Computers in Biology and Medicine 43 (2013) 705–716

Contents lists available at SciVerse ScienceDirect

Computers in Biology and Medicine

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

Segmentation of microarray images using pixel classification—Comparison with clustering-based methods

Nikolaos Giannakeas [a,b], Petros S. Karvelis [c], Themis P. Exarchos [b], Fanis G. Kalatzis [b], Dimitrios I. Fotiadis [b,*]

[a] Laboratory of Biological Chemistry, Medical School, University of Ioannina, GR 45110 Ioannina, Greece
[b] Unit of Medical Technology and Intelligent Information Systems, Department of Materials Science and Engineering, University of Ioannina, PO Box 1186, GR 45110 Ioannina, Greece
[c] Department of Computer Science, University of Ioannina, GR 45110 Ioannina, Greece

ARTICLE INFO

Article history:
Received 10 January 2012
Accepted 14 March 2013

Keywords:
Microarray image segmentation
Support vector machines
Microarray experiment

ABSTRACT

Objective: DNA microarray technology yields expression profiles for thousands of genes, in a single hybridization experiment. The quantification of the expression level is performed using image analysis. In this paper we introduce a supervised method for the segmentation of microarray images using classification techniques. The method is able to characterize the pixels of the image as signal, background and artefact.

Methods and material: The proposed method includes five steps: (a) an automated gridding method which provides a cell of the image for each spot. (b) Three multichannel vector filters are employed to preprocess the raw image. (c) Features are extracted from each pixel of the image. (d) The dimension of the feature set is reduced. (e) Support vector machines are used for the classification of pixels as signal, background, artefacts. The proposed method is evaluated using both real images from the Stanford microarray database and simulated images generated by a microarray data simulator. The signal and the background pixels, which are responsible for the quantification of the expression levels, are efficiently detected.

Results: A quality measure (q_{index}) and the pixel-by-pixel accuracy are used for the evaluation of the proposed method. The obtained q_{index} varies from 0.742 to 0.836. The obtained accuracy for the real images is about 98%, while the accuracies for the good, normal and bad quality simulated images are 96, 93 and 71%, respectively. The proposed classification method is compared to clustering-based techniques, which have been proposed for microarray image segmentation. This comparison shows that the classification-based method reports better results, improving the performance by up to 20%.

Conclusions: The proposed method can be used for segmentation of microarray images with high accuracy, indicating that segmentation can be improved using classification instead of clustering. The proposed method is supervised and it can only be used when training data are available.

1. Introduction

Microarray technology provides the quantification of gene expression profiles under different experimental conditions [1]. As a result, microarrays have been extensively applied for the prediction, staging and diagnosis of several types of cancer. Changes in the transcription level of almost all genes in the whole genome taking place in a specific cell type or tissue can be estimated, during different developmental stages of the disease and in response to external stimuli, such as drug treatment. Thus, the response patterns can be used to explore disease mechanisms, to predict disease progression and assess activities of new compounds, such as drugs.

A microarray image is a dual-channel image; each channel represents the intensity level at each microarray location for both wavelengths that fluorescence dyes emit. The intensity level is correlated with the absolute amount of RNA in the original sample, and by extension, the expression level of the gene associated with this RNA. Typically, a microarray image contains several blocks (or subgrids) which consist of a number of spots, placed in rows and columns. The level of intensity of each spot represents the amount of sample hybridized with the corresponding gene [2].

The efficiency of the microarray image processing directly affects the precision of the microarray data analysis [3]. Microarray image processing usually follows three main stages: (i) spot addressing or gridding, (ii) segmentation and (iii) intensity extraction. Initially, the

* Corresponding author. Tel.: +30 26510 08803; fax: +30 26510 08889.
E-mail address: fotiadis@cs.uoi.gr (D.I. Fotiadis).

0010-4825/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2013.03.003

locations of each spot in the image are found by the spot addressing. During the segmentation, the pixels of the image are divided into foreground pixels, called signal pixels, and background pixels. The intensities of the background pixels are used to adjust the foreground intensities for local noise, resulting in corrected red and green intensities for each spot [4]. The intensity extraction stage calculates the ratio of the background corrected values between the two channels, for each spot in the image.

