Automatic Method to Compare the Lanes in Gel Electrophoresis Images

Chih-Yang Lin, Yu-Tai Ching, *Member, IEEE*, and Yun-Liang Yang

***Abstract*—Gel electrophoresis (GE) is an important tool in genomic analysis. GE results are presented using images. Each image contains several vertical lanes. Each lane consists of several horizontal bands. Two lanes are identical if the relative positions of the bands are the same. We present a computer method designed to compare the lanes and identify identical lanes. This method, developed using many image-processing techniques, is applied to segment the lanes and bands in GE images. The lanes are then converted into "position vectors" that describe the positions of the bands. Comparing lanes becomes equivalent to comparing the position vectors. This method can accurately identify identical lanes, helping biologists to identify the identical lanes from many lanes with much less effort.**

***Index Terms*—Dynamic programming, gel electrophoresis (GE), lane comparison, matched filter, segmentation, watershed method.**

I. Introduction

GEL ELECTROPHORESIS (GE) was developed as a means for separating biological macromolecules, such as DNA, RNA, and protein molecules [1]. There are several different types of GE based on their resolution ranges. One major application of GE is to separate DNA molecules from 0.5 kbp to approximately 10 Mbp or larger. GE is an invaluable tool for gene and genomic analysis and it is routinely used in many applications, such as gene identification, isolation, and purification. GE is used in various fields, such as biology, molecular biology, biochemistry, biotechnology, medicine, and clinical diagnosis.

This technique produces images that consist of several vertical *lanes*, each lane corresponding to one sample. Each lane contains a number of horizontal *bands*. Each band represents a part of the sample. The positions of the horizontal bands in the lane represent the molecular weights of that part of the sample. Two samples are considered to be the same if their lanes have the same pattern. In this paper, we present a computer method that automatically identifies lanes with the same pattern among many lanes.

Previous work regarding the study of this problem can be found in [2]–[8], the lanes in a GE image were first segmented and converted into a chain code representation. The lane comparison was performed by calculating the longest common subsequence (LCS) in two chain codes. This method did not segment the bands in each lane so it could not produce an exact comparison result. It could only eliminate those very different lanes and reduce the number of lanes to be compared. In [3], a clustering algorithm for detecting lanes in GE images was proposed. In this method, only the lanes were segmented. The accuracy was less than 70%. In [4], an iterative algorithm for segmenting lanes in GE images was proposed. This method improved the accuracy for lanes segmentation to 96%. The bands were not segmented in this method. There are four existing software packages for GE band extraction [5]–[8]. None of them provides a "one click" operation to do everything including accurate lanes and bands segmentation, as well as lane comparison. Many user interfaces such as lane selection are required.

In this paper, we present a method that accurately identifies identical lanes. Segmentation is the first task. Lanes and bands segmentation is difficult due to the quality of the GE images. The images acquired in our system contain a grid texture that is contaminated during the images acquisition step. Furthermore, there are many factors, such as the applied voltage, field strength, pulse time, reorientation angle, agarose type, concentration, and buffer chamber temperature, that affect the image quality and the patterns in the lanes [9], [10]. Especially the agarose type affects the lane pattern significantly. Agarose is a polysaccharide consisting of 1,3-linked-beta-d-galactopyranose and 1,4-linked 3,6-anhydro-alpha-L-galactopyranose. The basic repeat unit forms long chains with an average mass of 120 kD. The gelation process eventually leads to formation of mass fractals like a spider web with a pore size about 52 ± 5 nm when the concentration of agarose is between 0.7% and 1.5%, the range of conventional applications. With the increase of concentration, the pore size reduces. In electrophoresis, DNA or other charged molecules are forced to move through the maze formed by the polymers. The mobility is guided by two factors, the mass and the shape of the molecules. The smaller the mass, the fast it moves. With a homogenous sample, such as DNA, the mass is the dominant factor. However, due to the flexibility of the biological molecules, the shape is not consistently maintained rigidly without any change. Therefore, as the samples move farther away from the original starting point, the effects of the shape on the molecules starts to show and the bands become blur. Thus, the segmentation task is difficult. To overcome this problem, we present a sequence of procedures, including the time-variant matched filter, watershed, and dynamic programming techniques, to segment the bands and lanes. Each lane is then converted into a normalized "position vector" denoted PV, that specifies the positions of the bands in the lane. Two lanes

Manuscript received October 28, 2002; revised August 13, 2005 and January 10, 2006. This work was supported by the National Science Council, Taiwan, R.O.C., under Grant NSC 93-2213-E-009-037 and in part by the NCHC.

C.-Y. Lin is with the Department of Electrical Engineering, Ta-Hwa Institute of Technology, Hsinchu 307, Taiwan, R.O.C (e-mail: richard@cs.nctu.edu.tw).

Y.-T. Ching is with the Department of Computer Science, National Chiao-Tung University, Hsinchu 300, Taiwan, R.O.C. (e-mail: ytc@cs.nctu.edu.tw).

Y.-L. Yang is with the Department of Biological Science and Technology, National Chiao-Tung University, Hsinchu 300, Taiwan, R.O.C. (e-mail: yyang@cc.nctu.edu.tw).

Digital Object Identifier 10.1109/TITB.2006.875661

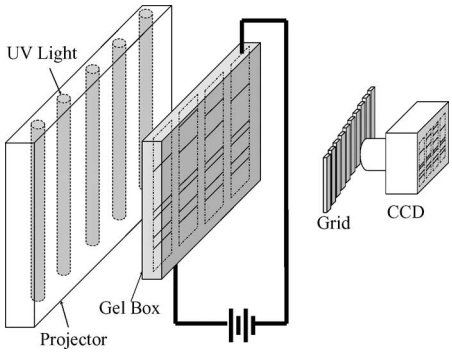


Fig. 1. Imaging system. There is a grid between the CCD camera and the gel box.

that have the same PVs have the same pattern. All these operations can be done with the least possible human intervention.

The first step in the proposed method is the preprocessing step. The grid texture is first eliminated. The background is then removed. In the next step, the bands are enhanced. The bands and the lanes are then extracted. The positions of the bands are normalized and converted into PVs. Comparison of lanes then becomes a comparison of the PVs of the lanes. The proposed method is described in Section II. The results are shown in Section III. In Section IV, we present the accuracy verification of the method. A software tool that was developed based on the proposed method is briefly described in Section V. An application to a biological study is presented in Section VI. Finally, we have the conclusion and discussion in Section VII.

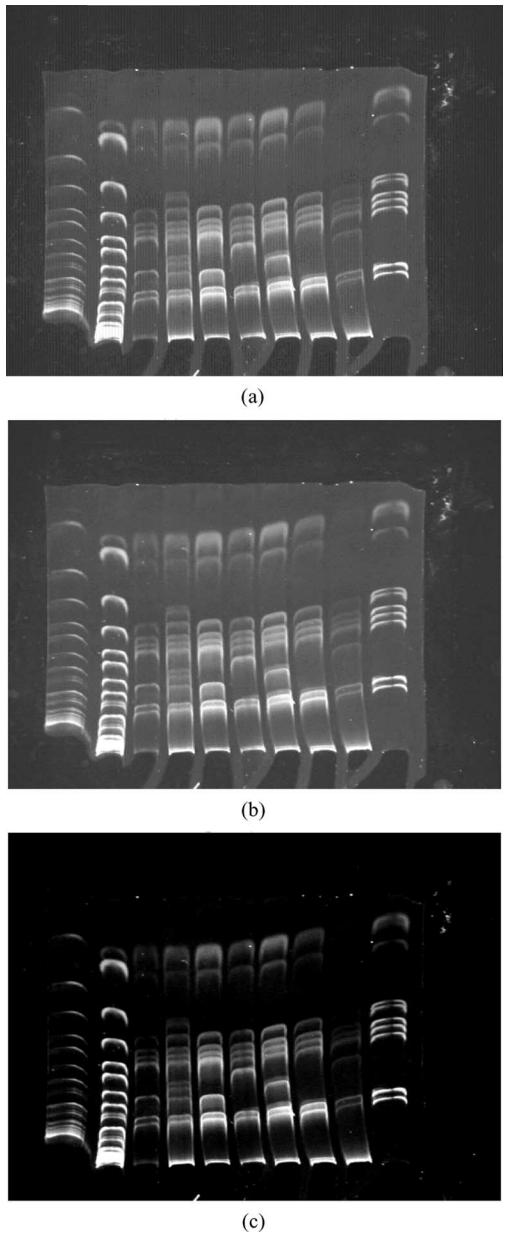
II. Method

A. Preprocessing

There are two tasks in the preprocessing step. The first is the grid-texture removal. The grid-texture is contaminated from the image acquisition system. Fig. 1 shows a typical GE image acquisition system. Because the bands and lanes in gel are not visible under visible light, the gel box is illuminated using a fluorescent ultraviolet (UV) light source. There is a grid located between the gel box and the charge-coupled device (CCD) camera. The grid is used to collimate the light to prevent scattering. The grid improves the sharpness of the GE images by trapping most of the scattered light but also causes grid-texture artifacts in the GE image.

