The Computational Techniques Developed to Analyze DNA Gel Images

Riham M. Alawdi[1], Rania B. M. Amer[2], Ahmed M. Alzohairy[3], Wael M.Khedr[4]

[1,4] Mathematics department (computer science), Faculty of science, Zagazig University, Egypt.
[2]Physics & Mathematics Engineering Department, Faculty of Engineering, Zagazig University, Egypt.
[3] Genetics department, Faculty of agriculture, Zagazig University, Egypt.

Abstract— *The analysis of gel electrophoresis images is very crucial for molecular biologists to comprehend and interpret their experimental results. Thus, enhancing current mathematical methods and developing new accurate ones is very important and challenging task for bioinformaticians. For example, enhancing the commonly used mathematical method in gel analysis known as "Fitting method estimation" and proposing a new efficient method entitled "Ruler estimation" for preprocessing a given image and detecting lanes and bands automatically. Both mathematical methods implemented in our newly developed software. Three mathematical models namely, linear, quadratic and cubic fitting are tested for the accuracy of detecting the bands and lanes in the gel image to determine the best fitting model. A friendly user interface is developed for this new program using MATLB GUI to extract useful bimolecular information accurately and automatically. The new software has the ability to manually add or delete any band(s) and estimate the size of any unknown band(s) on the gel. Moreover, the similarity and (dis)similarity between lanes "samples" are estimated based on comparing the numbers and sizes of bands to generate a phylogram tree.*

Keywords— Clustering, Electrophoresis (GE), Fitting Data, Gel Image Preprocessing, MATLAB.

I. INTRODUCTION

Gel electrophoresis (GE) is an important technique for many molecular biology analysis's. DNA and protein gel images [1, 2] are obtained through the gel electrophoresis separations techniques of DNA and protein fragments. The separation of the polymorphic bands is based on the sizes of the negatively charged DNA fragments running from the negative cathode toward the positive anode. Each Image has some vertical lanes; each lane corresponds to one sample and has a number of horizontal bands. The band's position in a lane represents the molecular length/sizes of each band which results from its speed of immigration on the gel [2].The basic principle of band distribution in the gel image is that the larger pieces are run slower than smaller ones and staying in the upper position of the lanes [3]. Meanwhile, the smaller fragments migrate faster through the gel and occupy the lower position of the lanes.

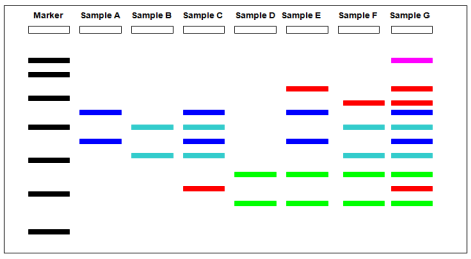
There are many factors that could affect the image quality, such as voltage, field strength, time, reorientation angle, agarose type, and concentration, the buffer chamber temperature, etc….. [4]. Image quality could affect the accuracy of extracting right information from these images. Thus, the need for enhancing and analyzing software the images is essential for biologist. The available Commercial softwares are very expensive and most free softwares are very complicated with limited options. The most famous softwares are the molecular imager Gel Doc XR, ImageJ and PyElph. For instance, the molecular imager Gel Doc XR is fairly easy but it is not free beside thatit has limited options (e.g. the user can't get build a phygram trees based on the gel image data) [5]. On the other hand, ImageJ is famous free open source software [6]that depends on the user to detect lanes and bands manually; which also do not allow the user to build or get trees from the image data. In addition, PyElph is a free software [7]to analyze DNA gel images but the user does not have the privilege add or delete any band manually.

In the current research work we tryingto introduce a software program that avoids the drawbacks of previous softwares(e.g. easy to use, draw phygram trees, add or delete any band manually or automatically, using new mathematical algorithms and MATLB GUI) [8, 9],to extract useful molecular in formation automatically from gel images. The newly developed program, "Image Analyzer", analysis is based on new bands mathematical calculation method in gel analysis known as "Fitting method estimation" and proposed new efficient method entitled "Ruler estimation". In addition, two new mathematical methods to determine molecular weight of DNA were described. Moreover, three mathematical models were tested to identify marker weights from evaluating coefficient of determination R2, residual values and RMS values for each band to determine the best fitting model. The Implantation of Gel analyzer flow chart is described and perspectives are presented.

II. BIOLOGICAL DATA BACKGROUND

Some DNA gel images were used to apply the new software analysis.[10].All the selected gel images contain a marker lane in the far left side of the image followed by different DNA samples as outlined on (Figure1)

A)



B)

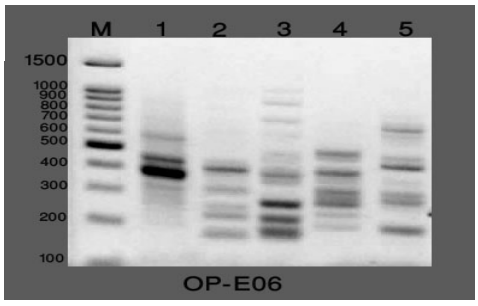


Fig.1: An outline of gel image shows. A) One of these lanes which are called Marker (M) and followed by lanes of the samples. Marker lane is usually containing bands of known size that are used to measure and compare the sizes of sample's bands. B) An example of Gel image contains a range of sizes in the marker lanes and different samples.

The distribution and immigration of the band on the gel is based on the size. The larger fragments run slower than the smaller fragments on the gel as outlined on (Figure 2).

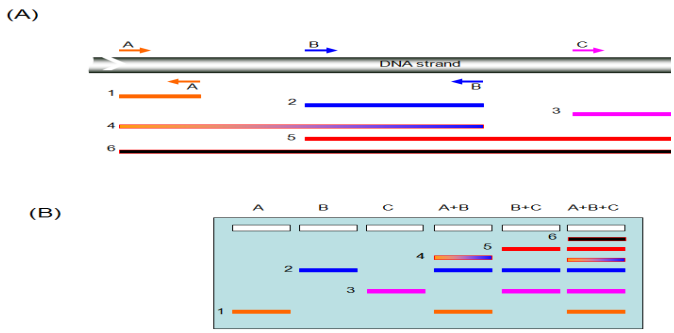


Fig.2: The amplification and immigrations of bands on the gel. A) Amplified fragments with different sizes produced by biological experiments. B) Smaller fragments run faster than the larger fragments on the gel.

III. GEL ANALYZER METHODOLOGY

There is much important information for biologist that can be extracted from gel images. The accuracy of those information is depending on the accurate detection of some parameters. The current proposed program implemented many mathematical methods to analyze an image. The user has the ability to make the analysis through five key steps starting with Image enhancement preprocessing followed byLane detection, Band detection &Length estimation, Lanes comparison& determination of Bands type and drawing a Phylogram tree (Figure3). Many samples of DNA gel images were used in this analyses in which mostly contain a marker lane and at least a five different lanes with different samples.

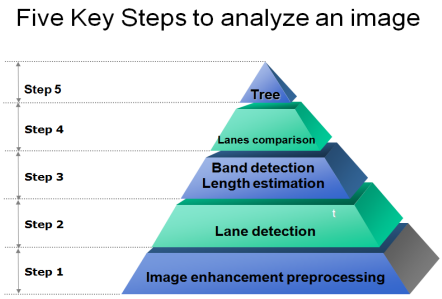


Fig.3: The methodology of gel analysis outline of the five key steps of the analysis.

