Bioinformatics and Biosystems 2006, Vol. 1, No. 3, pp. 146-151

An Integrative Analysis Program for Two-dimensional Gel Electrophoresis

Dongil Chung,[1] Juhyun Jung,[1] Jaeduck Jang,[1]
Jongchul Ye,[1] Jaeseung Jeong,[1] Kwang H. Lee[1,*]

[1]Department of BioSystems, Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, Republic of Korea 305-701

Abstract

Two-dimensional gel electrophoresis (2DE) is one of the most popular way to separate thousands of proteins simultaneously with quantification. To analyze the position of each gel spot containing the protein profile, i.e. protein isoelectric point (pI) and molecular weight (Mw), several computer assisted tools are suggested, including MELANIE and PDQuest. Those softwares provide various and performant methods for image processing and spot detection, but their handling is time-consuming and requires the assistance of user guide manual. In this study, we propose an automatic gel image processing program with a friendly graphic user interface based on MATLAB®. Our tool not only processes spots detection, including preprocessing of noise and distortion of the data, but also provides an automatic association to the corresponding reference spot of the dataset, or landmarks. The latest operation is also called the 'landmark matching' process. In this paper we present an example of application of our integrative analysis program with to the data set of human leukemia. We think that this program will be able to gently assist biologist in the analysis of high-throughput proteomics.

Introduction

One of the main purposes of proteomics is to study structure and functions of proteins in biological systems. The current combination of mass spectrometry, which readily identifies an individual protein, and 2D page gel has been widely used for a better understanding of proteomics. (Kalia and Gupta, 2005), propulsing the two-dimensional gel electrophoresis (2DE) as a major analysis tool for the characterization of protein.

The 2DE first separates the protein along one direction using an isoelectric focusing and then along a perpendicular direction using a sodium dodecyl sulfate electrophoresis. Those two separations are able to characterize the protein by its isolectric point (pI) and its molecular weight (Mw), respectively.

Although the 2DE has serious inconveniencies such as limited number of spots, low load ability and poor separation of highly hydrophobic proteins, it is still widely utilized because of its high throughput capability.

To deal with such analysis of high-throughput data, i.e. tens of thousand spots detection, numerous computer tools have been proposed, such as MELANIE (Wilkins *et al.,* 1996) and PDQuset (Garrels, 1988). However, previous existing programs uncomfortable to handle and are hard to understand without heavy documentations. The origin of the problem resides in the fact that the noise due to experimental protein expression rates and the intensity of background, and distortions such as blur and edge artifacts caused to the picture after being digitalized, are delicate to handle for a non expert. The use of automatic spot detection on 2DE with user-friendly interface, picture preprocessing, high precision of spot localization, is of the best interest for biologists working on proteomics. (*Dowsey et al., 2003*)

Our integrative analysis program for 2DE provides an automatic spot detection, and also displays a validation of detection over known references of the data base, the whole analysis is supported by a user-friendly interface based on MATLAB®.

Image Analysis

Test data description

For a given dataset composed of N gel samples (S_1, S_2, ⋯ S_n), a reference gel sample (S_r) is designated by the provider of the dataset. The reference gel will contain reference spots designated by the provider, also called landmark, which will be used for the normalization and referencing of both intensity and distortion of the gel. The analysis of the remaining N-1 gels will be operated by pair comparison with the reference gel and its landmark points. Thanks to this method, the multiple analysis of gel is well orchestrated and its complexity is dramatically reduced.

To test our spot detection algorithm, we collected raw 2D gel

Corresponding Author: Kwang H. Lee (Email: khlee@kaist.ac.kr)

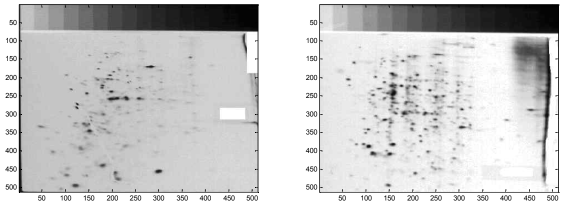
Published Jan 01, 1970 | doi:10.4051/ibce.2009.3.0017

images obtained from the CCR Nanobiology Program (http://www.lecb.ncifcrf.gov/2DgelDataSets/#HEME-MALIG). We chose human leukemia data set (Lester and Lipkin, 1984), which dataset is composed of 170 gel images from different experiment conditions. The gel No.19 was pointed as a reference gel, as specified in the previous research. (Lester, Lipkin and Cooper, 1980)

For each dataset to be analyzed, we proceed similarly using a two step procedure, including a preprocessing and spot detection (Fig.1). The following sections expose in detail each important operation provided by our program.