Fig. 2(a) shows one of the gel images containing grid-texture. The grid texture has a fixed frequency in the frequency domain, so it can be removed easily from the frequency domain. Let $f(x,y)$, $0 \le x \le M-1$ and $0 \le y \le N-1$, denote an M by N GE image. Let $f_r(x)$ denote a row in $f(x,y)$. The one-dimensional (1-D) discrete Fourier transformation pairs for $f_r(x)$) are

$$F_r(u) = \sum_{k=0}^{M-1} f_r(x) \exp(-j2\pi kx/M), \quad 0 \le u \le M-1 \quad (1)$$



(a) (b) (c)

Fig. 2. One of the gel images. The shape of the lane is distorted. (a) The original image containing a grid artifact. (b) The same image with the grid textures removed from the frequency domain. (c) The background is removed.

$$f_r(x) = \sum_{k=0}^{M-1} F_r(u) \exp(j2\pi kx/M), \quad 0 \le x \le M-1. \quad (2)$$

The spectrum is

$$|F_r(u)| = (F_r(u)F_r^*(u))^{0.5} \quad (3)$$

where * means conjugate.

The power spectrum is shown in Fig. 3(a). In this figure, a double-sided spectrum was used. The left half is the complex conjugate reflection of the right half. The grid-texture in the spatial domain is transformed into a specific frequency that causes observable peaks on both sides within the vertical line

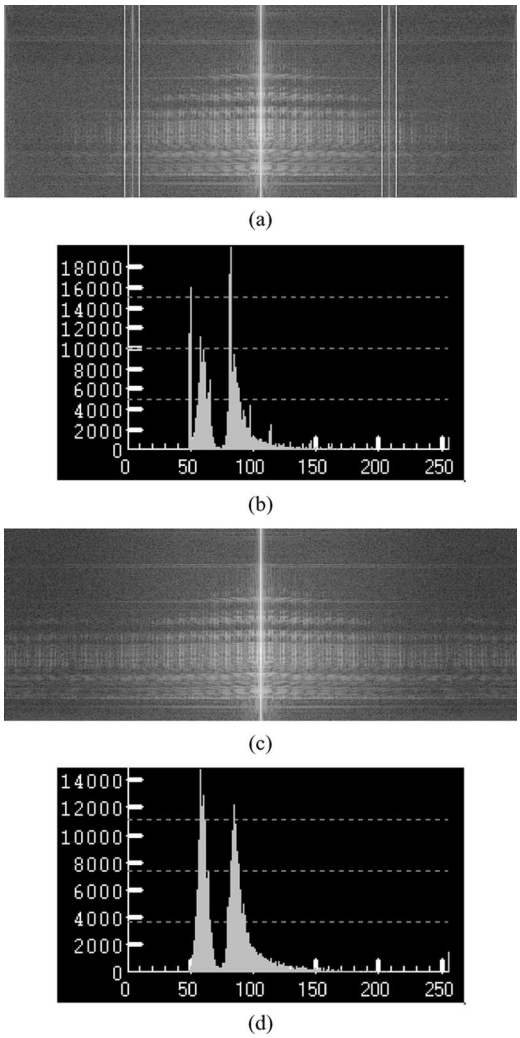


Fig. 3. Spectra and histogram of Fig. 2(a) in the frequency domain. (a) The spectrum of $f\ (x, y)$. There are two peaks within the two pairs of lines. The gray scale is logarithmic for visualization purposes. (b) Histogram of $f\ (x, y)$. (c) The spectrum after eliminating the grid-texture frequency. The gray scale is also logarithmic. (d) Histogram after eliminating the grid texture.

pairs [Fig. 3(a)]. The power of the peak frequencies is many times that of the other frequencies except for those near the dc term. Such peaks f_t can be easily located in the spectrum. We approximate the spectrum of the grid artifact by a Gaussian distribution (4)

$$G(u) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(u-\mu)^2}{2\sigma^2}\right). \tag{4}$$

By estimating μ and σ of the Gaussian distribution, we can construct the band-stop Gaussian filter $B(u)$. μ should be equal to the peak frequencies f_t. σ is obtained by calculating the standard deviation over the interval from $f_t - (f_t/10)$ to $f_t + (f_t/10)$ in the spectrum as shown in (5)

$$\sigma = \frac{\sum_{i=1}^{N} |F_r(i) - \mu|}{N}, \qquad F_r(i) \in f_t - (f_t/10) \text{ to } f_t(f_t/10). \tag{5}$$

Substituting the mean and the standard deviation, we have the Gaussian band-stop filter shown in (6)

$$B(u) = \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(u-\mu)^2}{2\sigma^2}\right), \quad u = 1 \ldots M. \tag{6}$$

We multiply $F_r(u)$ by $B(u)$to obtain $F_r'(u)$ shown in (7)

$$F_r'(u) = F_r(u)B(u), \qquad u = 1 \ldots M. \tag{7}$$

Taking the 1-D inverse Fourier transform (2) of $F_r'(u)$, we obtain a grid-texture free row, $F_r'(x)$. The process is applied to each row in $f(x, y)$ to obtain an image, $f'(x, y)$. $f'(x, y)$ is free of grid-texture, as shown in Fig. 2(b). Fig. 3(a)–(d) shows the spectrums and histograms before and after removing the grid-texture frequency.

The second task in the preprocessing step is to set the intensities of the pixels not in the bands to zero. These background pixels generally have a lower intensity than the pixels in the bands. In Fig. 3(d), the histogram approximately consists of two normal distributions. A threshold is set to be the closest gray-level corresponding to the minimum probability between the maxima of two normal distributions, which results in minimum error segmentation. Such optimal threshold was solved by using the method proposed by Glasbey [11]. The result after the preprocessing step is shown in Fig. 2(c).

B. Lanes and Bands Segmentation

Some observable properties of bands and lanes are presented before presenting the proposed method.

1. The bands closer to the top are wider than those closer to the bottom of the image.
2. The shapes of the bands are concave downward and similar in the same lane.
3. The bands in different lanes in the same image may have different shapes.
4. A band could break into several fragments due to noise.

The method for segmenting bands and lanes was designed based on the properties stated above. Bands and lanes segmentation consists of several steps. The first step is to enhance the bands. The skeleton for each band is then found. Lane segmentation is based on the band skeletons.

To enhance the blurred bands, the matched filter technique [12]–[14] is employed. Matched filter is designed based on the shape and intensity distribution of the bands. A large response can be obtained if the matched filter is applied to a place where there is a band. The intensity profile along the vertical line (y-direction) passing through a lane is shown in Fig. 4.