Image Enhancement Preprocessing

Enhancement of an Image is very important to make an information extraction easier. The original Gel image passes through many preprocessing steps to enhance it. .For example, if the image is very small, it must be enlarged without missing important information. In current work, the tested image was cropped to be used and MATLAB toolbox was used to Convert RGB image to gray scale, which result of measuring the intensity of light at each pixel. Then, background was subtracted to extract object recognition.

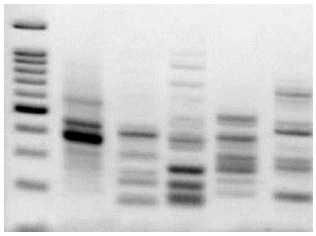


Fig.4: The original test image to go through enhancement preprocessing.

The mathematical algorithms for Background subtraction procedure can be indicated as [11]:

Background estimation:

$$\text{Background}(x,y) = (f \bullet B)(x,y) \quad \forall\ (x,y) \in \quad F$$

Background subtraction:

$$f'(x,y) = f(x,y) - \text{Background}(x,y)(x,y)$$
$$= \big(f - (f \bullet B)\big)(x,y) \quad \forall\ (x,y) \in F$$

Where, B is a size image of structuring element which is computed as in Ref. [12, 13], f(x, y) is a raw image and f. B is a dilation.

The image filtering processes are applied on image by using Gaussian low pass filter(3x3) to make image smoothing, approximate out of focus blur and to remove the different types of noise that are either present in the image during capturing or introduced into the image. Moreover, Image deblurring is used ordinary filter to refer to the procedures, which attempts to reduce the blur amount in a blurry image to be clearer and sharper [14, 15, 16].The erosion of (f) by a flat structuring element (B) at any location $(x, y)$is the set of all points in the image, where the structuring element fits into, the origin of the structuring element at every pixel location in the image is placed. The erosion is the minimum value of(f) from all values of (f) in the region of (f)coincident(B). This knows as the morphological transform to extract details from the image [17].The following figure shows the resulting histogram from enhancing an image (Figure5a). The steps of the processes is illustrated in the following flowchart (Figure5b)

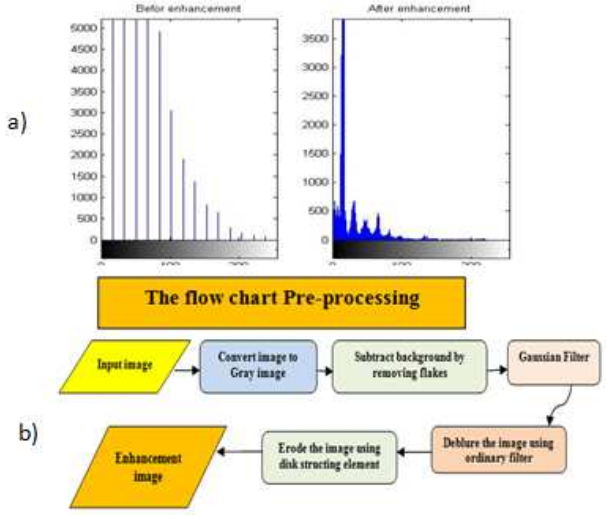


Fig. 5: The steps for erosion and deblurring image. a) The histogram of results to enhancing image. b) Flowchart illustrating the processes of Enhancement.

Lane Detection

There are many approaches to handle the lane detection problems. Troy, Z .et al described a general strategy for band mapping that uses background banding patterns to facilitate lane calling and size calibration without the detection of individual background bands and without requiring the presence of dedicated marker lanes[18].This was achieved by detecting local maxima of intensity profiles formed by vertically integrating pixel values over, either a set of horizontal sectors or the entire image, using Gel Buddy which is a practical PC and Macintosh computers tools. Also, Akbari, A. et al in Ref. [19] applied low-pass filter followed by equivalent width algorithm to locate and separate the lanes in DNA gel images and to enhance the image followed by an edge preserved noise filtering algorithm that is based on the one dimensional signal obtained by averaging the intensity for each column in the gel image into the horizontal axis.

In this research, the previous two strategies in Refs [18,19] were applied by using Wiener Low pass filter [20]which is called Minimum Mean Square Error (MMSE) or Least-Square (LS) filtering as general filter to detect lanes for any type of gel image to give the best reconstruction of the original image. Then, Nominal spacing (δ) [21, 22]is estimated between lanes in the following steps:

1- Compute the mean for every column in the image,
2- Determine the auto-covariance to the mean profile ,
3- Calculate the difference between auto-covariance profiles to get the location of the peaks in the lanes and get the mean for difference between these centers (δ). So the boundary of the first lane from 1 to δ column and the second lane from δ to 2 δ and so on as shown in Figure 6[a, b& c].

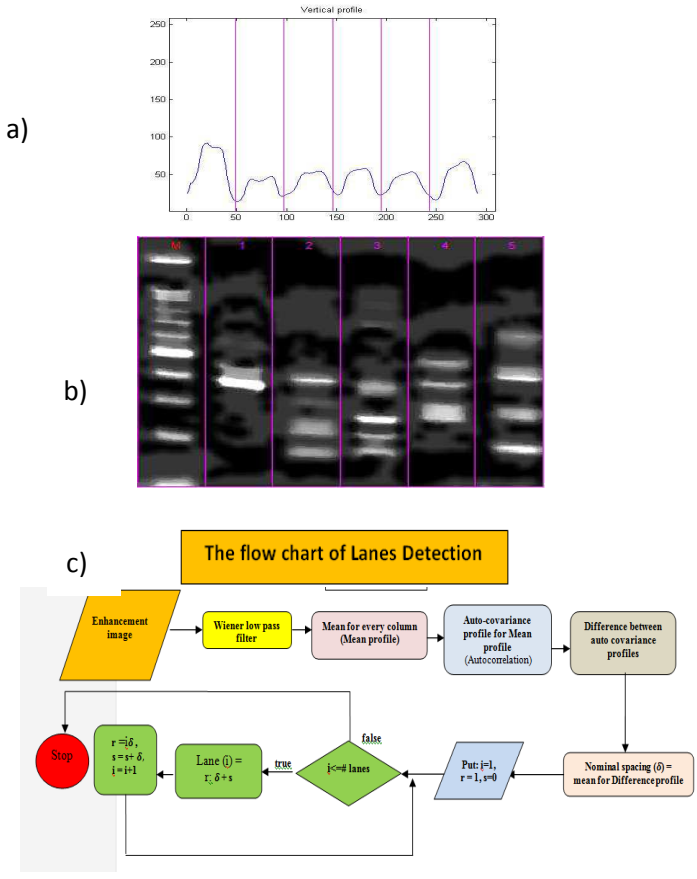


Fig. 6: Lane Detection from a gel. a) Vertical profile that shows lanes with blue color and the spacing between them with vertical pink line for the number of lanes in the image. b) Lane's detection using winner filter. c) Flow charts presenting the steps for detecting lanes.

The number of lanes is computed by using the following equation:

$$\text{number of lanes} = \text{round}\left(\frac{\text{number of the image's columns}}{\delta}\right) = \text{Real number}$$

Bands Detection

A band is an area of high density of pixels on the binary image ((starting from left and from right)) with semi-rectangular shape. Caridade, C.M.R. et al [23] presented the band as a local maximum in the histogram function obtained for the number of pixels on a line (for one lane). This function is calculated using only the central 2/3of the lane's width. A margin of 1/6of the lane width is used at both sides of the lane. In this research, the image was divided into sub images and every sub image is a lane.

For every lane:

1- Convert the gray enhancement image into binary enhanced image.
2- Labeled every white region [24].
3- Locate centers of every labeled region, these centers detect the bands.