Preprocessing

Our image analysis tool is performed based on MATLAB® program. The analysis of 2D gel image demands such spot information as center position and number of spots. As specified previously, there are numerous artifact on the image to analyze, which will disturb the spot detection. The following explains our preprocessing method to cope with the variety of contrast, noise and distorsion of each image.



(a) (b)

Figure 2. Two 2DE sample gel images. (a) gel-HM-001. (b) gel-HM-019, the reference gel.

Adjustment of size and intensity

The 2DE gel images are stored as digital images using a CCD camera or a scanner after experiments. Digitalization of these gel images creates blur effect and artifacts in the edge area. We first resized all the gel images corresponding to a reference size and normalized the intensity. As shown in fig.2 (a), objects such as the colormap bar on the top of the gel, the deformed edges and other minor artifacts were removed manually. The resulting resized image is shown in fig.3.

After the size adjustments, we transform the gel image to obtain its negative for a convenient calculation. In this manner, spots are expressed with bright colors, i.e. value near 1, and background pixels are expressed with dark colors, i.e. value near 0.

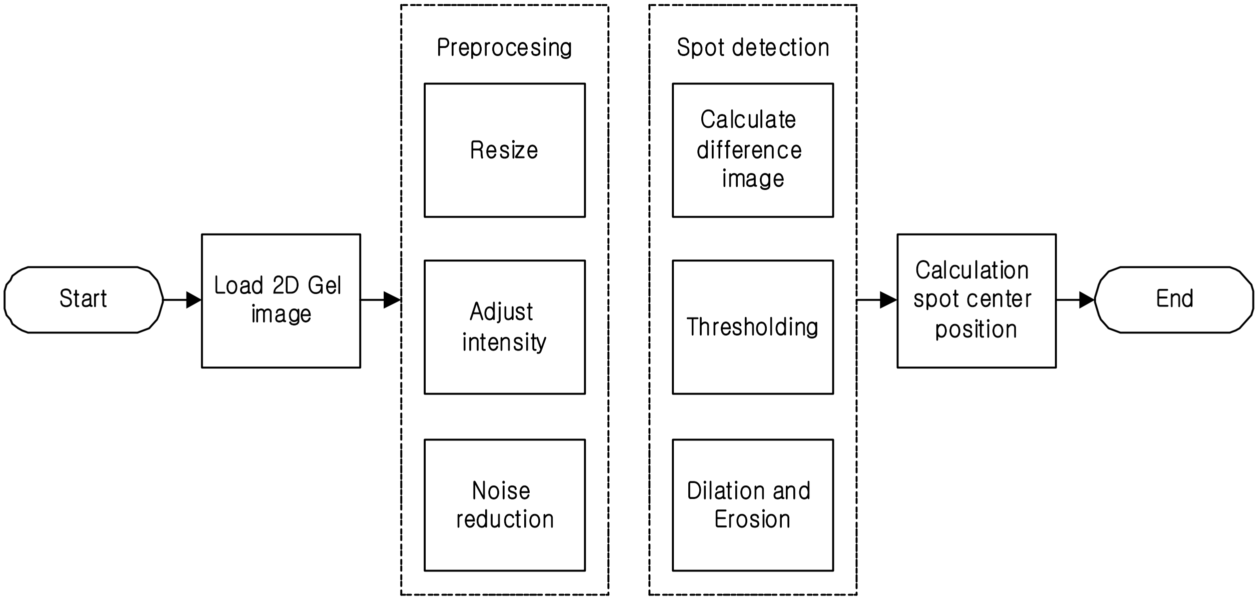


Figure 1. Image processing flow chart. In the flow chart, there are 2 level of image processinge, i.e. preprocessing and spot detection.

