cDNA マイクロアレイ画像からのスポット認識および定量の自動化に関する研究（２００１．６．２６）

長嶋 剛史[1]、 高橋 勝利[2]、坊農 秀雅[34]、 岡崎 康司[34]、 小長谷 明彦[135]

[1] 北陸先端科学技術大学院大学 知識科学研究科
[2] 産業技術総合研究所 生命情報科学研究センター
[3] 理化学研究所 ゲノム科学総合研究センター
[4] 遺伝子構造・機能研究グループ
[5] ゲノム情報科学研究グループ

概要

本稿では、一度に多数の遺伝子の発現量を調べることができる cDNA マイクロアレイ実験によって得られた画像から、各遺伝子の発現量を求める際に必要となる一連の操作を自動で行う手法について述べる。マイクロアレイ画像の解析においては、各遺伝子の画像中における位置の決定とそれらの発現量の測定が必要である。我々はこれをローリングボール法、ラプラシアンフィルタ等の画像処理技術を組み合わせることで実現した。実際に実験で得られた画像を用いて従来法と比較を行った結果から、我々の手法は従来法に比べ、より高速に解析が可能であること、定量結果の精度が従来法とほぼ同程度であることが分かった。

Fully-Automated Spot Recognition and Quantification from cDNA MicroArray Images

Takeshi Nagashima[1], Katsutoshi Takahashi[2], Hidemasa Bono[35], Yasushi Okazaki[35], Akihiko Konagaya[145]

[1] School of Knowledge Science, Japan Advanced Institute of Science and Technology
[2] Computational Biology Research Center,
National Institute of Advanced Industrial Science and Technology (AIST)
[3] Genome Exploration Research Group,
[4] Bioinformatics Group,
[5] Genomic Sciences Center (GSC), RIKEN (The Institute of Physical and Chemical Research)

Abstract

We have developed a powerful image analysis tool for cDNA microarray images. cDNA microarray is one of the methods to simultaneously monitor a large number of gene expressions under various environmental conditions. The essential problems of image analysis for microarray are to locate spots and measure their respective intensity. Our tool uses a combination of some image processing operators including rolling ball, laplacian filter and so on, are used for solving these problems. Experimental result shows that our method can find not only normal spots but also overlapped spots as flagged spots with no human interventions. Moreover, it takes only eight minutes for one image set, while it takes one or two hours for visual inspection in interactive systems.

1 Introduction

Recently, a large number of genomic sequence data is made available by the ongoing world-wide genome projects. It is commonly recognized that, due to the rapid growth of sequence data, it is necessary to develop methods that can handle functions of genes [1]. cDNA microarray, the latest breakthrough in experimental molecular biology, is well cited for this purpose.

cDNA microarray means a collection of cDNAs stamped on slide glass systematically with high density [6]. This useful technology allows us to monitor and measure thousands of gene expression level simultaneously. For example, developmental and metabolic pathways are delineated by expression profile using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays [5].

In general, the raw results of microarray experiments are given by two images. They are labeled different fluorescent dye which are introduced in reverse transcription reaction to denote which is target or reference DNA. Using different labeled DNA and comparing their amount of expression, difference of mRNA expression level between different two conditions can be measured. To deal with gene expression data, these images have to be quantified, because their intensity reflects the amount of gene expression. In quantification, to raise data quality, irregular spots (e.g. contaminated, injured, and missrecognized spots) should be removed. Although there exist software packages to do this, most of these softwares require human interventions. It is hence desirable to automate the process. For instance, a user often has to correct spot locations when he or she uses ScanAlyze [2]. Because, mismatch occurs between grid model and spot locations in a real image. In grid model, spots locates exactly on lattice, but in real images spots are not always arranged lattice precisely. It requires considerable time and effort to correct this difference. In addition to being time-consuming and laborious, human interventions also increase the risk of introducing artifacts. So, the ob-

jectivity of results obtained this way remains questionable. Furthermore, even enough commercial softwares are generally expensive, they do not actually facilitate image analysis [8].

To overcome the above problems, we have developed a suite of programs for cDNA microarray image analysis. These programs can automatically recognize and quantify spots.

2 Methods

In our method, microarray image analysis consists of two parts ; spot recognition and quantification. Figure 1 shows the flowchart of the two parts.

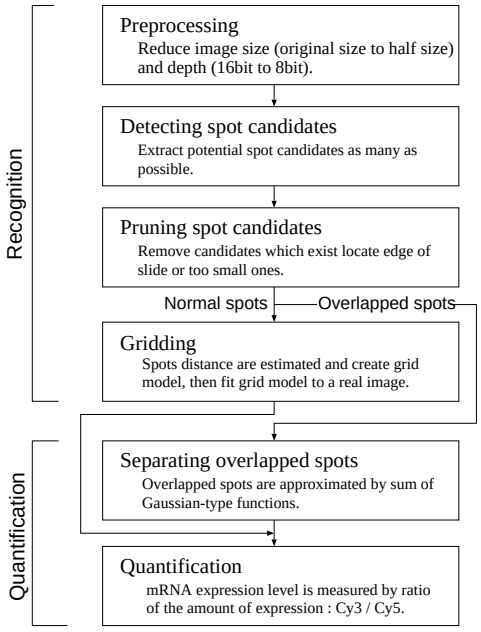

Figure 1: Flowchart of cDNA microarray image analysis.

