IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 2, 2019, accepted December 22, 2019, date of publication January 3, 2020, date of current version January 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2962319

EGBIOIMAGE: A Software Tool for Gel Images Analysis and Hierarchical Clustering

MAHMOUD H. ALNAMOLY[1], AHMED MANSOUR ALZOHAIRY[2], AND IBRAHIM MAHMOUD EL-HENAWY[1]

[1]Computer Science Department, Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Egypt
[2]Genetics Department, Faculty of agriculture, Zagazig University, Zagazig 44519, Egypt

Corresponding author: Mahmoud H. Alnamoly (mahmoudalnamoly2014@gmail.com)

ABSTRACT This paper presents a novel software called EGBioImage that is implemented for analyzing gel Electrophoresis images, computing the molecular weights of the bands of the unknown size, detecting the types of the bands, grouping the bands according to their molecular weights and labeled each band with its group number in the gel image and generating the corresponding phylogenetic tree using the data extracted from the analyzed gel image. Students and Researchers working in molecular biology and genetics would benefit greatly from EGBioImage. This software is designed and developed under windows operating system using a C-sharp programming language. It uses Emgu CV platform to extract contours objects from an image that uses the green theorem to detect the bands in each detected lane. It uses polynomial regression to calculate the molecular weights of the bands of unknown size. It uses K-means clustering algorithm to cluster bands according to their molecular weights. It uses each of upgma algorithm and the matching bands amongst lanes and to construct the phylogenetic tree. EGBioImage provides a very friendly Graphical User Interface that guides the user from the step of uploading a gel image toward getting the molecular weights of the bands of unknown size and generating the Phylogenetic Tree. Moreover, EGBioImage provides the user with the ability of processing the gel image using a completely manual processing and automatic processing with the ability of manual correction of lanes and bands ''Semi-automatic process'' and it is more accurate than comparable software in some respects and the only software that divides the bands into groups and labeled them with numbers in the gel image based on their group by implementing k-means algorithm and these claims are supported with experiments. EGBioImage is suitable for students and researchers who do not have access to commercial software.

INDEX TERMS Gel electrophoresis, image processing, phylogenetic tree, band matching, polynomial regression, UPGMA algorithm, k-means algorithm, software.

I. INTRODUCTION

Gel Electrophoresis (GE) is an important technique which widely used in the molecular biologist experiments to separate the DNA based on their size or weights and generating the DNA and protein gel images [1]. The gel image which produced from the gel electrophoresis on the (issr-pcr, rapd-pcr, and sds-pages) experiments is one of the most important and essential sources of information for any molecular biologists as they draw their conclusions based on the results getting from analyzing the gel image. Each generated electrophoresis gel image consists of vertical tracks called lanes which represent a DNA or Protein sample and horizontal fragments called bands which are loaded into the gel from something called wells from cathode (−) to anode (+), from top to bottom based on its weights as shown in Fig. 1.

Bands on the upper have a larger weight than ones on the bottom. The larger segments are run slower than smaller ones and staying in the upper position of the lanes so that bands on the upper of Lane is the larger, has the biggest weights and the smaller fragments migrate faster through the gel and occupy the lower position of the lanes. The separation process of the polymorphic bands is done based on the size of the DNA fragments which running form the negative cathode toward the positive anode. The smaller fragments of DNA migrate faster through the gel and occupy the lower position of the gel and The fragments of DNA with larger size will be appear on the top of the gel as it blocked from passing through the

The associate editor coordinating the review of this manuscript and approving it for publication was Navanietha Krishnaraj Krishnaraj Rathinam.

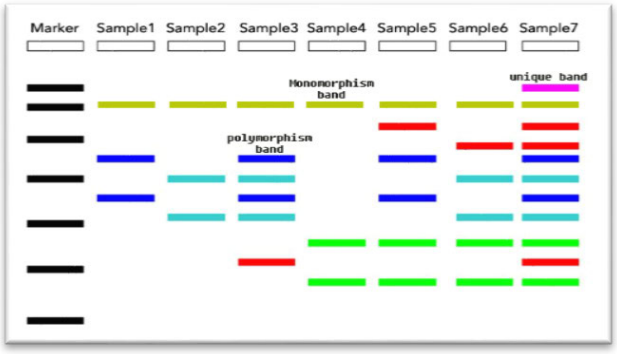


FIGURE 1. Gel image which is generated from the gel electrophoresis.

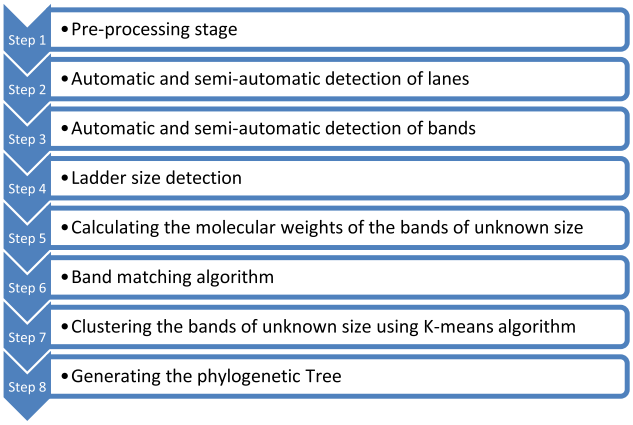


FIGURE 2. The core eight steps of EGBioImage software.

FIGURE 3. Gel image enhancement.

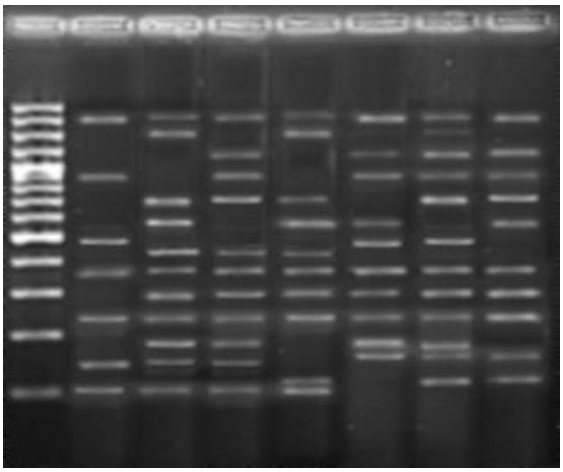


FIGURE 4. Original gel image.

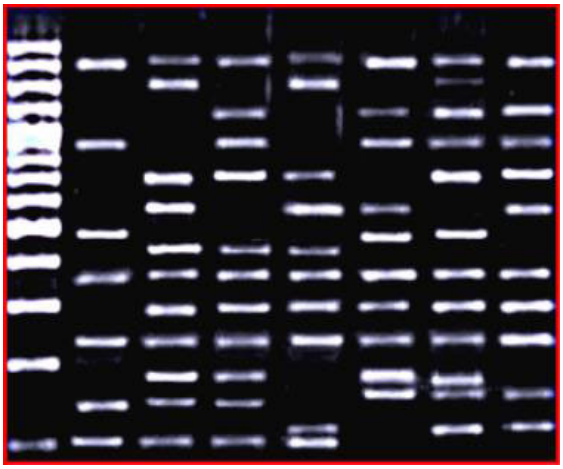


FIGURE 5. Enhanced gel image.

small pores on the gel, for this reason, the DNA fragments with large size move less slowly than other DNA fragments with small size. Gel Electrophoresis has many different applications in molecular biology, forensics, microbiology, genetic and biochemistry. The process of DNA restriction is done in three steps, running segments on the gel, and analyzing the generated map, these two steps are done inside a laboratory but the third step used to be done by the human eye, but it consumes a lot of time and a lot of errors may be occurred by using only the human eye. Therefore, after the request for gel electrophoresis increased, computer science was used to speed up the processing, analysis, and clustering of samples based on the gel images.

With respect to above, some researchers have endeavored to develop their own non-commercial and simple tools for this aim such as GelAnalyzer [24], a java based but not open source product and doesn't generate a phylogenetic tree, GelClust [2], a c-sharp based product but doesn't display the molecular weights of bands of unknown size to the user, PyElph [5] which is an open-source product but the user does not have any privilege to add or delete any band manually and developed for educational uses and is not so accurate in the process of detection lanes and bands.

Having said all these, a need to accurate user-friendly software that provides most features the user needed is obvious. The main contribution of this paper is developing a novel open access software that will address some of the issues of the existing tools and provide other features that are not found in some other tools such as:

- Estimating, displaying, and saving the molecular weights of bands of unknown size.
- Clustering the bands into groups using the k-means algorithm.
- Fingerprint comparison.
- Detecting the matching bands between lanes.
- Detecting the number of primer dimmer bands in each lane.
- Saving and printing a report for the experiment.
- Inserting names of the samples by the user.

TABLE 1. Features & abilities of gel electrophoresis images analyzer software.

