*Ceylon Journal of Science* (*Bio. Sci.*) 44 (1):45-54, 2015
DOI: http://dx.doi.org/10.4038/cjsbs.v44i1.7340

# An Automated System for Analyzing Agarose and Polyacrylamide Gel Images

**A.M.K.W.K. Abeykoon[1], M.P.C.S. Dhanapala[2], R. D. Yapa[1] and S.D.S.S. Sooriyapathirana[2*]**

[1]Department of Statistics and Computer Science, University of Peradeniya, Peradeniya, Sri Lanka.
[2]Department of Molecular Biology and Biotechnology, University of Peradeniya, Peradeniya, Sri Lanka.
*Accepted March 22, 2015*

## ABSTRACT

Agarose or polyacrylamide gel electrophoresis is a common technique used to separate nucleic acids (DNA and RNA) according to their molecular weight, which generates images that can be analyzed automatically by a system. The separated DNA fragments of different molecular weights give a series of bands with positions in the gel image corresponding to their molecular weights. Automated image analysis of the gels removes much of the subjectivity of manual interpretation of band positions and sizes. In this paper, an automated system was designed for analyzing DNA bands using image and signal processing techniques. The proposed algorithm consists of four main steps: preprocessing to enhance the image, detecting lanes and bands, identifying band lengths and clustering; in which the similarities of biological samples based on shared bands can be identified. The proposed technique eliminates defects due to noise and double bands in images.

**Keywords:** DNA, gel electrophoresis, image-processing, signal-processing, cluster analysis

## INTRODUCTION

Electrophoresis is an electrochemical separation process of biological molecules such as DNA and proteins (Meyers *et al*., 1976). The electric field charge causes individual DNA or protein molecules of the same size to migrate to discrete positions within the bed of electrophoretic matrix made up of agarose or polyacrylamide. The result of the electrophoresis can be presented as a gel image which contains several vertical lanes, each corresponding to a sample of analysis which has horizontal bands where each band is representing a mass of individual molecules of the same size (Mickel *et al*., 1977). Analysis and interpretation of gel images followed by electrophoresis is cumbersome (Ho *et al*., 2004) especially when large number of samples are subjected to analysis (Brunello *et al*., 2001). Even when few samples are being electrophoresed, it is difficult to guess the molecular weight of individual bands on the gel compared to a standard sample (i.e. ladder). An automated system to read the bands in lane of the gel and then to estimate the individual band sizes is therefore very useful (Bailey and Christie, 1994; Kaabouch *et al*., 2007; Akhter *et al*., 2008). However, contemporary algorithms for identifying, processing and analysis of gel images not only lack all the applications required for processing but also these are not freely available for molecular biologists (Pavel and Vaslie, 2012). Further these may take a longer time for the process (Stathopoulou *et al*., 2006). The objective of this paper was to develop an automated system for agarose and polyacrylamide gel image analysis for research in molecular biology.

## MATERIALS AND METHODS

### Structuarl design of the system

The architecture of the system is shown in Figure 1. It has nine steps beginning from the acquisition of the gel image to the final creation of outputs.

#### *Uploading gel image and preprocessing*

The preprocessing stage was started with uploading the original gel image (previously taken from a routine gel electrophoretic experiment) in Red, Green and Blue Color Model (RGB color) and then, converted into the grayscale format. After that, the image was resized to a dimension of 480 × 370 pixel area. A moderate sized image is required for processing. Smaller image is not good as some information could be lost while larger image can cause slowing down of the process, therefore, 480 × 370 pixel size was chosen. Since it could contain noises due to practical imaging problems, a median noise filter was applied to remove the noises (Maheswari and Radha, 2010). To the resultant image, a histogram was drawn using averages of all the pixels in particular areas of grey and white. If that average value is greater than 100, it was considered as the means of the most of pixels are

*Corresponding author's email: sunethss09@gmail.com

closer to white and bands contain a lighter background. The average value 100 was selected as it was the most appropriate threshold for a variety of gel images used to check the performance of the developed system. For further processing, the band should be in lighter coloured than its background. Thus, a system was employed using this average value to check whether it would require the image inversion in cases where backgrounds would be darker than the bands.

***Identifying the sample lanes and the ladder lane***

After processing of the gel image, the lanes (which were produced due to the migration of DNA fragments from the negatively charged electrode to the positively charged electrode) were identified (Lin *et al*., 2007). Here, it was assumed that lanes migrate linearly in vertical passion. The lanes were displayed as rectangles (Akbari *et al*., 2010). There is a special lane called ladder (*i.e*. marker), which could be placed at any lane in the image.