Four categories of methods for microarray image segmentation have been proposed: (a) fixed [5] or adaptive circle segmentation [6–8] which are based on the assumption that all the spots are circular. Thus, a circular mask is placed on the location of each spot, considering all pixels inside the mask as foreground pixels. Adaptive circle segmentation methods modify the radius of the circular mask manually [6] or automatically [7] for each spot, while the radius of the circular mask is fixed for the fixed circle segmentation case. (b) Histogram-based techniques [9–12] estimate a threshold such that pixels with intensity lower than the calculated threshold are characterized as background pixels, whereas pixels with higher intensity as signal pixels. The most known method for the estimation of the appropriate threshold uses the Mann–Whitney test [10]. (c) The adaptive shape segmentation is more sophisticated and they do not include any assumption about the size and the shape of the spot. Algorithms such as seed region growing [11], watershed transform [12] and Markov random field [13] have been employed. (d) The fourth category is based on machine learning techniques. More specifically, methods in this category employ clustering algorithms, such as K-means [14–18], fuzzy C-means (FCM) [17–20], expectation–maximization [21] and partitioning around medoid (PAM) [22] for unsupervised segmentation. Other clustering-based methods, which use the kernel-based estimator [23], and model-based clustering [24], have been also presented. In the machine learning category classification-based methods are also included [25,26]. Apart from the above methods, there are several hybrid methods which employ both image processing and machine learning techniques. Weng et al. [27] presented a segmentation method using both the cosine discrete transform and the K-means clustering algorithm. Battiato et al. [28,29] employed both the statistical region merging (SRM) algorithm and the K-means algorithm. A method which is based on self-organizing map (SOM) and the Fuzzy C-means has been presented by Battiato et al. [30].

The current work introduces a novel pixel-by-pixel supervised segmentation method which is based on classification techniques. The method results in a trained system, which segments the spot of the image by classifying their pixels into signal, background, and artefact pixels. More specifically, the method classifies the pixels of the image into two (foreground and background) or three categories (signal, background and artefacts) using support vector machines (SVM). Due to the fact that the training of the method requires training data, a set of spots with pixel-by-pixel information is extracted. We use the fixed circle technique [5] to extract pixel-by-pixel information for the real images, while for the simulated images the training data are extracted directly during the production of the images. The proposed method is advantageous compared to the clustering-based methods, due to the direct characterization of each pixel to the designated category. Otherwise, using clustering techniques different clusters are generated but no distinction exists between them unless a set of rules is applied to separate them.

For the evaluation of the proposed method real and simulated images are employed. Two classes (signal and background pixels) are used for the evaluation using real images from the Stanford microarray database (SMD), while three classes of pixels are produced from the simulated images. Apart from the signal and the background pixels, the third class includes pixels of artefacts, pixels of the contour of the spot, and pixels of inner holes which exist in donut spots [24]. A set of features from each pixel is used as input for the classification. Principal component analysis (PCA) is applied to reduce the dimensionally of the input feature vector and reduce the computational effort of the method. The proposed method is advantageous since the segmentation method uses information from both channels to segment the multichannel image, whereas most of the previous methods segment each one of the two channels (red and green) separately.

The paper is organized as follows: in the next section a brief description of previous works for the segmentation of microarray images is presented. In the third section the datasets which are employed for training and evaluation of our experiments are described. The five steps of the proposed method and the obtained results are presented in the fourth and fifth sections, respectively. Finally, a quantitative and qualitative comparison between segmentation methods based on supervised classification and unsupervised clustering is provided.

2. Datasets

Real microarray images from the SMD [31] are employed for the evaluation of the proposed approach. The blocks of this image consist of 576 spots, forming 24×24 rows and columns (112,896 pixels). The annotation file of the SMD is used to extract the pixel-by-pixel information, simulating the fixed circle segmentation. For this task, the known radius of the fixed circle and the coordinates of the centers of each spot are used. A binary map is generated for the whole image, characterizing the pixels inside the circle as signal pixels and the pixels outside of the circle as background.

We use also simulated images which are generated using a spot simulator [32]. To produce the simulated image, the annotation of the SMD is used. The mean intensities of each spot are given as an input in the simulator in order to provide a block, similar to the real one. This spot simulator uses several tuning parameters: (i) general options control the behavior of the model, (i.e. what kind of data are simulated, such as cDNA microarray, affymetrix, etc.), (ii) noise options determine the statistical properties of the data, (iii) slide options determine the structure of the slide (for example the number of the spots in a block), (iv) hybridization options determine the quality of the slide, (v) scanner options include parameters which control the quality of the image (i.e. saturation, hue, etc.). The simulator proposes three different sets of parameters for the slide, the hybridization and the scanner options to categorize the generated images into good, normal, and bad. For instance, a good image consists of circular spots with homogeneous intensity, while the normal and the bad images consist of non-circular spots in which the pixel intensities follow the Gaussian distribution.

To extract the pixel-by-pixel information for each simulated image, the simulator is improved in order to build the annotation during the generation of the original image. During the simulation, an image with zeros is generated having the same size as the original image. When the simulator adds a spot on the original image, the corresponding pixel is marked on the annotation image representing the signal pixels. Accordingly, when an artefact is added in the original image, the corresponding pixels in the annotation image are marked as artefacts. The simulator adds several types of artefacts in the image. Apart from the background noise, the model adds (i) spot bleeding effect which causes the dyes to be outside the printed spot area, (ii) scratches which may introduce to the slide surfaces due to the careless handling of the slides, and (iii) air bubbles and inner holes in donut spots. The pixels of all these artefacts are marked on the initial binary image, in order to create the third class for the classification problem.