Microarray Image Processing for Real Time Scanning with Reduced Dimensional Variables

Deok Hee Nam
Engineering and Computing Science, Wilberforce University, Wilberforce, Ohio, USA

Abstract - The image processing of microarrays is like analyzing the arrays of gene fragments from the sampled images. In addition, microarrays provide the information about the detection of differences in genome sequences so that they can be used to detect and classify genetic characteristics very precisely. The experiments of microarray analyses can produce huge amounts of data because thousands of genes require to be processed in a single experiment. In order to obtain significant information from the various experiments of microarray, it is necessary to develop a compact and simpler technique, which can handle the large number of data. This paper provides a technique to perform the real time scanning of microarray images using embedded variables with applying multivariate analysis to reduce the dimension of the original images in vertically as well as horizontally.

Keywords: dimensional reduction, image processing, microarray, multivariate analysis, data mining

1 Introduction

There are many compounds and drug targets with the advent of genomics and advanced combinatorial chemistry. Simultaneously, various methods are introduced to detect and quantify gene expression levels, including northern blots, differential display, and sequencing of cDNA libraries, S1 nuclease protection, and serial analysis of gene expression. Over the last several years there has been a huge expansion of microarray technology [1] in the fields of biosciences including medical sciences, biotechnology, and pharmaceutical industry.

The most important technology which has been a focus on is how to provide a platform for determining in a single experiment, the gene expression profiles of various kinds of genes in tissue, tumors, cell, or biological fluids. The quick and comprehensive implementation about the bio-scientific technology has been evaluated and predicted on the reduction of its complexities with providing large amounts of highly relevant data successfully.

Among the various biomedical scientific technologies, the gene expression about microarrays is the most important technology to improve the blockage for the discoveries of drugs and disease diagnostics. Using the pattern recognition of microarrays [10], the expression of various genes, can be easily identified. Due to the increments of genomic and advanced biomedical scientific analyses, the technologies for enhanced data extraction and analysis are demanded more and more. Hence, there are now hundreds of papers related to the topics of biomedical scientific or genomic science fields and the application of microarray technology in biological research.

All microarray images can be recorded by microarray scanning technology [8] and represent the visual information [9]. In the fields of engineering and science, image processing is applied to examine properties of objects or processes encoded in images. Hence, after recording the objects or processes for the studies in the images, the image analysis [6] is applied to extract and quantify the characteristics of the objects and processes about the studies by the statistical analyses [4][5][7][12]. For examples, the pattern recognition [2][3] of characters, fingerprints, or various face images performed by computer systems are image analysis tasks. Especially, detecting, outlining, and measuring bio-scientific objects, like boundaries of blood vessels, structural deformations of the heart, degeneration of the retina, in medical applications, and extracting spots and detecting their boundaries, are very important studies to estimate parameters which quantify gene expression levels in microarray applications.

For the microarray image processing, the fundamental goal is to measure the intensities of the arrayed spots, and based on these intensities to quantify the spots expression levels. In a more sophisticated and complete approach, the array image processing will also assess the reliability of the quantified spot data and generate warnings to the possible problems during the array production and/or hybridization phases.

For the instrument point of view, any scientific instrument must perform consistently over time so that results can be compared from day to day and week to week. At the same time, the instrument needs to show whether its performance is in the real time or not. But sometimes the instrument may need more capabilities to perform the required processes due to the limitation of the instrument. For example, scanning the microarray images using a scanner may not be successfully done whenever it needs because of the processing speed of the scanner. To avoid the delayed scanning process, the scanner which can perform the process in the real time as the human eye is scanning the images of the objects may need to be used.
In order to achieve the required performance like a human eye, there are some options to approach the real time processing such as using the smaller or simpler microarray images with the less resolution or the faster instrument which can scan the images much faster. However, when the less resolution microarray images are used, the microarray image quality [10] is the key point to perform the image processing not to lose any meaningful information from the sampled microarray images. Hence, if there is a method which can satisfy the condition such as keeping all the information from the sampled microarray images with relatively less dimension of the microarray images, then the results of scanning the extracted, reduced, or transformed microarray images will be same as the results of scanning the original microarray images. To reduce or extract only the required the microarray images from the sampled microarray images with relatively more dimension, the proposed algorithm with multivariate analyses [7] will be used. Scanning the extracted microarray images may give more realistic time for scanning the microarray images without losing the meaningful information from the sampled microarray images and the time delay for the scanning procedure.