Software name	Feature code			
	c1	c2	c3	c4
EzQuant	×	Win	2005	×
Dolphin 1D	×	Win	2006	×
GelScan	×	Win	2007	✓
Laneruler	✓	All	2007	×
Quantity One	×	Win, Mac	2008	✓
GelAnalyzer	✓	All	2010	×
gelQuest	×	Win	2010	✓
Jelmarker	×	Win, Mac	2010	✓
BioDocAnalyze	×	Win	2011	✓
Gel-Pro Analyzer	×	Win	2011	×
GelQuant	✓	Win	2011	×
ImageQuant	×	Win	2011	✓
Intelligent Quantifier	×	Win, Mac	2011	×
Advanced Quantifier	×	Win, Mac	2012	✓
Molecular Imaging	×	Win, Mac	2012	×
myImage Analysis	×	Win	2012	×
Gelclust	✓	Win	2013	✓
GelComparII	×	Win	2013	✓
GelQuant Pro	×	Win	2013	✓
GeneTools	×	Win, Mac	2013	✓
Phoretix 1D Pro	×	Win	2013	✓
PyElph	✓	All	2013	✓
TotalLab	×	Win	2013	✓
Un-Scan-it	×	Win, Mac	2013	×
ImageJ	✓	All	2014	×
Gel-Quant	×	Win	2014	×
Image Lab	×	Win, Mac	2014	×
ImageStudio	×	Win, Mac	2014	×
LabImage	×	All	2014	×
Ultraquant	×	Win	2014	×
VisionWorks	×	Win	2014	✓

c1:- Free
c2:- OS
c3:- Year "Last release"
c4:- Fingerprint comparison

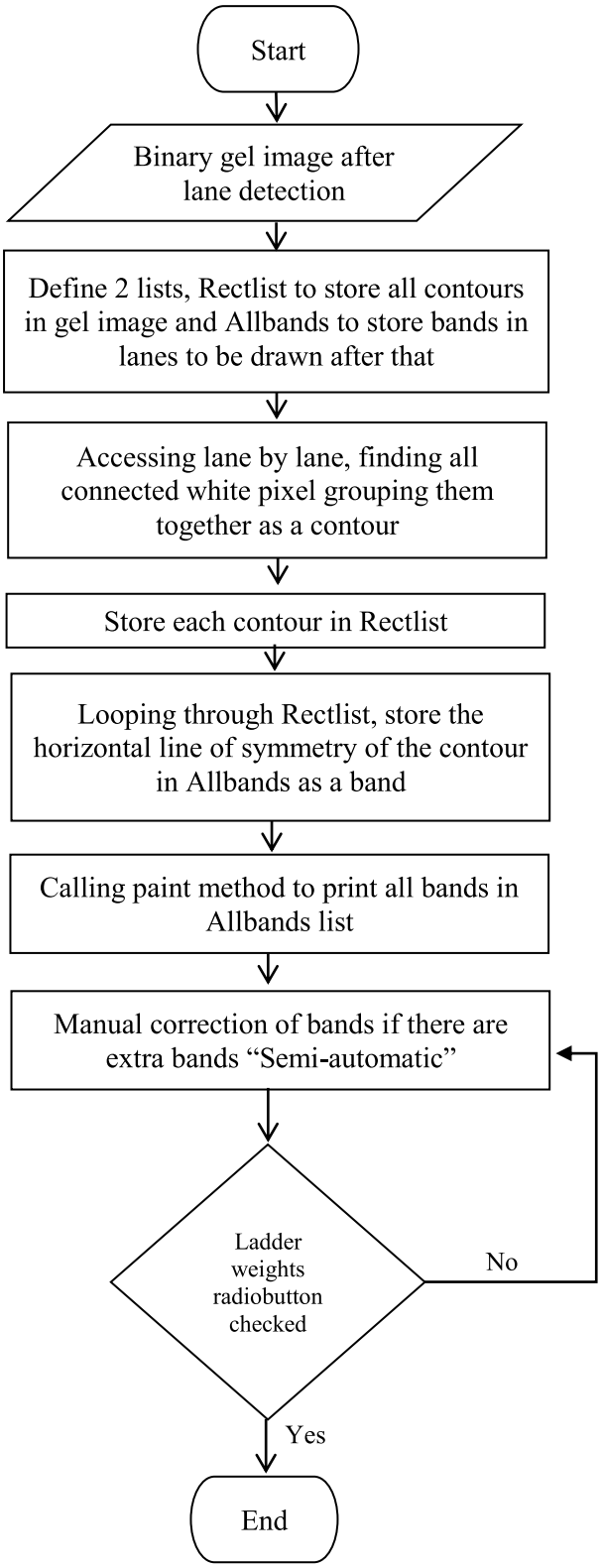


FIGURE 6. Flowchart of automatic and semi-automatic band detection.

- The ability to do experiment as a manual process or automatic process with the ability of manual correction of lanes and bands.

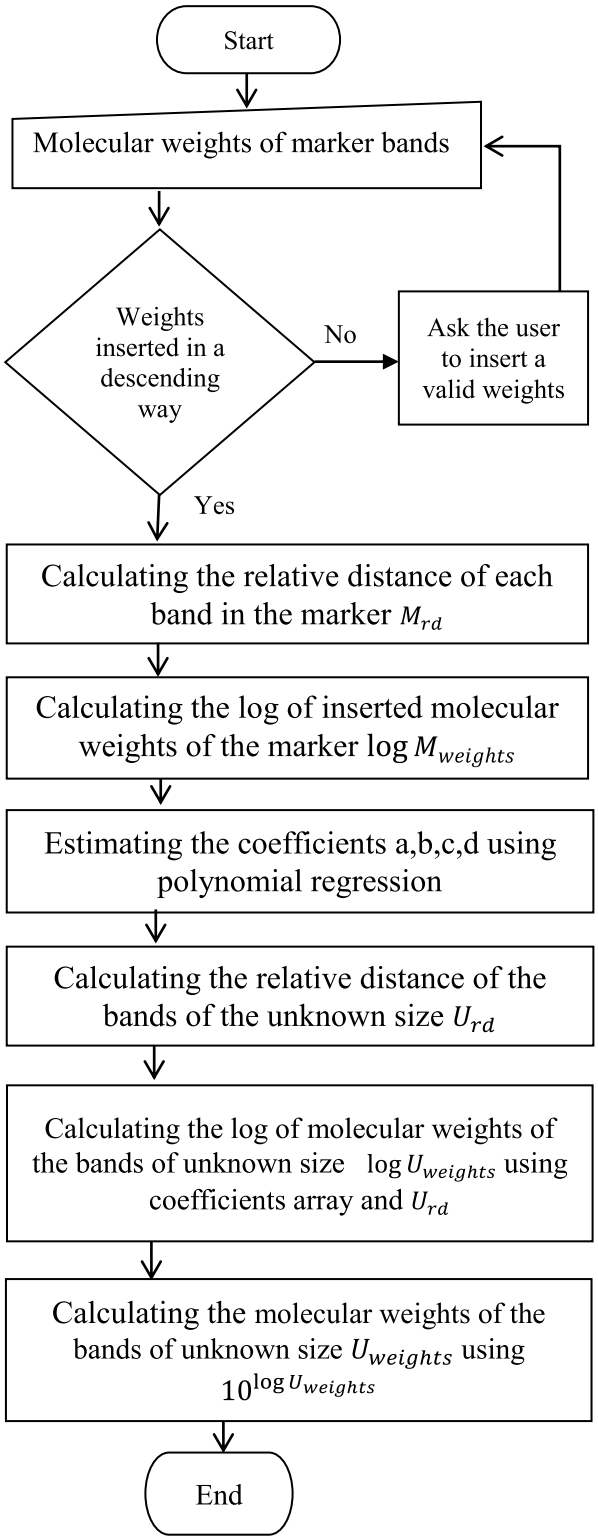


FIGURE 7. Flowchart of calculating the molecular weights of the bands of unknown size.

This research is organized as follows. Section II is the related work that presents all the current gel images analyzer software tools. Section III is the methodology of the proposed software tool. Section IV is the results and discussion, in this section, the capabilities of this proposed software tool are compared with some other software tools and presenting the proposed software tool in the runtime. Section V is the conclusions and further research that illustrate the conclusions based on the results and future work.

II. RELATED WORK

There are some available software which is used to analysis the gel images such as PyElph, GelJ, GelClust, GelAnalyzer, GELect, QIAxcel System, BioDocAnalyze, Gel Doc EZ, Gelquest & ClusterVis, Phoretix 1D, GelScan and Quantity One [2], [3] but they are not free and most other free software are very complex for user and don't give many options. For example, QIAxcel System [4], has the ability to quantize the images but it cannot cluster the samples, it cannot generate the dendrogram. PyElph [5] is a free software to analyze DNA gel images and cluster the results but the user does not have any privilege to add or delete any band manually and developed for educational uses and is not so accurate in the process of detection lanes and bands. ImageJ [6], [7] is a free open source software in which the process of lanes and bands detection is done manually but it doesn't provide the option of phylogenetic tree or dendrogram generation.

Gelquest & ClusterVis [8] has the ability to generate the phylogenetic tree of the samples using unweighted pair group method with Arithmetic Mean (UPGMA) by comparing the similarity between samples but however it is easy to work with GelQuest, most of the processes should be executed manually and this decreases the agility. GelScan, Phoretix 1D and Quantity are three commercial software which however they have the ability to make manual and automatic detection and generate the dendrogram, but the core problem of them is their high price, and the user must be exercised enough to be able to deal with them [9]. All available gel electrophoresis image analyzer tools [10] up to now are shown in Table 1 ordered by the last release of the gel image analyzer tool.

Finally and from Table 1 we observed that there are some free tools or gel image analyzer software but also don't support dendrogram generation such as Laneruler, GelAnalyzer, GelQuant and ImageJ which is very important for the biologist and on the other side there are other tools which free and generate dendrogram such as Gelclust and PyElph but as we mentioned earlier, in PyElph tool the user does not have any privilege to add or delete any band manually and it is developed for educational uses and is not so accurate in the process of lanes and bands detection so that we are in need for more accurate user-friendly software that facilitates the user work and reduces the need for intervention of the user in the gel image analysis process.

