Gridding and Segmentation Method for DNA Microarray Images

B. Sivalakshmi1*, N.Naga Malleswara Rao2

1 Research Scholar, Department of CSE, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.
2 Professor, Department of IT, RVR and JC College of Engineering, Guntur, India
bslakshmi85@gmail.com

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: This article mainly explores meshing and segmentation techniques for microarray image analysis. The term "grid" refers to dividing an image into subgrids of dots and then dividing them into point detection. Most of the existing methods depend on input parameters such as the number of rows / columns, the number of points in each row / column, the size of the subarrays, etc. This article proposes a fully automatic mesh generation algorithm. This can remove any initialized parameter without any manual intervention. In the segmentation step, clustering algorithms are used because they do not consider the size and shape of the spots, do not depend on the initial state of the pixels, and do not require post-processing. In this article, a new method is proposed to estimate the initial parameters (centroid and number of clusters) required by any clustering algorithm. Qualitative and quantitative analysis shows that the algorithm can perform grid processing on microarray images well, and improves the performance of the clustering algorithm.

Keywords: Microarray image, Image segmentation, Mathematical morphology, clustering algorithms.

1. INTRODUCTION

Microarray technology is used to analyze gene expression values of thousands of points in parallel [1]. The output of the microarray experiment is the image extracted from the hybrid microarray slide using a sensor with two different wavelengths, Cy3 and Cy5. The analysis of the microarray image is carried out in three stages, namely gridding, segmentation and quantification. The gridding process is divided into two stages. First, divide the image into sub-arrays, called sub-grids, and then divide these sub-arrays into gene point regions, called point detection [1]. Segmentation is the process of grouping the spot pixels and background pixels in each spot area. Quantification is the process of finding the value of gene expression, which is equal to the logarithmic ratio of the green and red intensity values of a point [2].

The gene expression value depends on the intensity values of foreground pixels (spot areas) and background pixels. Most of the existing algorithms for microarray image analysis are semi-automatic, which means that manual intervention is required to initialize the parameters to execute the gridding algorithm [3]. This article offers a fully automatic mesh generation algorithm. Microarray technology is widely used in genetics, disease diagnosis, drug development, and pharmacology [4]. The microarray image analysis mechanism is shown in Figure 1.

In this article, in the second part, a grid generation algorithm is presented, in the third part, an algorithm for estimating the necessary parameters of the clustering algorithm is presented, in the fourth part, a clustering algorithm is presented, and in the fifth part, experimental results and conclusions are presented.
2. MICROARRAY GRIDDING:

The grid is the most important step in microarray image analysis. For sequential analysis of microarray images, a fine mesh can improve the efficiency of the segmentation and quantification steps. To capture each spot of a gene from a microarray image, the image is first divided into sub arrays, and then these sub arrays are divided into spot regions. The first step is called global meshing (sub-meshing), and the second step is local meshing (point detection). The end result of the mesh is an area with one spot and a background. The existing meshing algorithms are divided into manual and semi-automatic. When building a mesh manually, the user must specify all the parameters necessary for building the mesh, such as the number of sub arrays, the number of points and the size of the points. In a semi-automatic program, the user part provides parameters such as the number of rows/columns and the number of points in each row/column. This paper proposes an algorithm to automatically divide the grid using the horizontal and vertical contours of the microchip image. Figures 2 and 3 show the algorithm used to detect sub-grids and points.