Figure 7[a& b] shows the band detection.

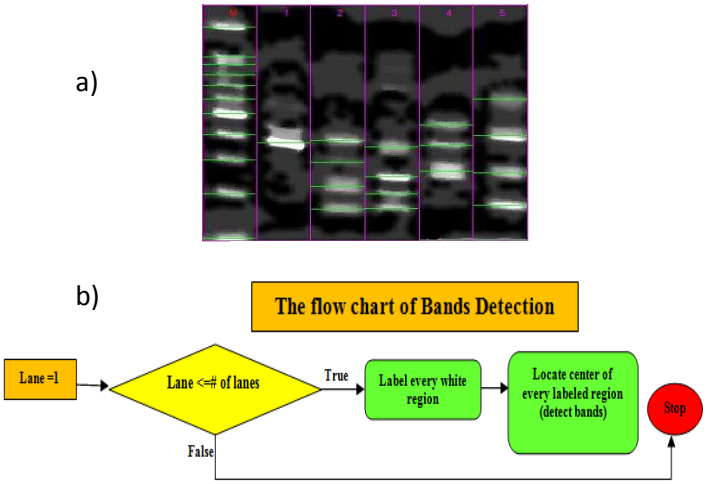


Fig. 7: Detect the bands in the image. a) Horizontal profile for the number of bands in each lane in the image. b) Flow charts presenting the steps for detecting bands.

Length Estimation

The DNA molecular weight markers[25] is known as a set of DNA fragments of known molecular sizes that are used as a standard to determine the sizes of unknown fragments. During the experiment, the user has to choose one of the lanes to load his molecular marker. Marker bands are pieces of DNA with known sizes that are used to identify the approximate size of other parallel bands that runs on the same gel during the same time of electrophoresis [26].Thus, Molecular weight is proportional to migration rate through a gel matrix comparing to the marker. In this software, determination of DNA molecular weight markers of unknown bands is estimated by two different methods entitled "Fitting estimation" and "Ruler estimation".

From these methods will get the type of every band (mono, poly or unique) which the term Polymorphism (poly) in biology happens when two or more obviously dissimilar phenotypes exist in the same population of a species. In other words, the occurrence of more than one form or morphs [27]. In same regards, monomorphism(mono) means having only one form and Dimorphism means having only two forms. However, unique morph means that there is only single individual have a distinct phenotype from his population. In same manner, the same concepts can be applied in gel analysis. Polymorphic bands mean that the bands existence is different between different lanes (samples.).This means that the bands (markers) could present in an individual or some individuals but is absent in another individual(s). Monomorphism means having only one band with the same size in all the samples and unique band is band with a unique size that only can be found in this lane (sample) as shown in figure 8 [10].

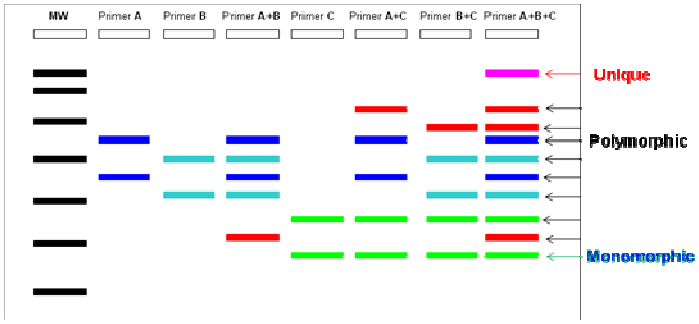


Fig.8: The type of bands mono, poly or unique.

Fitting Estimation

Advanced fitting method estimation is a famous method that is used in many papers to estimate unknown band's size .It is based on band's position. It doesn't use only linear equation but also use non linear regression. Nouh, E.A [28] was used linear equation to get unknown band's size, but Istvan, L. [29] was applied four curve types (Linear, Quadratic, Linear log and exponential) for the fitting process to be chosen the best fit. In this paper, advanced fitting method is used non linear regression (cubic equation) for all proteins and DNA gel images to detect the size of the unknown bands inside the lanes. The band of DNA gel is analyzed to obtain relative distance (x_{RD}) values for each band in the gel. The relative distance is defined as the migration distance from the top of the gel, loading well, of band of interest measured with

reference to a marker DNA lane or to a tracking dye in the bottom of the gel.

Relative distance (x_{RD}) of the DNA gel standards is calculated as follows:

$$x_{RD}=\frac{\text{distancemigratedbyDNAgel}}{\text{distancemigrated by dye}}$$

If dye line is not found in gel image, relative distance (x_{RD}) can be computed from this equation:

$$x_{RD}=\frac{\text{position of the unknowen band}}{\text{number of the rows in the gel image}}.$$

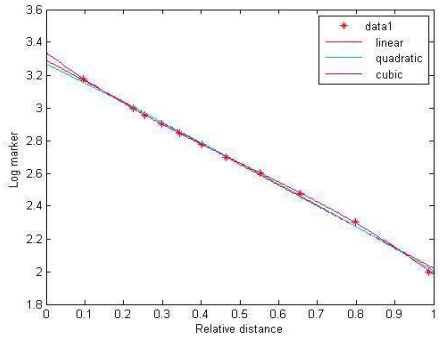
Then, relative distance is used to calculate the log of unknown band's size from cubic equation as follows:

$$y_{log}=ax_{RD}^{3}+bx_{RD}^{2}+cx_{RD}+d$$

Where y_{log} is alog of the unknown size , x_{RD} is the Relative distance of a chosen band and a,b, c, d are coefficients .

Plotting the results from first order to ten order polynomials with the actual DNA molecular weight marker. Then, the cubic equation is chosen to estimate the unknown band weights as shown in figure 9[a & b].

a)



b)

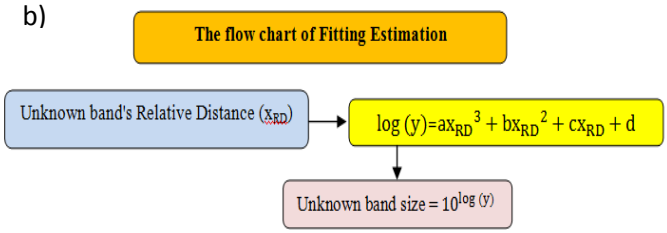


Fig. 9: Comparison of models with given data. a) The relative distances between linear, quadratic and cubic. b) Flow charts presenting the steps of Fitting Estimation

The Following tables shows the mathematical comparison between those linear, quadratic and cubic models to show the best model for identifying unknown bands of DNA molecular weight marker.

Table 1 shows the predicted molecular weights produced by linear, quadratic and cubic models and known molecular weights of bands, where,

x_i the position of the band in the lane

x_{RD} is the relative distance

y_{ex} is the exact value of marker's weights

y_{log_ex} is a log of the marker's weights

$\hat{y}_{lin}$ is approximate values from the linear equation

y_{log_lin} is log of the weights from the linear equation.

$\hat{y}_{qu}$ is approximate values from the quadratic equation.

y_{log_qu} is log of the weights from the quadratic equation.

$\hat{y}_{cu}$ is approximate values from the cubic equation

y_{log_cu} is log of the weights from the cubic equation.

$\bar{y}_{ex}$ is the mean for the exact values (y_{ex})

Table 1: Comparison of linear, quadratic and cubic models for every band with known molecular weight.