A band profile is bell-shaped and can be approximated using a Gaussian distribution in the y direction, as shown in (8)

$$D(y) = e^{\frac{-y^2}{2\sigma^2}}, \qquad -\infty \le y \le \infty. \tag{8}$$

Because the bands are not horizontal line segments but rather concave downward curves, a rectangular-shaped (two-dimensional) 2-D matched filter shown in (9) is needed

$$D(x, y) = e^{\frac{-y^2}{2\sigma^2}} - \frac{d_y}{2} \le y \le \frac{d_y}{2}, -\frac{d_x}{2} \le x \le \frac{d_x}{2}. \tag{9}$$

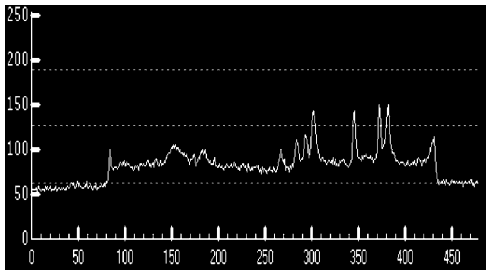


Fig. 4. Intensity profile of a vertical scan line passing through the middle of a lane.

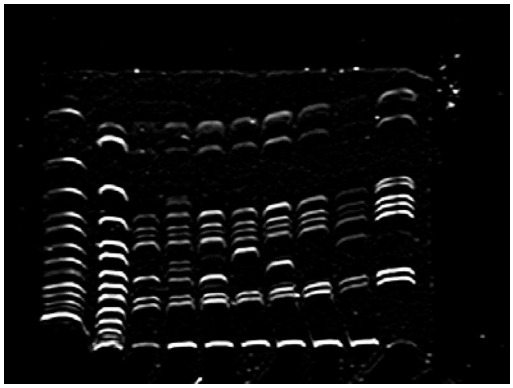


Fig. 5. Result from applying the matched filters.

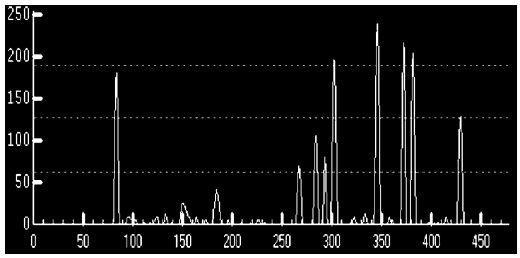


Fig. 6. Intensity profile of a lane scan line after filter matching.

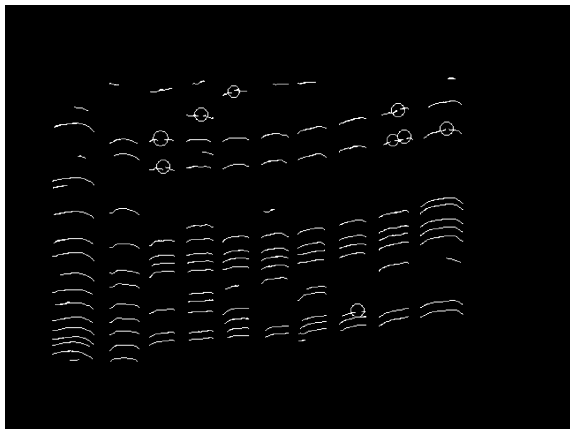


Fig. 7. Result from applying the 1-D watershed segmentation algorithm to Fig. 5. The break points are highlighted with the circles.

Two parameters, the width d_x and the height d_y, for the match filter must be determined. In determining the width d_x, since the bands are not perfectly straight lines, d_x should not be as wide as the length of the bands. In our experiment, $d_x = 5$ is an appropriate value for most of the cases. The height d_y of the matched filter depends on the variance σ^2 in (9). Since the bands closer to the top are wider than those closer to the bottom of the image, σ should vary depending on the location of the band. In the case of different σ, a time-variant matched filter is needed. We set σ as a linear function of y, as shown in (10)

$$\sigma = 1 + c * y/N \tag{10}$$

where y is the distance between the band and the bottom side of the image. For a small σ, the Gaussian quickly drops to zero and so d_y should be small. Conversely, for a large σ, the Gaussian slowly becomes zero and so d_y should be large. We used the method in [15] to determine d_y from a given σ. The height is between $-(4\sigma + 3)/2 \leq y \leq (4\sigma + 3)/2$. Based on the above discussions, the matched filter is shown in (11)

$$D(x,y) = e^{\frac{-y^2}{2\sigma^2}} - \frac{d_y}{2} \leq y \leq \frac{d_y}{2}, \quad -\frac{d_x}{2} \leq x \leq \frac{d_x}{2}, \quad \sigma = 1 + c * y_1/N. \tag{11}$$

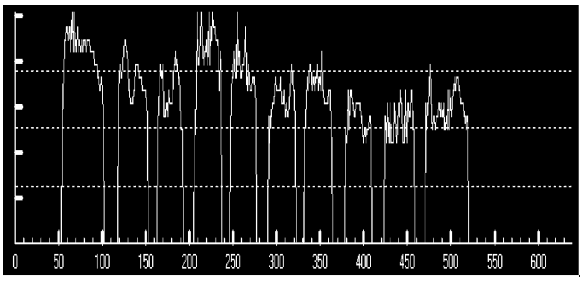
where $d_y = 4(\sigma + 3)$ and $d_x = 5$. Images convolved with the matched filter have the bands enhanced. The result after applying the matched filters is shown in Fig. 5. The intensity profile along a vertical line passing through a lane is shown in Fig. 6. Compared to the image shown in Fig. 4, the bands in Fig. 6 are much easier to identify.

In Fig. 6, the peak of a bell-shaped curve is a point on the centerline of a band. Thus, the center of a band can be found by determining the local maxima (peaks) on the profile. The intensity threshold is not applicable because the peaks do not have the same height. The *watershed algorithm* [16], [17] was used to segment the peaks.

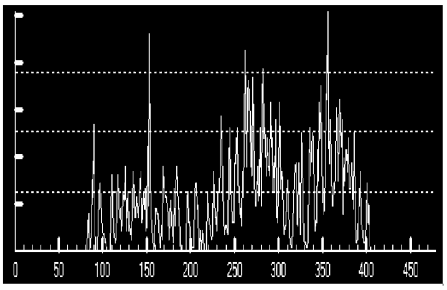
The 1-D watershed algorithm is applied to determine the peaks of all the vertical scan lines in the image. Since the peaks are candidate points on the center line of bands, the peaks tend to form a set of connected components. Among all of these connected components, most are at the center of the bands while some are just noise. A size filter with a proper threshold is then applied to remove small connected components that are generally noise. The threshold was set to 3. The resulting image is a binary image, as shown in Fig. 7.

To segment the lanes, three parameters $\mathrm{TH_{LO}}$, $\mathrm{TH_{HI}}$, and $\mathrm{TH_{Lane}}$ are set. $\mathrm{TH_{LO}}$ is the smallest possible number of bands in a lane. This parameter is set to remove empty lanes from the image. $\mathrm{TH_{HI}}$ is a parameter that represents the largest possible number of bands on a lane. The parameter $\mathrm{TH_{Lane}}$ defines the smallest lane width.

We count the number of nonzero pixels on each column in Fig. 7 to obtain the profile shown in Fig. 8(a). In Fig. 8(a), the high-rising portion along the curve corresponds to a lane. To segment the lanes, we start with the threshold $\mathrm{TH_{LO}}$. The horizontal line $y = \mathrm{TH_{LO}}$ cuts the curve into several connected components formed by the high-rising parts. A connected component is considered a lane if the width of the connected component is greater than $\mathrm{TH_{Lane}}$. The number of lanes obtained



(a)



(b)

Fig. 8. (a) Vertical projection of the points in Fig. 7. (b) Horizontal projection of the points in Fig. 7.