In this research, to develop a new algorithm of microarray image processing using the multivariate analyses for the real time processing. The new proposed method can be implanted to the robust of the microarray image scanners whose scanning time is close to the real scanning time like human eyes.

2 Principal Component Analysis (PCA)

Principal component analysis (PCA) [4] is a technique for dimensionality reduction, by diminishing the number of dimension with grouping relatively highly correlated variables among original variables after extracting principal components. Among the newly identified principal components, the first principal component implies the highest variability in the data and each succeeding component accounts for as much of the remaining variability as possible for the original data. In order to identify the extracted components, the eigenvalue or eigenvector approach of the covariance matrices is often used with singular value decomposition (SVD) of the data matrix. [13]

Consider a random vector \( X = [X_1, X_2, X_3, \ldots, X_n, X_p] \) from a multivariate distribution, and let \( X_1, X_2, X_3, \ldots, X_n \) be a random sample of \( n \) observations from the distribution \( X_p \) is the corresponding output for \( X \). Assume that it had been centered to zero mean. Since the scales of the different variables will not be proportionate, it requires the normalization with a common scale by dividing them by their standard deviations.

The mean value of a variable is defined as the sum of the observed values divided by the number of values. It can be shown as following

\[
\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (1)
\]

where \( \bar{X} \) is a mean of the variable and \( n \) is the number of observations. The variance is determined by dividing the sum of squared deviations from the mean by \( n - 1 \). It is represented as

\[
v_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1} \quad (2)
\]

where \( v_x^2 \) is the variance of a variable. The standard deviation is found from the variance. The square root of variance is called the standard deviation. Thus

\[
v_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}} \quad (3)
\]

where \( v_x \) is the standard deviation. When two variables problem is considered, the covariance can be defined as the sum of the products of the deviations of two variables from their respective means divided by the number of the observation in a variable. Assume if there is the same number of observation for each variable. Thus

\[
s_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{n - 1} \quad (4)
\]

where \( X \) and \( Y \) are variables, and \( \bar{X} \) and \( \bar{Y} \) are means of each variables.

The correlation is a measure of the degree of agreement between two sets of scores from the same individuals. The correlation coefficients between two variables can be expressed as the cosine values between any two variables. Therefore, the correlation coefficients measure the tendency of the points to cluster along a line.

Suppose variables can be grouped by their correlations. All variables within a particular group are highly correlated among themselves but have relatively small correlations with variables in a different group. It is conceivable that each group of variables represents a
single underlying construct, or factor, that is responsible for the observed correlations.

Let \( R \) be the correlation matrix with \( q \) by \( q \), from the data matrix with \( n \) by \( p \), \( A \), can be expressed as following

\[
R = \begin{bmatrix}
r_{11} & r_{12} & \cdots & r_{1q} \\
r_{21} & r_{22} & \cdots & r_{2q} \\
\vdots & \vdots & \ddots & \vdots \\
r_{n1} & r_{n2} & \cdots & r_{nq}
\end{bmatrix}
\]

(5)

where \( q = \max\{n, p-1\} \)

and

\[
r_y = \frac{s_y}{\sum_{i=1}^{n} d_{ii}} = \frac{\sum_{i=1}^{n} d_{ii}^2}{\sqrt{\sum_{i=1}^{n} d_{ii}^2}}
\]

(6)

with unit diagonals. Note that \( s_y \) is the covariance of the matrix \( A \) for any two variables \( i \) and \( j \).

After calculating the correlation, then estimate the eigenvectors with eigenvalues. From the matrix of the eigenvectors, rearrange the eigenvectors for each eigenvectors with eigenvalues. From the matrix of the cumulative value of the eigenvalues, or variances, are variables from the original data set. (Normally the part of the matrix \( R \) with eliminating the closely related components of each observation will be scored by greater or equal to 0.9. [14]) Therefore, the principal

After that, extract the vector which covers the majority part of the matrix \( R \) with eliminating the closely related variables from the original data set. (Normally the cumulative value of the eigenvalues, or variances, are greater or equal to 0.9. [14]) Therefore, the principal components of each observation will be scored by

\[
Y_1 = A_1^T X, \quad Y_2 = A_2^T X, \quad \ldots, \quad Y_p = A_p^T X
\]

(7)

3 Fuzzy C-Means (FCM) Clustering Analysis [2]

Fuzzy C-Means Clustering (FCM) algorithm attempts to partition a finite collection of \( n \) elements of the given data into a collection of \( c \) fuzzy clusters with respect to some given criterion. Given a finite set of data, for \( 2 \leq c \leq n \), consider a set of \( n \) vectors \( X = \{x_1, x_2, \ldots, x_n\} \) to be clustered into \( c \) groups of data. Each of the groups, \( x_i \in R^S \), is a feature vector consisting of \( s \) real-valued measurements describing the features of the object represented by \( x_i \). The features could be length, width, color etc. Fuzzy clustering of the objects can be represented by a fuzzy membership matrix called as a fuzzy partition.