#band	x_i	x_{RD}	y_{ex}	y_{log_ex}	y_{log_lin}	$\hat{y}_{lin}$	y_{log_qu}	$\hat{y}_{qu}$	y_{log_cu}	$\hat{y}_{cu}$
1	25	0.0965	1500	3.1761	3.1673	1470.1	3.1539	1425.4	3.1783	1507.8
2	58	0.2239	1000	3.0000	3.0054	1012.4	3.0022	1005.1	2.9958	990.4
3	66	0.2548	900	2.9542	2.9661	924.9	2.9649	922.4	2.9553	902.2
4	77	0.2973	800	2.9031	2.9121	816.8	2.9133	819	2.9015	797.0
5	89	0.3436	700	2.8451	2.8532	713.2	2.8566	718.7	2.8447	699.4
6	104	0.4015	600	2.7782	2.7796	602.0	2.7850	609.5	2.7761	597.2
7	120	0.4633	500	2.6990	2.7010	502.4	2.7079	510.4	2.7046	506.5
8	143	0.5521	400	2.6021	2.5881	387.4	2.5957	394.1	2.6025	400.5
9	170	0.6564	300	2.4771	2.4556	285.5	2.4618	289.6	2.4796	301.7
10	207	0.7992	200	2.3010	2.2740	187.9	2.2746	188.2	2.2957	197.6
11	256	0.9884	100	2.0000	2.0335	108	2.0201	104.7	2.0016	100.4

To measure the errors of the obtained results, root mean squares of errors (RMS) are calculated such as[30]:

$$\text{RMS of errors} = \left[\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}{n}}\right],$$

$$\text{Residual } \hat{e}_i(\text{Absolute error}) = |\hat{y}_i - y_i|$$

where $\hat{y}_i$ is the obtained approximate solution (predicted value) while yi is the exact (observed value) solution and n is the number of bands.

Table 2: Calculating the root mean square of errors (RMS) and absolute errors for three models

No. of bands	$\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_{i\_lin}-y_i)^2}{n}}$	$\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_{i\_qu}-y_i)^2}{n}}$	$\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_{i\_cu}-y_i)^2}{n}}$
1	9.0207	22.5064	2.3407
2	3.7436	1.5333	2.8873
3	7.5049	6.7461	0.6663
4	5.0544	5.7363	0.9066
5	3.9712	5.6455	0.1738
6	0.5900	2.8785	0.8567
7	0.7162	3.1329	1.9597
8	3.8066	1.7652	0.1359
9	4.3731	3.1416	0.5233
10	3.6409	3.5637	0.7339
11	2.4151	1.4247	0.1141
RMS	44.8367	58.0741	11.2983

There are three regression formulas [31]:

1- The measure of explained variation (Regression sum of squares),

$$SSR = \sum_{i=1}^{n}(y_{ex}-\bar{y}_{ex})^2.$$

2- The measure of explained variation (Error sum of squares),

$$SSE = \sum_{i=1}^{n}(y_{ex}-\hat{y})^2.$$

3- The measure of total variation(Total sum of squares),

$$SST = SSR + SSE = \sum_{i=1}^{n}(y_{ex}-\bar{y}_{ex})^2.$$

Coefficient of Determination $R^2 = 1 - \frac{SS_E}{SS_T} = \frac{SS_R}{SS_T}$ is computed for three models, where $R^2 \geq 0$.

Table 3: Calculating Error sum of squares SSE and R^2 for linear, quadratic and cubic models.

$(y_{ex} - \bar{y}_{ex})^2$	$(y_{ex} - \hat{y}_{lin})^2$	$(y_{xg} - \hat{y}_{qu})^2$	$(y_{ex} - \hat{y}_{cu})^2$
745870	895.0990	5571.9	60.2694
132230	154.1585	25.9	91.7019
69500	619.5612	500.6	4.8831
26780	281.0192	362.0	9.0408
4050	173.4751	350.6	0.3321
1320	3.8288	91.1	8.0726
18600	5.6423	108.0	42.2461
55870	159.3939	34.3	0.2031
113140	210.3603	108.6	3.0127
190410	145.8155	139.7	5.9250
287690	64.1610	22.3	0.1433
$SS_T = \sum_{i=1}^{n}(y_{ex} - \bar{y}_{ex})^2$ = 1645500	$SS_{E\_lin} = \sum_{i=1}^{n}(y_{ex} - \hat{y}_{lin})^2$ =2712.5	$SS_{E\_qu} = \sum_{i=1}^{n}(y_{ex} - \hat{y}_{qu})^2$ =7314.9	$SS_{E\_cu} = \sum_{i=1}^{n}(y_{ex} - \hat{y}_{cu})^2$ = 225.8302
R^2	R^2_{lin} = 0.9984	R^2_{qu} = 0.9956	R^2_{cu} =0.9999

Table 4: Calculating Regression sum of squares SSR and R^2 for linear, quadratic and cubic models.

$(y_{ex} - \bar{y}_{ex})^2$	$(\hat{y}_{lin} - \bar{y}_{ex})^2$	$(\hat{y}_{qu} - \bar{y}_{ex})^2$	$(\hat{y}_{cu} - \bar{y}_{ex})^2$
745870	695090	622510	759340
132230	141420	135960	125360
69500	83250	81800	70670
26780	32540	33370	25800
4050	5900	6780	3980
1320	1180	720	1540
18600	17950	15870	16860
55870	62000	58670	55650
113140	123110	120260	111980
190410	201100	200870	192540
287690	279160	282640	287280
$SS_T = \sum_{i=1}^{n}(y_{ex} - \bar{y}_{ex})^2$ = 1645500	$SS_{R\_lin} = \sum_{i=1}^{n}(\hat{y}_{lin} - \bar{y}_{ex})^2$ = 1642700	$SS_{R\_qu} = \sum_{i-1}^{n}(\hat{y}_{qu} - \bar{y}_{ex})^2$ = 1559400	$SS_{R\_cu} = \sum_{i=1}^{n}(\hat{y}_{cu} - \bar{y}_{ex})^2$ = 1651000
R^2	R^2_{lin} = 0.9983	R^2_{qu} = 0.9477	R^2_{cu} = 1.0034

Residual $\hat{e}_i = |y_i - \hat{y}_i|$.

Table 5: Estimate the residual errors for three models

$\hat{e}_i$	$\lvert y_{ex} - \hat{y}_{lin} \rvert$	$\lvert y_{xg} - \hat{y}_{qu} \rvert$	$\lvert y_{ex} - \hat{y}_{cu} \rvert$
$\hat{e}_1$	29.9182	74.6453	7.7633
$\hat{e}_2$	12.4161	5.0852	9.5761
$\hat{e}_3$	24.8910	22.3742	2.2098
$\hat{e}_4$	16.7636	19.0252	3.0068
$\hat{e}_5$	13.1710	18.7240	0.5763
$\hat{e}_6$	1.9567	9.5469	2.8412
$\hat{e}_7$	2.3753	10.3908	6.4997
$\hat{e}_8$	12.6251	5.8544	0.4506
$\hat{e}_9$	14.5038	10.4195	1.7357
$\hat{e}_{10}$	12.0754	11.8193	2.4341
$\hat{e}_{11}$	8.0101	4.7251	0.3785

From figure 9 and tables 1 to 5, show that the cubic model is to be the best fitting and it is used to identify unknown bands of DNA molecular weight marker. Cubic model is best because:

1- Coefficient of Determination R2 is near than 1.
2- Get smaller residual values $\hat{e}_i$.
3- RMS ≤1.053013102 .

Table 6: Final approximated marker weight of unknown bands using cubic model.

