Automated Subspots Acquiring and Merging Detection Algorithm for 2DEG Images

Shu-Ying Chen, Tung-Shou Chen, *Member*, *IACSIT*, Wien Hong, Jeanne Chen, *Member*, *IACSIT*, Hui-Fang Tsai, Mingli Hsieh, and Chun-Wei Tsai

***Abstract*—The optimal and accurate detection result of two-dimensional electrophoresis gel (2DEG) image is necessary for accurate diagnostic of diseases. In this paper, we proposed a novel auto-detection algorithm for accurate detection of protein spots which do not required complicating parameters. The method requires copies of the original image to be represented as binary images in different thresholds ranging from 255 to 0. The binary images are used to select suitable states of the protein spots which were merged so that the spots are enhanced and easy to detect. The method is simple and heuristic. The proposed method easy to use without requiring any parameter and useful in applications that are feature oriented, such as proteomics, medical images and others with similar requirements.**

***Index Terms*—Two-dimensional electrophoresis gel (2DEG), protein spots , detection, merging .**

I. Introduction

The two-dimensional electrophoresis gel (2DEG) image is important protein research [5]. Protein spots detection is the first and important step in 2DEG image analysis [6]. The optimal and accurate detection result is necessary for efficient execution of later steps. There are many commercial software packages that have been developed for 2DEG images analysis to get optimal detection result.

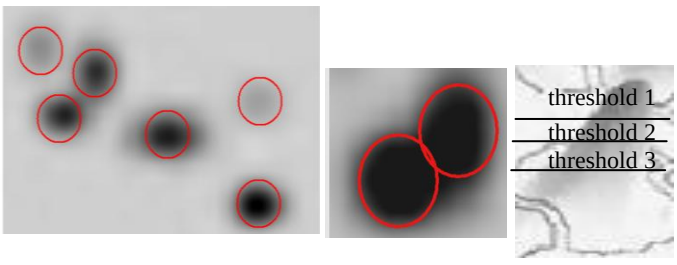
The protein spot maybe divided into two kinds of protein spot in 2DEG images. One kind is the independent spots (see Fig. 1(a)) in low abundance (high intensity and lighter) or high abundance (low intensity and darker) and the second kind is the overlapped spots (see Fig. 1(b)).

Currently, evaluation of the detection rate and quality are by the statistics of number of spots detected successfully. The errors will influence the quality of comparison and analysis. Software developed for detection such as ImageMaster often uses default settings for various types of 2DEG images. However, the various 2DEG images have different background and spot intensities which makes the default settings unsuitable. On the other hand, using user-adjustable parameters and deciding which ones to chose can be difficult for some end-users. Furthermore, parameters are based on intensities where some spots and noises are sometimes not effectively separated.

This paper proposes a novel detection technique for protein spots in 2DEG images called Automated Subspots Acquiring and Merging (ASAM) detection which is very simple and heuristic. ASAM requires the 2DEG image to be divided in multiple copies which is similar to Tsai *et al.*'s [4] content based technique where multiple images were generated from bit-slicing of the JPEG wavelet coefficients. In ASAM, multiple binary images were generated by converting pixel values to thresholds in the range of 255 to 0. The images are evaluated for suitable state of each spot which will then be selected and merged with others also having suitable states to enhance detection of spots.

II. Proposed Method

The proposed ASAM does not require any parameters for auto-detection of protein spots. Each spot in 2DE images is a gradient element, which can be presented as a mountain (see Fig. 1(c)) [1]. The height of the hill is the intensity of the spot. Several binary images will be generated based on a detecting image and several threshold values ranging from 255 to 0. In the different binary images, the same protein spot is located the same position but has different size. Fig.2 shows the image difference in different thresholds. Each spot is normalized based on three important roles set for the binary images. The selection of suitable state for each spot and merging of the selection is repeatedly performed to improve the original state of the protein spots for optimal detection.



(a) (b) (c)

Fig. 1. (a) (a) Independent spots, (b) overlapped spots (c) Spot presented as a mountain [1]

A. Intensity Normalization

Different types of 2DEG images have different background

Manuscript received April 10, 2012; revised May 11, 2012

S. Y. Chen is with Hungkuang University, Taichung 443, Taiwan (e-mail: csy731@sunrise.hk.edu.tw).

T. S. Chen and J. Chen are with the National University of Science and Technology, Taichung, 404 Taiwan (e-mail: tschen@nutc.edu.tw, jeanne@nutc.edu.tw).