The set of all \( c \times n \) non-degenerate constrained fuzzy partition matrices is denoted by \( M_{fcn} \), and is defined as

\[
U = [u_{ij}]_{i=1,2,\ldots,c; j=1,2,\ldots,n} = M_{fcn}
\]

where \( u_{ij} \) expresses the degree to which the element \( x_j \) belongs to the \( i^{th} \) cluster and is a numerical value in \([0,1]\) such that the constraints in

\[
\sum_{i=1}^{c} u_{ij} = 1 \quad \text{for all } j = 1, 2, \ldots, n
\]

and

\[
0 < \sum_{j=1}^{n} u_{ij} < n \quad \text{for all } i = 1, 2, \ldots, c.
\]

For the Fuzzy C-Means algorithms, the objective is to find \( U = [u_{ik}] \in M_{fcn} \) as the fuzzy \( c \)-partition matrix and \( V = (v_1, \ldots, v_c) \) with \( v_i \in R^d \) as the cluster center such that

\[
J_m(U, V) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m \|x_k - v_i\|^2
\]

(8)

is minimized, where \( u_{ik} \) is the value of the \( i^{th} \) membership function on the \( k^{th} \) data point \( x_k \), \( n \) is the number of samples in \( X \), and \( m \in (1, \infty) \) is a weighting constant. The distances, \( d_{ik} = \|x_k - v_i\|^2 \), are weighted with the membership values \( u_{ik}^m \), where \( \|x_k - v_i\|^2 \) is any inner product induced norm on \( R^d \) which is called the square of Euclidean distance for \( i^{th} \) cluster center and \( j^{th} \) data point. The Euclidean distance formula will be

\[
d_{ik} = \left\|x_k - v_i\right\|^2 = \left(\{x_k - v_i\}^2\right)^{1/2}
\]

(9)

where \( d_{ik} \) will be the distance between \( i^{th} \) cluster center and \( k^{th} \) data point \( x_k \).

The assumption is that the distance between their corresponding data vectors measures the similarity between objects. Then the necessary conditions to minimize the objective function, \( J_m(U, V) \), can be

\[
u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}}\right)^{m-1}}^{1/2}
\]

(10)
and

\[ v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m} \tag{11} \]

where \( 1 < i < c \), and \( 1 < k < n \).

Therefore, the Fuzzy C-Means algorithm is an iterated procedure through those two necessary conditions to minimize the objective function, \( J_m(U, V) \).

The following steps summarize the Fuzzy C-Means Clustering algorithm.

Step 1. Initialize the partition matrix, or membership matrix such that \( U^{(0)} \in M_{cn} \) randomly or prior knowledge.

Step 2. Calculate the cluster centers, \( v_i \), using the equation (11).

Step 3. Compute the distance, \( d_{ik} \).

Step 4. Update the partition matrix \( U^{(new)} \) using the equation (10) for \( u_{ik} \).

Step 5. Until \( \| U^{(new)} - U^{(old)} \| < \varepsilon \) where \( \varepsilon \) is the termination tolerance \( \varepsilon > 0 \). If this condition is not satisfied, then go back to Step 2.

4 Proposed Multivariate Analysis

There are various techniques to reduce or diminish the huge data set into an appropriate size of the data without losing any significant meaning from the original huge data. Among the frequently used techniques, multivariate analysis such as Principal Component Analysis or Factor Analysis is refocused by the scientists in these days. To develop the suitable technique for extracting the required microarray image, Fuzzy C-Means Clustering Analysis is used as a postprocessing procedure after the images are processed by Principal Component Analysis. The following algorithm summarizes the proposed algorithm using the Principal Component Analysis followed by Fuzzy C-Means Clustering Analysis.