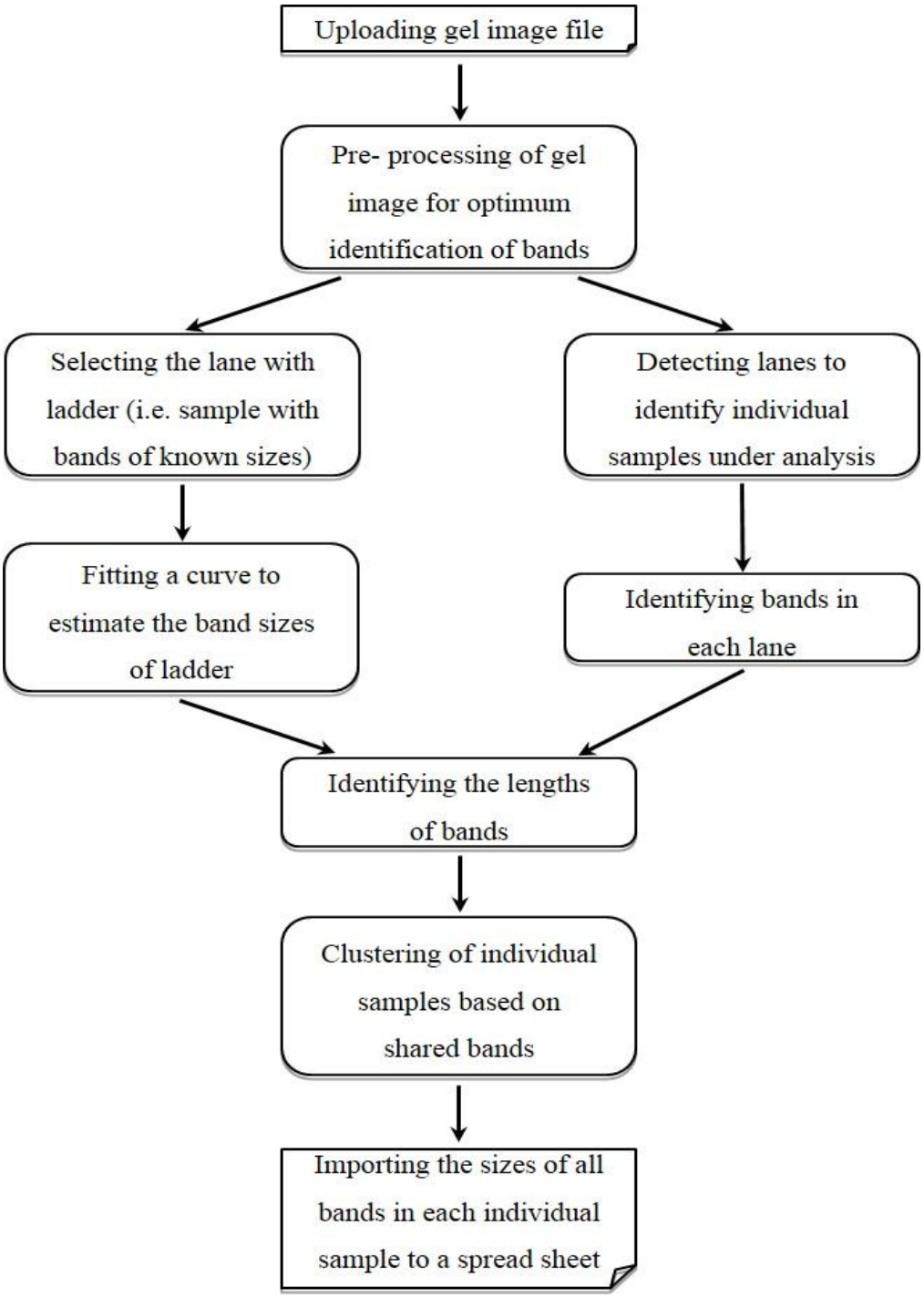


**Figure 1**. Architecture of the system of gel image analysis

The lane of bands is used as a standard to measure/estimate the size of the bands in other lanes. Thus, the ladder should be selected first. For identifying the lanes, column sums were calculated (Equation 1), then they were averaged and plotted. The image with light bands can be separated by finding minima of that plotted curve. Since bands contained lighter backgrounds and other areas contained black background, consecutive minima (*i.e.* black places) sandwiching the band would separately identify the lanes.

$$\text{TotalBand}[x] = \frac{\sum \text{Intensity}(x, y)}{\text{Count}(x)} \quad \textbf{(Equation 1)}$$
$$x = \text{column number}$$
$$y = \text{row number}$$

Count was the total number of white pixels per lane and the intensity was the summation of whiteness values per lane. Minima were obtained by getting the second derivatives of the curve (Figure 2A), unfortunately due to noises some unwanted minima were also received. Thus, the proper minima were filtered by using a variation of gray level to current minimum and by assigning a minimum pixel difference (Figures 2B and 2C).

***Identifying the bands and their sizes***

To identify bands in each lane, the row average was calculated using the same way as maximum values found for bands of a lane. To remove the noises, bands were filtered in the points where average values were greater than the average of lane curve. It was considered to be of greater than 0.1 portion (10%) of maximum height in the curve (Figure 3). The value 0.1 was selected as it was the most appropriate threshold for a variety of sample gel images to remove the noise to read the bands clearly.

Then, to measure the weight of the bands, migration distance of all the bands of unknown samples were compared relative to the bands in the ladder lane. Thus, the ladder bands were calibrated first. The relationship with weight and the gap can be represented using an exponential distribution (Stellwagen, 1998), so as to calculate the distribution parameters 'a' and 'b' of Equation 2 by mapping location with the known weights (Figure 4).

$$t(a, b, x) = a \times e\,(b \times x) \quad \textbf{(Equation 2)}$$

***Identifying the similarity of tested samples***

A clustering method was used to find the related samples. Dissimilarity in the band profiles between two lanes i and j were measured according to the following procedure. The numbers of bands in two lanes were identified separately. The lane with the highest number of bands was labeled as i and the other lane as j. By travelling through each band in i (as a) find closest band with the other layer j (as b). Two methods were used to check the similarity of two samples. The absolute difference with band value (size of the band in base pairs) of i and band value j should be less than the marker range (known band sizes in the ladder) (Equation 3).