In the next sections, EGBioImage is presented and its common and unique capabilities are discussed minutely. EGBioImage will address some of the issues of the current tools in Table 1.

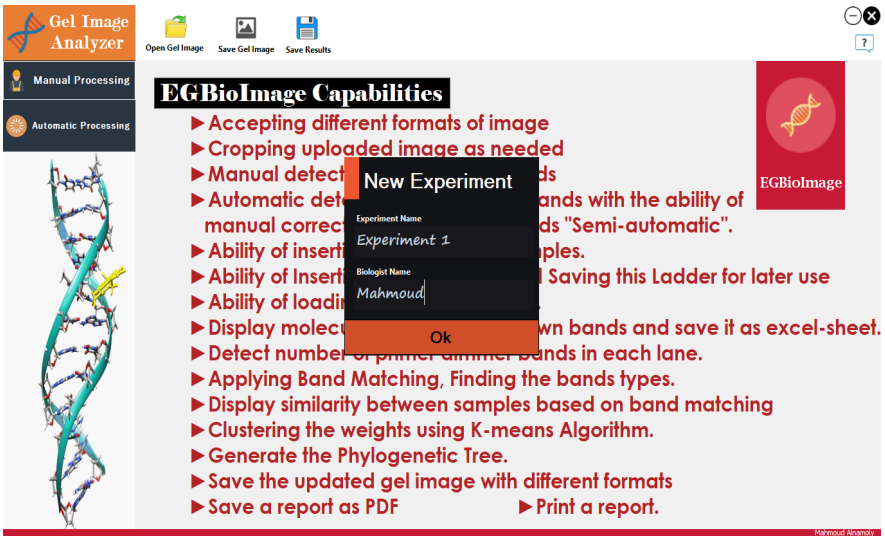


FIGURE 8. Start a new experiment.

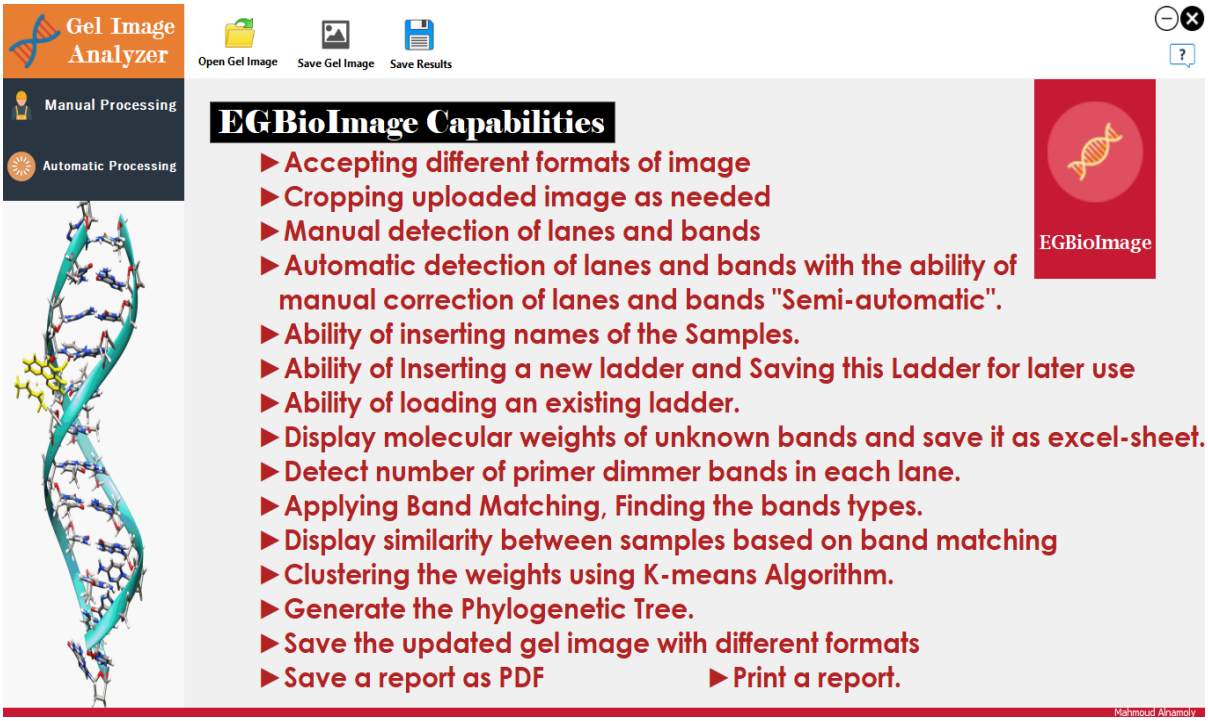


FIGURE 9. EGBioImage homepage.

III. METHODOLOGY

EGBioImage processes the uploaded gel image in eight steps as shown in Fig. 2. The software contains three panels. The top panel contains three buttons for uploading the gel image, saving gel image after the detection process and saving the molecular weights of bands of unknown size. The main panel that allows the user to insert, view and edit. A left-side panel that contains two buttons, one for the fully manual processing and the second for automatic processing. First of all, the user has the ability to process the gel image in two ways by choosing the type of process from the left-side panel, manual process or automatic process. After that, the eight steps are followed during processing the gel image. First step for uploading gel image, cropping and enhancement, two steps for lanes and bands detection, two steps for calculating the molecular weights of bands of unknown size, one step for detecting the type of band by determining the matching bands in the samples, one step for clustering the molecular weights

TABLE 2. Comparing EGBioImage with other similar non-commercial software.

Software name	Capability code										
	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11
EGBioImage	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓
GelClust	✓	✓	×	×	×	×	×	✓	✓	×	×
GelAnalyzer	✓	×	×	✓	×	×	×	×	×	×	✓
PyElph	×	×	×	×	✓	×	×	✓	✓	×	✓

c1:- Accuracy of column detection
c2:- Accuracy of band detection
c3:- Insert names of the samples by the user
c4:- Display the molecular weights of bands of unknown size to the user and saving them as a spread sheet.
c5:- Band matching algorithm
c6:- Detect the number of primer dimmer bands in each lane
c7:- Clustering bands using the k-means algorithm
c8:- Fingerprint comparison
c9:- Variety in clustering methods & similarity coefficient
c10:- Save a report as pdf & Print a report
c11:- Save molecular weights of the marker as a .txt file for later use and export an existing ladder

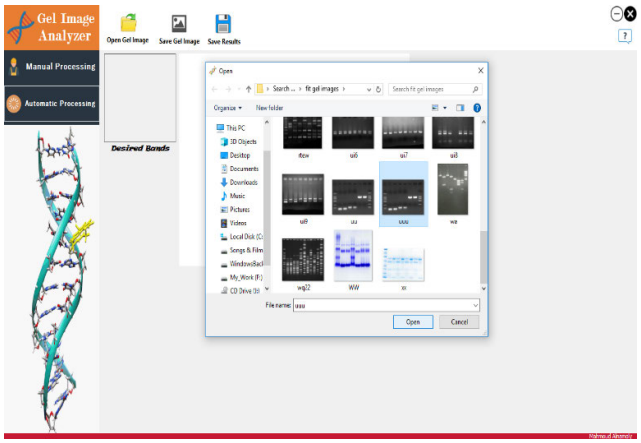


FIGURE 10. Upload a gel image.

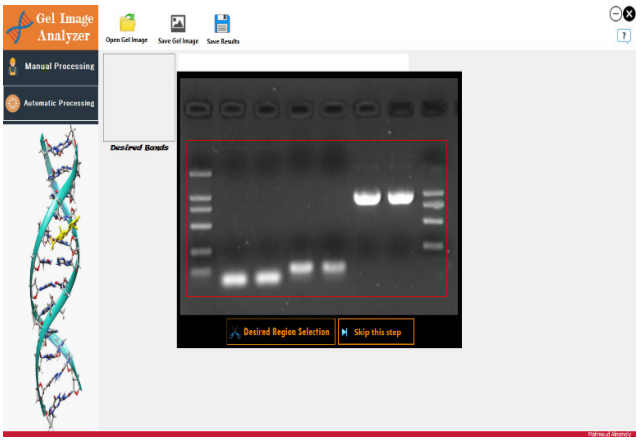


FIGURE 11. Selecting the region of interest.

of bands of unknown size using K-means algorithm, the last step for drawing the corresponding phylogenetic tree. In the next sections, we discuss these eight steps minutely.

A. PRE-PROCESSING STAGE

This stage starts with uploading the original gel image and selecting the region of interest from the gel image then enhancing the gel image. There are many factors that affect the image quality such as (agarose type, the buffer chamber temperature, voltage, time, reorientation angle, field strength, etc. . . .) which could affect the accuracy of extracting the right information from the gel image, so that the enhancement operation is a very important step to enhance the equality of the gel image. The workflow of gel image enhancement operation as shown in Fig. 3 is first, convert the RGB gel image to a gray scale image by accessing all the pixel of RGB image and get the average of the three values of pixel color red, green and blue then change the pixel color by the new average values of RGB. Then complement the grayscale gel image as the grayscale image complement operation is very useful for enhancing the brightness visibility of exact differences between gray levels where the fine details are obscured. After that applying some filters on the complemented grayscale gel image such as HistogramEqualization filter [11], ColorRemapping filter, ContrastCorrection filter, BrightnessCorrection filter, and GaussianSharpen filter. Finally applying gamma correction [12] which used to encode and decode luminance values in image systems. The enhanced gel image is shown in Fig. 5 after doing this preprocessing step on the gel image shown in Fig. 4.

