Adaptive thresholding noise elimination and asymmetric diffusion spot model for 2-DE image analysis

Kwan-Deok Choi*, Young-Woo Yoon**
*Dept. of Health Information Management, Taegu Science College
**Dept. of Computer Engineering, YeungNam University, Korea

Abstract

In this paper we suggest two novel methods for an implementation of the spot detection phase in the 2-DE gel image analysis program. The one is the adaptive thresholding method for eliminating noises and the other is the asymmetric diffusion model for spot matching. Remained noises after the preprocessing phase cause the over-segmentation problem by the next segmentation phase. To identify and exclude the over-segmented background regions, if we use a fixed thresholding method that is choosing an intensity value for the threshold, the spots that are invisible by one's human eyes but mean very small amount proteins which have important role in the biological samples could be eliminated. Accordingly we suggest the adaptive thresholding method which comes from an idea that is got on statistical analysis for the prominences of the peaks. There are the Gaussian model and the diffusion model for the spot shape model. The diffusion model is the closer to the real spot shapes than the Gaussian model, but spots have very various and irregular shapes and especially asymmetric formation in x-coordinate and y-coordinate. The reason for irregularity of spot shape is that spots could not be diffused perfectly across gel medium because of the characteristics of 2-DE process. Accordingly we suggest the asymmetric diffusion model for modeling spot shapes. In this paper we present a brief explanation of the two methods and experimental results.

1. Introduction

Two-dimensional electrophoresis (2-DE) is an important tool in the molecular biology and is a technique for separating large numbers of proteins from biological samples. 2-DE is used in many different fields, like toxicology, clinical chemistry, cancer research, etc[1,2]. 2-DE is composed of two biochemical separations, the one (the first dimension) is iso-electric focusing (IEF) electrophoresis and the other (the second dimension) is SDS-PAGE(Sodium Dodecyl Sulfate - Polyacrylamide Gel Electrophoresis). After 2-DE, the result (called the gel) is stained for the visibility. The result of 2-DE is two dimensional spot patterns and each spot in the patterns means individual protein. Therefore identifying a spot to a protein is necessary for further applications.

There are two techniques for identifying a spot to a protein. The one is extracting a spot by robot and identifying it by the mass spectroscope. The other is the method using visual cross comparison between already analyzed gel image and new comer. The former can analyze precisely but spend a lot of time because one spot can be analyzed at one time. If the latter method is processed by human eyes, it is very difficult because complexity of spot patterns, so the researches on the automatic identification of spots using a computer are widely performed. For identifying a spot to a protein automatically by a computer, the spot patterns are digitized by a very high resolution scanner or a laser photometer and the digitized image is analyzed by the gel image analysis program[1,2].

The gel image analysis program consists of three phases in general that are the spot detection phase, the gel matching phase, and quantitative comparison phase. The spot detection phase detects regions that have high possibility for protein spot and quantifies spot information such as location, intensity, volume and parameters for spot model. The spot detection phase segments a gel image into individual spot regions using an image segmentation algorithm and fits each region to the specific spot model. Therefore when we implement a spot detection phase, the most important things are designing a suitable image segmentation algorithm and a spot model for spot matching. The watershed algorithm[3] is generally used for the gel image segmentation algorithm because of its robustness for noises, but it has over-segmentation problem that is caused by background inhomogeneity. There are Gaussian model[1] and diffusion model[4] for the spot shape model. The Gaussian model is a model assumes that a protein spot is diffused from a point and the diffusion model is a model assumes that a protein spot is diffused from a disc. The diffusion model is closer to real spot shapes than the Gaussian model, but spots in

** Corresponding Author : Young-Woo Yoon

the gel image have very various and irregular shapes and especially asymmetric formation in x-coordinate and y-coordinate. The reason for irregularity of spot shape is that spots could not be diffused perfectly across gel medium because of the characteristics of 2-DE process.

This paper aims at design and implementation of a spot detection phase of a gel image analysis program and especially focuses on solving two problems. The first is solving over-segmentation problem of watershed algorithm and the second is designing a model for asymmetric diffused spot shapes. For solving the two focused problems we suggest the adaptive thresholding method and the asymmetric diffusion model.

2. Existing techniques

2.1 Noises elimination

A gel image is made by complex processing procedures for a very long time. So there are lots of noises besides protein spots and background (i.e. gel itself) in the gel image. Noises are black specs, streaks and the very small amount of non-protein elements. Particularly the very small amount of non-protein elements makes background inhomogeneous so that it makes difficult to eliminate noises [5]. The most of gel image analysis program have a preprocessing phase before a segmentation phase for reducing influences of noises. The typical preprocessing methods are smoothing method and background subtraction method.