$$\textbf{Distance}(\mathbf{a}, \mathbf{b}) = |(\textbf{Band Value}(\mathbf{a}) - \textbf{Band Value}(\mathbf{b})| \quad \textbf{(Equation 3)}$$

According to the absolute difference, a weight value using fuzzy membership function was given. But before mapping to a weight, the distance must be normalized and minimum-maximum normalization procedure was employed. Minimum value was set as 0 and maximum value was set as marker range (Equation 4).

$$\textbf{Normalized Distance} = \frac{\textbf{Distance}(\mathbf{a}, \mathbf{b})}{\textbf{Marker Range}} \quad \textbf{(Equation 4)}$$

The normalized distance converged to an interval between 0 and 1 and it got mapped with the fuzzy membership function like an S-function. It was denoted as the difference of two sigmoidal functions. The sigmoidal membership function used depends on the two parameters a and c and is given in Equation 5.

$$\mathbf{f}(\mathbf{x}; \mathbf{a}, \mathbf{c}) = \frac{\mathbf{1}}{\mathbf{1} + \mathbf{e}^{-\mathbf{a}(\mathbf{x}-\mathbf{c})}} \quad \textbf{(Equation 5)}$$

If the distance was closer, the bands were very similar and mapped to value 1. If the value was higher, the bands were dissimilar and map to value 0, and there was a range that cannot say exactly similar or dissimilar. In that range, a weight value was assigned between 0 and 1 (Figure 5).

By comparing each band in i with closest band with the other lane j, the summation was obtained of that given weights and they were averaged by using the number of comparisons done. This value represented the distance between the compared lanes (i.e. biological samples). As shown below a similarity matrix was created by comparing each and every lane pairs. It was n × n matrix which contain diagonal element with 0 values (Figure 6).

By calculating the City Block Distance using the similarity matrix, the degree of similarity among the samples were identified. Given an m-by-n data matrix X, which was treated as m (1-by-n) row vectors x1, x2, ..., xm, the various distances between the vector xr and xs were defined as given in Equation 6.

$$d_{rs}^2 = \sum_{j=1}^{n} |x_{rj} - x_{sj}| \quad \textbf{(Equation 6)}$$

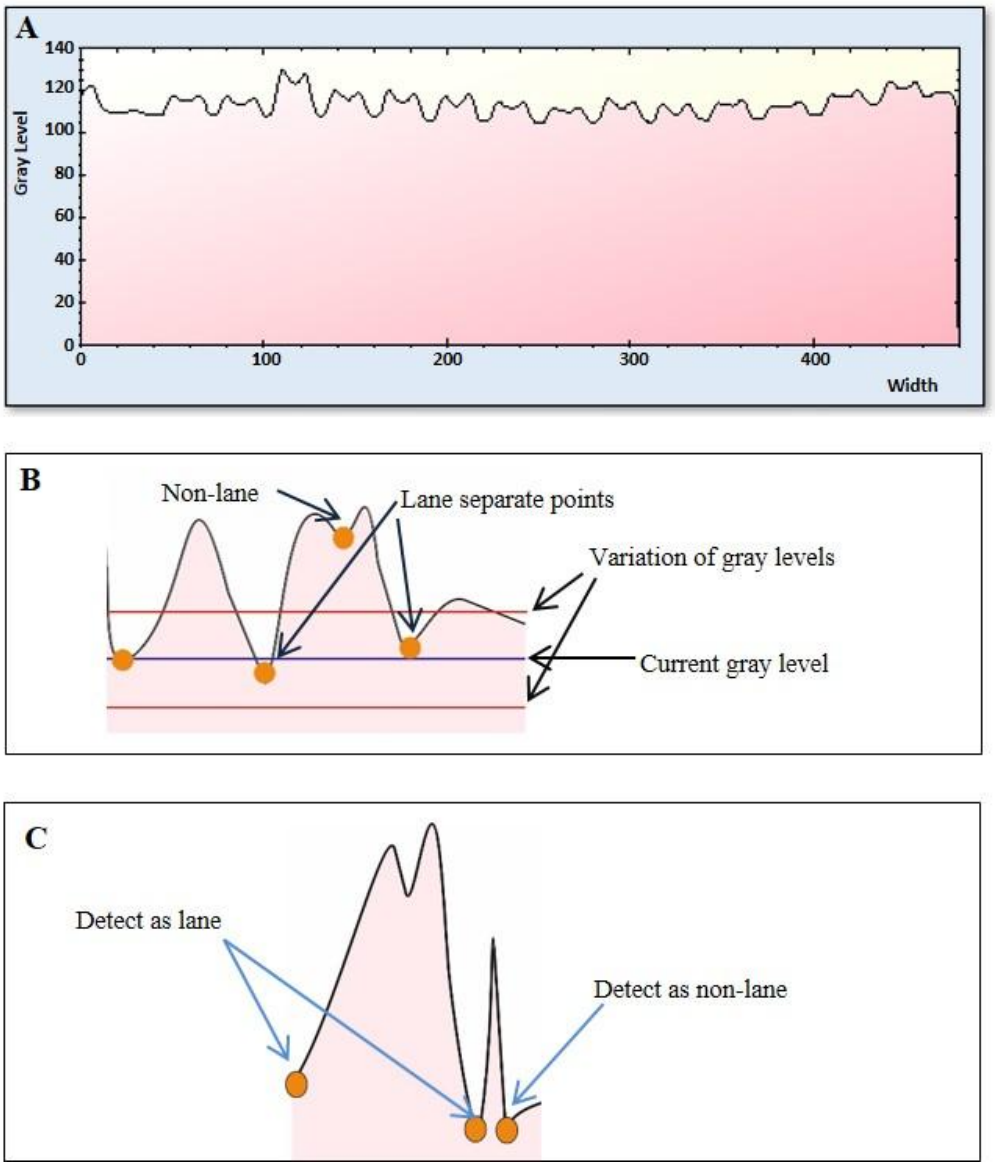


**Figure 2.** A: Average of columns, B: Filtering the lane separation points using gray levels, C: Filtering the lane separation points using pixel difference