In the rest of this section, we will describe each operation shown in Figure 1.

2.1 Preprocessing

Preceding the image analysis, DNA spotted slide glass should be digitized with an image scanner. In preprocessing, to reduce processing time, scanned images file was re-sampled so as to give half size of original image. Here, to enhance weak signal, gray level of the image is transformed by logarithmic function. Then, to reduce the effect of noise, standard smoothing and background normalization (rolling ball operator [7]) were applied. This preprocessed image is used in the following steps.

2.2 Detecting spot candidates

To detect spot candidates, we applied two image processing operators, binarization and laplacian filter, to preprocessed image. Then, the resultant two images are superimposed. In this image, we can find possible spot candidates easily.

We employed binarization operator to detect spot regions. Laplacian filter is used to extract outline of spots. It is important to obtain spot candidates as many as possible, in this step. This requirement was implemented by using mode of intensity histogram of inversed image as a threshold of binarization.

2.3 Pruning spot candidates

Before performing grid assignment, we pruned spot candidates. If a spot is within the edge of slide or too small, then the spot is removed from spot candidates. Here, we regarded the edge of slide as the region within 10 pixels from the edge. When a spot consists under 4 pixels, it regarded as small spot. Then, contaminated, injured or miss-recognized spots are marked and recognized as irregular spots. This process is called flagging and marked spots are called flagged spots.

After pruning was finished, we flagged irregular shape spots. First, we calculated three spot feature values, (1) the number of pixel included in a spot, (2) ratio of spot height and width, and (3) circularity (plausibility of round shape), for each spot. Third feature value is calculated by dividing square of spot circumference by its area. Then, if a spot have at least one spot feature value which is out of $\pm 4\sigma$, the spot is flagged.

2.4 Gridding

To decide spot address on the slide glass, gridding was performed. In gridding, we assumed that spots are arranged lattice on a slide glass. Here, grid means a circle that encloses a spot. Then, we construct grid model by using (1) spot candidates found in above step, (2) the number of block rows and columns, and (3) the number of spot rows and columns included a block. Second and third values are specified by user. Then, we fitted this model to real image.

Gridding process consists of two steps ; estimating spot distance and creating grids in each block.

Estimating spot distance To estimate distance between a spot and adjacent one, we took into account that distance between blocks is larger than distance between spots. Then, we estimated distance between spots. Estimation was performed as below.

1. Create two index histogram along with horizontal and vertical orientation of the image.

2. Detect peak of both index histogram.
3. Create histogram of distance between peaks, then the mode value of this histogram is employed as distance between spots.

Creating grids in each block With the spot distance calculated above procedure, we can easily create grid model. Preceding the grid creation, an image is splitted to $bx \times by$ regions with a little overlap. Here, bx and by are the number of block rows and columns specified by user.

After the grid model was constructed, we had to fit this model to a real image. We detected block locations as follows. First, fitting score between a block in constructed grid model and each location in a splitted region are calculated. If the center point of a spot candidate equals to grid center, fitting score is added up. Then, the location which gives maximum fitting score is employed as a block location. After block location is fixed, spot locations corresponding to each grid position are determined by same way. If multiple spots are corresponding to one grid, the nearest spot is assigned to this grid. When gridding is finished, spots outside the block are removed.

The result of applying procedures described above to raw image is shown in Figure 2.

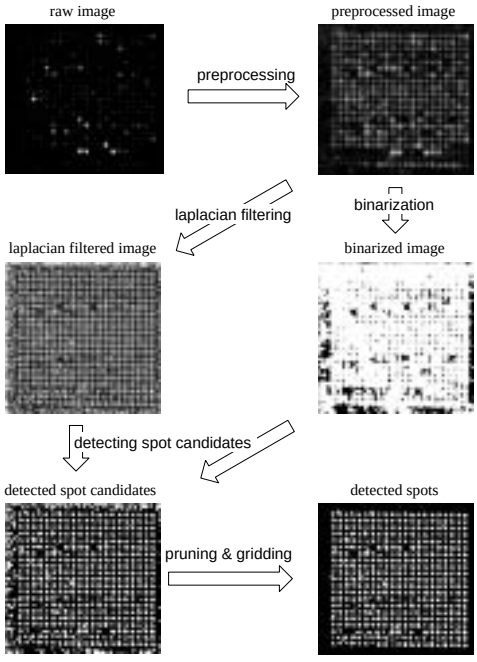

Figure 2: The flow of spot recognition with typical intermediate images.