	A	B	C	D	E	F	G
1	M	lane1	lane2	lane3	lane4	lane5	
2	1500	365(Uniqu	376(Uniqu	350(Uniqu	448(Uniqu	597(Unique)	
3	1000	0	292(Uniqu	244(Uniqu	357(Uniqu	396(Unique)	
4	900	0	218(Uniqu	198(Uniqu	262(Uniqu	259(Unique)	
5	800	0	159(Uniqu	161(Uniqu	0	168(Unique)	
6	700	0	0	0	0	0	
7	600	0	0	0	0	0	
8	500	0	0	0	0	0	
9	400	0	0	0	0	0	
10	300	0	0	0	0	0	
11	200	0	0	0	0	0	
12	100	0	0	0	0	0	

From table 6, cubic model is used to estimate the unknown bands weight and is very useful to get the type of every band that is mono, poly or unique.

Table 7: The bands converging in distance take the same values in different lanes.

	A	B	C	D	E	F	G
1	M	lane1	lane2	lane3	lane4	lane5	
2	1500	366(Poly)	366(Poly)	358(Poly)	448(Uniqu	597(Unique)	
3	1000	0	292(Uniqu	244(Uniqu	358(Poly)	396(Unique)	
4	900	0	218(Uniqu	198(Uniqu	261(Poly)	261(Poly)	
5	800	0	160(Uniqu	164(Poly)	0	164(Poly)	
6	700	0	0	0	0	0	
7	600	0	0	0	0	0	
8	500	0	0	0	0	0	
9	400	0	0	0	0	0	
10	300	0	0	0	0	0	
11	200	0	0	0	0	0	
12	100	0	0	0	0	0	

From figure 7 and table 7, the software program can take the same values of bands converging in distance with different lanes.

Example 1:

From cubic model,

$$y_{log} = ax_{RD}^3 + bx_{RD}^2 + cx_{RD} + d$$

where: a, b, c, d are coefficients such that a = -0.8387, b= 1.2457, c = -1.7636, d= 3.3377.

1- At x =149 is the position of the unknown band (the first unknown band in the second lane),
2- Since the number of rows in the gel image is 259
3- Then compute relative distance $x_{RD} = \frac{149}{259} = 0.5753$.
4- Calculate ylog = log of band's length, then $y_{log} = 2.5757$
5- So unknown band's length= $10^{2.5757}$ = 376.4401 ≅ 376 bp as seen in table 6.

Ruler Estimation

It's a new method applied to evaluate unknown bands between any two marker's bands respectively. For a non equal marker's sizes distances between bands according to their length in the marker, we have found the bands closer to the top are wider than those closer to the bottom

of the image. So a new ruler scale is applied between every two respective bands based on changing the scale between bands in the marker by using this equation:

Scale = $\frac{wb-wa}{da-db}$, wb > $wa, da > db$.

Where

wa, da are the length and centered of the band a respectively, and wb, db are the length and centered of the band b respectively

Let Wc is the weight of the unknown band and dc is the position of the unknown band, then the below equation is introduced to evaluate the weight of the unknown band:

$$Wc = wa + (da - dc) * scale$$

Or, we can write:

$$Wc = wb - (dc - db) * scale$$

Table 8: Application of a ruler scale between two bands respectively.

da	db	wa	wb	Scale
58	25	1000	1500	15.1515
66	58	900	1000	12.5000
77	66	800	900	9.0909
89	77	700	800	8.3333
104	89	600	700	6.6667
120	104	500	600	6.2500
143	120	400	500	4.3478
170	143	300	400	3.7037
207	170	200	300	2.7027
256	207	100	200	2.0408

Example 2:

Let the position of the unknown band dc =149. Compute the weight marker of this unknown band Wc.

Solution:

By using the scale equation :

$$scale = \frac{wb-wa}{da-db}$$

We can get the values of da, db, wa and wb from table 8,

Then da =170, db =143, wa =300 and wb = 400 .

Applying equation Scale = (400-300)/(170-143) = 3.7037.

Then, the weight marker of unknown band is found:

Wc = 300 + (170 –149)* 3.7037=377.777 378 or

$\cong 378$

Wc = 400 – (149–143)*3.7037=377.777

Table 9: Final approximated marker weight of unknown bands using ruler model estimation.

	A	B	C	D	E	F	G
1	M	lane1	lane2	lane3	lane4	lane5	
2	1500	367(Uniqu	378(Uniqu	352(Uniqu	448(Uniqu	600(Unique)	
3	1000	0	292(Uniqu	249(Uniqu	359(Uniqu	396(Unique)	
4	900	0	222(Uniqu	200(Uniqu	265(Uniqu	262(Unique)	
5	800	0	165(Uniqu	167(Uniqu	0	173(Unique)	
6	700	0	0	0	0	0	
7	600	0	0	0	0	0	
8	500	0	0	0	0	0	
9	400	0	0	0	0	0	
10	300	0	0	0	0	0	
11	200	0	0	0	0	0	
12	100	0	0	0	0	0	

Ruler estimation based on making a scale between every two marker's weight bands respectively. From table 9, ruler estimation is very important to get the type of every band that is mono, poly or unique.

Table 10: The bands converging in distance take the same values in different lanes

	A	B	C	D	E	F	G
1	M	lane1	lane2	lane3	lane4	lane5	
2	1500	368(Poly)	368(Poly)	360(Poly)	448(Uniqu	600(Unique)	
3	1000	0	292(Uniqu	249(Uniqu	360(Poly)	396(Unique)	
4	900	0	222(Uniqu	200(Uniqu	264(Poly)	264(Poly)	
5	800	0	166(Uniqu	170(Poly)	0	170(Poly)	
6	700	0	0	0	0	0	
7	600	0	0	0	0	0	
8	500	0	0	0	0	0	
9	400	0	0	0	0	0	
10	300	0	0	0	0	0	
11	200	0	0	0	0	0	
12	100	0	0	0	0	0	

From figure 7 and table 10, the software can take the same values of bands converging in distance with different lanes as similar to fitting method.

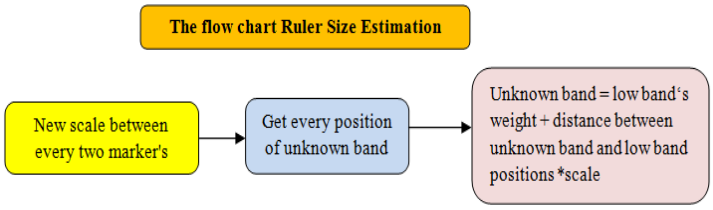
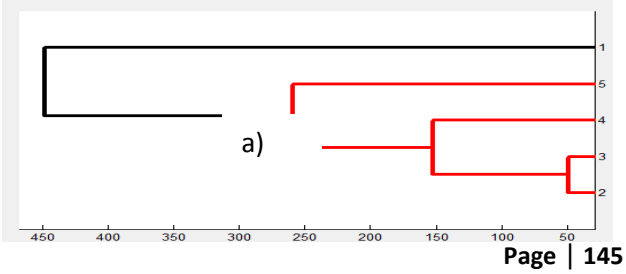


Fig. 10: Flow charts presenting the steps of Ruler Estimation

IV. Phylogenetic tree

The similarity/ dissimilarity between the DNA lanes (samples) were computed and used to generate a tree based on the number and sizes of band in each lane. This kind of estimation is very useful for molecular and population genetics studies [32].

Both the ruler and fitting could be used optionally to generate the tree with merely similar results as seen in figure 11.