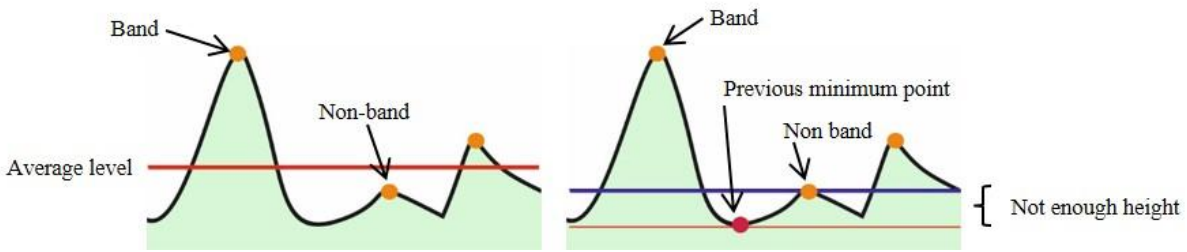


**Figure 3**. Filtering noise band using average level and given height

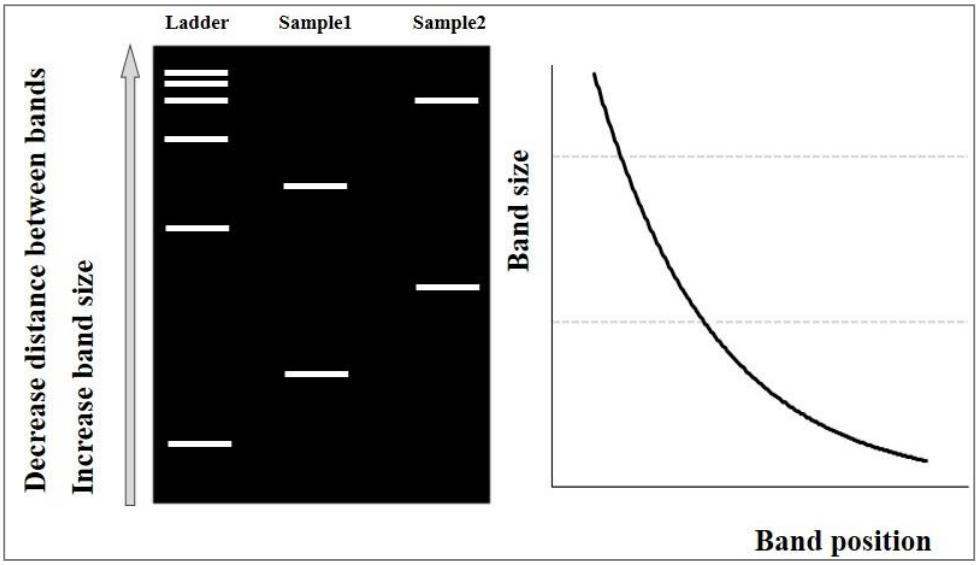


**Figure 4.** Relationship with size of the band and the gap between adjacent bands

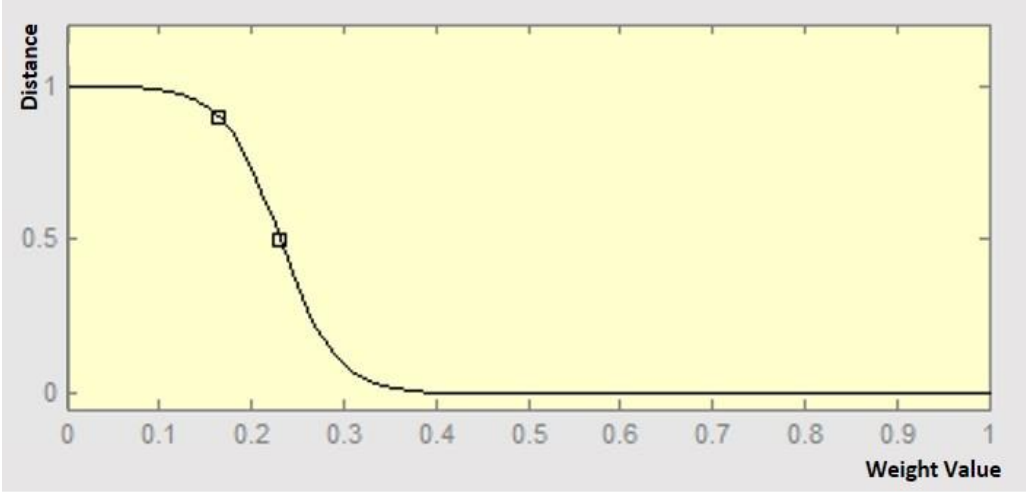


**Figure 5.** Fuzzy membership function for mapping

$$\begin{bmatrix} 1 & \cdots & 0 & 0.3 \\ 0.4 & \ddots & 1 & 0.1 \\ 0 & \cdots & 0.1 & 1 \end{bmatrix}$$