However remained noises after the preprocessing phase influence the next phase (i.e. the segmentation phase). The watershed algorithm is generally used for the gel image segmentation algorithm because of the robustness for noises. But one major drawback of watershed is over-segmentation due to noise creating false minima. There are two methods for solving over-segmentation. One is a method that removes unwanted catchment basins before segmentation [6] and the other is a method that removes unwanted watersheds after segmentation [7].

2.2 Spot models

There are two classes of techniques for modeling spot shape that are the non-parametric technique and the parametric technique. In this paper we treat the parametric one. The parametric technique uses a model to parameterize spots. The existing models are the Gaussian model and the diffusion model.

The Gaussian model is a model assumes that a protein spot is diffused from a point. The Gaussian model is defined by (1).

$$s(x,y) = B + I\exp\left(-\frac{(x-x_0)^2}{2\sigma_x^2}\right)\exp\left(-\frac{(y-y_0)^2}{2\sigma_y^2}\right) \quad (1)$$

In (1), B is background intensity, and I is spot intensity (i.e. spot height), and x_0 and y_0 are spot location, and σ_x and σ_y are diffusion factors (i.e. spot width). The parametric expression for Gaussian model is defined by (2).

$$G_s = (B, I, x_0, y_0, \sigma_x, \sigma_y) \quad (2)$$

The Gaussian model cannot model a spot shape accurately, especially a flat-top shaped spot. Bettens [4] suggest the diffusion model for covering the problem. The diffusion model is a model assumes that a protein spot is diffused from a disc and defined by (3)

$$s(x,y) = B + \frac{c_0}{2}\left[erf\left(\frac{(a'+r')}{2}\right) + erf\left(\frac{(a'-r')}{2}\right)\right] + \frac{c_0}{r'\sqrt{\pi}}\left[\exp\left(-\left(\frac{(a'+r')}{2}\right)^2\right) - \exp\left(-\left(\frac{(a'-r')}{2}\right)^2\right)\right] \quad (3)$$

with

$$r' = \sqrt{\frac{(x-x_0)^2}{D'_x} + \frac{(y-y_0)^2}{D'_y}}$$

In (3), B is background intensity, and c_0 is an initial concentration of peak, and *erf* is the error function (i.e. $erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x \exp(-u^2)du$), and a' is the area of the disc containing the protein material, and D_x and D_y are diffusion factors. If a' is 0 then the diffusion model is the same as the Gaussian model. The parametric expression for diffusion model is defined by (4).

$$D_s = (B, C_0, x_0, y_0, D_x, D_y) \quad (4)$$

3. Spot detection system

(Fig.1) shows the spot detection system that is designed and implemented in this paper. The system processes the preprocessing for an input gel image, and segments the preprocessed image into regions by watershed algorithm, and applies the adaptive thresholding method for the regions and creates a candidate spot list, and finally fits the regions to the asymmetric diffusion model and adds parameter values in the candidate spot list.

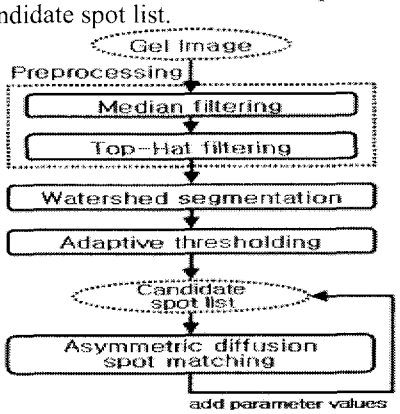


Figure 1.Spot detection system

4. Adaptive thresholding method

We suggest an adaptive thresholding method to identify and exclude over-segmented background regions. The adaptive thresholding method works as followings. We determine the prominence value for each region, and calculate an average prominence value curve for height of region, and fit it to the exponential function so that we can get parameters for the exponential function. And then we calculate a threshold value by using the parameters and the probability distribution of errors. Lastly we apply the threshold value to the region for determining the region is a noise or not.

We suggest the prominence as a measuring unit for discriminating a peak that has a distinct three-dimensional shape from highly but widely spreading inhomogeneous background. The prominence p_r is defined by (5).

$$P_r = \log\left(h_r^2 / A_r\right) \qquad (5)$$

In (5), h_r is the height of region, and A_r is the area of ah_r, $0 < a < 1$.