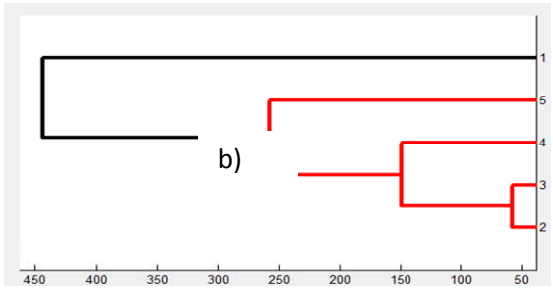


Fig. 11: Two mathematical methods to produce Phylogram tree a) Proposed Phylogram based on ruler method b) Phylogram tree based on the Fitting method

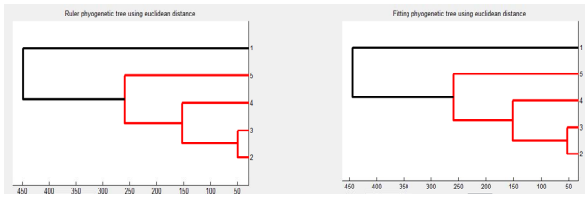
Different distance measures are integrated in the program to generate phylogram trees by pairwising distance between lanes. The choice of distance measure should be based on the application area. Calculation of the distance between two clusters is based on the pairwise distances between members of the clusters. Pairwise distance between two sets of observations (vectors)xs = (xs1, …,xsn), xt = (xt1, …, xtn) as shown in figure 12 are:

1- Euclidean distance: .

It measures absolute differences between vectors.

a)

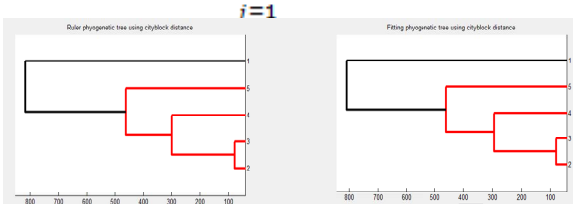
$$d_{st} = \sqrt{\sum_{i=1}^{n} (x_{si} - x_{ti})^2}$$



2- Cityblock distance (= Manhattan distance):

$$d_{st} = \sum_{i=1}^{n} |x_{sj} - x_{tj}|$$

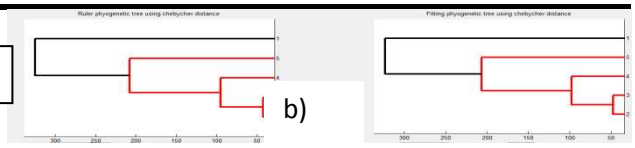
b)



3- Chebyshev distance : The distance equals to the maximum coordinate difference of the attributes. It used if the worst case must be avoided:

$$d_{\infty}(x_s, x_t) = \lim_{p \to \infty} \left(\sum_{j=1}^{n} |x_{sj} - x_{tj}|^p \right)^{1/p}$$

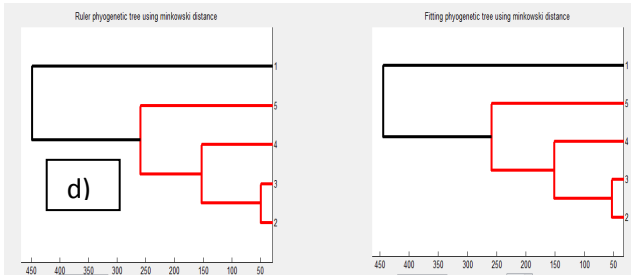
c)



4- Minkowski distance in n-dimensions: It is a generalization of Euclidean Distance

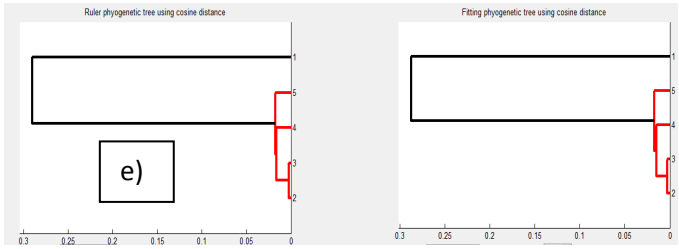
$$d_{st} = \sqrt[p]{\sum_{j=1}^{n} |x_{sj} - x_{tj}|^p}$$

p = 2 gives Euclidean distance
p = *1* gives city-block distance
p= =gives Chebyshev distance



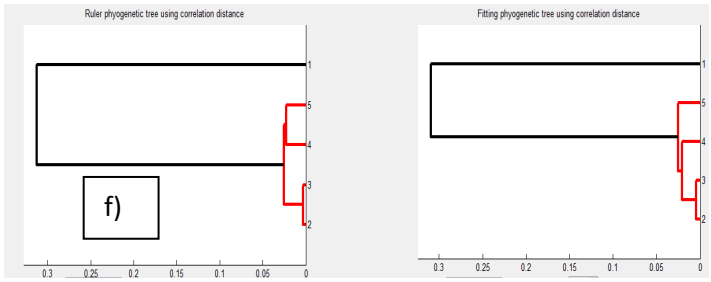
5- Cosine distance:One minus the cosine of the included angle between points (treated as vectors).

$$d_{st} = 1 - \frac{\sum_{i=1}^{n} x_{si} \times x_{ti}}{\sqrt{\sum_{i=1}^{n} (x_{si})^2 \times \sum_{i=1}^{n} (x_{ti})^2}}$$



6- Correlation distance:One minus the sample correlation between points (treated as sequences of values) . It measures trends/relative differences:

$$d_{st} = 1 - \frac{(x_s - \bar{x}_s)(x_t - \bar{x}_t)}{\sqrt{(x_s - \bar{x}_s)^2}\sqrt{(x_t - \bar{x}_t)^2}} \quad where \ \bar{x}_s = \frac{1}{n}\sum_{j} x_{si} \ and \ \bar{x}_t = \frac{1}{n}\sum_{j} x_{ti}$$



7- Spearman distance:One minus the sample Spearman's rank correlation between observations (treated as sequences of values).

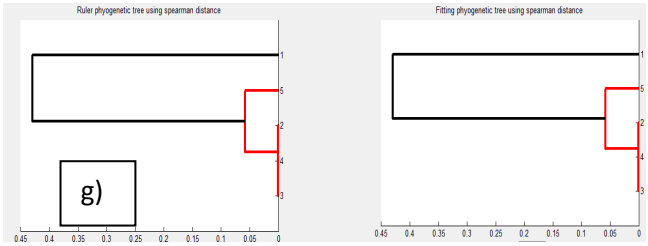
$$d_{st} = 1 - \frac{(r_s - \bar{r}_s)(r_t - \bar{r}_t)'}{\sqrt{(r_s - \bar{r}_s)(r_s - \bar{r}_s)'}\sqrt{(r_t - \bar{r}_t)(r_t - \bar{r}_t)'}}$$ where

$r_{sj}$is the rank of x_{sj} taken over $x_{s1j}, x_{s2j}, \dots x_{snj}$,

$r_{tj}$is the rank of x_{tj} taken over $x_{t1j}, x_{t2j}, \dots x_{tnj}$,

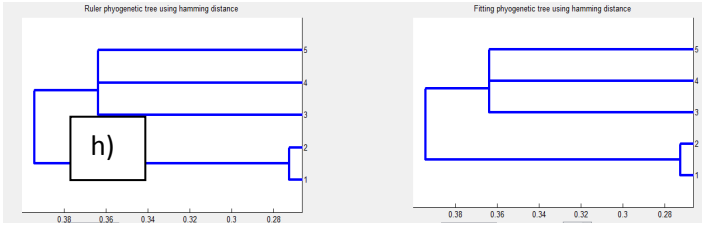
$r_s$andr_t are the coordinate wise rank vectors of $x_s$and x_t and

$$\bar{r_s} = \frac{1}{n}\sum_j r_{sj} = \frac{(n+1)}{2}, \quad \bar{r_t} = \frac{1}{n}\sum_j r_{tj} = \frac{(n+1)}{2}$$