**Figure 6.** Similarity matrix created by comparing all the lane pairs

The hierarchical clustering procedure was used as the clustering method. Agglomerative hierarchical processing consists of repeated cycles where the two closest remaining samples were joined by a node/branch of a tree, with the length of the branch set to the distance between the joined samples. The two joined samples were removed from list of samples being processed and replaced by a sample that represented the new branch. The distances between this new sample and all other remaining samples were computed, and the process was repeated until only one sample remained. The average linkage method that get the distance between two samples x and y was the mean of all pairwise comparisons (Equation 7).

$$\frac{1}{|A||B|}\sum_{a\in A}\sum_{b\in B} d(a,b) \quad \textbf{(Equation 7)}$$

A dendrogram was used to show how the samples were related to each other by using the results of the clustering.

***Graphical User Interface of the system***

The main Graphical User Interface (GUI) of the system is given in the Figure 7. The icons are provided ('**a**' to '**o**' in Figure 7) to apply the changes when necessary and to conduct the required analyses of the uploaded image. Adding, editing and removing of lanes and bands supported by another GUI (Figure 8A). At the same time marker bands were detected and labeled from bottom reference to top with increasing marker range (Figure 8B). If an error is coming for a rare unique ladder, the manual adjustment of the marker bands could also be done. After correctly assigning to map reference value with the labeled value, an exponential distribution was used to identify parameters call to MATLAB (The MathWorks Inc., Natick, MA, USA). For unknown samples, the band values were calculated by using Equation 8.

$$Cf(x) = a \times \exp(b \times x) \quad \textbf{(Equation 8)}$$

Where a and b are the coefficients with 95% confidence bounds [a = 834.4 (669.9, 999), b = -0.01039 (-0.01232, -0.008467)].

The detected bands of the samples were then can be applied to hierarchical clustering method to identify the related samples (Figure 9). The degree of similarity is deduced according to the bands shared among the samples.For saving and downstream computational analysis, the result can be exported as an Excel spread sheet.

***Verification of the results***

The system was tested for its accuracy by comparing its outputs for 10 diverse gel images to the outputs generated for the same gels by three trained human subjects worked as a team to read the bands.