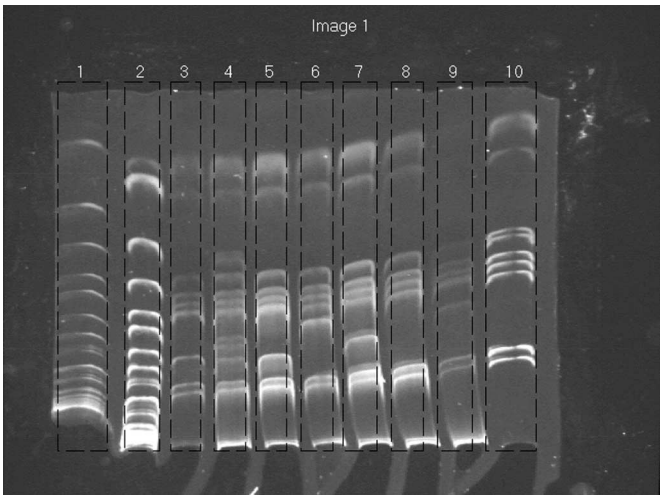
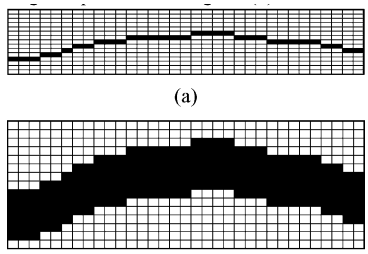


Fig. 9. Result of segmentation of lanes.

is then counted using this given threshold. The threshold is increased and the steps iterated until the threshold reaches TH_{HI}. The number of lanes obtained is counted each time. The largest threshold that maximizes the number of lanes is the best threshold.

To determine the top and bottom lane boundaries, we count the nonzero pixels along each row in Fig. 7 to obtain a profile shown in Fig. 8(b). A threshold is set so that the left-most and right-most peaks satisfy the thresholds corresponding to the top and bottom sides. The threshold was set to 20% of the mean of the histogram. The lane segmentation result is shown in Fig. 9.



(a)

(b)

Fig. 10. (a) Average shape of the bands in a lane. (b) Dilation of (a).

C. Calculate the Lane PVs

On each segmented lane, the bands in the lane will be converted into horizontal line segments. A normalized PV that describes the position of the bands will then be established. The first step is to recover the broken bands by connecting the break points. The break points are highlighted by the circles in Fig. 7. Break points are the boundary points of connected components. They can be easily determined. Recall that the shapes of the bands in a lane are similar. An "average shape" of the bands in a lane can be calculated. A weighted directed graph based on the average shape is then created. Recovering the bands becomes a problem of finding the shortest path in the weighted graph. The average shape in a lane is presented by a mask containing binary values. The method for calculating the average shape is stated in the following.

Suppose that there are n vertical lines $l_i, i = 1, \ldots, n$ from left to right passing through a lane. Let S_i be the set of intersection points of l_i and the skeletons of the bands. Consider a pair of intersection points p and q on a band, p in S_i and q in S_{i+1}. $\overrightarrow{p,q}$ points in one of the following directions: northeast, east, or southeast. The "average direction" from the points in S_i to the points in S_{i+1} is determined using the majority of the directions from the points in S_i to the points in S_{i+1}. The average shape is the sequence of points determined by the sequence of major directions. An average shape of the bands in a lane is shown in Fig. 10(a). The dilation operation was applied to the skeleton in Fig. 10(a) to obtain a template mask for the average shape shown in Fig. 10(b).

The width of the template mask is the same as the lane width. If we slide the template mask in the lane to a place where there are an even number of break points covered by the mask, a weighted graph $G = (V, E)$ is established.

Let s and t be the break points in S_i and in $S_j, i \leq j$, to be connected. Let $V_k, k = i+1, \ldots, j-1$, be the set of points on l_k covered by the template mask. V is the set of vertices corresponding to the points in the sets $V_k, k = i+1, \ldots, j-1$. E is the set of edges that is the union of the following three sets of edges, which are shown in (12) at the bottom of the next page. There are weights on the vertices. The weight of a vertex is the inverse of the intensity of its corresponding pixel in the image after the matched filter enhancement (Fig. 5). Given the weighted directed graph G, we can find the shortest path from s to t [18]. Connecting s and t using the shortest path recovers the broken band. The result is shown in Fig. 11.

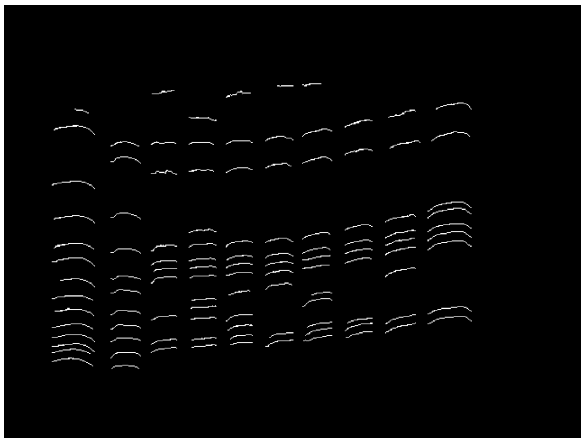


Fig. 11. Broken band result from recovering and removing the areas that are not in the lanes.

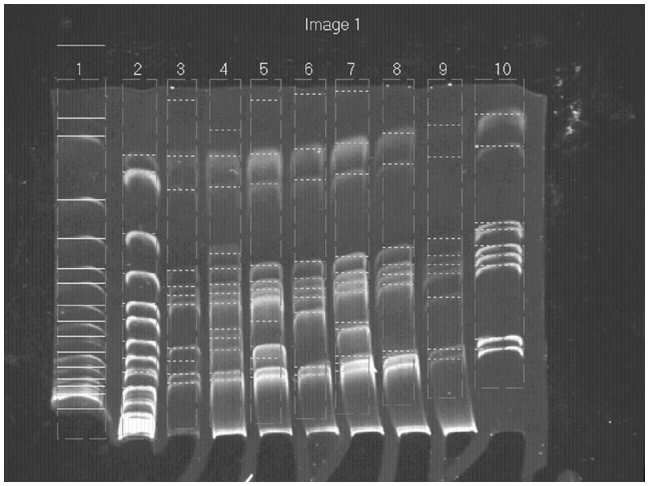


Fig. 12. Result is shown superimposed on the original image.

Each band is converted into a horizontal line segment that passes through the middle point of a recovered band. The result is shown in Fig. 12. The position of the horizontal line segment is regarded as the position of the band. The PV of a lane is obtained.

D. PVs Normalization

Since there are many factors that affect the relative position between bands and the height of the lane, the PVs of two identical lanes can be very different. Before the lanes can be compared, the PVs of the lanes must be normalized. There are two parameters that need to be estimated for PV normalization: the "offset" o, and the "scaling factor" s. The offset and scaling factors for the jth lane in the ith image are denoted o_{ij} and s_{ij}.

All of the images contain a "marker" (the left-most lane) and a "vector" (the right-most lane) that provide important information for PV normalization. A marker is a characteristic of a molecule or cell that is known. The vector is used to refer to a carrier nucleic acid molecule into which a nucleic acid sequence can be inserted for introduction into a cell where it can be replicated. All of the lanes except the marker contain the sample of the vector. We propose to use the marker as a reference to justify the variations between two images, i.e., the marker is used to normalize the PVs between images to obtain o_i and s_i. Because all of the lanes contain the samples in the vector, the vector is used to normalize the lanes within an image.

The procedure starts with normalizing a PV of the marker in an image to the range 0–999 so that the bottom and top sides are set at 0 and 999, respectively. For all the other markers in image i, o_i and s_i are calculated by using the algorithm stated in the following and applied to adjust image i. In image i, the vector is normalized to the range 0–999. All the lanes are then normalized by using the same method.

For a PV, $\mathbf{v} = \langle b_1, b_2, \ldots, b_m \rangle$, the operations "$+$" and "$*$" are defined as follows.

1) Let o be a constant, $\boldsymbol{v} + o = \langle b_1 + o, b_2, +o, \ldots, b_m + o \rangle$.
2) Let s be a constant, $\boldsymbol{v} * s = \langle b_1 * s, b_2 * s, \ldots, b_m * s \rangle$.