By investigating experimental results statistically, we found out the average p_r curve has the exponential characteristics except the case that the height of region is very low. The possibility of the region which has very big height is turned out to be background region is very low, so that we do curve fitting p_r to the exponential function defined by (6) for h_r is greater than the height threshold $thBG$ defined by (7).

$$f(h) = a[1 - \exp(-bh)] + c \qquad (6)$$

$$\left|\{r \mid h_r > thBG\}\right| = \frac{|R|}{2} \qquad (7)$$

From the curve fitting, we get values for parameters (i.e. a, b, c) of the exponential function and use them for determining a threshold. Determining a threshold is based on the probability distribution of errors. The error between average p_r curve and $f(x)$ is defined by (8), and the average of error is defined by (9), and the standard deviation of error is defined by (10).

$$e_r = \left|p_r - f(h_r)\right|, r \in R \qquad (8)$$

$$\bar{e} = \frac{1}{|R|}\sum_{r \in R} e_r \qquad (9)$$

$$\sigma = \sqrt{\frac{1}{|R|}\sum_{r \in R}\left(e_r - \bar{e}\right)^2} \qquad (10)$$

Mistaking a background region to a peak (i.e. spot) has a second chance for correcting it in the next time, but the reverse case has not a chance for correcting in anytime. Therefore we determine a threshold based on the false positive policy rather than the false negative policy. According to the probability distribution of errors, 99.7 percent of regions which are not included in background exists in $\bar{e} - 3\sigma < e_i \le \bar{e} + 3\sigma$. Accordingly the regions satisfying $p_i \le f(h_i) - \bar{e} - 3\sigma$ can be eliminated because those have very high possibility of included in background. The probability of error of the adaptive thresholding is 0.075 percent so that the reliability of the adaptive thresholding method is 99.925 percent.

5. Asymmetric diffusion model

The diffusion model is the closer to the real spot shapes than the Gaussian model, but spots have very various and irregular shapes. The reason for irregularity of spot shape is that spots could not be diffused perfectly across gel medium because of the characteristics of 2-DE process.

(Fig.2) shows a typical gel image. Inspecting closely (Fig.2) we can see that the most of spots have asymmetric formation rather than symmetric formation in x-coordinate and y-coordinate.

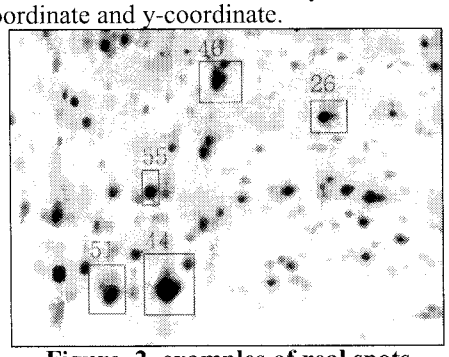


Figure 2. examples of real spots

The asymmetric diffusion model is a model assumes that a protein spot is diffused from a disc at initial time but is diffused asymmetric in x-coordinate and y-coordinate as time goes on. The formula (11) defines the asymmetric diffusion model.

$$s(x,y) = B + \frac{c_0}{2}\left[erf\left(\frac{(a'+r')}{2}\right) + erf\left(\frac{(a'-r')}{2}\right)\right] + \frac{c_0}{r'\sqrt{\pi}}\left[\exp\left(-\left(\frac{(a'+r')}{2}\right)^2\right) - \exp\left(-\left(\frac{(a'-r')}{2}\right)^2\right)\right]$$

with

$$r' = \sqrt{\frac{(x - x_0)^2}{D'_{xp}} + \frac{(y - y_0)^2}{D'_{yp}}}, x > 0 \text{ and } y > 0$$

$$r' = \sqrt{\frac{(x - x_0)^2}{D'_{xm}} + \frac{(y - y_0)^2}{D'_{yp}}}, x < 0 \text{ and } y > 0$$

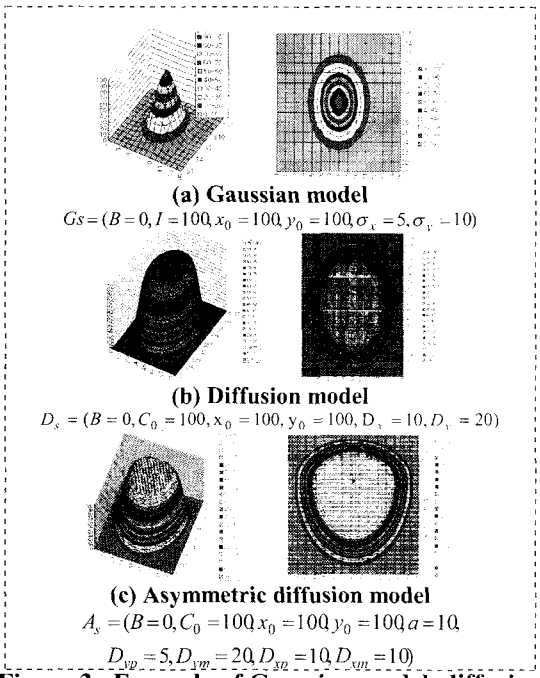
$$r' = \sqrt{\frac{(x - x_0)^2}{D'_{xm}} + \frac{(y - y_0)^2}{D'_{ym}}}, x < 0 \text{ and } y < 0$$

$$r' = \sqrt{\frac{(x - x_0)^2}{D'_{xp}} + \frac{(y - y_0)^2}{D'_{ym}}}, x > 0 \text{ and } y < 0 \qquad (11)$$