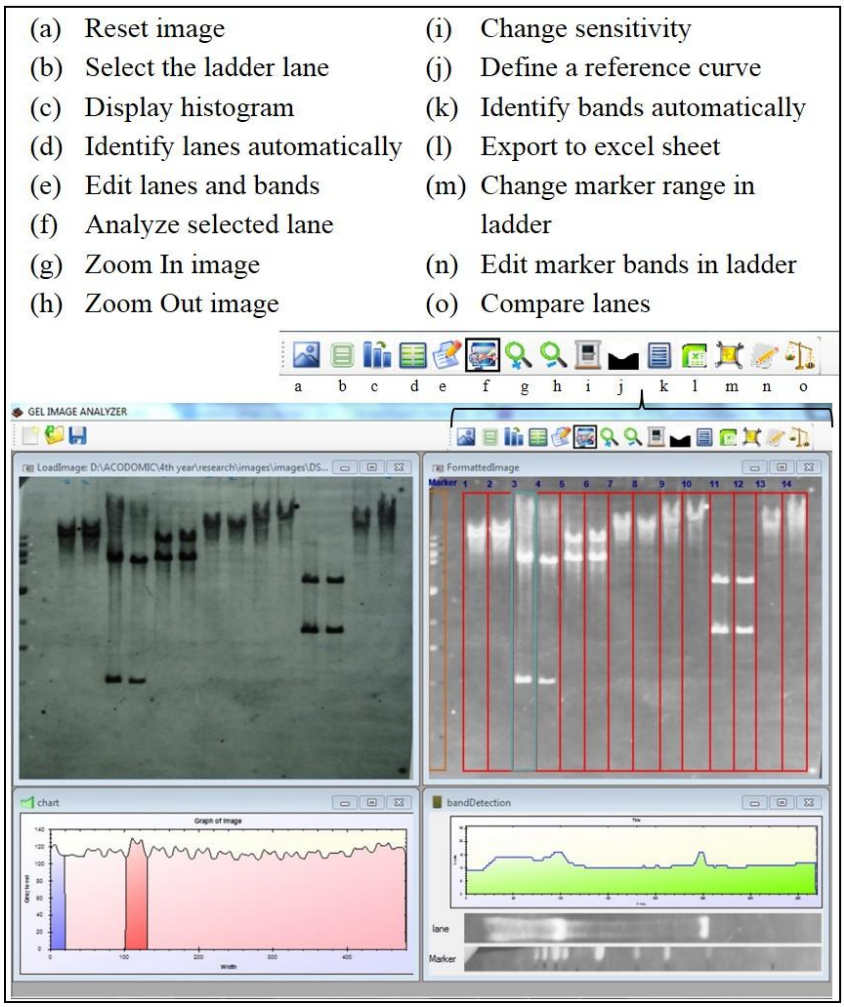


**Figure 7**. The main Graphical User Interphase

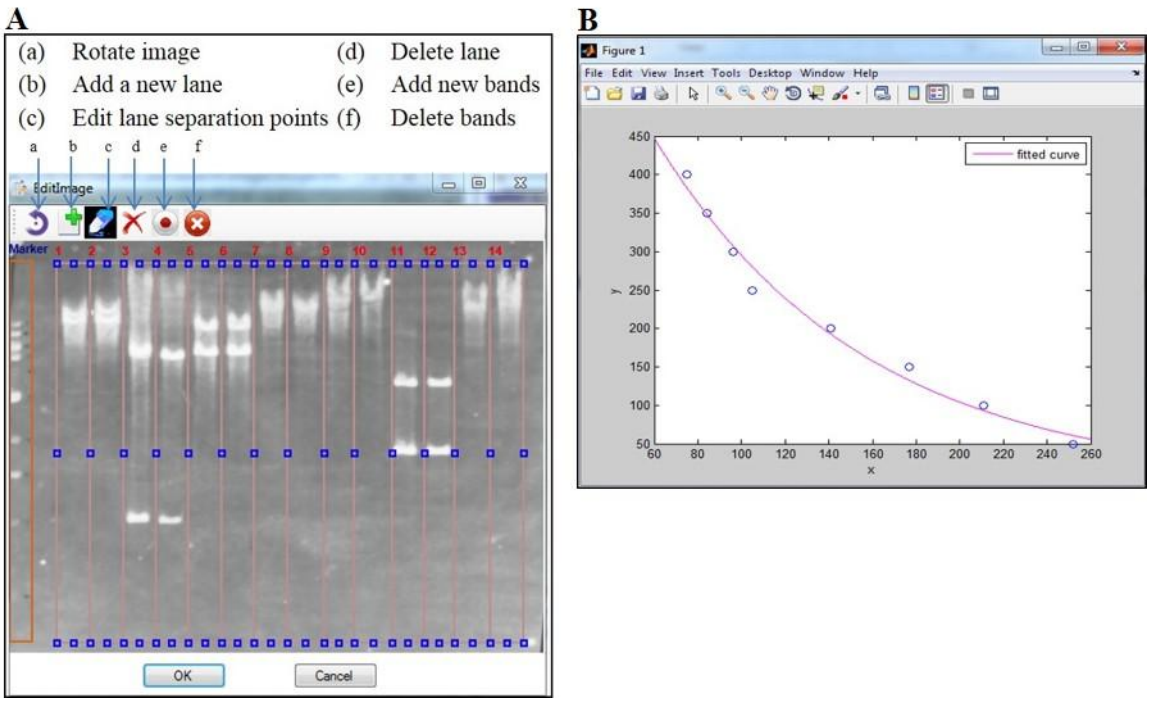


**Figure 8.** A: The Graphical User Interphase for editing lanes and bands. B. Fitting a curve to size the marker bands in ladder