Step 1. Read the original data set as a matrix format.
Step 2. Normalize the original data from Step 1.
Step 3. Find the correlation matrix of the normalized data from Step 2.
Step 4. Find eigenvalues and eigenvectors of the correlation matrix from Step 3 using characteristic equation.
Step 5. Define a matrix that is the eigenvectors from Step 4 as the coefficients of principal components using the criterion for extracting components.
Step 6. Multiply the standardized matrix from Step 2 and the coefficients of principal components from Step 5.

Step 7. Using the result from Step 6, find the centers of clusters for each clustering technique.
Step 8. To clustering the components from Step 7, initialize the partition matrix, or membership matrix randomly such that \( U^{(0)} \in M_{cn} \).
Step 9. Calculate the cluster centers, \( v_i \), using the equation (11).
Step 10. Compute the distance, \( d_{ik} \).
Step 11. Update the partition matrix \( U^{(new)} \) using the equation (10) for \( u_{ik} \). If \( d_{ik} > 0 \), for \( 1 \leq i \leq c \), and \( 1 \leq k \leq n \), then get the new \( u_{ik} \). Otherwise if \( d_{ik} > 0 \), and \( u_{ik} = [0, 1] \) with \( \sum_{i=1}^{c} u_{ik}^{(new)} = 1 \), then \( u_{ik}^{(new)} = 0 \).
Step 12. Until \( \| U^{(new)} - U^{(old)} \| < \varepsilon \) where \( \varepsilon \) is the termination tolerance \( \varepsilon > 0 \). If this condition is not satisfied, then go back to Step 9.

5 Analysis and Results

To develop the real time scanning for the microarray image, the reduced-dimensional images from the original image with applying the proposed algorithm are analyzed. To reduce the dimension into the less number of pixels, the appropriate number of new components needs to be determined without losing any significant meaning from the original image. In Fig. 1, the relation between the evaluated eigenvalues and the number of newly extracted components are plotted to determine the appropriate number of newly extracted components from the original image.

![Extracted components vs Eigenvalues](image)

Fig. 1 The number of extracted components vs. the corresponding eigenvalues

In Fig.2, the images are shown how the original images can be reduced by applying selected proposed algorithms. To analyze the reduced dimensional images comparing with the original image to present a method of simplifying the analysis of large amounts of microarray image data by reducing information without losing
validity, the consecutive sampled rows of pixels from each extracted image are compared by evaluating its correlation coefficients [11]. As shown above, when the image (b) is reduced by 11 by 9 using Principal Component Analysis, the correlations between the original image and the extracted image are -0.704, -0.594, -0.914, and 0.036, in TABLE 1, respectively. Since the images are shown that the lighter parts extracted as the darker part and the darker parts are extracted by the lighter parts after applying the procedure. There appears to be some loss of intensity information, as represented by the diagrams shown in the analysis.

From Fig. 2, the extracted image with reducing the dimension of the height and width using the factor analysis with principal components and Fuzzy C-Means Clustering Analysis is shown that the microarray spot can be recognizable even though the image of the bottom-right portion is not clearly recognized as comparing to the original image. From (c) in Fig. 2, the extracted image loses the characteristics of the right-middle portion for the original image with reducing the dimensions in the vertical and horizontal directions. From (d) and (e) in Fig. 2, the images loses the characteristics of the original image as comparing with the image (a). In (f) in Fig. 2, even though the image is shown with a less intensity image as comparing with the image (a), but the characteristic of the original image can be captured.