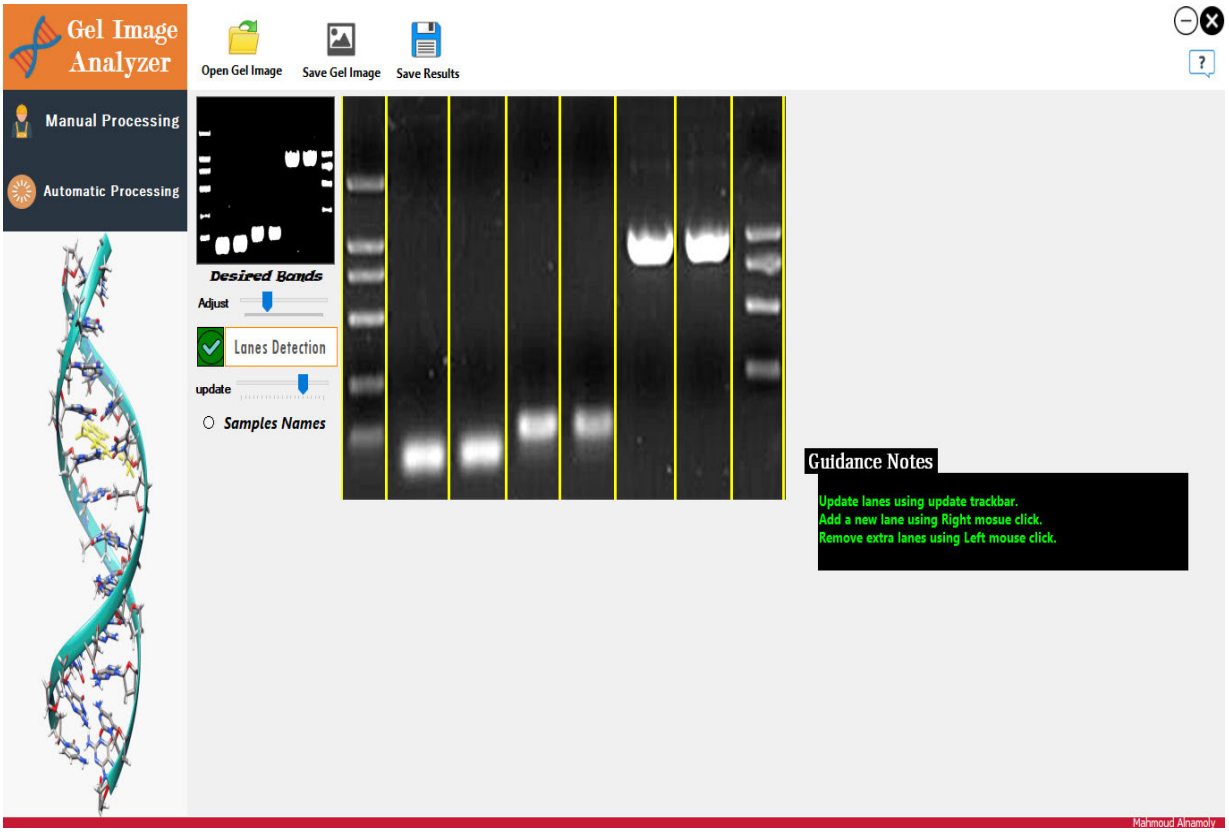


FIGURE 12. Detection of lanes.

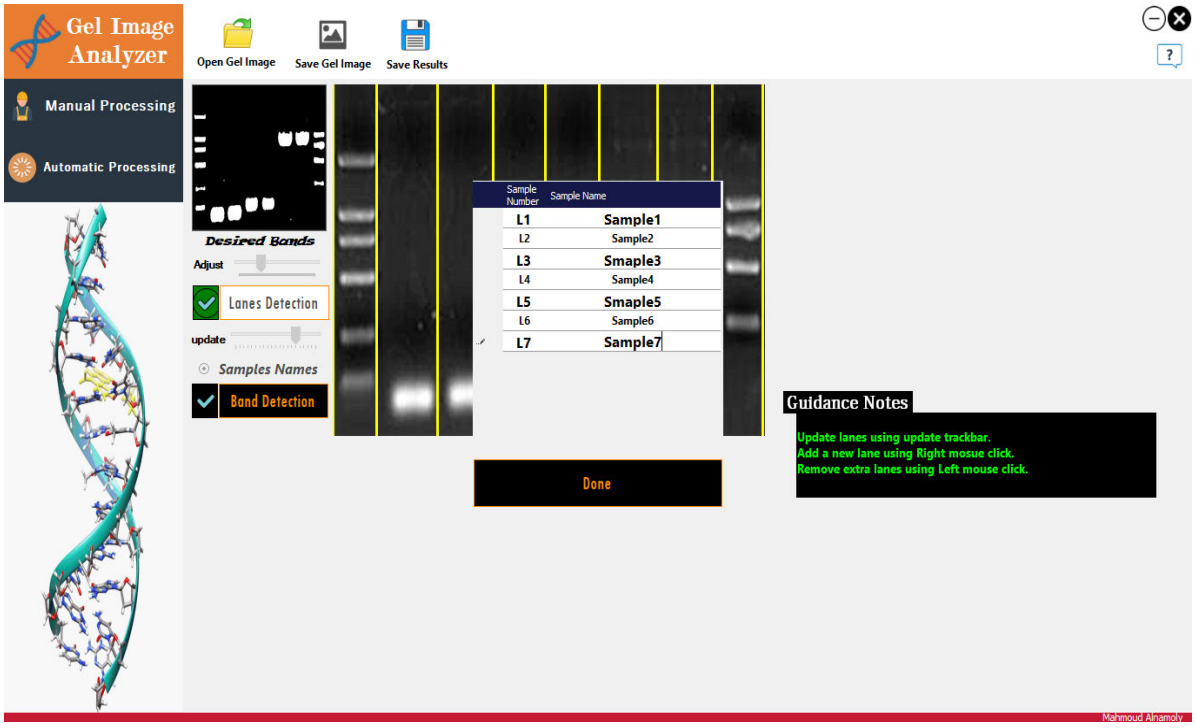


FIGURE 13. Inserting names of lanes.

B. AUTOMATIC PROCESSING

Based on the intensity of pixel, EGBioImage detects lanes and bands. The first step is the image binarization. EGBioImage converts the pixel image to a binary image in which each pixel will take only one value 0 for black pixel and 1 for white pixel based on specific threshold that is controlled

Algorithm 1 Automatic Lane Detection

```
Define "Lane" line object,"AllLanes" list, sum,
distBlanes, laneNum and binaryimg
for i=0 to binaryimg.width-1
  Sum=1
  for j=0 to binaryimg.height-1
    Color cx = binaryimg.GetPixel(i, j)
    if cx = Color.FromArgb(0, 0, 0)
      sum++
    else
      break
    end
  end for
  if sum = binaryimg.Height
    Lane = new Lines()
    Lane.startPoint.X = i+1
    Lane.startPoint.Y = 0
    Lane.endPoint.X = i+1
    Lane.endPoint.Y = binaryimg.Height
    if Lane.startPoint.X-
        AllLanes[laneNum]. startPoint.X >
                distBlanes
      AllLanes.Add(Lane)
      laneNum ++
    end if
  end if
 end for
for each lane in AllLanes
  Calling paint method that used to draw on
      the image.
  Draw lane over the gel image based on the
      startpoint and endpoint of them.
end for
```

by the user through a track bar in which if the intensity value of the pixel is smaller than the threshold value, convert the current pixel to black pixel with value 0 but if the intensity value of the pixel is larger than the threshold value, convert the current pixel to white pixel with value 1. In the next section, we will discuss automatic and semi-automatic lane and band detection after converting the gel image to a binary image.

1) AUTOMATIC AND SEMI-AUTOMATIC DETECTION OF LANES

After converting the gel image to binary image, passing this binary image "binaryimg" to algorithm 1 to start the process of lane detection. In algorithm 1 using Lane object to store each lane location, AllLanes list to store all Lane objects that will be drawn, sum variable as counter increased in case of the current pixel is black, and distBlanes variable expresses the distance between lanes to reduce the number of the drawn lane after that. After the process of automatic lanes detection, the user has the ability of manual correction of lanes by using a track bar and mouse-click for adding or remove any lane. Finally, after finishing the process of lanes detection, the user will insert the names of the samples.

2) AUTOMATIC AND SEMI-AUTOMATIC DETECTION OF BANDS

After finishing automatic lane detection and do a manual correction of lanes "adding a new lane or removing extra lanes, the software detects the bands in each lane automatically by processing each lane row by row, finding the connected white pixel in each lane grouping them together as a contour using green theorem [13] and image moments [14] which are used in Emgu CV platform to extract contours objects from an image. The user has the ability of manual correction of bands as needed by using left mouse-click to remove the extra band and right mouse-click to add a new band. The process of automatic detection of bands is shown in Fig. 6. After finishing the process of automatic and semi-automatic detection of bands, the user inserts the marker weights "ladder" in a specific order as described in the following section.