Published Jan 01, 1970 | doi:10.4051/ibce.2009.3.0017

Noise Reduction

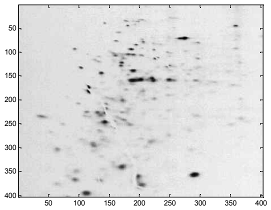


Figure 3. Resized image.

The main goal of image processing emphasize the contrast between the background and the spot for a more accurate detection. We first apply a 2-dimensional Wiener filter (Thompson and Shure, 1995) to the image for reduction of both noise and blurring effect. The Wiener filter first estimates the local mean and variance around each pixel, as showed in eq. (1) and (2). respectively

$$\mu = \frac{1}{NM} \sum_{x,y \in \eta} \mathbf{X}(x,y), \tag{1}$$

$$\sigma^2 = \frac{1}{NM} \sum_{x,y \in \eta} \mathbf{X}^2(x,y) - \mu^2, \tag{2}$$

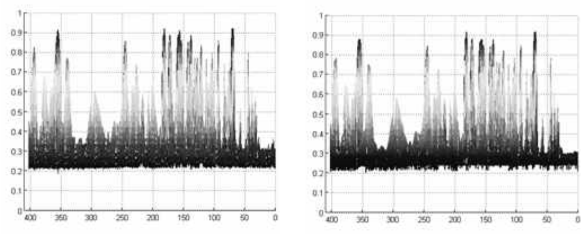
where η is the $N \times M$ local neighborhood of each pixel in the image, $\mathbf{X}$, and N and M both have value of 3. 2-dimensional Wiener filter then creates a pixel-wise Wiener filter using following estimations,

$$B(n_1,n_2) = \mu + \frac{\sigma^2 - \nu^2}{\sigma^2}(X(n_1,n_2) - \mu), \tag{3}$$

where, μ and σ^2 are local mean and variance respectively. In eq. (3), ν^2 is the noise variance. If the noise variance is not given, 2 dimensional Wiener filter in MATLAB® uses the average of all the local estimated variances. The resulting image B should be free from noise and blur, as shown in the comparison of frequency component of raw and filtere image, Fig.4 (a) and (b), respectively.

Spot Detection

In 2DE gel images, aliasing effects, including spots overlapping and diffused, are hard to analyze with naked eyes. For easy and precise spot detection, we used difference between succeeding pixels. (*Matlab Function Reference*, 2000) By using this method, we can analyze gel images with various peculiarities. The difference can be calculated following Eq. (4)



(a) (b)

Figure 4. Comparing the intensity of raw image with that of filtered image. The raw image (a) has high noise ripples compared to the filtered image (b). Through the noise reduction process, the level of high frequency noise has been decreased.

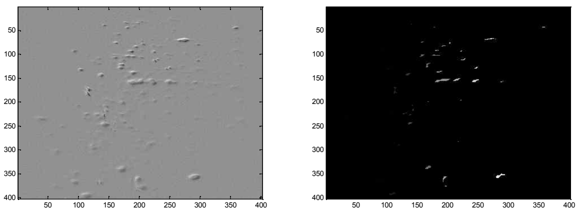
$$\begin{aligned} &diff(\mathbf{X}) = [\mathbf{X}(2:m,:) - \mathbf{X}(1:m-1,:)], \\ &for\ m = 1,\ldots,n \end{aligned} \tag{4}$$

where n is number of pixels in a row or column of the preprocessed image B. The differences are calculated both in row and column direction because spots which are stuck together can be vertically or horizontally located. Two direction averaged differences are shown as figure 5(a).

Using this difference method, the background contrast such as noise or gradient of background color should be extracted. Using an additional threshold, Eq. (5), we can operate an extraction of information.

$$\mathbf{X}(x,y) = \begin{cases} 0 & , if\ \mathbf{X}(x,y) < mean(\mathbf{X}(x,y)) \\ \mathbf{X}(\mathbf{x},\mathbf{y}), otherwise \end{cases} . \tag{5}$$

As we can see from figure 5(a), lots of small spots caused by noise are remained. We used dilation and erosion to deal with small remaining noise or to connect the spots which are accidently divided into several small spots. We then applied mask sizes of 3 x 3 and 4 x 4 to discard spots smaller than 8 pixels. The remaining spots are labeled with monotone color, as shown in Fig 5(b). The center points of these spots are calculated, and yellow circles are displayed on the raw gel image to guide the user's analysis. (Fig 6(a), Fig 6(b)) Figure 6(b) and (d) show the landmarks and references spots displayed with red circles on the raw image.