**TABLE 1 Correlations between sampled pixels**

<table>
<thead>
<tr>
<th></th>
<th>AP1</th>
<th>AP2</th>
<th>AP3</th>
<th>AP4</th>
</tr>
</thead>
<tbody>
<tr>
<td>bP1</td>
<td>-0.704</td>
<td>-0.623</td>
<td>-0.84</td>
<td>-0.104</td>
</tr>
<tr>
<td>bP2</td>
<td>-0.853</td>
<td>-0.594</td>
<td>-0.925</td>
<td>-0.035</td>
</tr>
<tr>
<td>bP3</td>
<td>-0.76</td>
<td>-0.64</td>
<td>-0.914</td>
<td>0.007</td>
</tr>
<tr>
<td>bP4</td>
<td>-0.713</td>
<td>-0.603</td>
<td>-0.905</td>
<td>0.036</td>
</tr>
<tr>
<td>cP1</td>
<td>-0.662</td>
<td>-0.707</td>
<td>-0.886</td>
<td>-0.064</td>
</tr>
<tr>
<td>cP2</td>
<td>-0.853</td>
<td>-0.594</td>
<td>-0.925</td>
<td>-0.035</td>
</tr>
<tr>
<td>cP3</td>
<td>-0.713</td>
<td>-0.603</td>
<td>-0.905</td>
<td>0.036</td>
</tr>
<tr>
<td>cP4</td>
<td>-0.704</td>
<td>-0.623</td>
<td>-0.84</td>
<td>-0.104</td>
</tr>
<tr>
<td>dP1</td>
<td>-0.715</td>
<td>-0.605</td>
<td>-0.906</td>
<td>0.034</td>
</tr>
<tr>
<td>dP2</td>
<td>-0.832</td>
<td>-0.468</td>
<td>-0.774</td>
<td>-0.06</td>
</tr>
<tr>
<td>dP3</td>
<td>-0.877</td>
<td>-0.39</td>
<td>-0.713</td>
<td>-0.079</td>
</tr>
<tr>
<td>eP1</td>
<td>-0.726</td>
<td>-0.607</td>
<td>-0.909</td>
<td>0.028</td>
</tr>
<tr>
<td>eP2</td>
<td>-0.884</td>
<td>-0.521</td>
<td>-0.89</td>
<td>0.042</td>
</tr>
<tr>
<td>eP3</td>
<td>-0.721</td>
<td>-0.673</td>
<td>-0.905</td>
<td>-0.029</td>
</tr>
<tr>
<td>fP1</td>
<td>-0.644</td>
<td>0.166</td>
<td>-0.321</td>
<td>-0.129</td>
</tr>
<tr>
<td>fP2</td>
<td>0.121</td>
<td>0.266</td>
<td>0.251</td>
<td>0.298</td>
</tr>
<tr>
<td>fP3</td>
<td>0.026</td>
<td>0.607</td>
<td>0.503</td>
<td>0.275</td>
</tr>
<tr>
<td>fP4</td>
<td>0.278</td>
<td>0.208</td>
<td>0.393</td>
<td>0.06</td>
</tr>
</tbody>
</table>

(a) aP1, aP2, aP3, and aP4 are sampled pixels from the original images. (b) bP1, bP2, bP3, and bP4 are sampled pixels from the 11 by 9 extracted image extracted by Principal Component Analysis. (c) cP1, cP2, cP3, and cP4 are sampled pixels from the 11 by 9 extracted images processed by Principal Component Analysis followed by Fuzzy C-means Clustering Analysis. (d) dP1, dP2, and dP3 are sampled pixels from the 8 by 9 extracted images processed by Principal Component Analysis followed by Fuzzy C-means Clustering Analysis. (e) eP1, eP2, and eP3 are sampled pixels from the 6 by 9 extracted images processed by Principal Component Analysis followed by Fuzzy C-means Clustering Analysis. (f) fP1, fP2, fP3, and fP4 are sampled pixels from the 11 by 9 extracted image extracted by Factor Analysis.

6 Conclusion

The microarray analysis is an important research topic in the biomedical image processing. Moreover, the amount of the processing data is a very important issue with the processing time. To improve the data processing and its execution, the proposed algorithm is developed to help the real time scanning and reduce the amount of the time to process the analysis of the microarray images. Through the various extracted images with reducing the dimensions in vertical and horizontal directions, the characteristics of the original image can be recognized by the dimension-reduced images without losing its validities if the dimension of the image is kept in the reasonable intensity.

![Fig. 2 Original image and extracted images](image)

Fig. 2 Original image and extracted images

(a) Original image of a single microarray spot
(b) 11 × 9 extracted image of a single microarray spot with Principal Component Analysis
(c) 10 × 9 extracted image of a single microarray spot with Principal Component Analysis followed by Fuzzy C-means Clustering Analysis
(d) 8 × 9 extracted image of a single microarray spot with Principal Component Analysis followed by Fuzzy C-means Clustering Analysis
(e) 6 × 9 extracted image of a single microarray spot with Principal Component Analysis followed by Fuzzy C-means Clustering Analysis
(f) 11 by 9 extracted image of a single microarray spot with only Factor Analysis.

Acknowledgement

This project has been supported by United Negro College Fund (UNCF) through Henry C. McBay Fellowship.

7 References


[13] Hairong Qi, Tsei-Wei Wang, and J. Douglas Birdwell, “Global Principal Component Analysis for Dimensionality Reduction in Distributed Data Mining,” University of Tennessee, Knoxville, TN 37996, USA