Given a PV, $\boldsymbol{v}_{\mathrm{i}}$, the jth band in $\boldsymbol{v}_{\mathrm{i}}$ is denoted as b_{i_j}. Given two PVs, $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, two bands b_{1_i} and b_{2_j} are matched if $|d(1_i, 2_j)|$ is less than a given threshold where $d(1_i, 2_j)$ is the difference between b_{1_i} and b_{2_j}. A best match for v_1 and v_2 is obtained using an offset o, such that the number of the matched bands of v_1 and $v_2 + o$ is maximized and the sum of the differences $|d(1_i, 2_j)|$ is minimized. If scaling is allowed, the best match for v_1 and v_2 is determined by o and s such that the number of matched bands for v_1 and $v_2 * s + o$ are maximized and the sum of the differences is minimized. To determine the best possible combination of s and o, a brute force approach was used that evaluated all possible combinations of s and $o \cdot s$ is a real number ranging between $1000/N$ and $3000/N$ (N is the number of pixels in the y-direction). The interval is $0.01 \cdot o$ is a set of real numbers between -500 and 500. The interval is 1. Recall that bands closer to the top have a larger variance than bands closer to the bottom. The tolerance for a match varies according to the position of the bands. The following error tolerance function was used to define the threshold:

$$\mathrm{th} = \mathrm{th}_M + W * \frac{b_{l_i}}{1000}. \tag{13}$$

The error tolerance varies from top to bottom and ranges from th_M to $\mathrm{th}_M + W$. We used 14 and 15 for th_M and W, respectively, in our experiment.

$$\begin{cases} \{\langle s, v\rangle \mid v \in V_{i+1}\}, \\ \{\langle v, t\rangle \mid v \in V_{j-1}\}, \text{ and} \\ \left\{\left\langle v_{(p,q)}, v_{(p+1,q-1)}\right\rangle, \left\langle v_{(p,q)}, v_{(p+1,q)}\right\rangle, \left\langle v_{(p,q)}, v_{(p+1,q+1)}\right\rangle : v_{(x,y)}\right. \\ \left.\text{is the } y\text{th vertex in } V_x \text{ and } p = i+1, \ldots, j-2\right\} \end{cases} \tag{12}$$

III. Results

The band and lane segmentation results were shown in the previous section. In this section, the vector removal results are presented first. The lane and band segmentation result using a PFGE with an *Escherichia coli* genomic sample image is also presented. The lane comparison result is presented last.

The vector shown in the last lane is common to all other lanes (except the Marker lane). Removing the vectors from a lane produces a new lane that contains only the samples of interest. The threshold defined in (13) is used to identify the bands in a lane belonging to the vector. Three vector removal results are shown in Fig. 13. After the vector is removed, only a very few bands are left. The resulting image makes the comparison job for the biologist easier.

The proposed method was also applied to a PFGE image. This kind of image has a different appearance. The image was obtained from an *E. coli* genomic study experiment. The set of images is shown in Fig. 14(a)–(e). Fig. 14(a) shows the original image. In Fig. 14(b), the image was obtained by applying a time-variant matched filter. The watershed algorithm was applied to Fig. 14(b) to obtain Fig. 14(c). Fig. 14(d) shows the result from superimposing the segmented results onto the original image. Note that the bands in the fourth lane from the left are hard to recognize by a human. The computer method can still segment the bands well.

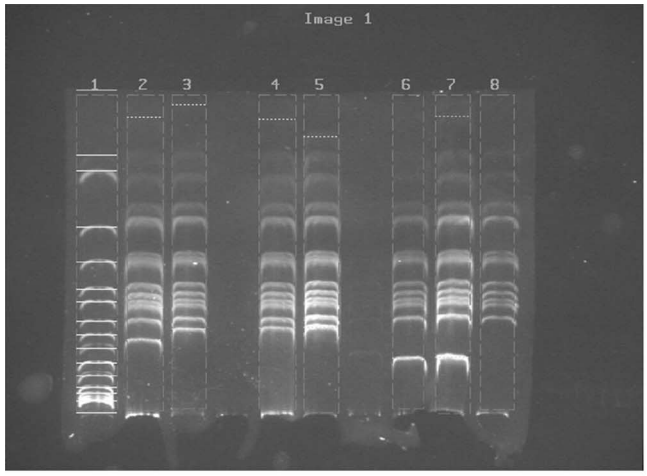
The main goal of this work was to identify identical lanes. Two lanes are identical if they have an identical PV. After comparing all pairs, our method generates a report, shown in Fig. 15. In this report, the number of different bands between the lanes is listed.

In Fig. 15, "*i*-*j*" means the jth lane in image i. The report shows that "1-5," "2-3," and "3-6" are exactly the same and "2-8" and "3-8" are also identical. "1-2@3-3=1" means that there is one different band between "1-2" and "3-3". "1-4@3-3=1,3-4=1" means that there is one different band between the pairs "1-4" and "3-3," as well as "1-4" and "3-4."

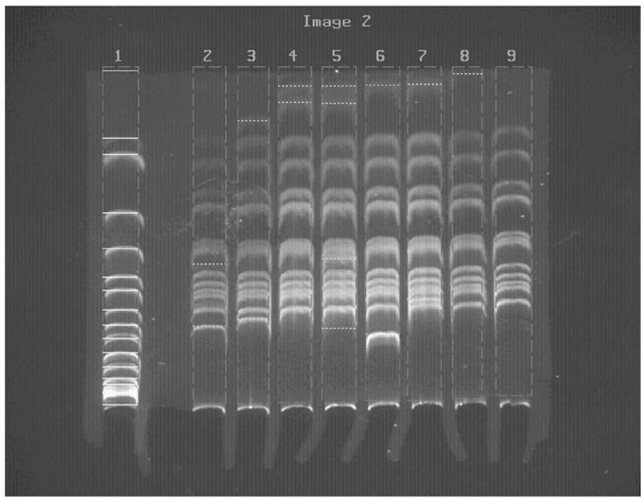
IV. Accuracy Verification

We verified the accuracy of the bands segmentation by comparing the results obtained by the proposed method, a commercially available software (Molecular Analyst Software [8]), and an experienced biologist. Molecular Analyst Software [8] has been developed by Bio-Rad Laboratories.

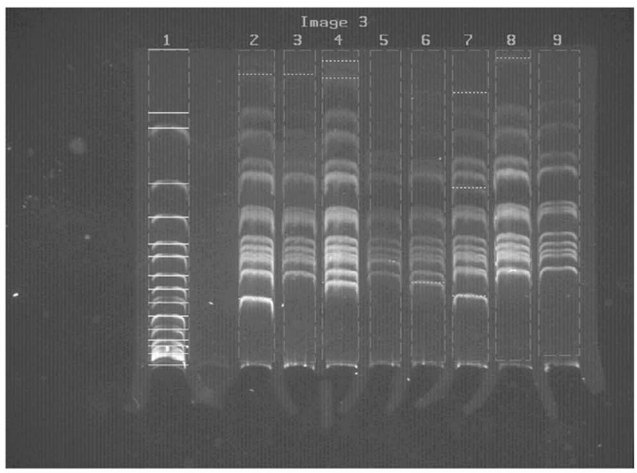
We randomly selected ten images. There were a total of 774 bands in these ten images. A human expert manually drew horizontal bars on the bands. These images were also segmented by using the proposed method and Molecular Analyst Software. The human expert then reviewed the results obtained by the proposed methods. Both the human expert and the proposed methods made mistakes. In the proposed method, the computer interpreted some artifacts as bands. However, the computer was also much more sensitive to the bands than the human expert. There were cases that the human expert agreed with the computer after reviewing the computer's results. These cases occurred mostly when there were light bands.