W. Hong is with Yu Da University, Miaoli, 361 Taiwan (e-mail: wienhong@ydu.edu.tw).

H. F. Tsai is with Chung Shan Medical University, Taichung, 402 Taiwan (e-mail: HFtsai@csmu.edu.tw).

M. L. Hsieh is with Tunghai University, Taichung, 407 Taiwan (e-mail: mhsieh@thu.edu.tw).

C. W. Tsai is with the National Taichung Institute of Technology, Taichung, 404 Taiwan (e-mail: chen.corrs@gmail.com).

and spots intensities. In order to establish a benchmark between different types of 2DEG images, the background intensity must normalized for detecting image *R*. Since the 2DEG images are gray images, the *R* intensities are translated into pixel values ranging from 255 to 0. First, ASAM locates the maximum pixel value of row *i* in *R* from Eq.(1).

$$Max_i = \max(R(i, j)|, 0 \le j < n) \quad (1)$$

where *n* is the height of *R*, and $R(i, j)$ is the pixel value of position at (i,j) located in *R*. Next *R*' is the normalization result of *R* and Max_i.

$$R'(i, j) = R(i, j) - Max_i . \quad (2)$$

B. Binary Image Generation

Here, ASAM generates binary image B_t based on *R*' and threshold *t*. The main idea of this step is to use a threshold to filter pixels of *R*' every time and *t* will be decreased by a *g* value gradually from 255 to 0.

$$t = \{255, (255 - g), (255 - 2 \times g), ..., 0\}. \quad (3)$$

Each spot can be drawn as a three dimensional mountain as in Fig. 1(c). The value *g* is decided by gradient level of each detecting image *R*.

$$count\_2 = \sum_{m}^{i=0} \sum_{n}^{j=0} R(i, j) \quad \text{if } R(i, j) > 100 \text{ and } R(i, j) < 201 \quad (4)$$

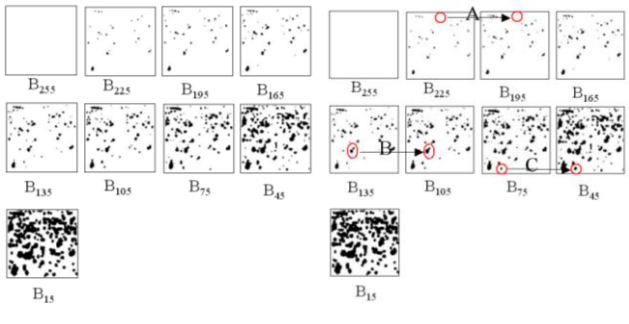
$$count\_3 = \sum_{m}^{i=0} \sum_{n}^{j=0} R(i, j) \quad \text{if } R(i, j) > 200 \quad (5)$$

where the size of *R* is $(m \times n)$.

$$g = \begin{cases} 25 & \text{if } (count\_2 / count\_3) \ge 0.6 \\ 20 & \text{if } (count\_2 / count\_3) \ge 0.5 \\ 15 & \text{if } (count\_2 / count\_3) \ge 0.4 \\ 10 & \text{if } (count\_2 / count\_3) \ge 0.3 \\ 5 & \text{if } (count\_2 / count\_3) \le 0.2 \end{cases} \quad (6)$$

All pixels of *R*' are scanned pixel-by-pixel. If $R'(i, j)$ is bigger than *t* and *R*' (i, j) is one pixel of any assumed spot in B_t, then $R'(i, j)$ is set 0. On the other hand, if $R'(i, j)$ is smaller than *t* and *R*' (i, j) is one pixel of background in B_t, and this pixel is set to 255. Fig. 2 shows B_t with g set at 30.

$$B_t(i, j) = \begin{cases} 0 \\ 255 \end{cases} \begin{matrix} \text{if } R'(i, j) > t \\ \text{if } R'(i, j) < t \end{matrix} \quad (7)$$



(a) (b)

Fig. 2. Samples of binary images B_t (*g*=30) with different thresholds *t*

C. Spots Classification

In each B_t, spots are assumed to be the black regions by human eye where many pixels are clustered together. The pixels are classified recursively to construct spots [2]. The B_t image is scanned pixel by pixel from top-left and bottom-right. If the scanned pixel value equals 0, this position is recorded. Its neighbor eight directions are scanned to find the next black pixel (pixel = 0). This process is iterated. The pixel value of the located pixel is set to 255 to cause a loop. The recursive cycle will stop after all pixels in the same black region are recorded. The next recursive loop is started after B_t is scanned for the next cluster of black pixels. The position and numbers in the group will be recorded.

