes are used. Only the observed bands inside the reference range plus a margin of 10% of the range in each side are tested. The relative error is computed for each band as $|bp_e - bp_o|/bp_e$, where bp_e and bp_o are the expected and observed values of *bp*.

Each of the 12 GEI was processed separately. The results for image *G1b* (figure 1) are presented in Table I. The first element in Table I (using *B* both as reference and observed substances) is obtained as the average of 21 bands measured using lane 1 as reference and lane 5 as observation (1.8%)

TABLE I

Relative error in *bp*, for image *G1b*. Number of bands tested in ().

Ref \ *Obs*	*B*	*C*	*D*	*E*
B	1.9% (42)	12.4% (16)	11.8% (27)	10.4% (31)
C	8.6% (68)	3.3% (16)	4.4% (14)	6.0% (32)
D	7.5% (24)	0.6% (8)	9.3% (16)	4.4% (12)
E	6.8% (68)	7.0% (16)	9.8% (28)	5.0% (16)

TABLE II

Relative error in *bp* for image *G2b*. No. bands in ().

Ref \ *Obs*	*B*	*C*	*D*	*E*
B	5.0% (39)	17.3% (16)	18.9% (22)	11.5% (29)
C	12.2% (72)	8.7% (16)	2.7% (12)	8.8% (32)
D	17.0% (19)	3.4% (8)	7.0% (11)	6.4% (12)
E	10.6% (68)	7.2% (15)	11.3% (24)	1.5% (17)

TABLE III

Relative error in *bp* for image *G3b*. No. bands in ().

Ref \ *Obs*	*B*	*C*	*D*	*E*
B	2.1% (41)	10.8% (14)	10.2% (29)	14.2% (31)
C	7.7% (76)	5.6% (13)	2.7% (14)	5.1% (32)
D	12.6% (22)	3.3% (7)	14.2% (16)	6.4% (8)
E	6.6% (68)	10.1% (12)	11.3% (28)	7.2% (17)

TABLE IV

Relative error in *bp* for image *G4b*. No. bands in ().

Ref \ *Obs*	*A*	*B*	*C*	*D*
A	4.4% (38)	2.2% (84)	16.1% (8)	18.4% (24)
B	2.7% (76)	1.0% (42)	13.5% (8)	18.9% (24)
C	15.8% (63)	14.1% (73)	8.5% (17)	8.0% (14)
D	14.6% (20)	17.3% (19)	6.3% (8)	13.0% (16)

and the 21 bands measured using lane 5 as reference and lane 1 as observation (2.0%). The second value in the table, for reference *C* and observed substance *B*, is an average of 4 values: using lane 3 as reference and lanes 1 and 5 as observation (relative errors of 6.8% and 7.1%) and using lane 7 as reference and lanes 1 and 5 as observation (relative errors of 9.1% and 11.4%). In each of these four cases 17 bands were tested, resulting in a total of 68 bands, as presented in table I (inside brackets). For this GEI (*G1b*) a total of 434 bands were tested. The average relative error overall is 6.8%. The relative error are lower when the same DNA is used for reference and observation. With different DNA used for reference and observation the average error varies between 3.3% and 12.4%.

The results for images *G2b*, *G3b* and *G4b* are presented in Tables II, III and IV. For image *G2b* a total os 412 bands were tested, with an average relative error of 10.0%. For image *G3b* (428 bands) the average relative error is 8.1%, and for image *G4b* (534 bands) 10.9%.

The same processing was done for the GEI obtained with the other two levels of exposure (*a* and *c*). The results for over exposure (*c*) are overall comparable to those for normal exposure: 7.2% for *G1c*, 9.2% for *G2c*, 8.5% for *G3c* and 10.4% for *G4c*. For under exposure (*a*), the detection of bands is much harder and therefore the errors in computing *bp* are generally higher: 8.4% for *G1a*, 10.2% for *G2a*, 11.5% for *G3a* and 9.0% for *G4a*.

V. Conclusions

The proposed methods for the automatic processing of DNA Gel Electrophoresis Images (GEI) is very efficient in the detection of the number of gels, as well as the number and location of lanes. The quantitative calculation of the molecular weights for an observed DNA depends on the experimental conditions, including the reference substance used, but also on the quality of the resulting GEI (exposure, amount of drag and noise). In the experiment carried out, the average error in the estimation of *bp* for GEI with standard exposure was found to be 9.0%, with 1808 bands evaluated for 224 reference / observed lane pairs. For overexposure the error in *bp* was 8.9%, (1834 bands) and for underexposure 9.7% (1801 bands), both also for 224 reference / observed lane pairs. The error in the estimation of *bp* are obviously lower when an appropriate reference is used. Considering all 12 test images, the average error was 9.2%, with 5443 bands tested in total.

More test images with reference substances will be prepared to further evaluate the method. Plans for future work also include the automatic adjustment of rotation between the gel grid and the image, and the computation of the mass present on each band.

References

[1] P. Borst, ”Ethidium DNA Agarose Gel Electrophoresis: How it Started”, *IUBMB Life*, vol. 57, 2005, pp. 745-747.

[2] P.A. Sharp, B. Sugden, and J. Sambrook, ”Detection of two restriction endonuclease activities in haemophilus parainfluenzae using analytical agarose-ethidium bromide electrophoresis”, *Biochemistry*, vol. 12, 1973, pp. 3055-3063.

[3] R.B. Helling, H.M. Goodman, and H.W. Boyer, ”Analysis of endonuclease R-EcoRI fragments of DNA from lambdoid bacteriophages and other viruses by agarose-gel electrophoresis”, *J. Virol.*, vol. 14, 1974, pp. 1235-1244.

[4] Sambrook and Russel, *Molecular cloning: A laboratory manual (3rd edition.)*, Cold Spring Harbor Laboratory Press. 2001.

[5] N. Otsu, ”A threshold selection method from gray-level histograms”, *IEEE Trans. on Systems Man Cybernetics*, vol. 9(1), 2002, pp. 62-69.

[6] R.C. Gonzalez and R.E. Woods, *Digital Image Processing*, Prentice Hall, Upper Saddle River, New Jersey, 3rd edition, 2008.

[7] C.M.R. Caridade, A.R.S. Marçal, and T. Mendonça, ”Automatic extraction and classification of DNA profiles in digital images”, *Computational Vision and Medical Image Processing*, 2007.