(a)



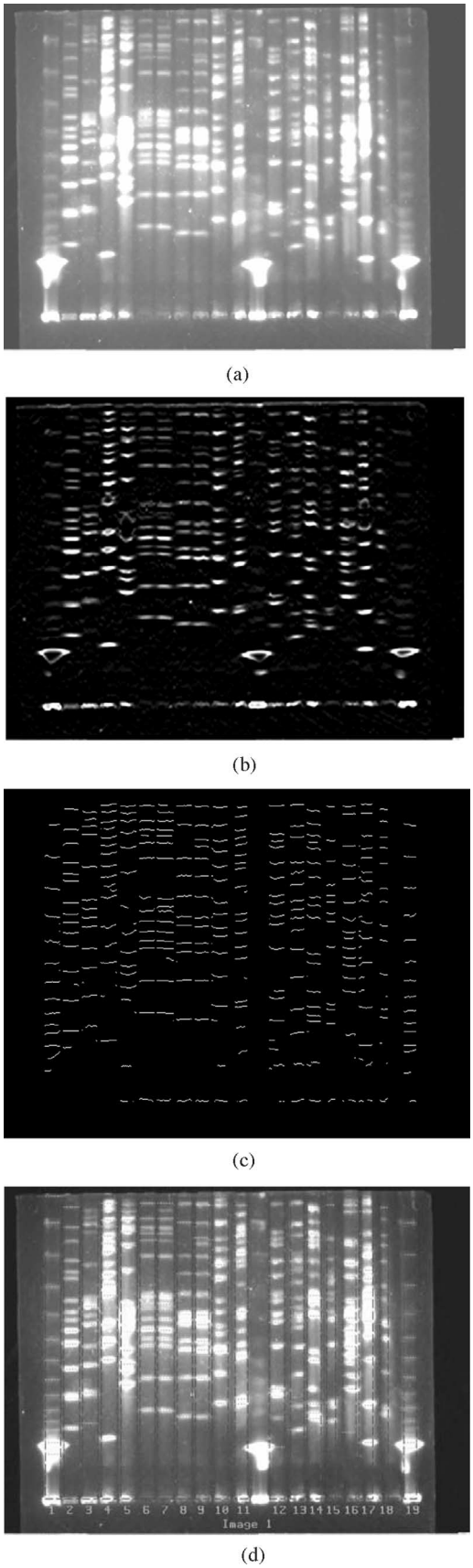
(b)



(c)

Fig. 13. Vector is removed. The samples of interest are shown using bright bars superimposed onto the original image.

The comparison results are shown in Table I. For each band detected by the computer methods but not by the human expert, the human expert had a false negative case. Otherwise the computer had a false positive case. In contrast, for each band



(a) (b) (c) (d)

Fig. 14. (a) Original image. (b) Result from applying a time-variant matched filter. (c) Result from applying the 1-D watershed segmentation algorithm to (b). (d) The segmented result is shown superimposed onto the original image.

```
less than 0 differenc(es) :
1-5,2-3,3-6,
2-8,3-8,

less than 1 differenc(es) :
1-2 @ 3-3=1,
1-4 @ 3-3=1,3-4=1,
1-5 @ 2-3=0,3-6=0,
1-6 @ 3-5=1,
2-3 @ 1-5=0,3-6=0,
2-4 @ 2-8=1,3-2=1,3-3=1,3-4=1,3-8=1,
2-6 @ 3-2=1,3-8=1,
2-7 @ 2-8=1,3-8=1,
2-8 @ 2-4=1,2-7=1,3-5=1,3-8=0,
3-2 @ 2-4=1,2-6=1,
```

Fig. 15. Report generated by the proposed method.

TABLE I
COMPARISON OF THE BAND SEGMENTATION RESULTS OBTAINED BY THE PROPOSED METHOD, A HUMAN EXPERT, AND MOLECULAR ANALYST SOFTWARE

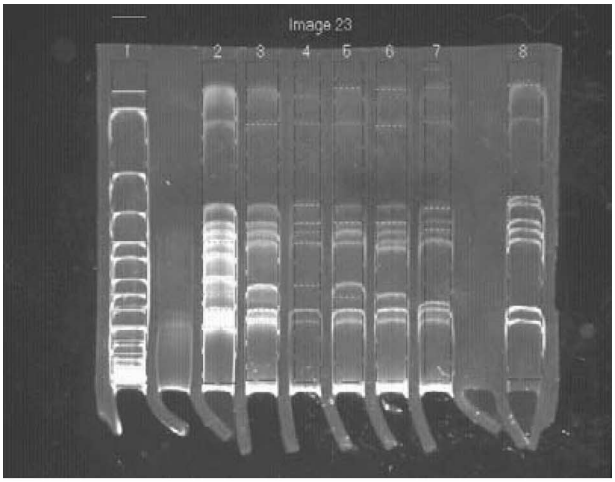
	# of Bands	False Positive	False Negative	Total Error	Error Rate
Proposed Method	774	10	10	20	2.6%
Experienced Human Expert	774	3	19	22	2.8%
Molecular Analyst Software	774	44	27	71	9.2%

detected by the human expert but not by the computer methods, the human expert had a false positive case if the human expert agreed with the computer. Otherwise, the computer had a false negative case. The error rate was obtained by dividing the total errors by the total number of bands. The error rate for band detection was 2.6% using the proposed method. The error rate for band detection was 9.2% using the Molecular Analyst Software. Fig. 16 shows segmentation results obtained by using the proposed method and the Molecular Analyst Software. There are many bands detected by using the proposed method but not by the Molecular Analyst Software. The proposed method could achieve better accuracy. With regards to the accuracy of the lane segmentation, all lanes detected by the human expert were detected by the computer methods. The error rate for lane detection was 0% using the proposed method.

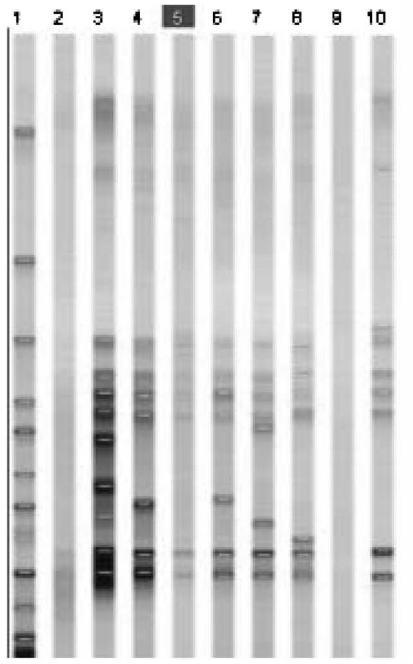
Comparing the proposed method with the method using the Molecular Analyst Software, there were many user interventions involved when the Molecular Analyst Software was used to segment bands. The proposed method provides a "one click" operation to do everything including lanes and bands segmentation, as well as lane comparison. It took more than 1 h to segment the bands in the ten images using Molecular Analyst Software for comparison purpose in the accuracy analysis. The proposed method segmented the bands in more than 1000 lanes and identified the identical lanes in less than 4 min. The proposed method saves biologist effort greatly.

V. SOFTWARE TOOL

A software tool based on the proposed method was developed. The software tool provides a "one click" operation; i.e.,



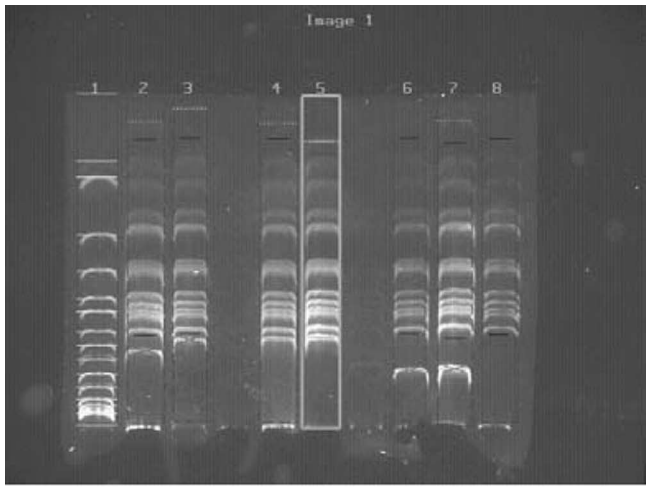
(a)



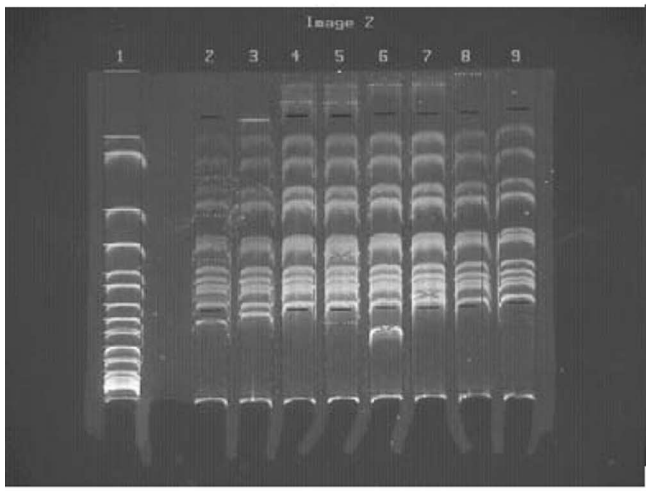
(b)

Fig. 16. (a) Segmentation result obtained by using the proposed method. (b) The segmentation result of the same image obtained by using the Molecular Analyst Software. There are bands detected by the proposed method but not by the Molecular Analyst Software.