In (11), B is background intensity, and c_0 is an initial concentration of peak, and erf is the error function (i.e. $erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x \exp(-u^2)du$), and a' is the area of the disc containing the protein material,

and D_{xm}, D_{xp}, D_{ym} and D_{yp} are diffusion factors. The parametric expression for the asymmetric diffusion model is defined by (12).

$$D_s = (B, C_0, x_0, y_0, D_{yp}, D_{ym}, D_{xp}, D_{xm}) \quad (12)$$



(a) Gaussian model

$Gs = (B = 0, I = 100, x_0 = 100, y_0 = 100, \sigma_x = 5, \sigma_y = 10)$

(b) Diffusion model

$D_s = (B = 0, C_0 = 100, x_0 = 100, y_0 = 100, D_{\cdot} = 10, D_{\cdot} = 20)$

(c) Asymmetric diffusion model

$A_s = (B = 0, C_0 = 100, x_0 = 100, y_0 = 100, a = 10,$

$D_{yp} = 5, D_{ym} = 20, D_{xp} = 10, D_{xm} = 10)$

Figure 3. Example of Gaussian model, diffusion model and asymmetric diffusion model

6. Experiments and results

The object system for experiments is a PC system based on Microsoft Windows XP and the whole algorithm is written in Microsoft Visual C++ 6.0. In experiments we use the silver-stained gel images from the SWISS-2DPAGE[8] as test gel images.

To verify the adaptive thresholding method, we define the mark ratio that is a ratio of 'number of identified spots in candidate spot lists after the adaptive thresholding phase' to 'the number of the original identified spots'. For 1312 of the original identified spots in 19 gel images, we got 97.71 percent as mark ratio. The main factors for lowering the mark ratio were duplicated identified spots in same region (25 identified spots) and identified spots on watershed line (5 identified spots). The most important result is that there was no identified spot missed by the adaptive thresholding method itself in all of 19 gel images. Therefore we could confirm the correctness of the suggested method.

To verify the asymmetric diffusion model, we fit regions in the candidate spot list to three spot models respectively so that we get parameter values and determine a SNR which is defined by (13) for each spot model.

$$SNR = 10 \log_{10} \frac{\sum_{p \in R'} I(p)^2}{\sum_{p \in R'} \{I(p) - f(x)\}^2} \quad (13)$$

In (13), $I(p)$ is the intensity value of pixel p, and f is a spot model, and x is a parameter vector, R' is regions that those height is over than middle height of region R. As averages of SNR for 19 gel images, we got 14.126426 for the Gaussian model, 20.662714 for the diffusion model and 22.77242 for the asymmetric diffusion model. Therefore we could confirm the asymmetric diffusion model is the best model among the three models.

7. Conclusions

In this paper we have suggested and verified two novel methods for an implementation of the spot detection phase in the 2-DE gel image analysis program. The one (the adaptive thresholding noise elimination) is a good method for eliminating noises and the other (the asymmetric diffusion spot model) is a good model for spot matching.

8. References

[1] A. W. Dowsey, M. J. Dunn and G.-Z. Yang, "The role of bioinformatics in two-dimensional gel electrophoresis," Proteomics 2003, pp. 1567-1597, 2003.

[2] L. Addario-Berry, "2D Gel Electrophoresis - An Overview," http://www.mcb.mcgill.ca/~hallett/GEP/PLectures/Plecture3.pdf, May 9, 2002.

[3] L. Vincent and P. Soille, "Watershed in digital spaces: An efficient algorithm based on immersion simulations," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, No. 7, July, pp. 583-598, 1990.

[4] E. Bettens, "Peak characterization using parameter estimation methods," PhD thesis, University of Antwerpen, 1999.

[5] A. Roy, K. R. Lee, Y. Hang, M. Marten and B. Rauman, "Analyzing two-dimensional Gel Images," Technical Reports, Dept. of Mathematics and Statistics, University of Maryland, 2003.

[6] L. Pedersen, "Analysis of two-dimensional electrophoresis gel images," PhD Thesis, Informatics and Mathematical Modelling, Technical University of Denmark, 2002.

[7] K.-P. Pleissner, F. Hoffman, K. Kreigel, C. Wenk, S. Wegner, A. Sahlstrom, H. Oswald, H. Alt and E. Fleck, "New algorithmic approaches to protein spot detection and pattern matching in two-dimensional electrophoresis gel databases," Electrophoresis 1999, 20, pp. 755-765, 1999.

[8] C. Hoogland, K. Mostaguir, J.-C. Sanchez, D. F. Hochstrasser and R. D. Appel, "SWISS-2DPAGE, ten years later," Proteomics 2004, 4(8), pp. 2352-2356, 2004.