C. LADDER SIZE DETECTION

The term ladder comes from the appearance of DNA fragments on the maker lane which looks like rungs of a ladder. The molecular weights of the marker are inserted in a descending way as the fragment or band on the top of the gel image has the largest intensity compared to the other bands on the marker as we mentioned before. The software gives the user the option for using a new ladder by inserting the weights in data gridview or uploading existing molecular weights. The user inserts the molecular weights of the marker into a data gridview then the software tests them if they are not in the correct form, it pops up a message box for the user telling him that he must insert valid weights with a descending way then try again.The user can save the current inserted weights for later use.

D. CALCULATING THE MOLECULAR WEIGHTS OF THE BANDS OF UNKNOWN SIZE

EGBioImage uses the molecular weights of the marker to calculate the molecular weights of bands of unknown size using polynomial regression. Polynomial regression is one of the most important used approaches of regression analysis to model a not linear relation between two vectors or arrays X and Y. By using the approach of ordinary least squares, polynomial regression models are usually fit. The ordinary least squares is one of the methods that is used to estimate the unknown coefficients or parameters of the equation by minimizing the distinction between the unbiased estimator of the parameters or coefficients and all of this happens using the gauss markov theorem conditions. The best fitting data on the model is done by using the smaller distinction or differences between unbiased estimator of the parameters or coefficients calculated by the ordinary least squares. The sequence of calculating the molecular weights of bands of unknown size as shown in Fig. 7 is, first preparing the two arrays x and y that are used by polynomial regression. The first array is

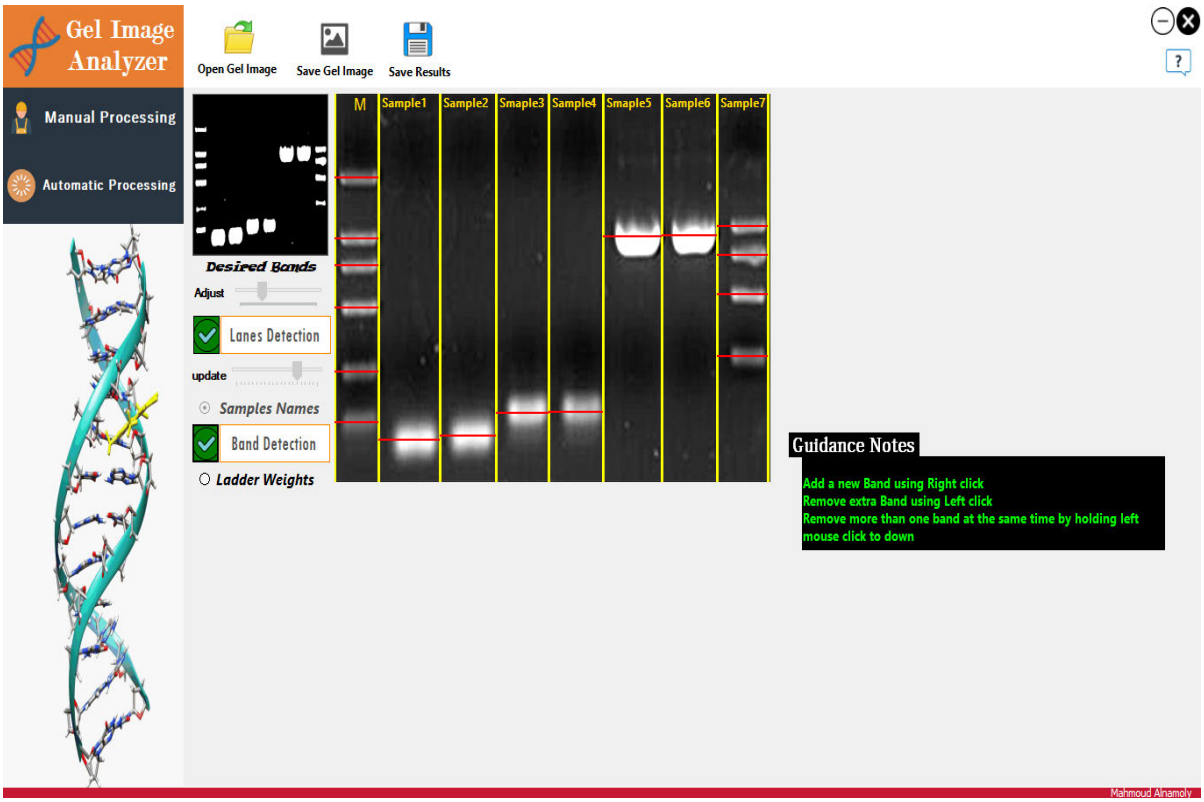


FIGURE 14. Detection of bands.

calculated using equation 1 and the second array is the common logarithm of the molecular weights of the marker.

$$M_{rd} = \frac{position\ of\ band\ on\ y - axis}{Gel\ image\ height} \quad (1)$$

where M_{rd} is the relative distance of each band in the marker and it is considered as x array.

These two arrays are used in equation 2 to estimate the best coefficients a, b, c, d using the polynomial regression.

$$\log M_{weights} = a\,(M_{rd})^3 + b\,(M_{rd})^2 + c\,(M_{rd}) + d \quad (2)$$

where log $M_{weights}$ is the common logarithm of the molecular weights of the marker and it is considered as y array.

After calculating the best coefficients, the software calculates the relative distance of the bands of unknown size U_{rd} in the same way of calculating M_{rd} in equation 1. Finally, nonlinear regression is used to calculate the common logarithm of molecular weights of the bands of unknown size as mentioned in equation 3. Then using equation 4 to get the molecular weights of the bands of unknown size $U_{weights}$.

$$\log U_{weights} = a\,(U_{rd})^3 + b\,(U_{rd})^2 + c\,(U_{rd}) + d \quad (3)$$

where log $U_{weights}$ is the common logarithm of the molecular weights of the bands of unknown size.

$$U_{weights} = 10^{\log U_{weights}} \quad (4)$$

where $U_{weights}$ is the molecular weights of the bands of unknown size.

E. BAND MATCHING ALGORITHM

After calculating the molecular weights of bands of unknown size, EGBioImage implements the band matching algorithm by finding the matching band in each lane to find the type of band. There are three types of band unique band, polymorphism band, and monomorphism band. If the band is founded in only one lane, it means that the type of this band is unique and if the band is founded in more than one lane, it means that the type of this band is polymorphism but if the band is founded in all the lanes, it means that the type of this band is monomorphism [15], [16]. Unique bands are marked by a "unique" string in the gel image. The software connects the matching bands in lanes by a line drawn in the gel image. Two bands in different lanes are matched even if their weights are not equal but close to each other, the difference between their molecular weights can be determined as a tolerance value. This tolerance value can be fixed or determined by the user, the software provides the user with the ability to determine the tolerance value. This functionality of enabling the user to determine the tolerance value is provided by 16 of the 25 software which supports band matching.

After finishing this step, the software provides the user with the ability to find the count of primer dimmer bands in each lane. The user determines the threshold of the primer dimmer band in a textbox then all bands under this value will be detected as primer dimmer band.

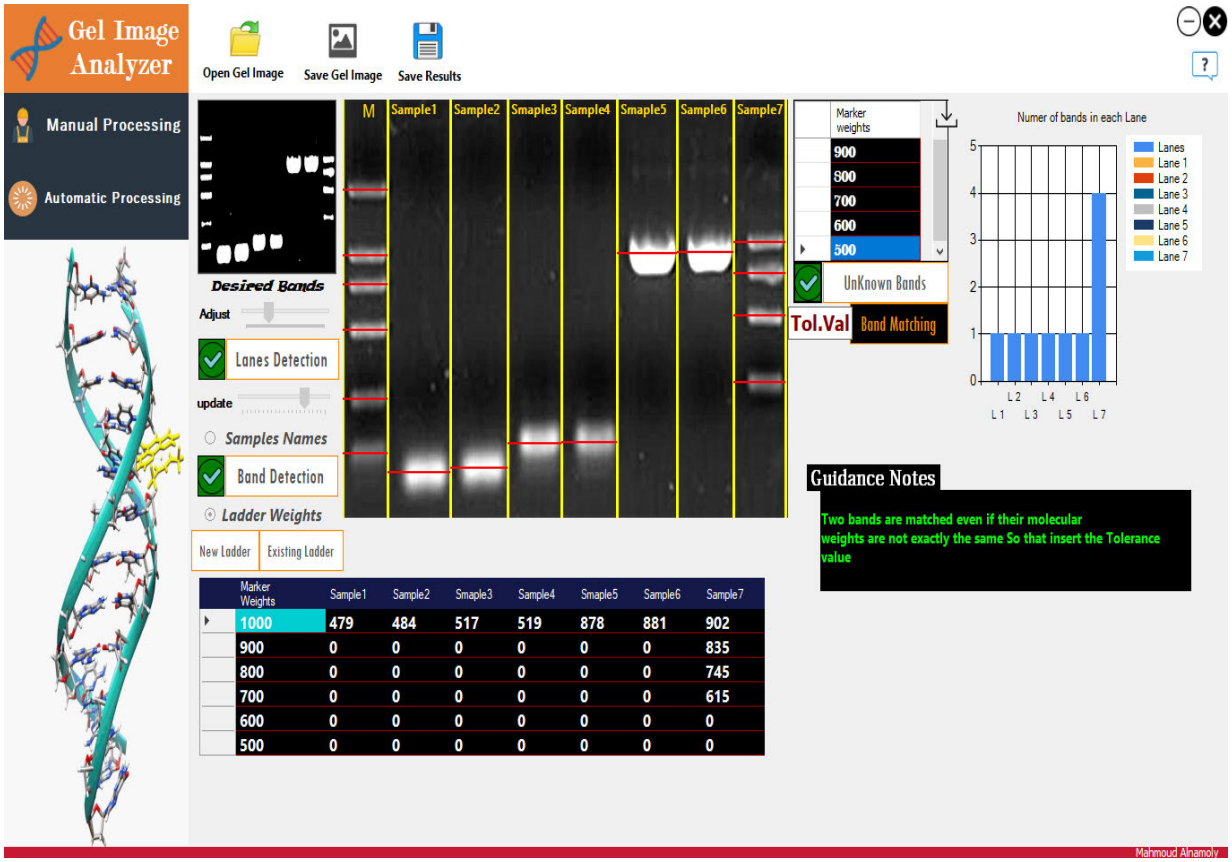


FIGURE 15. Inserting the molecular weights of the marker and display the molecular weights of the bands of unknown size.