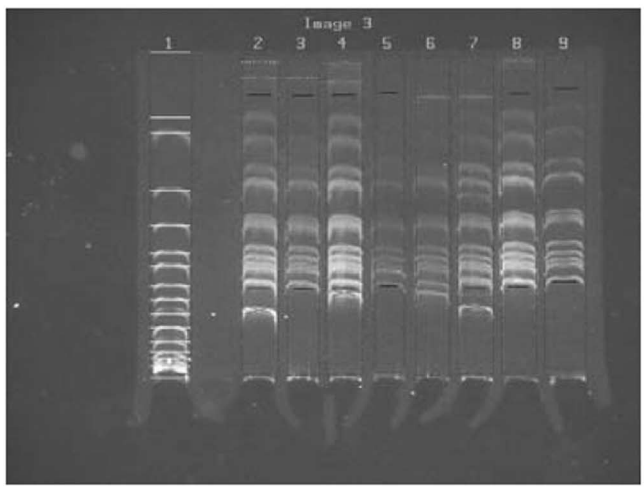
click a "Go" button, all images are processed and the report shown in Fig. 15 is generated. Since the computer could make a mistake, the software tool provides operations for the biologist to insert bands or delete detected bands. This software system also provides a way to graphically display the differences. One can use a mouse to select a lane. The software system will show the differences between the other lanes and the selected lane in the images. The matched bands are shown with bright bars. The bands that do not match are shown with an X over the bands. If there are bands on the selected lane that do not appear in the other lanes, these bands are shown by a dark bar. For example, in Fig. 17, lane 5 in image 1 is selected. The selected lane is highlighted with a bold yellow rectangle. Bands in all of the



(a)



(b)



(c)

Fig. 17. Software system shows the differences between the selected lane and the other lanes in the same or different images. (a) Using the mouse to select lane 5. (b) Lane 3 is identical to lane 5 in (a). (c) Lane 6 is also identical to lane 5 in (a).

other lanes are marked by bright or dark bars or an X. We can easily see that "1-5," "2-3," and "3-6" are exactly the same. For the other lanes, because there are marks on the bands, we

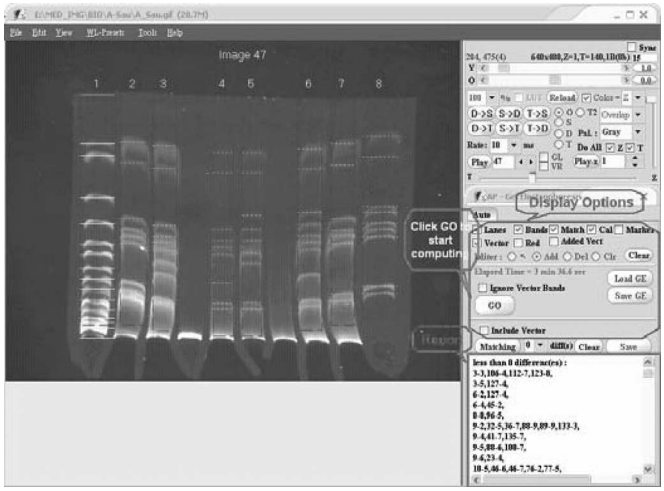


Fig. 18. Click "Go" to start the computation including lane and band segmentation, as well as lane comparison.

can easily verify the correctness of the results obtained by the computer. Using the editing tool provided by the computer, we can modify the computer's segmentation results.

There are parameters, for example, d_x in (11), the constant c in (11), and the threshold for the size of components, that can be inputted to the software system. These parameters need to be adjusted only once for a study. A screen shot of the developed software system is shown in Fig. 18. The comparison result can be seen in the lower right corner in the display window of the software.

VI. Application to a Real Case

The proposed method has been applied to analyze *Candida albicans* [19]. *C. albicans* is the most frequently isolated fungal pathogen in humans and has caused morbidity in seriously debilitated and immunocompromised hosts [20]. Coinciding with the increased usage of antifungal agents, the incidences of drug resistance have also increased [22], [23]. The cells of *C. albicans* can switch from the unicellular yeast form into either one of the two distinct filamentous forms, pseudohyphae or hyphae. This ability to switch is associated with its virulence, or its ability to cause diseases (5). Hence, we are interested in knowing the mechanism of this process. The first step is to identify genes involved in the morphogenesis. We have performed transcriptome analysis by means of suppression subtractive hybridization [24] to isolate all cDNA fragments related to the morphogenesis under our experimental condition. In total, more than 1000 cDNA plasmids have been obtained, of which the mRNA levels were different between the unicellular form and the filamentous form. To unveil the DNA sequences in those cDNA fragments and to reduce the cost of sequencing, all the plasmids were then digested with *Hae* III to cleave the cDNA fragments into several smaller ones. Then, the cleaved products were run on agarose gel in electrophoresis to resolve the cleaved DNA fragments. Each unique DNA sequence will produce a unique set of cleaved DNA fragments, known as restriction patterns, which reflect the contents of the DNA sequences. The restriction patterns were then used to classify the cloned cDNA fragments. If the restriction patterns of two cDNA clones were identical, we assumed that they had the same DNA sequences. To facilitate the categorization of the candidates, the software tool was used to replace the time-consuming label-intensive eye-balling process of comparing more than 1000 sets of recoded restriction patterns. Representative candidates from each category were subjected to sequencing and database comparison to reveal the identities of the genes. One of the candidate genes whose expression was higher in yeast form cells than in filamentous form cells was the *ERG 3* gene, known to be important for drug resistance to azoles in *C. albicans*. Hence, the regulatory pathways of morphogenesis and drug resistance are connected to each other.

VII. Conclusion and Discussion

In this paper, an accurate lanes and bands segmentation method for GE images was presented. The method converted a lane to a normalized PV. Comparing the PVs identifies the same lanes among many lanes. The accuracy for lanes and bands segmentations is, respectively, 100% and more than 97%. A software system was developed based on the presented method. The required user interface is reduced to the simplest possible. This software tool was developed on a PC with a Pentium 4 (2.2 GHz) CPU running on the Windows XP operating system. The overall execution time for a 640×480 image took less than 2 s. For a case of 140 images (about 1200 lanes), the total computing time was less than 4 min. The time needed for comparing the PVs can be ignored and the final report can be generated in seconds. This system helps biologists to save a great deal of effort in comparing GE images.

In the normalization step in the presented method, it is required that there is a vector in every image. This is not the case for many other similar experiments using GE technique. Designing a more general method is our future work.

Appendix

Summary of the Presented Method

The method is summarized in the pseudocode.

Procedure FindIdenticalLanes()

{

1. Preprocessing the images, removing the grid-texture and the background;
2. Applying the matched filter to enhance the bands;
3. Applying the watershed algorithm to calculate the centerline of the bands;
4. Broken bands are recovered based on the average shape using the shortest path algorithm;
5. Normalized PVs are computed based on the marker and vector;
6. Generate the comparison result.

