Global Journal of Computer Science and Technology
Volume 11 Issue 19 Version 1.0 November 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Segmentation of Microarray Image Using Information Bottleneck

By S.Raghavarao, M.S.Madhanmohan, Dr.G.M.V.Prasad

BVC Engineering College, Odalarevu

Abstract - DNA microarrays provide a simple tool to identify andquantify the gene expression for tens of thousands of genessimultaneously. The DNA microarray image analysis includes three tasks: gridding, segmentation and intensity extraction.Spots segmentation, which isto distinguish the spot signals from background pixels,is a critical step in microarray image processing. In this paper, new image segmentation algorithm based on the hard version of the information bottleneck method is presented. The objective of this method is to extract a compact representation of a variable, considered the input, with minimal loss of mutual information with respect to another variable, considered the output. The input variable here, is the histogram bins and the output variable is the set of regions obtained from the split and merge algorithm. The proposed method is compared with existing segmentation methods such as k-means and Fuzzy C-means. The experimental results show that the proposed algorithm has segmented spots of the microarray image more accurately than other segmentation methods.

Keywords : *Image Processing, Microarray, Image Segmentation.*

GJCST Classification : *I.4.5, I.4.4*

Segmentation of Microarray Image Using Information Bottleneck

S.Raghavarao [α], M.S.Madhanmohan [Ω], Dr.G.M.V.Prasad [β]

Abstract - DNA microarrays provide a simple tool to identify andquantify the gene expression for tens of thousands of genessimultaneously. The DNA microarray image analysis includes three tasks: gridding, segmentation and intensity extraction.Spots segmentation, which isto distinguish the spot signals from background pixels,is a critical step in microarray image processing. In this paper, new image segmentation algorithm based on the hard version of the information bottleneck method is presented. The objective of this method is to extract a compact representation of a variable, considered the input, with minimal loss of mutual information with respect to another variable, considered the output. The input variable here, is the histogram bins and the output variable is the set of regions obtained from the split and merge algorithm. The proposed method is compared with existing segmentation methods such as k-means and Fuzzy C-means. The experimental results show that the proposed algorithm has segmented spots of the microarray image more accurately than other segmentation methods.

Keywords : Image Processing, Microarray, Image Segmentation.

I. Introduction

Microarrays, widely recognized as the next revolution in molecular biology, enable scientists to analyze genes, proteins and other biological molecules on a genomic scale [1]. A microarray is a collection of spots containing DNA deposited on the solid surface of glass slide. Each of the spot contains multiple copies of single DNA sequence [2].

Microarray expression technology helps in the monitoring of gene expression for tens and thousands of genes in parallel. During the biological experiment, the mRNA of two biological tissues of interest is extracted and purified. Each of the mRNA samples are reverse transcribed into complementary DNA (cDNA) copy and labeled with two different fluorescent dyes resulting in two fluorescence-tagged cDNA (red Cy5 and green Cy3). The tagged cDNA copies, called the sample probe, are hybridized with the slide's DNA spots. The hybridized glass slides are fluorescently scanned at different wavelengths (corresponding to the different dyes used), and two digital images are produced, one for each population of mRNA. Each digital image contains a number of spots of various fluorescence intensities. The intensity of each spot is proportional to the hybridization level of the cDNAs and the DNA dots, the gene expression information is obtained by analyzing the digital images [3].

The processing of the microarray images usually consists of the following three steps: (i) gridding, which is the process of assigning the location of each spot in the image. (ii) Segmentation, which is the process of grouping the pixels with similar features and (iii) Intensity extraction, which calculates red and green foreground intensity pairs and background intensities.

Nowadays, segmentation algorithms such as K-means and Fuzzy C-Means have been used for the segmentation of spots of the microarray images. In this paper, we present a histogram clustering algorithm for segmentation of spots of the microarray image. The proposed algorithm is based on the minimization of the mutual information loss, where now the input variable represents the histogram bins and the output is given by the set of regions obtained from the split and merge algorithm.

The rest of the paper is organized as follows.

Section II presents K-Means Algorithm, Section III presents Fuzzy C-Means Algorithm, Section IV presents present Histogram Clustering algorithm for segmentation of spots in Microarray image, Section V presents experimental results and finally Section VI reports conclusion.

II. K-means Clustering Algorithm

K-means is one of the basic methods in clustering introduced by Hartigan et al. in 1979 [3]. This method is applied to microarray image segmentation in recent years [21]. K-means clustering algorithm implemented in this paper aims to group the pixels into two clusters. Given x = $\{x_1,x_2 ,...,x_N\}$ and c = $\{c_1, .. \ c_j \}$ representing the pixels of microarray image and clusters respectively, the objective is to minimize the sum of squares of the distances given by the following:

$$d_{ij} = || x_i\text{-}c_j ||. \quad arg\ min \sum_{i=1}^{N} \sum_{j=1}^{C} d_{ij}^2 \qquad (1)$$

First two cluster centers c_1 and c_2, the centroid of spots and background have to be initialized at the outset. Iteratively, the pixels are assigned to the closest cluster and the new centroid of a cluster is calculated by the following: The k-means algorithm to segment microarray image is summarized as below:

Author [α Ω β] : BVC Engineering College, Odalarevu.