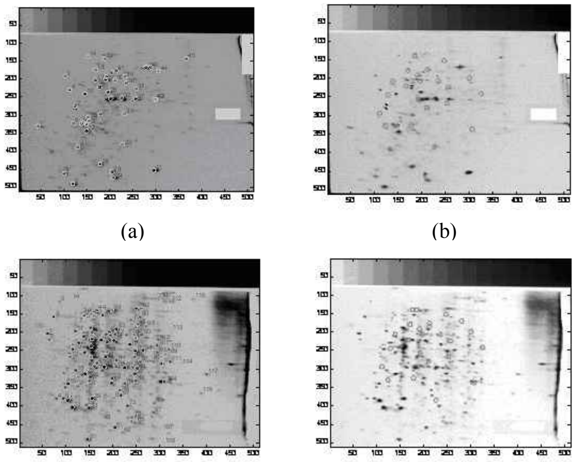


(a) (b)

Figure 5. Processing images. After the noise reduction processing, finding the row and column difference value of the filtered image and the labeling are processed as (a) and (b), respectively.

Database construction

In order to compare our gel spots with known protein pro-files, which are already validated by previous experiments, we built a database containing protein characteristics and expression data. Using the Universal Modeling Language (ULM) class database diagram, we can present a conceptual model of our database (Fig. 7). We constructed our database retrieving information from the Human Protein Reference Database (HPRD, http://www.hprd.org) (Peri *et al*., 2003). HPRD contains information about location of expression of each protein, which data is rigorously validated by expert biologists.



(a) (b)

(c) (d)

Figure 6. Calculated and reference spots displayed on raw images. (a) Calculated spots in gel-HM-001. (b) Landmarks in gel-HM-001. (c) Calculated spots in gel-HM-019. (d) Reference spots in gel-HM-019.

We calculate each protein's pI and molecular weight using biojava (http://www.biojava.org) (Pocock *et al*., 2000). Biojava provides CalcMassand and IsoelectricPointCalc classes which calculate the mass and pI of a peptide from a given sequence.

The database system was supported by the relational database system MySql (http://mysql.com) (Mysql, 2004) on a Linux platform. The database mainly consists of two parts, known data

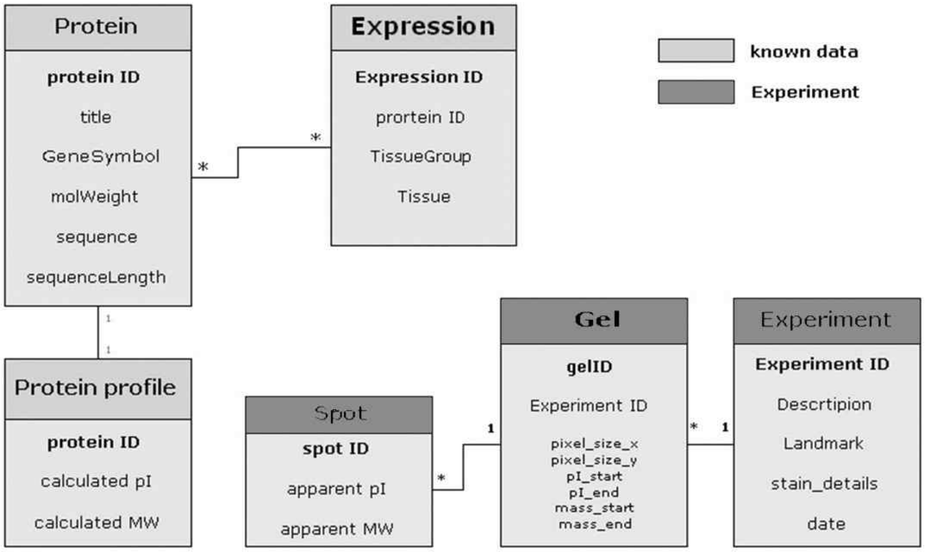


Figure 7. UML diagram of our relational database scheme.