}

The images contain the grid textures. The grid texture has a fixed frequency in the frequency domain. Thus, it can be removed easily from the frequency domain. Clean images are obtained so that the optimal threshold can be computed by using the method proposed by Glasbey [10].

Steps 2–5 were designed based on the observable properties of the lanes and bands. If the shape and intensity of the object of interest are known, the matched filter technique is the best to enhance the objects. Since the bands should be horizontal bars but are actually concave downward curves, a rectangular matched filter with smaller width is employed. In Step 3, the watershed algorithm was applied to determine the centerlines of the bands. Since the matched filter has been applied, the watershed method can determine accurate centerlines of the bands. The shape of the bands is similar in a lane. This provides important information to recover the broken bands. In Step 4, we compute the average shape. A weighted directed graph is established based on the average shape and the image after application of the matched filter. We recover the broken band by finding the shortest path in the graph. The bands in a lane are converted into the PV. In Step 5, since we know that there are an identical marker and vector in each image, we use this information to calculate the normalized PVs. Finally, lanes comparison is carried out by comparing the normalized PVs in Step 6.

REFERENCES

[1] J. Sambrook and D. W. Russell, *Molecular Cloning: A laboratory Manual*, 3rd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 2001.

[2] W. Z. Cheng, K. S. Yen, C. Y. Lin, Y. T. Ching, and Y. L. Yang, "Comparing lanes in the pulsed-field gel electrophoresis (PFGE) images," in *Proc. 23rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Oct. 25–28, 2001, vol. 3, pp. 2911–2913.

[3] A. Machado, A. Siqueira, O. Carvalho, and M. Campos, "Automatic lane detection in gel electrophoresis images," Univ. Fed. Minas Gerais, Belo Horizonte, Brazil, Tech. Rep. DCC-013/97, 1997.

[4] A. M. C. Machado, M. F. M. Campos, A. M. Siqueira, and O. S. F. D. Carvalho, "Iterative algorithm for segmenting lanes in gel electrophoresis images," in *Proc. 10th Braz. Symp. Comput. Graph. Image Process*, Oct. 14–17, 1997, pp. 140–146.

[5] *BioNumerics,* BioSystematica, Ceredigion, U.K, 2004.

[6] *Phoretix 1D,* Nonlinear Dynamics Ltd., Durham, NC, 1999.

[7] *1-D AAB,* ABBI Inc., Adv. Amer. Biotechnol. Imaging, Fullerton, CA, 2002.

[8] *Molecular Analyst Software,* Bio-Rad Lab., Hercules, CA, 2000.

[9] R. A. J. Van Daelen and P. Zabel, "Preparation of high molecular weight plant DNA and analysis by pulsed-field gel electrophoresis," in *Plant Molecular Biology Manual*, S. B. Gelvin, R. A. Schilperoort, and D. P. S. Verma, Eds. Dordrecht, The Netherlands: Kluwer, 1991, pp. A15/1–A1/25.

[10] B. Birren, L. Hood, and E. Lai, "Pulsed field gel electrophoresis: Studies of DNA migration made with the programmable, autonomously-controlled electrode electrophoresis system," *Electrophoresis*, vol. 10, pp. 302–309, 1989.

[11] C. A. Glasbey, "An analysis of histogram-based thresholding algorithms," *CVGIP: Graph. Models Image Process.*, vol. 55, pp. 532–537, 1993.

[12] L. A. Wainstein, V. D. Zubakov, and D. Hildereth, *Extraction of Signals from Noise*. Englewood Cliffs, NJ: Prentice-Hall, 1962.

[13] G. L. Turin, "An introduction to matched filter," *IRE Trans. Inf. Theor.*, vol. 6, pp. 311–329, Jun. 1960.

[14] D. Middleton, "On new classes of matched filters and generalizations of the matched filter concept," *IRE Trans. Inf. Theor.*, vol. 6, pp. 349–360, Jun. 1960.

[15] G. Lohmann, *Volumetric Image Analysis*, New York: Wiley, 1998.

[16] J. B. T. M. Roerdink and A. Meijster, "The watershed transform: Definitions, algorithms and parallelization strategies," in *Fundamenta Informatic*. Groningen, The Netherlands: Inst. Math. Comput. Sci. Univ., 2000, pp. 187–228.

[17] L. Vincent and P. Soille, "Watersheds in digital space: An efficient algorithm based on immersion simulations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 6, pp. 583–598, Jun. 1991.

[18] T. Corman, C. Leiserson, and R. Rivest, *Introduction to Algorithms*, 2nd ed. Cambridge, MA: MIT Press, 2001.

[19] H. J. Lo, J. S. Wang, C. Y. Lin, C. G. Chen, T. Y. Hsaio, C. T. Hsu, C. L. Su, M. J. Fann, Y. T. Ching, and Y. L. Yang, "Efg1 involved in drug resistance through regulating the expression of ERG3 in *Candida albicans*," *Antimicrob. Agents Chemother.*, vol. 49, no. 3, pp. 1213–1215, Mar. 2005.

[20] E. J. J. Edwards, "Candida species," in *Principles and Practice of Infectious Diseases*, G. L. Mandell, R. G. Douglas, and J. E. Bennett, Eds. New York: Wiley, 1990, pp. 1943–1958.

[21] H. J. Lo, J. R. Kohler, B. DiDomenico, D. Loebenberg, A. Cacciapuoti, and G. R. Fink, "Nonfilamentous *C. albicans* mutants are avirulent," *Cell*, vol. 90, pp. 939–949, 1997.

[22] M. A. Pfaller, R. N. Jones, G. V. Doern, H. S. Sader, S. A. Messer, A. Houston, S. Coffman, and R. J. Hollis, "Bloodstream infections due to Candida species," *Antimicrob. Agents Chemother.*, vol. 44, pp. 747–751, 2000.

[23] B. H. Vanden, D. W. Warnock, B. Dupont, D. Kerridge, G. S. Sen, L. Improvisi, P. Marichal, F. C. Odds, F. Provost, and O. Ronin, "Mechanisms and clinical impact of antifungal drug resistance," *J. Med. Vet. Mycol.*, vol. 32, pp. 189–202, 1994.

[24] L. S. Diatchenko, L. Y. F. Lau, and P. D. Siebert, "Suppression subtractive hybridization: A versatile method for identifying differentially expressed genes," *Methods Enzymol.*, vol. 303, pp. 349–380, 1999.

Chih-Yang Lin received the M.S. degree in computer science and engineering from Yuan Ze University, Chungli, Taiwan, R.O.C., in 1996, and the Ph.D. degree in computer and information science from National Chiao-Tung University, Hsinchu, Taiwan, in 2005.

From 1988 to 1993, he was a Vice Manager in the R&D Department of Chilong Company Ltd. From 1993 to 1996, he was a Senior Engineer at Motorola Electronics Ltd., Taiwan. He is currently an Assistant Professor in the Department of Electrical Engineering, Ta-Hwa Institute of Technology, Hsinchu. His research interests include medical imaging, signal processing, and computer version.

Yu-Tai Ching (M'01) received the B.S. degree in industrial engineering from Tsing Hua University, Hsinchu, Taiwan, R.O.C, in 1980 and, the M.S. and Ph.D. degrees in computer science from Northwestern University, Evanston, IL, in 1983 and 1987, respectively.

He is currently an Associate Professor in the Department of Computer Science, National Chiao-Tung University, Hsinchu. His research interests are design and analysis of computer algorithms, medical image analysis, and computer graphics.

Yun-Liang Yang received the B.S. degree in plant pathology from National Chung Hsing University, Taichung, Taiwan, R.O.C., in 1985, and the Ph.D. degree in molecular, cellular, and developmental biology from Indiana University, Bloomington, in 1993.

He is currently an Associate Professor in the Department of Biological Science and Technology, National Chiao-Tung University, Hsinchu, Taiwan. His research focuses on microbiology and functional genomics.