Algorithm KM(x,n,c)
Input:
N=number of pixels to be clustered;
x = $\{x_1,x_2,...,x_N\}$ pixels of microarray image;
c=2: foreground and background clusters;
Output:
cl: cluster of pixels

Begin
Step_1: Cluster centroids are initialized,
Step_2: Compute the closest cluster for each pixel and classify it to that cluster,
Step_3: Compute new centroids after all the pixels are clustered,
Step_4: Repeat the Steps 2-3 till the sum of squares given in Equation
End.

III. Fuzzy C-Means Clustering

Algorithm Fuzzy C-Means(x,n,c,m)
Input:
N=number of pixels to be clustered;
x = {x1,x2 ,...,xN}: pixels of microarray image;
c=2: foreground and background clusters;
m=2: the fuzziness parameter;
Output:
u: membership values of pixels

Begin
Step_1: Initialize the membership matrix u_{ij} is a value in (0,1) and the fuzziness parameter m. The sum of all membership values of a pixel belonging to clusters should satisfy the constraint expressed in the following.

$$\sum_{j=1}^{c} u_{ij} = 1 \qquad (2)$$

For all i= 1,2,…….N, where c is the number of clusters and N is the number of pixels in microarray image.

Step_2: Compute the centroid values for each cluster c_j. Each pixel should have a degree of membership to those designated clusters. So the goal is to find the membership values of pixels belonging to each cluster. The algorithm is an iterative optimization that minimizes the cost function defined as follows:

$$F= \sum_{j=1}^{N} \sum_{i=1}^{c} u_{ij}^{m} \, || x_j - c_i ||^2 \qquad (3)$$

Where u_{ij} represents the membership of pixel x_j in the i^{th} cluster and m is the fuzziness parameter.

Step_3: Compute the updated membership values u_{ij} belonging to clusters for each pixel and cluster centroids according to the given formula.

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\|x_j-v_i\|}{\|x_j-v_k\|}\right)^{2/(m-1)}},$$

and

$$v_i = \frac{\sum_{j=1}^{N} u_{ij}^{m} x_j}{\sum_{j=1}^{N} u_{ij}^{m}}. \qquad (4)$$

Step_4: Repeat steps 2-3 until the cost function is minimized.

End.

IV. Histogram Clustering Algorithm

We present a greedy histogram clustering algorithm that takes as input partitioned image and obtain histogram clustering based on the minimization of the loss of Mutual Information. The Mutual Information between two random variables X and Y is defined by

$$I(X,Y)=H(X)-H(X|Y)$$

Where $H(X) = -\sum_{x\in X} p(x)\log p(x)$ *and*

$$H(X|Y) = -\sum_{x\in X} p(x) \sum_{y\in Y} p(y|x)\log p(y|x) \qquad (5)$$

That is we group the bins of the histogram so that the mutual Information is maximally preserved. From the perspective of the information bottleneck method the binning process is controlled by a given partition of the image. The histogram clustering algorithm is presented in [9].

Our Clustering algorithm is based on the channel G→R, and is defines by the conditional probability matrix p(R|G) which expresses how the pixels corresponding to each histogram bin are distributed into regions of the image . Bayes' theorem, expressed by p(g)p(r|g)=p(r)p(g|r), establishes the relationship between the conditional probabilities of both channels G→R and R→G. The basic idea underlying our histogram clustering algorithm is to capture the maximum information of the image with the minimum number of histogram bins. In general, if the two bins are very similar the channel can be simplified by substituting these two bins by their clustering, without a significant loss of information. The algorithm proceeds by merging the two bins so that the loss of information is minimum. During the clustering process H(R)=H(R|G) + I(G,R), where H(R) is the entropy of p(R) and H(R|G) and I(G,R) represent, respectively, the successive values of conditional entropy and MI obtained after successful clusterings. Observe also that H(R|G) is the average entropy of the bins and increases at each iteration.

V. Experimental Results

Segmentation steps of the microarray image processing are performed on a sample microarray slide that has 48 blocks, each block consisting of 110 spots. A sample block has been chosen and 108 spots of the block have been cropped for simplicity. The sample image is a 154*200 pixel image that consists of a total of 30800 pixels. The RGB colored image microarray image have been converted to grayscale image to specify a single intensity value that varies from the darkest (0) to the brightest (255) for each pixel shown in figure1.

