Computerized Medical Imaging and Graphics 33 (2009) 40–49

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

An automated method for gridding and clustering-based segmentation of cDNA microarray images

Nikolaos Giannakeas[a,b], Dimitrios I. Fotiadis[b,c,*]

[a] Laboratory of Biological Chemistry, Medical School, University of Ioannina, Ioannina, Greece
[b] Unit of Medical Technology and Intelligent Information Systems, Department of Computer Science, University of Ioannina, Ioannina, Greece
[c] Biomedical Research Institute - FORTH, Ioannina, Greece

ARTICLE INFO

Article history:
Received 24 March 2008
Received in revised form
18 September 2008
Accepted 6 October 2008

Keywords:
Microarray image processing
Gridding
Segmentation
K-means
Fuzzy C means

ABSTRACT

Microarrays are widely used to quantify gene expression levels. Microarray image analysis is one of the tools, which are necessary when dealing with vast amounts of biological data. In this work we propose a new method for the automated analysis of microarray images. The proposed method consists of two stages: gridding and segmentation. Initially, the microarray images are preprocessed using template matching, and block and spot finding takes place. Then, the non-expressed spots are detected and a grid is fit on the image using a Voronoi diagram. In the segmentation stage, K-means and Fuzzy C means (FCM) clustering are employed. The proposed method was evaluated using images from the Stanford Microarray Database (SMD). The results that are presented in the segmentation stage show the efficiency of our Fuzzy C means-based work compared to the two already developed K-means-based methods. The proposed method can handle images with artefacts and it is fully automated.

1. Introduction

The analysis of gene expression has become simple and efficient using microarray technology [1]. The cycle of the microarray experiment begins with the biological question and ends with the data analysis results along with the biological conclusions, which might lead to a new question and so on. Microarrays can yield expression profiles for thousands of genes simultaneously in a single hybridization experiment.

During a biological experiment, two messenger RiboNucleic Acid (mRNA) samples are reverse transcribed into complementary DeoxyriboNucleic Acid (cDNA). Most experiments compare a normal sample with a cancer sample in order to find, which genes are related to the current type of cancer. DeoxyriboNucleic Acid (DNA) obtained from thousands of known genes of interest is printed on a glass microscope slide by a robotic arrayer. The cDNA samples, labeled with two different fluorescent dyes, are hybridized with the known genes on the slide at the same time. The most common dyes for tagging cDNA are the red fluorescent dye Cy5 (emission from 630 to 660 nm) and the green-fluorescent dye Cy3 (emission from 510 to 550 nm) [2].

The microarray images are generated using a scanner. The slide is scanned in the two wavelengths, producing two 16-bit images corresponding to the two fluorescent dyes. In terms of microarray image analysis the two channels are referred to as red and green channels. Thus, the image can be represented as an RGB image with blue being zero. The digitization of the image is modified in order to generate spots with 10 pixels diameter (2–5 mm). The images contain several blocks (or subgrids), which consist of several spots, placed in rows and columns. The level of intensity of each spot represents the amount of sample hybridized with the corresponding gene.

The processing of microarray images [3] provides the input for further analysis of the extracted microarray data [4]. It includes the following stages:

- Spot addressing and gridding, which are the processes of assigning the location of each spot and fit a grid on the image.
- Segmentation, which is the process of grouping the pixels with similar features (this step results in the separation of foreground and background pixels).
- Intensity extraction, which calculates red and green foreground fluorescence intensity pairs and background intensities.

An ideal microarray image must have the following properties [5]:

- All the blocks are of the same size.
- The spacing between the blocks is regular.

* Corresponding author at: Unit of Medical Technology and Intelligent Information Systems, Department of Computer Science, University of Ioannina, PO Box 1186, GR 451 10 Ioannina, Greece. Tel.: +30 26510 98803; fax: +30 26510 98889.
E-mail address: fotiadis@cs.uoi.gr (D.I. Fotiadis).

0895-6111/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compmedimag.2008.10.003

- The spots are centered on the intersections of rows and columns.
- The size and shape of the spots is perfectly circular and identical for all spots.
- The location of the blocks is fixed in the images for a given slide type.
- No dust or other contamination is present on the slide.
- The background intensity is minimal and uniform.

However, a scanned microarray image has none of the above properties. Thus, automated gridding and spot addressing is crucial for the microarray image processing.