D. Information Observation for Merging

The same protein spot in different B_t may be located in the same position but slightly off placed. Fig. 3(b) shows the three important roles for any spot in B_t, which are the new create spots marked A, two merge spots merged marked B and increasing spot marked C. For creating spots in A, the information on the spots is recorded and B_{195} is chosen from between B_{225} and B_{195}. For two spots merging in B, the information of two spots are recorded individually before merging into B_{135}. The spot with suitable state will be selected before merging. For increasing spot, the information of the bigger spot is recorded for the spot. The bigger area in B_{45} will be the selected spot. The recorded positions of each protein spot in B_t are used to build a relative table of spots between B_t. If a spot numbered *j* in $B_{(t-g)}$ can cover a whole spot numbered *i* in B_t, then $spot_{(t-g)\_i}$ is the father of $spot_{t\_i}$.

$$spot_{t\_i} \subset spot_{(t-g)\_j} \quad (8)$$

E. Best State for Each Spot Combination

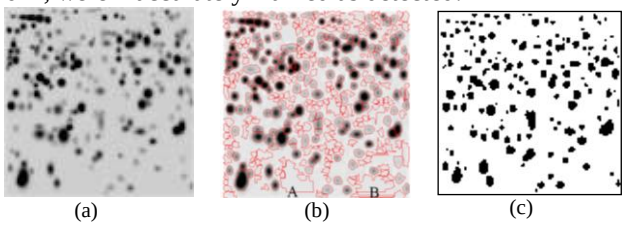
The suitable state for each spot is decided based on spot relative table and the three roles. These states are combined to create an accurate detection result for *R*. The merging process starts from 255 to 0. The relations between spots in B_t and $B_{(t-g)}$ are then checked. If $spot_{(t-g)\_j}$ does not have any child in B_t, this means $spot_{(t-g)\_j}$ is a new spot in $B_{(t-g)}$; record the state of $spot_{(t-g)\_j}$ for this position. If $spot_{(t-g)\_j}$ has one more children in B_t then $spot_{(t-g)\_j}$ is an overlapped spot merging from one more spots in B_t. The last state in this position is deleted and the state is recorded before merging in B_t. If $spot_{(t-g)\_j}$ has only one child $spot_{t\_i}$ in B_t, then this $spot_{(t-g)\_j}$ is an increasing area spot from $spot_{t\_i}$. The state in this position is deleted and the bigger area is recorded in $B_{(t-g)}$. Each suitable state of each spot is merged to get an accurate detection result.

III. Experimental Results

Six experiments were conducted with ImageMaster and ASAM for comparison purposes. The datasets for the experiments were downloaded from different labs [7] and were chosen for different background intensity, spots intensity, many low abundance proteins, and many overlapped protein spots. The ImageMaster software and default settings were used in the experimental tests.

A. Experiment 1

The original image in Fig. 3(a) is chosen for higher background intensity 2DEG image. Fig. 3(b) shows the detected result by ImageMaster, and Fig. 3(c) shows detection results by ASAM. Noises from background were not accurately detected by ImageMaster. In fact, these areas, A and B, were inaccurately marked as detected.

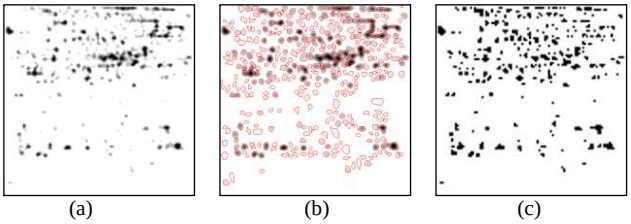


(a) (b) (c)

Fig. 3. (a) Original image (b) Detection by ImageMaster (c) Detection by ASAM

B. Experiment 2

The original image chosen in Fig. 4(a) has higher background intensity, and many overlapped spots. Both ImageMaster and ASAM obtained high detection results for overlapped spots, but ImageMaster has more sensitive detection for low abundance proteins. Fig. 4(b) shows more detection for both noise and low abundance protein.