2.5 Separating overlapped spots

By applying the operators described above to raw image, we could obtain spot locations which stamped clearly. But some spots could not be recognized well. In most cases, these spots have an outlier feature value. Consequently, they were flagged in the pruning step. To separate these overlapped spots, gaussian-type function fitting [9] was done by least square error minimization.

2.6 Quantification

Now, we have spot locations and its radius. The next step is quantification. In this step, we returned to the original scanned image.

For the quantification, the circle method [3] is employed. In this method, first spot domain is defined as a square with middle point of spot center and center of adjacent four spots. Then, average intensity value of all pixel contained in spot circle is regarded as foreground intensity. Here, spot circle is defined by its location and radius determined in the spot recognition step. Background intensity is estimated as median value of pixel intensity which is located outside the spot circle but located inside the spot domain. The use of median value avoids effect of noise. Finally, spot intensity is calculated by subtracting background intensity from foreground intensity.

3 Results and Discussions

To validate our method, we compared our system with ScanAlyze which is widely used for cDNA microarray image analysis. In the following experiment, we used two images. These are derived from different fluorescent dye, Cy3 and Cy5.

3.1 Processing time

To examine the effect of automation, we measured processing time required for analyzing two images. The results are shown in Table 1.

Table 1: Processing time for two images.

Method	Time (seconds)
Our method	450
ScanAlyze	> 7200

From Table 1, we see that automation reduces processing time to less than one fifteenth of what can be achieved with ScanAlyze.

3.2 The accuracy of background estimation

To evaluate the accuracy of background estimation, we measured correlation coefficient between spot intensity and background intensity. The reason why

we use this value is that there should be no correlation between these two values. Otherwise, intensities could be dependent on factors other than the hybridization of the target to the probe [4]. Therefore, it is considered that the estimation gives near zero correlation coefficient is better than the estimation gives far from zero. In calculation of correlation coefficient, we used only lower half of intensities. Because weak signals are more sensitive to background intensity than strong ones.

Correlation coefficient is shown in Table 2.

Table 2: Correlation coefficient between lower half of spot intensity and its background intensity.

	Our method	ScanAlyze
Cy3	0.1455	0.1370
Cy5	0.0225	0.0517

From Table 2, spot intensity and background intensity are correlated very weakly and we confirmed that our method can estimate background intensity accurately as well as ScanAlyze.

3.3 The effect of flagging

To examine the effect of fully-automated flagging, we counted the number of flagged spots. The number of flagged spots assigned by our method and ScanAlyze are 140 and 129. The number of spot type was also counted. This value is shown in Table 3. In Table 3, for example, second column denotes that 109 spots are recognized as normal spots in our method and recognized as flagged spots in ScanAlyze.

Table 3: The number of spot type.

Our method	ScanAlyze	
normal	normal	6807
normal	flagged	109
flagged	normal	120
flagged	flagged	20

From the number of flagged spots, it seems that about same number of spots are flagged by two methods. But from Table 3, it is also clarified that these methods flagged different spots.

It is considered that these results are derived from difference between two methods. In ScanAlyze, most of flagged spots are tended to be overlapped. Because, these spots are seem to be contaminated spots in visual inspection. On the other hand, our method tends to flag not only overlapped spots but also small or not round shape spots. This is the reason why two methods flag different spots.

4 Summary

In this paper, we described the method which realizes fully automated recognition and quantification of a large number of spots found in cDNA microarray images. It provides rapid and objective way to analyze numerous spots without any human interventions. In fact, it takes less than only eight minutes for one image set, while it takes more than one or two hours by visual inspection. It also provides the way to accurate background estimation as well as visual inspection.

As a conclusion, our method highly contributes to the automation of high-speed microarray image analysis which is one of the major bottlenecks using current technology. We believe that our method is feasible and could be used for a systematic analysis of gene function and network.

References

[1] Brown, P.O., Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, 21:33–37, 1999.

[2] Eisen, M.B. ScanAlyze Ver.2.44. available from `http://rana.lbl.gov/`.

[3] Eisen, M.B. *ScanAlyze User Manual*. Stanford University, 1999.

[4] Eisen, M.B., Brown, P.O. DNA Arrays for Analysis of Gene Expression. In *Methods in Enzymology*, volume 303, pages 179–205. Academic Press, 1999.

[5] Miki, R., Kadota, K., Bono, H., et al. Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays . *Proc. Natl. Acad. Sci. USA*, 98:2199–2204, 2001.

[6] Schena, M., Shalon, D., Davis, R.D., Brown, P.O. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270(5235):467–470, 1995.

[7] Sternberg, S.R. Biomedical Image Processing. *IEEE Computer*, January:22–34, 1983.

[8] Strehlow, D. Software for Quantitation and Visualization of Expression Array Data. *BioTechniques*, 29(1):118–121, 2000.

[9] Takahashi, K., Nakazawa, M., Watanabe, Y., Konagaya, A. Fully-Automated Spot Recognition and Matching Algorithms for 2-D Gel Electrophoretogram of Genomic DNA. *Genome Informatics*, 9:161–172, 1998.