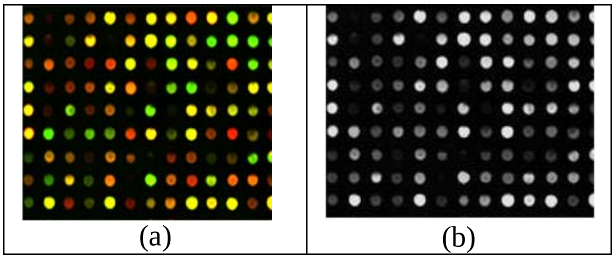
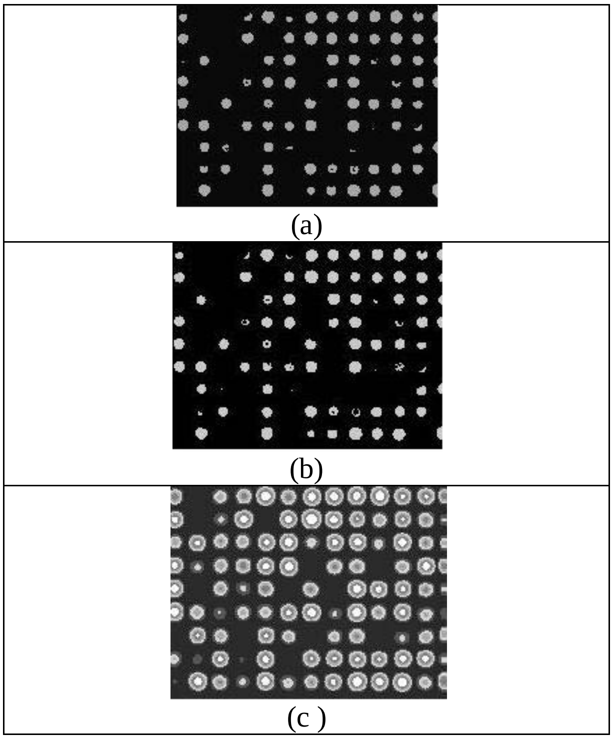

(a) (b)

Fig1 : a) RGB Color microarray image b) Grayscale Image

The segmented microarray image using three different segmentation algorithms (K-means, Fuzzy c-Means and Histogram Clustering algorithm) is shown in figure 2.


(a)

(b)

(c)

Fig2 : a) K-means b) Fuzzy c-means c) Histogram Clustering Algorithm

The histogram gives the distribution of intensity values for each cluster. The K-means have calculated mean of the spots as 25.32 and the mean of the background as 74.68 for this sample and clustered 7800 pixels as fore and 23,000 pixels as back. The Histogram Clustering has calculated mean of the spots as 40.64 and the mean of the background as 59.35 for this sample and clustered 12,520 pixels as fore and 18,280 pixels as back. The proposed algorithm have calculated mean of the spots as 49.35 and the mean of the background as 50.64 for this sample and clustered 15,200 pixels as fore and 15,600 pixels as back.

VI. Conclusion

Histogram clustering algorithm constitutes a valid tool to segment the spots of microarray image. Even though the mathematical bases for these techniques are complex, their implementation is simple, quick and easier on the user. The proposed segmentation algorithm has the advantage of processing spots of variable shapes and being insensitive to variations. In order to process the images of low intensity background correction is necessary. The proposed algorithm provides a more efficient way of segmenting the microarray image when compared with the segmentation achieved by K-Means and Fuzzy c-Means.

References References Referencias

1. M.Schena, D.Shalon, Ronald W.davis and Patrick O.Brown, “Quantitative Monitoring of gene expression patterns with a complementary DNA microarray”, Science, 270,199,pp:467-470.
2. Wei-Bang Chen, Chengcui Zhang and Wen-Lin Liu, “An Automated Gridding and Segmentation method for cDNA Microarray Image Analysis”, 19th IEEE Symposium on Computer-Based Medical Systems.
3. Tsung-Han Tsai Chein-Po Yang, Wei-ChiTsai, Pin-Hua Chen, “Error Reduction on Automatic Segmentation in Microarray Image”, IEEE 2007.
4. E.Erguit, Y.Yardimci, E.Mumcuoglu, O.Konu, “Analysis of microarray imagesusing FCM and k-means Clustering Algorithm”, in Proc IJCI, pp.116-121, 2003.
5. Volkan Uslan, Ihsan Omur Bucak, Clustering based Spot Segmentation of cDNA Microarray Images, IEEE 2010.
6. Rafael C.Gongalez, Richard E.Woods,” Digital Image Processing “,Third Edition, Pearson Education.
7. T.Deng and H.Heijmans, “ Grey-Scale Morphology Based on Fuzzy Logic”, Journal of Mathematical Imaging and Vision, Springer Netherlands, vol 16, no 2, pp. 155-171, 2002.
8. M.A.Wirth, D.Nikitento ,”Application of Fuzzy Morphology to Contrast Enhancement”, 2005 IEEE.
9. J.Rigau, M.Feixas and M.Sbert,” An Information Theoretic Framework for image segmentation”, IEEE 2004.