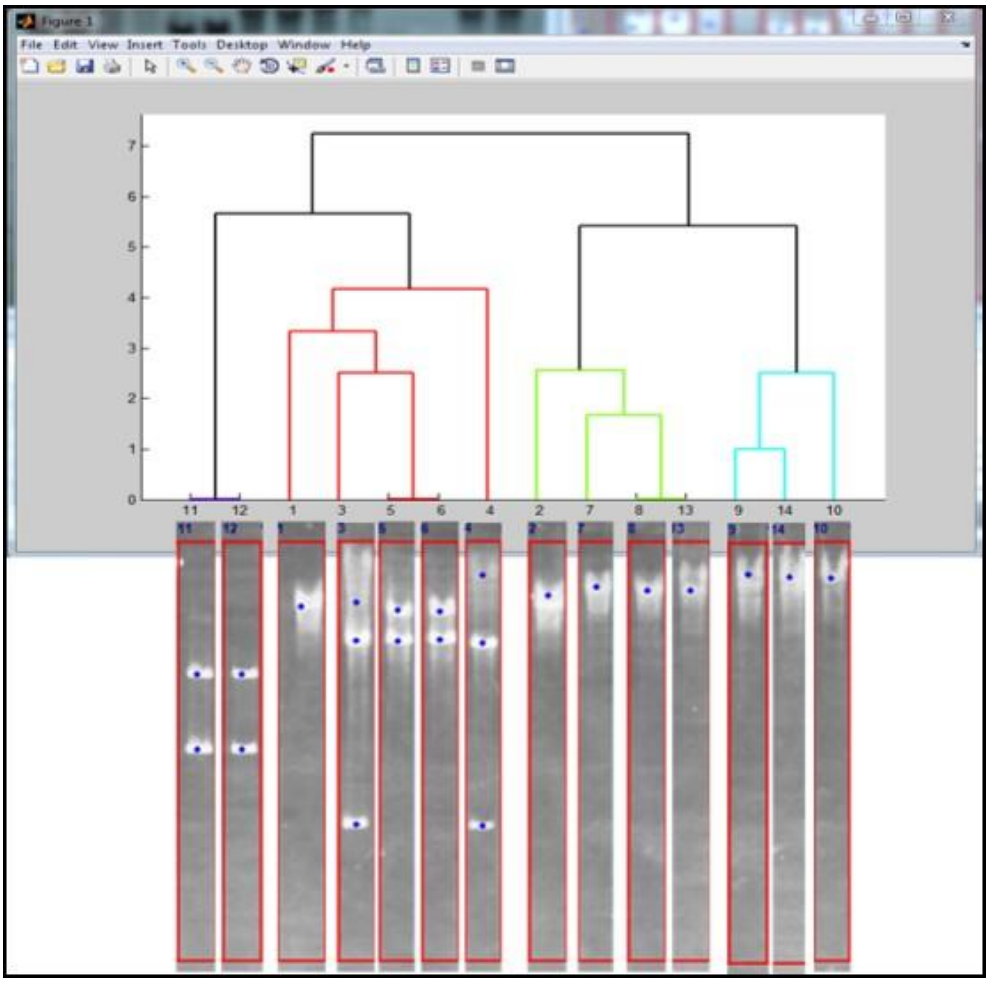


**Figure 9.** An example final result of clustering

## RESULTS

The comparison of system outputs and the outputs manually generated by three experts for 10 gel images is given in Table 1.In all three cases (Table 1) that the erors caused, were due to the presence of double bands (bands present next to each other) and no error was observed when there were no double bands present. The nearest band must be removed by considering it as a noise band as the second band could be an experimental artifact. The double bands are frequently observed for PCR based microsatellite DNA markers (Schlotterer and Tautz, 1992). In addition, the running time, easiness of uploading and subsequent processing were all found be smooth and quick enough for the use in routine experiments. The output as an excel spread sheet was mainly appreciated by the researchers who are doing large scale gel based genotyping in marker assisted breeding.

**Table 1**. Confirmation of reads by trained human subjects

| Gel Image | Total number of bands | Number of bands not identified by the GUI | Number of errors made by the GUI | Error percentage (%) |
|---|---|---|---|---|
| 1 | 16 | 0 | 0 | 0 |
| 2 | 22 | 0 | 0 | 0 |
| 3 | 20 | 0 | 2 | 10 |
| 4 | 8 | 1 | 0 | 13 |
| 5 | 24 | 4 | 0 | 17 |
| 6 | 22 | 0 | 0 | 0 |
| 7 | 19 | 0 | 0 | 0 |
| 8 | 30 | 0 | 0 | 0 |
| 9 | 22 | 0 | 0 | 0 |
| 10 | 25 | 0 | 0 | 0 |

## DISCUSSION

Accurate reading of bands in agarose and polyacrylamide gels in molecular research is essential to draw valid conclusions. Reading bands by human subjects could lead to countless errors and the error rate get even higher when relatively inexperienced researchers or technicians are employed to record bands (Elder *et al*., 1986). Manual reading of gels is time consuming and the biggest problem is the subjectivity when estimating the band size just by looking at the gel. Therefore, automated systems are useful in reading gels (Elder *et al*., 1986). With the introduction of Sanger based manual DNA sequencing procedures, early attempts have been recorded to automate gel reading and to generate sequence outputs (Elder *et al*., 1986; Stockwell, 1985), but they were not targeted to read the band sizes. Cooper *et al*., (1996) developed a UNIX based software with improved lane tracking capabilities but it was designed for fluorescently labelled / stained gel images. Promising patent-applied software, GelMaster, was developed by Bajla *et al*., (2005), which can read the gels and indicate the band sizes right on the bands of the gel. No tabular output with band size is indicated. Zerr and Henikoff, (2005) developed a system known as GelBuddy which was targeted to map mutations and polymorphisms in applications such as Amplified Fragment Length Polymorphism (AFLP) and Target Induced Local Lesions in Genomes (TILLING).

Another software known as GelClust, was developed by avoiding the ‘smiling’ effects of the bands and to create dendrograms. But tabular outputs were not indicated (Khakabimamaghani *et al*., 2013). PyElph, a software written based on a python code, which has most of the capabilities except band size output (Pavel and Vasile, 2012). To analyze the protein bands, similar tools could be noticed such as a tool to detect the Pulse Field Gel Electrophoresis (PFGE) patterns in *E.coli* (Yokohoma and Uchimura, 2006). Duck *et al*., (2003) developed a system to analyze PFGE macro restriction fragment patterns. It was found that when raw images were used, the deviation of the estimated size of bands compared to real band size was in the range of 1.14 to 99.00 kb. Therefore, they suggested image optimization before the analysis. Preprocessing of gel image was employed by all these gel reading applications (Bajla *et al*., 2005: Cooper *et al.*, 1996: Cardinali *et al.*, 2002), although Wu *et al*., (2010) highlighted the danger of adjusting the gel images because it can cause erroneous outputs. Only linear transformations are suggested not to deviate too much from the original image. However unalike band intensities are a frequent problem in these

tools (Elder *et al*., 1986).

It is evident that there is no universal or fully automated gel reading software could be developed given the variable nature of gels. Existing software must be updated frequently with the development of image capturing and gel electrophoresis technology (Cooper *et al*., 1996). Given the very high variability of gels developed in different labs by different people using different protocols, it is very much advisable to optimize and develop software for local in-house applications (Duck *et al*., 2003)

Because of the lack of all required capabilities such as tabular outputs of band sizes, expensive nature of software and difficulty in handling the software to generate required output, we developed the above mentioned GUI for reading and analyzing of bands in agarose and polyacrylamide gels. As explained double bands and unusual ladders can cause errors but the developed tool can identify similarities and can generate tabular outputs which would be very handy in downstream analysis.

**Concluding Remarks**

The developed system can be effectively used to read the gels correctly and save the time by avoiding the steps of manual reading and guessing bands sizes. The software is available with the Authors for free use in research and teaching. A video clip is available to visualize the procedure of using the developed software.