8- Hamming distance:It is used for number of different attributes values.It is the percentage of coordinates that differ.

$$d_{st} = (\#(x_{sj} \neq x_{tj})/n)$$



9- Jaccard distance:One minus the Jaccard coefficient, it is the percentage of nonzero coordinates that differ.

$$d_{st} = \frac{\#[(x_{sj} \neq x_{tj}) \cap (x_{sj} \neq 0) \cup) x_{tj} \neq 0)]}{\#[(x_{sj} \neq 0) \cup) x_{tj} \neq 0)]}$$

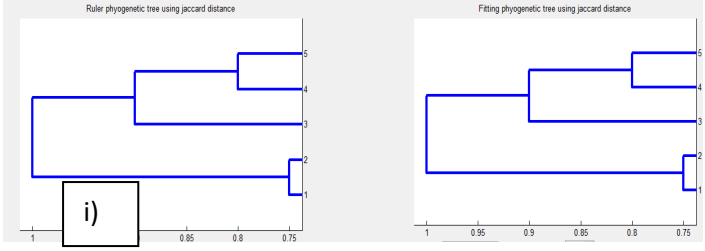
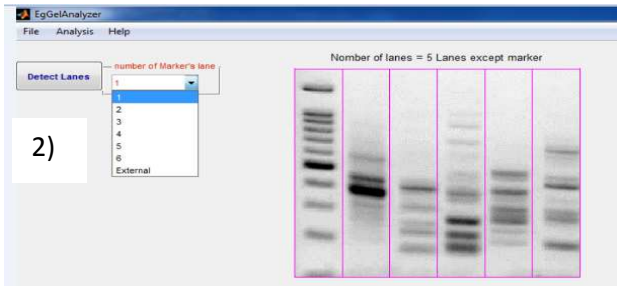


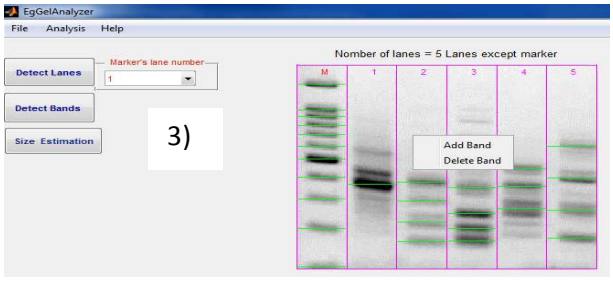
Figure 12: Different distance measures are used to generate phylogram trees. a)Euclidean distance. b) Cityblock distance. c) Chebyshev distance. d) Minkowski distance. e) Cosine distance. f) Correlation distance. g) Spearman distance. h) Hamming distance. i) Jaccard distance.

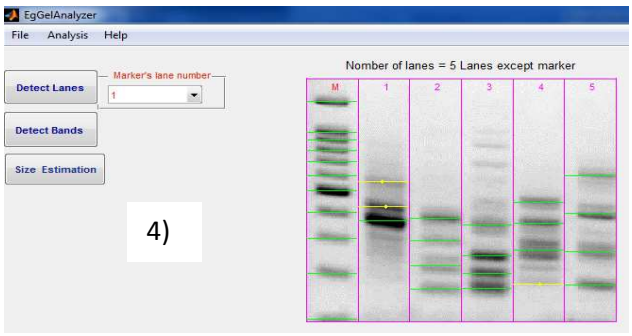
V. EXPERIMENTAL RESULTS AND EVALUATION

Databases of 30 random gel electrophoresis images have been tested for the analysis. However, only 22 images could be analyzed automatically and the other 8 GE images were poor quality, so user can detect bands manually to get the data. The following figures 13 show and example for the software results:

1- Enter the gel image file → open (extension to file jpg, jpeg, tiff, png or bmp).
2- To show number of every lane in the gel image, click : Detect lanes.
3- Select the lane of marker from the drop down list. Then, you can save final gel image.
4- Click: Detect bands, you can make right click to add or delete any band manually.
5- Click: Length Estimation to estimate the length of unknown bands
6- Write the length of the marker in table, then select from the two options fitting estimation or ruler estimation to estimate unknown band.
7- You can save the data from fitting estimation and ruler estimation in excel sheet.
8- You can save the data in zeros and ones after the work of the convergence of the distances bands in different lanes.
9- Phyogram tree can be generated







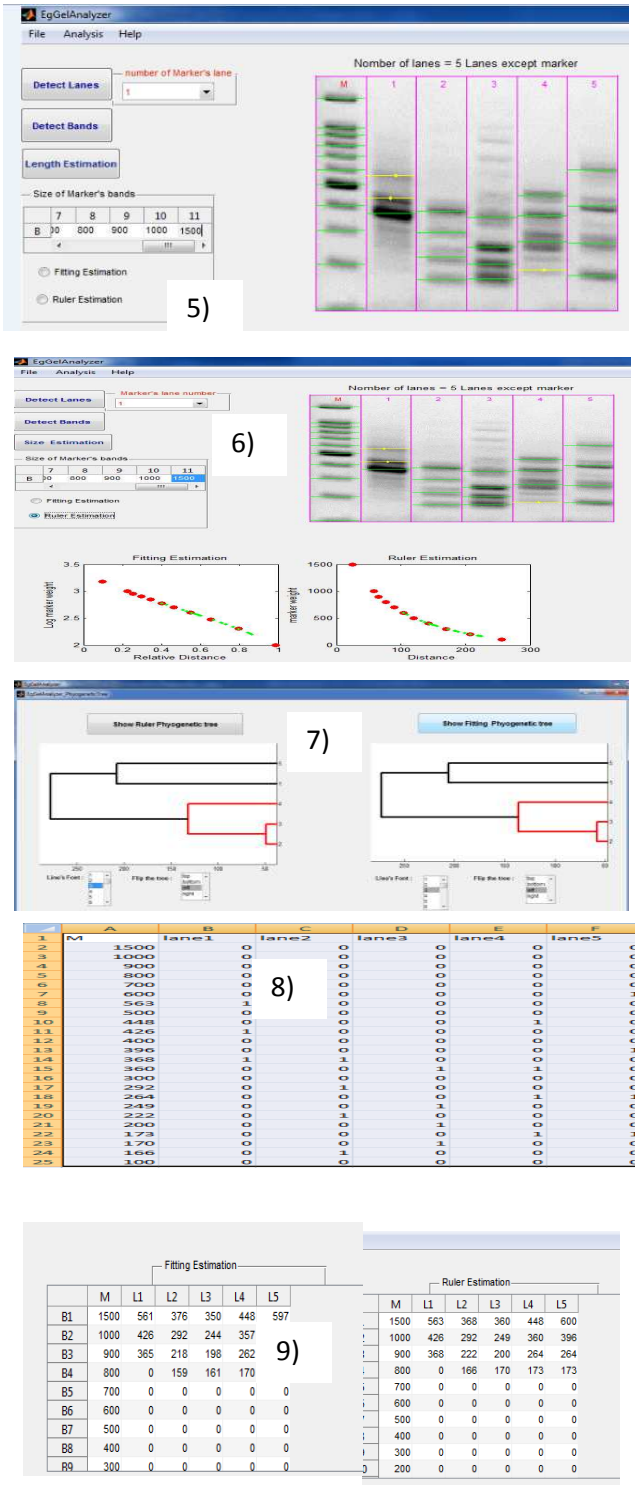


Fig. 13: All the data and images generated in each step and every step referred to in the program.