(a) (b) (c)

Fig. 4. (a) Original image (b) Detection by ImageMaster (c) Detection by ASAM

C. Spot detection analysis

TABLE I: SPOT DETECTION RESULTS ANALYSIS

Experiment No	Real protein	ImageMaster		ASAM	
		FN	FP	FN	FP
1	52	1	100	11	0
2	276	3	99	46	0
3	516	0	279	0	7
4	200	0	13	10	3
5	100	1	57	5	6
6	267	0	57	19	0

TABLE II: DETECTION ACCURACY RATES

Experiment No	ImageMaster		ASAM	
	Precision rate	Recall rate	Precision rate	Recall rate
1	-0.923	1.021	1.000	0.825
2	0.641	0.983	1.000	0.857
3	0.459	1.000	0.986	1.000
4	0.935	1.000	0.985	0.952
5	0.430	0.977	0.940	0.949
6	0.787	1.000	1.000	0.934
Average rate	0.388	0.997	0.985	0.920
Accuracy rate	0.692		0.95	

From the experiments, the total number of real proteins is identified manually for false negative (FN) and false positive (FP). Real proteins are detected visible proteins by human eye. In Table I, we can see the number of FPs of ImageMaster exceeds ASAM by a wide margin. This implies that ImageMaster detects everything from spots to noises in the background. The recall rate and precision rate are counted for estimating the quality of detection results by ImageMaster and ASAM [3].

$$Precision\ rate = TP/(TP+FP) \quad (9)$$

$$Recall\ rate = TP/(TP+FN) \quad (10)$$

where TP is true positive, and is the number of successfully detected spots. TP values in our experiments are calculated by the number of real protein – FP. Table II shows the rates of ImageMaster and ASAM with having the higher precision rate, lower recall rate and higher accuracy rate. The quality of detection is higher in ASAM.

IV. CONCLUSIONS

Protein spot detection is the first and an important step in 2DEG image analysis. In this paper, we have defined the condition of successful detection results for protein spots in 2DEG images, and have designed a novel detection algorithm, ASAM which is very simple and heuristic, and does not require complicated handling of parameters. The intensity is normalized for detecting varying types of 2DEG images to automatically compute the parameters needed for detection. Due to these features ASAM has a better detection rate, and is convenient to users. Also, ASAM has a higher precision rate of about 0.985 in comparison to ImageMaster at 0.388. ASAM may be used in images with varying backgrounds for accurate detection of objects.

REFERENCES

[1] T. R. Arsi, M. S. Jussi, A. Tero, W. Jan, L. Riitta, A. N. Tuula, and S. N. Olli, "Comparison of PDQuest and Progenesis Software Packages in the Analysis of Two-Dimensional Electrophoresis Gels," *Proteomics*, vol. 3, pp. 1936-1946, 2003.

[2] D. B. Gray, *A Recursive Application of a Support Vector Machine for Protein Spot Detection in 2-Dimensional Gel Electrophoresis*, Technique Report by University of Houston-Clear Lake, 2003, pp. 33-40.

[3] W. Jiang, L. Jimmy, M. L. Nicholas, Y. Haiyuan, and G. Mark, "Prediction of Regulatory Networks: Genome-Wide Identification of Transcription Factor Targets from Gene Expression Data," *BioInformatics*, vol. 19, no. 15, pp. 1917-1926, 2003.

[4] T. S. Tsai, J. Chen, T. S. Chen, S. F. Chiou, T. S. Wang, H. C. Wu, H. F. Tsai, M. Hsieh, M. C. Chou and S.Y. Li, "Content-based Progressive Transmission of 2-DGE Images," *Imaging Science Journal*, vol. 54, no. 3, pp. 192-199, 2006.

[5] T. Rabilloud, M. Chevallet, S. Luche, and C. Lelong, "Two-dimensional Gel Electrophoresis in Proteomics: Past, Present and Future," *Journal of Proteomics*, vol. 73, no. 11, pp. 2064-2077, 2010.

[6] F. P. Wayne, "Detection Technologies in Proteome Analysis," *Journal of Chromatography B*, vol. 771, pp. 3-31, 2002.

[7] K. P. Pleissner. (March 2012). Two-dimensional Polyacrylamide Gel Electrophoresis Database of Rat Heart.[Online]. Available: http://web.mpiib-berlin.mpg.de/bioinformatik/RAT-HEART/2d/2d.html.