## REFERENCES


Akbari, A., Algregtsen, F. and Jakobsen, K.S. (2010). Automatic lane detection and separation in one dimensional DNA gel images using continues wavelet transform. *Analytical Methods* **2**: 1360-1371.

Akhter, N., Khan, A.R., Talib, Y., Shadab, S. and Pater, R. (2008). Analysis of Gel Electrophoresis Images. *First International Conference on Emerging Trends in Engineering and Technology, IEEE Computer Society*, 106-109.

Bailey, D.G. and Christie, B.C. (1994). Processing of DNA and Protein Electrophoresis Gels by Image Analysis. *Proceedings of the second New Zealand Conference on Image and Vision Computing, Palmerstone North*, 2.2.1-2.2.8.

Bajla, I., Hollander, I., Fluch, S., Burg, K. and Kollar, M. (2005). An alternative method for electrophoretic gel image analysis in the Gel Master software. *Computer Methods and Programs in Biomedicine* **77**: 209—231.

Brunello, F., Ligozzi, M., Christelli, E., Bonora, E., Tortoli, E. and Fontana R. (2001). Identification of 54 *Mycobacterium* species by PCR-restriction fragment length polymorphism analysis of the *hsp65* gene. *Journal of Clinical Microbiology* **39**: 2799-2806.

Cardinali, G., Martini, A., Preziosi, R., Bistoni, F. and Baldelli, F. (2002). Multicenter comparison of three different analytical systems for evaluation of DNA banding patterns from Cryptococcus neoformans. *Journal of Clinical Microbiology* **40**: 2095–2100.

Cooper, M.L., Maffitt, D.R., Parsons, J.D., Hillier, L. and States, D.J. (1996). Lane tracking software for four- color fluorescence based electrophoretic gel images. *Genome Research* **6**: 1110-1117.

Duck, W.M., Steward, C.D., Banerjee, S.N., McGowan, J.E. and Tenover, F.C. (2003). Optimization of computer software settings improves accuracy of pulsed-field gel electrophoresismacrorestriction fragment pattern analysis. *Journal of Clinical Microbiology* **41**: 3035–3042.

Elder, J.K., Green, D.K. and Southern, E.M. (1986). Automatic reading of DNA sequencing gel autoradiographs using a large format digital scanner. *Nucleic Acids Research* **14**: 417-424.

Ho, H.T., Chang, P.L., Hung C.C. and Chang, H.T. (2004). Capillary electrophoretic restriction fragment length polymorphism patterns for the mycobacterial *hsp65* gene. *Journal of Clinical Microbiology* **42**: 3525-3531.

Kaabouch, N., Schultz, R.R. and Milavetz, B. (2007). An analysis system for DNA gel electrophoresis images based on automatic thresholding an enhancement. *IEEE, Engineering Information Technology Proceedings*, 26-31.

Khakabimamaghani, S., Najafi, A., Ranjbar, R. and Raam, M. (2013). *Computer Methods and Programs in Biomedicine* **3**: 512–518.

Lin, C.Y., Ching, Y.T. and Yang, Y.L. (2007). Automatic Method to Compare the Lanes in Gel Electrophoresis Images. *IEEE Transactions on Information Technology in Biomedicine* **11**: 179-189.

Maheswari, D. and Radha, V. (2010). Noise removal in compound image using median filter. *International Journal on Computer Science and Engineering* **2**: 1359-1362.

Meyers, J.A., Sanchez, D., Elwell, L.P. and Falkow, S. (1976). Simple agarose gel electrophoretic method for the identification and characterization of plasmid deoxyribonucleic acid. *Journal of Bacteriology* **127**: 1529-1537.

Mickel, S., Arena, V. and Bauer, W. (1977). Physical properties and gel electrophoresis behavior of R12-derived plasmid DNAs. *Nucleic Acids Research* **4**: 1465-1482.

Pavel, A.B. and Vasile, C.I. (2012). PyElph-a software tool for gel images analysis and phylogenetics. *BMC Bioinformatics*, **13**: 1-6.

Schlotterer, C. and Tautz, D. (1992) Simple synthesis of simple sequence DNA. *Nucleic Acids Research* **20**: 211-215.

Stathopoulou, I.O., Tsihrintzis, G.A., Gaitanis, G., Kollia, K. and Velegraki, A. (2006). Similarity measurement of electrophoresis strands for fungi fingerprinting. *13th International Conference on Signals, Systems, and Image Processing, Budapest, Hungary*.

Stellwagen, N.C. (1998) DNA gel electrophoresis. In: D. Tietz (Ed), *Nucleic Acid Electrophoresis Lab Manual*, Springer, Pp. 1-53.

Stockwell, P.A. (1985). VTUTIN: A full screen gel management editor. *CABIOS*, **1**, 253-259.

Wu, H.C., Yen, C.C., Tsui, W.H. and Chen, H.M. (2010). A red line not to cross: Evaluating the limitation and properness of gel image tuning procedures. *Analytical Biochemistry* **396**: 42–50.

Yokoyama, E. and Uchimura, M. (2006). Optimal settings of fingerprint-type analyzing computer software for the analysis of enterohaemorrhagic *Escherichia coli* pulsed-field gel electrophoresis patterns. *Epidemiol and Infection* **134**: 1004–1014.

Zerr, T. and Henikoff, S. (2005). Automated band mapping in electrophoretic gel images using background information. *Nucleic Acids Research* **33**: 2806–2812.