VI. CONCLUSION

In this work, new gel analyzer software with new mathematical methods were generated to facilitateand analyze DNA or Protein gel image. The new gel analyzer software extracts many biological data from an image (eg: number of lanes, detect its bands, estimate unknown band's length and determining type of the band in the lane). MATLAB GUI is presented to solve faint bands problems and enhancing the quality of gel figures. Determination of DNA molecular weight markers of unknown bands is estimated by two methods: fitting estimation and ruler estimation. The cubic model is presented a best model in compare with the linear and quadratic models to determine approximate marker weights from evaluating coefficient of determination R2, residual value and RMS values for each band. The cubic model shows no significant difference with ruler model estimation for preprocessing a given image and detecting lanes and bands automatically to approximate marker weight of unknown bands.

The program can be downloaded from the link: ftp://ftpserver.zu.edu.eg/

The future work on the current program will include regular enhancements to deal with multiple in the same time images. Moreover, adding more options to select the appropriate filter for an image and which mathematical modeling will be applied for analysis and comparison.

ACKNOWLEDGEMENTS

The Authors would like to thank the H3ABioNet project, NIH Common Fund project number U41HG006941, for supporting the developing Egyptian node and funding their scientific activities and research.

REFERENCES

[1] D.Tietz (1998)" Nucleic Acid Electrophoresis", Springer-Verlag Berlin Heidelberg New York.

[2] Ivan B., Igor H., Kornel B. (2001)," Improvement of Electrophoretic Gel ImageAnalysis", Measurement Science Review, Vol. 1, No. 1.

[3] Ashraf K. H, Ghada S. E (2012)" Semiautomatic detection of lanes and bands in DNA gel electrophoresis images ".J. Biomedical Science and Engineering, 2013, 6, 76-84 .

[4] Lin, C.Y., et al. (2002) "An automatic method to compare the lanes in Gel Electrophoresis (GE) images". IEEE Transactions on Information Technology in Biomedicine,11,179-189.

[5] " The molecular imager Gel Doc XR software" http://bti.cornell.edu/manuals/GelDox_SR_System_Manual.pdf

[6] Ferreira T and Rasband WS. (2012)"ImageJ User Guide — IJ 1.46", imagej.nih.gov/ij/docs/guide/, 2010–2012, Updated for v 1.46r, 2012, First edition: v 1.43, 2010.

[7] Ana Brânduşa Pavel and Cristian Ioan Vasile (2012) "PyElph - a software tool for gel images analysis and phylogenetics". BMC Bioinformatics, 13:9, doi:10.1186/1471-2105-13-9,

http://www.biomedcentral.com/1471-2105/13/91471-2105-13-9.

[8] Chapman & Hall, (2003),"Graphics and GUIs with MATLAB", 3'rd edition".

[9] A. B. M. Nasiruzzaman,(2010) " Matlab – Modeling, Programming and Simulations", Chapter2: user interface (GUI) software packages for educational purposes pp.17-40, Edited by Emilson Pereira Leite.ISBN 978-953-307-125-1,www.sciyo.com.

[10] Ahmed Mansour, Jaime A. Teixeira da Silva, SherifEdris, Rania A. A. Younis .(2010) Comparative assessment of genetic diversity in some tomato cultivars using IRAP, ISSR and RAPD molecular markers.Genes, genomes and genomics. (GGG_4(SI1)41-47o.

[11] Xiangyun.Ching Y. Suen, Mohamed C., Eugenia .W (1999) "A Recent Development in image Analysis of Electrophoresis Gels",pp. 432-438. Trois Rivieres, Canada.

[12] Adams, R., "Radial Decomposition of Discs and Spheres," Computer Vision, Graphics, and Image Processing: Graphical Models and Image Processing, Vol. 55, Number 5, pp. 325–332, September 1993.

[13] Jones, R., and P. Soille, "Periodic lines: Definition, cascades, and application to granulometrie," Pattern Recognition Letters, Vol. 17, pp. 1057–1063, 1996.

[14] Ankita.D, Archana.S(2013) "An Advanced Filter For Image Enhancement And Restoration" Journal Open Journal of Advanced Engineering Techniques OJAET. ISSN: 2336-0062

[15] Zohair ., Ghazali S. and Md. Gapar J. (2012)" A Comprehensive Study on Fast image Deblurring Techniques ". International Journal of Advanced Science and Technology. Vol. 44, pp. 1-10.

[16] Hansen, P., Nagy, O'Leary (2006)" Deblurring Images: Matrices, Spectra, and Filtering ", pages 130. Book News, Inc., Portland.

[17] Suresha,D., Ganesh,V.(2012) " A Survey-Mathematical Morphology operations on Images in MATLAB". International Journal of Advanced Scientific Research and Technology, PP. 2249-9954.

[18] Troy.Z, Steven.H(2005)" Automated band mapping in electrophoretic gel images using background information".Oxford Journals,Science&Mathematics,Nucleic Acids Research,Vol. 33, Issue 9, pp. 2806-2812

[19] Akbari, A.,Algregtsen, A.(2004)"Automatic Lane Detection And Separation In One Dimensional DNA Gel Images"Computational Structural And Functional Genomics.,.The Fourth International Conference on Bioinformatics of Genome Regulation and Structure(BGRS), Novosibirsk, Russia

[20] William H. Press"Computational Statistics withApplication to Bioinformatics " Unit 19:Wiener Filtering (and some Wavelets), The University of Texas at Austin, CS 395T, Spring 2008 .

[21] Anandhavalli,M., Chandan M., Ghose, M et al.(2009) "Analysis of Microarray Image Spots Intensity: A Comparative Study"International Journal of Computer Theoryand Engineering, 1793-8201.

[22] Robert Bemis(2010) "DNA MicroArray Image Processing Case Study"MATLAB® 7.

[23] Caridade, C.M.R., Marc¸ A.R.S. et al (2009)" An automatic Method to identify and extract information of DNA bands in Gel Electrophoresis Images" .31st Annual International Conference of the IEEE EMBS. Minneapolis, Minnesota, USA,

[24] Haralick, M., Linda G. S.(1992)" Computer and Robot Vision", Vol. I, Addison-Wesley, pp. 28-48.

[25] Lee S. L. (2006)"Molecular Marker Technologies", Training Workshop on Forest Biodiversity 5-16, Forest Research Institute Malaysia.

[26] Daniel M. and Stephen T. (2007) "Engineering a Program to Digitally Analyze Genetic Tests for HPV" ,Pages 1-18. Project & paper, ISEF Team Finalist.

[27] Ford E.B. (1965). Genetic polymorphism.Faber & Faber, London.

[28] Nouh.E.A.(2008) (GelAnalyzer 3: The first Arabic Bioinformatic software for Gel analysis) Jornal of Cell and molecular Biology.

[29] Istvan L. "GelAnalyzer 2010 User's manual"

[30] Yang W. Y. , Cao W, Chung T , Morris J, et al.(2005) "Applied numerical methods using Matlab". Hoboken, New Jersey: John Wiley & Sons.

[31] Larry W. (2003) "Advanced Statistical Techniques", Chapter 11 – Simple linear regression.

[32] GAO W., LI S., HONG D., DENG C., LU L.(2007) " Phylogenetic Classification As Revealed Based On Optimize ISSR-PCR System In The Osmanthus", Analele Ştiinţifice ale Universităţii „Alexandru Ioan Cuza", Secţiunea Genetică şi Biologie Moleculară, TOM VIII, Genetics and Molecular Biology.