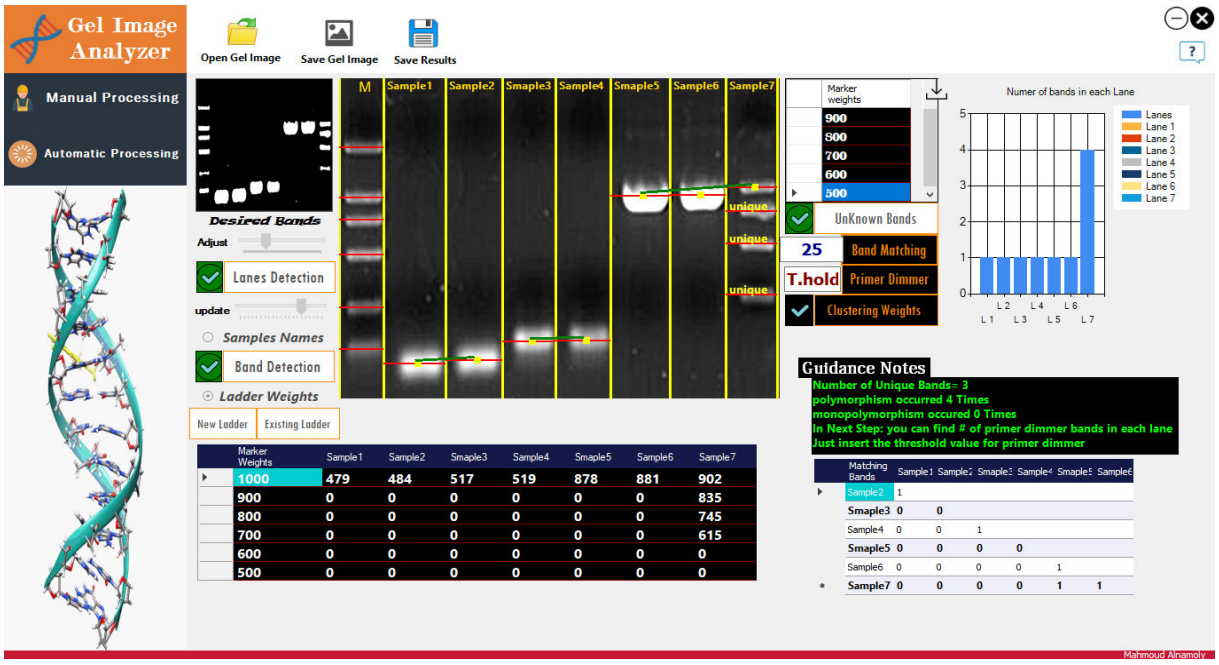


FIGURE 16. Detecting the matching bands between the lanes.

F. CLUSTERING THE BANDS OF UNKNOWN SIZE USING K-MEANS CLUSTERING ALGORITHM

The software divides the bands into groups according to their molecular weights by implementing the k-means clustering algorithm. K-means clustering algorithm [17] is one of the algorithms of machine learning, it is one of the simplest clustering algorithm and the most used clustering algorithm as it is a computationally faster than the other

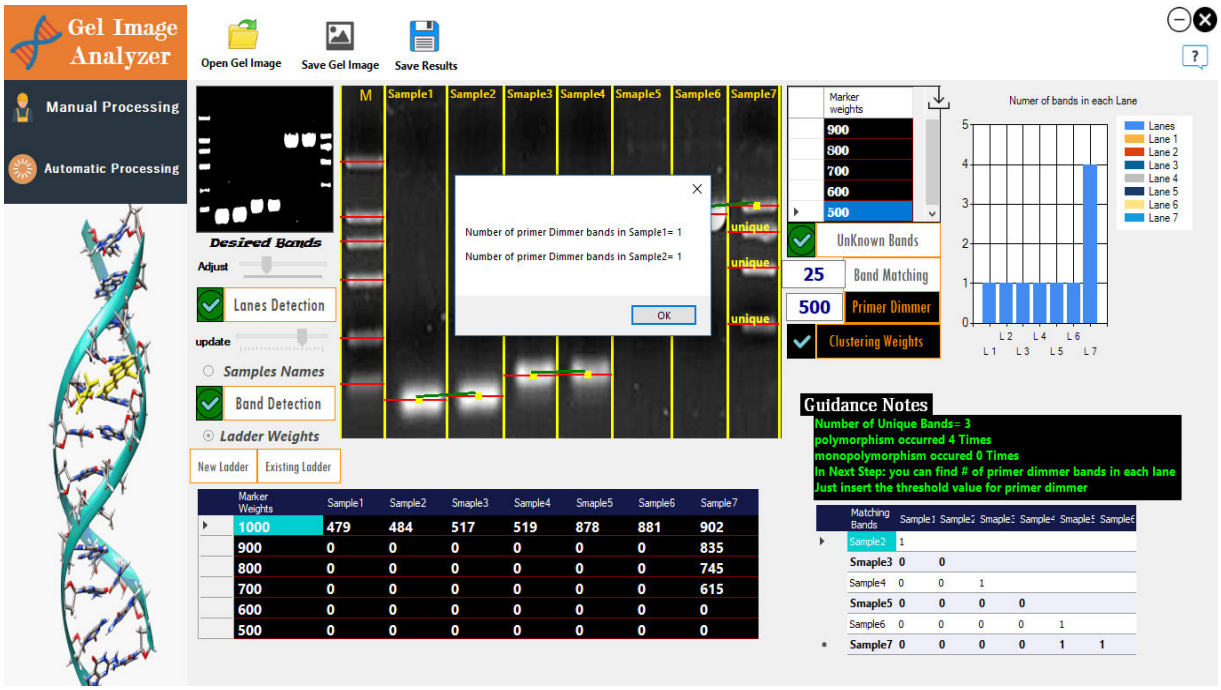


FIGURE 17. Detecting the number of primer dimmer bands in each lane.