Published Jan 01, 1970 | doi:10.4051/ibce.2009.3.0017

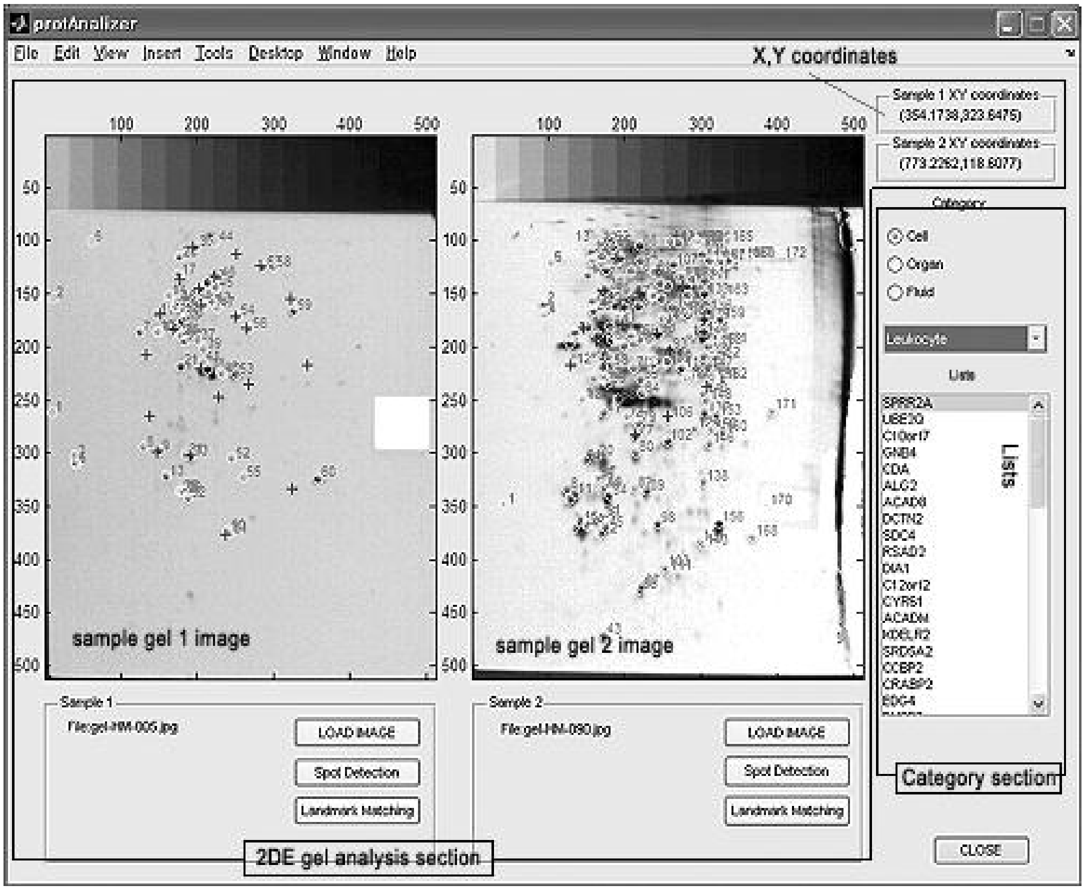
and experimental data, which are use as reference and test case, respectively (Fig 7).

Graphic User Interfaced Program

We constructed a friendly-user interface program using MATLAB® GUI. The program window is composed of two parts, 2DE figures analysis and categorization of proteins expressed in each cell line (Fig. 8), respectively detailed in the following sections.

2DE gel analyzing section

The gel analysis section is designed to perform two simultaneous gel study. The two gels can be loaded and examined independently from each other. Using the left click of the mouse, the user can obtain the $x-y$ coordinates of a selected point on the picture, which coordinates will be displayed on the top-right corner of the window. Using the previously explained image processing algorithm, the user can execute the automatic protein spot detection by pushing the 'Spot detection' button. Landmark information stored in the database can also be loaded and display on the same picture by pressing the 'Landmark matching' button.