The gridding methods proposed in the literature are manual, semi-automated and automated. Several software packages have been developed like Genepix [6] and ScanAlyze [7], which require inputs by the user. Towards automated methods, there exist several approaches, which process the vertical and horizontal projections of the image and generate lines in the image according to the valleys of the 1-D signal [8–10]. Morphological operators [11] and smoothing filtering [12] have also been used. Other methods preprocess the image using an initial segmentation and assign the centers of the objects in the image. Such techniques use histogram-based segmentation [13], template matching [14,16] combined with graph models [13], several types of transformations such as Affine [14], and Radon [15], or others, which take into account the symmetries of the microarray image [16]. The proposed gridding work attempts to engage the approaches, which are based on the template matching techniques with the approaches, where the projections of the image are processed. It consists of several steps to deal with the image rotation, the existence of low-intensity spots and the existence of artefacts and other types of noise. The current method is fully automated and does not require tuning of its parameters. In addition, our gridding approach is able to allocate effectively the non-expressed spots. The use of outlier detection can be considered as an advantage since artefacts existing in the image, are detected effectively.

The primary goal of microarray image processing is to extract the intensities of red and green channels. Thus, image segmentation must be applied, grouping the pixels of the image into foreground and background. The background intensities are used to adjust the foreground intensities for local noise, resulting in corrected red and green intensities for each spot [17].

The methods, which have been proposed for the segmentation of microarray images can be classified into four categories: (i) early approaches, which are based on fixed or adaptive circle segmentation. A circle is fit around each spot, characterizing the pixels in the circle as signal pixels and the pixels out of the circle as background pixels. Fixed and adaptive circle segmentation is used by ScanAlyze [7] and Dapple [18]. (ii) Histogram-based segmentation, which is used by QuantArray [19]. A threshold value is calculated [20], and pixels with intensity lower than the threshold are characterized as background pixels, whereas pixels with higher intensity as signal pixels. (iii) Adaptive shape segmentation methods, which are based on the watershed transform [21], seed region growing [22,23] and Marcov Random Field (MRF) [24]. (iv) The most recent techniques, which employ clustering algorithms like K-means [25–29], Fuzzy C means (FCM) [25], Expectation–Maximization (EM) [9], Partitioning Around Medoids (PAM) [30] and model-based clustering [31]. Lehmussola et al. [32] compared most of the above segmentation approaches and they have shown that the most effective are those based on K-means clustering. The proposed segmentation method also employs clustering techniques, including an innovative set of features to deal with artefacts and donut spots. Instead, most of the already developed methods for the clustering-based segmentation of the microarray image use only the intensities of the two channels as features. On the other hand, a set of features is used by the proposed method for each channel. In this way, the expression levels of each gene are accurately computed, according to the requirements of the biological experiment.

2. Material and methods

The proposed method consists of two stages. Fig. 1 demonstrates the flowchart of the proposed work. In the first stage, a gridding algorithm [34] is implemented to identify regions around each spot. Next, clustering techniques are used for the image segmentation in the above regions, which result in the identification of the signal and background pixels and artefacts. The first stage consists of five steps. Initially, the raw microarray image is preprocessed using a template matching technique and the image is rotated using least square fitting. In the second step, the blocks of the image are detected, using 1-D projections of the image. In the third step, the coordinates of the centers of each spot are addressed and in the fourth step the non-expressed spots are detected. Finally, in the last step a grid is fit on the image using a Voronoi diagram. In the second stage (segmentation), K-means and Fuzzy C means clustering is employed for grouping the pixels of the images into foreground, background pixels and artefacts. Note, that the above procedure is applied in parallel to both red and green channels.

2.1. Gridding

2.1.1. Preprocessing

The preprocessing of the image begins with the background removal. This kind of noise appears in most of the microarray images due to the emission of the slide, which affects the entire image. Since the number of background pixels is larger than that of the signal pixels, the median value of all pixel intensities is very close to the value of the global background. According to this assumption, the median value is used as the threshold for the removal of the noise, which is generated from the emission of the slide. The intensities of the pixels, which have intensities lower than the median intensity of the grayscale image, are set equal to zero [16]. All the other values of the image (i.e. the values, which have intensities greater the median value) are not changed during the background removal procedure.

In order to locate the objects of the image, which are similar to a spot, we perform an initial segmentation using template matching. The appropriate template is selected and placed at each pixel in the image. The size of the template is approximately equal to the size of the theoretical spot and the template follows a 2-D Gaussian distribution as it is shown in Fig. 2. Unfortunately, the size of the spots of the image varies due to the level of the hybridization. The aim of this procedure is to measure the similarity between the template and the image at each pixel. For this reason, we calculate the correlation coefficient of the values of the template with the corresponding values of the image. If all the spots have 11×11 size and their pixels follows the Gaussian distribution, the correlation coefficient of the center pixel of the spot will be equal to 1. As the correlation increases, the probability of a pixel to belong to a spot increases. Using a threshold value for the correlation coefficient, the image is converted to binary. This threshold was set to $r = 0.3$, which has been obtained heuristically after several experiments.

Since the microarray is often not parallel to the image boundaries, the estimation of the rotation angle is necessary. For this reason, we compute and afterwards eliminate the rotation between the microarray and the boundaries of the image. The slope between the four boundaries of the microarray and the corresponding boundaries of the image is found. The median value of these four slopes is calculated and the array is rotated accordingly. In some