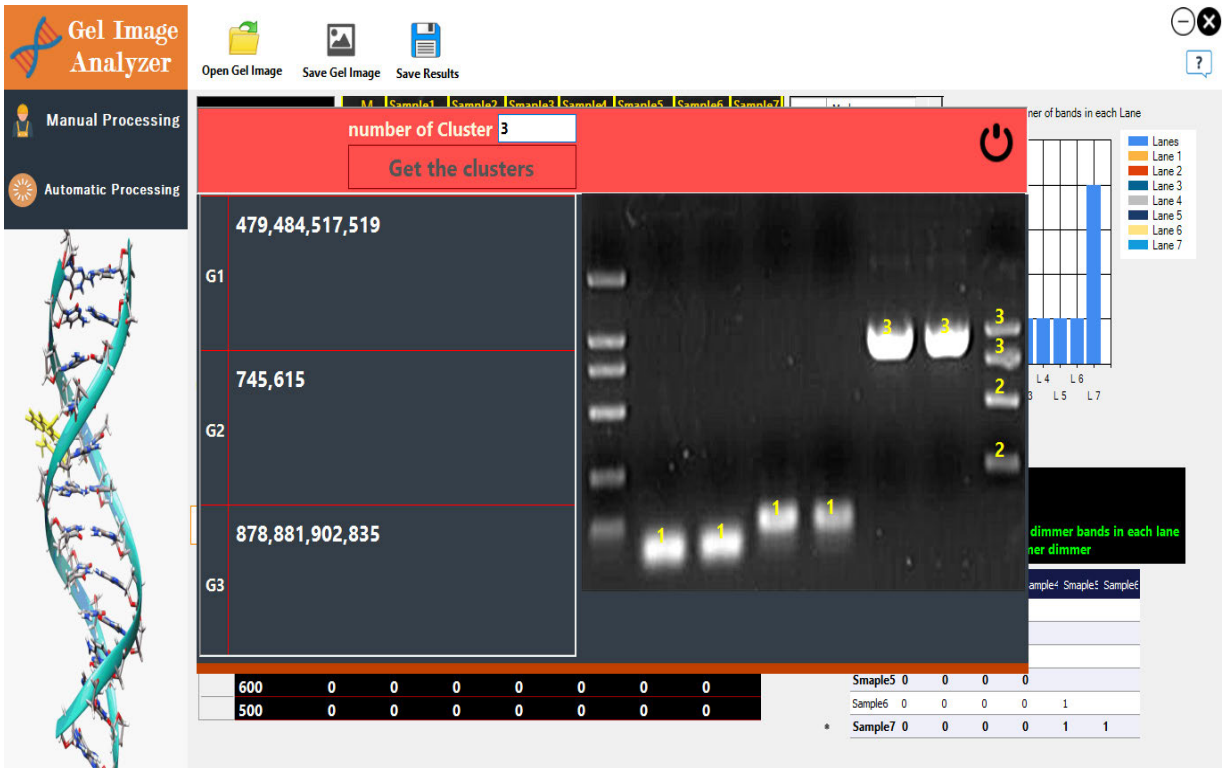
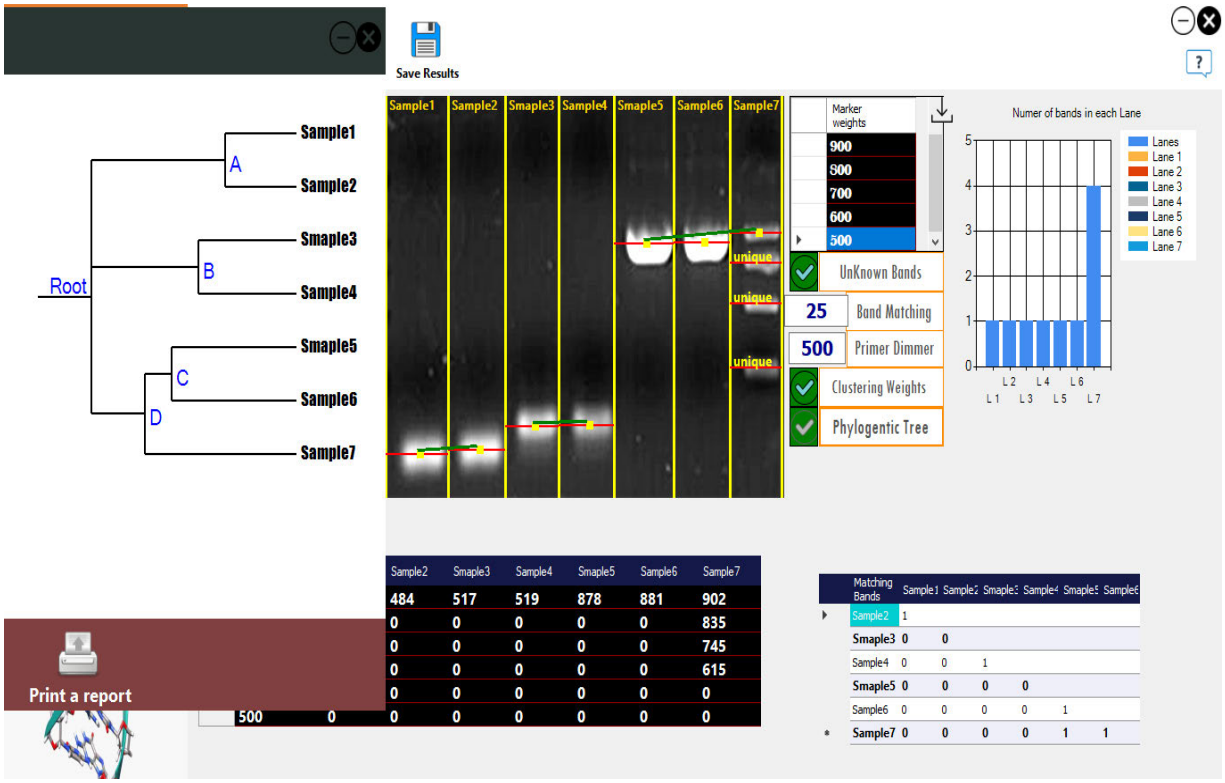


FIGURE 18. Clustering bands using k-means algorithm.

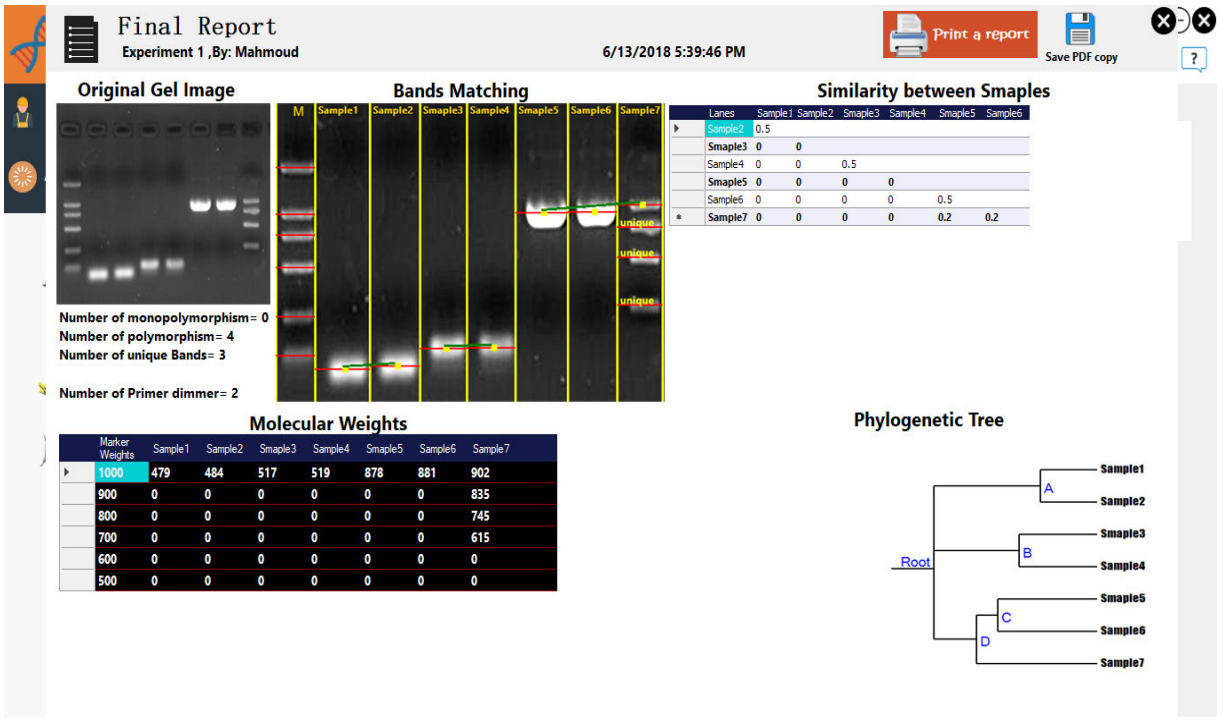
hierarchical clustering. The inputs of this algorithm are clusters ''K'' and the dataset. K means the number of the cluster or groups, the user controls the number of groups as he needed as the software gives him the option to determine the number of groups by entering the number of clusters in a textbox. The dataset refers to the molecular weights of bands.



Sample2	Smaple3	Sample4	Smaple5	Sample6	Sample7
484	517	519	878	881	902
0	0	0	0	0	835
0	0	0	0	0	745
0	0	0	0	0	615
0	0	0	0	0	0
500 0	0	0	0	0	0

Matching Bands	Sample1	Sample2	Smaple3	Sample4	Smaple5	Sample6
Sample2	1					
Smaple3	0	0				
Sample4	0	0	1			
Smaple5	0	0	0	0		
Sample6	0	0	0	0	1	
Sample7	0	0	0	0	1	1

FIGURE 19. Generating the phylogenetic tree.



Lanes	Sample1	Sample2	Smaple3	Sample4	Smaple5	Sample6
Sample2	0.5					
Smaple3	0	0				
Sample4	0	0	0.5			
Smaple5	0	0	0	0		
Sample6	0	0	0	0	0.5	
Sample7	0	0	0	0	0.2	0.2

Marker Weights	Sample1	Sample2	Smaple3	Sample4	Smaple5	Sample6	Sample7
1000	479	484	517	519	878	881	902
900	0	0	0	0	0	0	835
800	0	0	0	0	0	0	745
700	0	0	0	0	0	0	615
600	0	0	0	0	0	0	0
500	0	0	0	0	0	0	0

FIGURE 20. Printing a report for the current experiment.

Dividing bands into groups based on their molecular weights made the procedure of taking biological decisions simple and easy.

G. GENERATING THE PHYLOGENETIC TREE

Now and after finishing the previous seven steps we have all the information required for calculating the similarity

between the samples and generating the corresponding phylogenetic tree such as the number of the samples, the position of each band in the sample and molecular weights. The software implements UPGMA clustering algorithm which is considered as one of the most used methods for generating phylogenetic tree [18] in which all of GelClust, gelQuest [19], BioNumerics software [20], a laboratory technique called Pulsed-field gel electrophoresis (PFGE) [21], [22], GelJ [23] are used UPGMA algorithm to generate the phylogenetic tree. The process of generating the phylogenetic tree is done by calculating the centroid point for bands in each lane then using Euclidean distance to get the distance between each lane and calculate the distance matrix then passing this matrix to the implemented UPGMA clustering algorithm to construct the Phylogenetic tree. On the other hand, the software computing the similarity between samples as a similarity matrix. Calculating similarity matrices and construct the phylogenetic tree are the two methods used in fingerprints comparison.

IV. RESULTS AND DISCUSSION

According to all the above, EGBioImage software provides the user the ability to process the gel image in two ways, full manual process and automatic process with the ability of manual correction of lanes and bands. The capabilities of EGBioImage compared with other similar non-commercial software are illustrated in Table 2. Ten random gel electrophoresis images are selected for doing this comparison. EGBioImage is better than or equal to other software in most of the capabilities [2], [10] as illustrated in Table 2. Moreover, EGBioImage generates the phylogenetic tree unlike some other commercial software which does not generate phylogenetic such as EzQuant, Dolphin 1D, Gel-Pro Analyzer, Intelligent Quantifier, Molecular Imaging, myImage Analysis, Un-Scan-it, Gel-Quant, Image Lab, ImageStudio, LabImage and Ultraquant and some other non-commercial software such as Laneruler, GelAnalyzer, GelQuant and ImageJ.