Sample 1 XY coordinates
(354.1738,323.6475)

Sample 2 XY coordinates
(773.2262,118.6077)

Sample 1
File:gel-HM-006.jpg

Sample 2
File:gel-HN-090.jpg

Category
◉ Cell
○ Organ
○ Fluid

Leukocyte

Lists
SPRR2A
UBE2Q
C10orf17
GNB4
CDA
ALC2
ACAD8
DCTN2
SDC4
RSAD2
DIA1
C12orf12
CYRS1
ACADM
KDELR2
SRD5A2
CCBP2
CRABP2
EDG4

Figure 8. Our integrative analysis program detects spots automatically and yields systemic access to database which contains spots' information and experimental results.

The landmarks could be located in a different position for each gel picture, because each gel image is obtained from independent experiment and probably different conditions. Hence, the landmarks are automatically recalculated according to the properties of the original figure and displayed to the right.

Category section

In this part of the window, the cell line is divided into three categories: cell, organ, and fluid. A list of corresponding proteins to the selected cell line is displayed in the pop up box. Those candidate proteins could be expressed in the corresponding cell line, in accordance with the data base information. This list provides useful data, i.e. candidate proteins which can be detected from the cell line, and also pair of the spot position and the protein, for the experimenter. The association protein/spot position will be updated by user and manually saved to corresponding database.

Conclusion

In this paper, we presented our integrative analysis program for two-dimensional gel electrophoresis. This tool is designed to support automatic and comprehensive way to analyze high throughput protein experiments issued from certain tissues or cells.

Several accurate program for 2DE gel analysis have already been presented by preceding groups, but require an advance background knowledge to be used efficiently. Our tool provide a more convenient and intuitive analysis of spot detection and spot characterization with a correspondence to a database. We provide as an example an integrated database which is obtained from HPRD to help the interpretation the detected spots.

The current progress of proteomic experiments yet demands important follow-ups for integrating and standardizing the new data. We think that the accentuation of the analysis should pass through a more automatized and user-friendly procedures, short-cutting the time-consuming operations, and facilitate the interpretation work of high throughput data. We expect our tool to be of great help to biologists, providing a intuitive interface and powerful analysis options for the study of protein identification experiments such as mass spectrometry, western blot, or affinity chromatography.

References

[1] Wilkins MR, et al. (1996) Integrating two-dimensional gel databases using the Melanie II software. *Trends Biochem Sci*, 21:496-497

[2] Garrels JI. (1988) The QUEST System for Quantitative Analysis of Two-dimensional Gels. *THEJ OURNAOFL B IOLOGICACLHE MISTRY*, Val. 264:5269-5282,

[3] Lester EP LP, Lipkin LE.. (1984) Protein indexing in leukemias and lymphomas. *Ann N Y Acad Sci*, .428:158-172.

[4] Lester EP, et al. (1980) Computer-assisted analysis of two-dimensional electrophoreses of human lymphoid cells. *Clinical Chemistry*, 26:1392-1402

[5] Kalia A and Gupta RP. (2005) Proteomics: a paradigm shift. *Crit Rev Biotechnol*, 25:173-198

[6] Dowsey AW, et al. (2003) The role of bioinformatics in two-dimensional gel electrophoresis. *Proteomics,* 3:1567-1596

[7] Lester EP LP, Lipkin L, Cooper HL. (1980) Computer-assisted analysis of two-dimensional electrophoreses of human lymphoid cells. *Clin Chem,* 10: 1392-1402

[8] Thompson CM and Shure L. (1995) MATLAB Image Processing Toolbox User's Guide. *The MathWorks, Natick, Massachusetts.*

[9] *Matlab Function Reference* (2000) Vol. 2, FO, MathWorks. Inc,

[10] Peri S, et al. (2003) Development of Human Protein Reference Database as an Initial Platform for Approaching Systems Biology in Humans. *Genome Research*, 13:2363-2371

[11] Pocock M, et al. (2000) BioJava: open source components for bioinformatics. *ACM SIGBIO Newsletter*, 20:10-12

[12] MySql AB. (2004) MySQL Reference Manual.

Published Jan 01, 1970 | doi:10.4051/ibce.2009.3.0017