The flaw of this software is less variety of similarity coefficients and clustering algorithms used in drawing a phylogenetic tree compared to GelClust and PyElph. On the other hand, EGBioImage has a very good graphical user interface and much more user friendly. The user interact with the software using keyboard on six occasions, inserting name of experiment and name of biologist which will be added in the final report, inserting names of the samples, inserting the molecular weights of the marker bands, inserting the tolerance value in the case of detecting the matching bands in each lane, inserting the threshold value in case of detecting primer dimmer bands and finally inserting the number of groups or clusters in case of clustering the bands using K-means algorithm.

Furthermore, the main specific capability of EGBioImage software that is embedded only in this software is grouping the bands based on its molecular weights and labeled each band with its group number in the gel image using the k-means algorithm.

In the following section, we display EGBioImage software at the running time from Fig. 8 to Fig. 20. This experiment is executed on a laptop with the following capabilities: Intel core i5 with 6.0 GB Memory in windows 10 with 64 bit operating system.

V. CONCLUSION AND FURTHER RESEARCH

In this paper, after discussing the current tools that are used to analysis the gel images and detect the main lacks in them, a new software called EGBioImage is designed and developed to fill the shortage in the current software. According to the results, EGBioImage is better than a lot of existing tools in various aspects, the accuracy of lanes and bands detection, finding the matching bands according to a tolerance value taken from the user. Moreover, this software provides some specific feature or capabilities such as:

I. Grouping bands according to their molecular weights and labeled each band based on its group number in the gel image using the k-means algorithm.
II. Counting the number of primer dimmer bands in each lane based on a threshold value inserted by the user.

EGBioImage can be used in a lot of fields such as molecular biology, forensics, microbiology, genetic, biochemistry, etc...

This software is suitable for each of the researchers and students in their daily analysis of gel images.

The missing capabilities which could be added in the next versions of EGBioImage are:

(1) More clustering algorithms.
(2) Different algorithms for image processing.
(3) Database for saving all the experiments and displaying them at any time as needed.
(4) Mobile application version.

REFERENCES

[1] N. Kaabouch, R. R. Schultz, B. Milavetz, and L. Balakrishnan, "An analysis system for DNA gel electrophoresis images based on automatic thresholding an enhancement," in *Proc. IEEE Int. Conf. Electro/Inf. Technol.*, May 2007.
[2] S. Khakabimamaghani, A. Najafi, and R. Ranjbar, "GelClust: A software tool for gel electrophoresis images analysis and dendrogram generation," *Comput. Methods Programs Biomed.*, vol. 111, no. 2, pp. 512–518, 2013.
[3] R. M. Alawdi, R. B. M. Amer, A. M. Alzohairy, and W. M. Khedr, "The computational techniques developed to analyze DNA Gel images," *Int. J. Adv. Eng. Res. Sci.*, vol. 3, no. 7.
[4] X. Wang, T. A. Rinehart, P. A. Wadl, J. M. Spiers, "A new electrophoresis technique to separate microsatellite alleles," *Afr. J. Biotechnol.*, vol. 8, no. 11, 2009.
[5] A. B. Pavel and C. I. Vasile, "PyElph—A software tool for gel images analysis and phylogenetics," *BMC Bioinf.*, vol. 13, no. 1, p. 9, 2012.
[6] J. M. M. Pérez and J. Pascau, *Image Processing With ImageJ*. Birmingham, U.K.: Packt Publishing Ltd., 2013.
[7] T. Ferreira and W. Rasband, *ImageJ User Guide*. ImageJ/Fiji, 2012.
[8] S. Goraniya, A. R. Tusamda, G. R. Shirolkar, G. Rao, S. N. Murthy, and S. D. Pawar, "Molecular analysis of Manilkara hexandra Roxb. and Averrhoa carambola L. using RAPD markers helps to understand genetic variations," *Int. J. Pharmacy Pharmaceutical Sci.*, vol. 5, no. 3, pp. 626–628, 2013.
[9] P. Gerner-Smidt, L. M. Graves, S. Hunter, and B. Swaminathan, "Computerized analysis of restriction fragment length polymorphism patterns: Comparative evaluation of two commercial software packages," *J. Clin. Microbiol.*, vol. 36, no. 5, pp. 1318–1323, 1998.

[10] J. Heras, C. Domínguez, E. Mata, V. Pascual, C. Lozano, C. Torres, and M. Zarazaga, "A survey of tools for analysing DNA fingerprints," *Briefings Bioinf.*, vol. 17, no. 6, pp. 903–911, 2015.
[11] K. Singh and R. Kapoor, "Image enhancement using exposure based sub image histogram equalization," *Pattern Recognit. Lett.*, vol. 36, pp. 10–14, Jan. 2014.
[12] G. Jiang, C. Y. Wong, S. C. F. Lin, M. A. Rahman, T. R. Ren, N. Kwok, H. Shi, Y.-H. Yu, and T. Wu, "Image contrast enhancement with brightness preservation using an optimal gamma correction and weighted sum approach," *J. Mod. Opt.*, vol. 62, no. 7, pp. 536–547, 2015.
[13] L. Yang and F. Albregtsen, "Fast and exact computation of Cartesian geometric moments using discrete Green's theorem," *Pattern Recognit.*, vol. 29, no. 7, pp. 1061–1073, 1996.
[14] O. Tahri and F. Chaumette, "Point-based and region-based image moments for visual servoing of planar objects," *IEEE Trans. Robot.*, vol. 21, no. 6, pp. 1116–1127, Dec. 2005.
[15] A. K. Mukherjee, S. Ratha, S. Dhar, A. K. Debata, P. K. Acharya, S. Mandal, P. C. Panda, and A. K. Mahapatra, "Genetic relationships among 22 taxa of bamboo revealed by ISSR and EST-based random primers," *Biochem. Genet.*, vol. 48, nos. 11–12, pp. 1015–1025, 2010.
[16] A. Badr, H. H. El-Shazly, N. S. Helail, and W. El Ghanim, "Genetic diversity of *Artemisia* populations in central and north Saudi Arabia based on morphological variation and RAPD polymorphism," *Plant Systematics Evol.*, vol. 298, no. 5, pp. 871–886, 2012.
[17] N. Dhanachandra, K. Manglem, and Y. J. Chanu, "Image segmentation using K-means clustering algorithm and subtractive clustering algorithm," *Procedia Comput. Sci.*, vol. 54, pp. 764–771, 2015.
[18] G.-J. Hua, C. L. Hung, C. Y. Lin, F. C. Wu, Y. W. Chan, and C. Y. Tang, "MGUPGMA: A fast UPGMA algorithm with multiple graphics processing units using NCCL," *Evol. Bioinf.*, vol. 13, Oct. 2017, Art. no. 1176934317734220.
[19] R. Shah, M. M. Malgaonkar, A. R. Shirolkar, S. G. Sangvikar, S. N. Murthy, S. D. Pawar, and Y. R. Kulkarni, "Molecular analysis of mucuna Pruriens (L.) DC. and Mucuna cochinchinensis (Lour.) A. Chev. Using RAPD markers helps to understand genetic variations," *IOSR J. Pharmacy Biol. Sci.*, vol. 10, no. 4, pp. 1–5, 2015.
[20] G. R. Golding, J. Campbell, D. Spreitzer, and L. Chui "Pulsed-field gel electrophoresis of *Staphylococcus aureus*," in *Pulse Field Gel Electrophoresis*, New York, NY, USA: Springer, 2015, pp. 85–93.
[21] A. Alaidan, M. F. Bohol, S. I. Al-Thawadi, A. Al-Suwaine, A. A. Al-Qahtani, and M. N. Al-Ahdal, "Rapid PFGE method for fingerprinting of Serratia marcescens isolates," *J. Microbiol. Methods*, vol. 78, no. 2, pp. 238–241, 2009.
[22] T. Revazishvili and J. A. Johnson, "Pulsed-field gel electrophoresis of *Yersinia pestis*," in *Pulse Field Gel Electrophoresis: Methods and Protocols*, K. Jordan and M. Dalmasso, Eds. New York, NY, USA: Springer, 2015, 115-128.
[23] J. Heras, C. Domínguez, E. Mata, V. Pascual, C. Lozano, C. Torres, and M. Zarazaga, "GelJ—A tool for analyzing DNA fingerprint gel images," *BMC Bioinf.*, vol. 16, no. 1, p. 270, 2015.
[24] *GelAnalyzer*. [Online]. Available: http://www.gelanalyzer.com/

MAHMOUD H. ALNAMOLY received the B.S. degree from the Department of Computer Science, Faculty of Computer and Informatics, Zagazig University, Egypt, in 2015, where he is currently pursuing the master's degree. His research interests include computer science, bioinformatics, natural language processing, and machine learning.

AHMED MANSOUR ALZOHAIRY is currently a Professor of genetics and agricultural genetic engineering and the Head of the Department of Genetics, Faculty of Agriculture, Zagazig University. His research interests include plant genomics, biotechnology, and bioinformatics.

IBRAHIM MAHMOUD EL-HENAWY is currently a Professor of computer science and the Head of the Department of Computer Science, Faculty of Computer and Informatics, Zagazig University, His research interests include computer science, bioinformatics, deep learning, soft computing, and optimization.