3. MICROARRAY SEGMENTATION

Image segmentation is the process of dividing an image into regions, and for microarray images, it is the process of dividing a subnet into foreground and background regions. This foreground area corresponds to the spot area. Using this spot of the segmented image, gene expression levels are estimated. This segmentation is a difficult task, since the intensities in this area are heterogeneous, and the sizes and shapes of the points also differ from each other. Several statistical segmentation methods are proposed for segmentation of microarray images. Shape-based circular segmentation [6], in which a circular mask with a specific radius is used to identify a point. But all spots are not of equal shape. Region growing method [5] in this the spot area and background area are separated by selecting seed pixels in both areas and regions are extracted by using some predefined criteria. Selection of seed pixel is a difficult task. Thresholding method [6], by using the histogram of the image, a suitable threshold is estimated, and the image is divided into two regions. This estimation is done using Mann Whitney test. Morphology based segmentation [7], uses the morphological operations such as hit or miss transforms to segment the image. Here the process depends on the selection of mask used for morphological operations. Supervised learning-based segmentation [8] uses support vector machine algorithm to segment the image. In this paper, clustering algorithms are used to segment the image. These algorithms are efficient than other existing segmentation methods, as they do not depend on the shape and size of the spot. This algorithm does not require any mask and seed pixels. Every clustering algorithm depends on the number of clusters and initial values of centroids. If these values are estimated, the number of iterations required for any algorithm to segment an image can be minimized. The required number of centroids and clusters for microarray image segmentation are two, as every algorithm divides into two regions background and spot area. Existing algorithm used minimum value in the background area and maximum value in the foreground area for segmenting the image into two regions. This article presents an algorithm for estimating centroids and the number of clusters using empirical mode decomposition. A block diagram for estimating the number of clusters is shown in Figure 4. This article uses clustering algorithms such as k-mean, moving k-mean, and FCM.
Sub-gridding Algorithm:

1. Convert the original microarray image into grayscale image.

2. Compute the standard deviation of pixel intensities of every row, producing the profile of horizontal standard deviation
   \[ S_i = \frac{1}{M-1} \left( \frac{1}{N} \sum_{j=1}^{N} I(i,j) \right) \]
   for \( i = 1, 2, \ldots, M \) and \( I(i,j) \) is the intensity of pixel located at row \( i \) and column \( j \).

3. Compute the standard deviation of pixel intensities of every column, producing the profile of vertical standard deviation
   \[ S_j = \frac{1}{N-1} \left( \frac{1}{M} \sum_{i=1}^{M} I(i,j) \right) \]
   for \( j = 1, 2, \ldots, N \) and \( I(i,j) \) is the intensity of pixel located at row \( i \) and column \( j \).

4. Find the minimum value in \( S_i \) and \( S_j \).
   - MIN1 = minimum \( (S_i) \)
   - MIN2 = minimum \( (S_j) \)

5. These standard deviation profiles have to be thresholded for the estimation of row width and column width.
   - \( TS_i = 1 \), if \( S_i > \text{MIN1} + \text{Th} \times \text{MIN1} \)
   - \( TS_i = 0 \) otherwise
   - \( TS_j = 1 \), if \( S_j > \text{MIN2} + \text{Th} \times \text{MIN2} \)
   - \( TS_j = 0 \) otherwise

6. Calculation of row width and column width: The value in \( TS_i \) and \( TS_j \) looks like 1's followed by 0's again 1's and so on.

7. The procedure for row width calculation is as follows:
   i. for \( i = 1 \) to \( M \)
      - Count the number of 0’s in \( TS_i \) (between the elements in \( TS_i \) having 1)
      - Count the number of 1s in \( TS_i \) (between the elements in \( TS_i \) having 0)
   ii. for \( p = 1, 2, \ldots, k \)
      - \( RW_{sp} = RW_{sp} + (RW_{sp} + RW_{sp+1})/2 \)
   iii. Row width \( (RW_i) \) = median \( (RW_i) \)

8. The procedure for column width calculation is as follows:
   i. for \( j = 1 \) to \( N \)
      - Count the number of 0’s in \( TS_j \) (between the elements in \( TS_j \) having 1)
      - Count the number of 1s in \( TS_j \) (between the elements in \( TS_j \) having 0)
   ii. for \( p = 1, 2, \ldots, k \)
      - \( CW_{sp} = CW_{sp} + (CW_{sp} + CW_{sp+1})/2 \)
   iii. Column width \( (CW_j) \) = median \( (W_j) \)

9. Using the value \( CW_j \) draw horizontal grid lines at positions of \( j \), where \( j = 1, 2, \ldots, N \) with step increment \( CW_j \).

10. Map these grid matrix \( G_i \) onto the microarray image producing sub-gridding.

Figure 2: Sub-gridding Algorithm
Gridding and Segmentation Method for DNA Microarray Images

Spot-Detection Algorithm:

Consider the sub-gridded microarray image.

Perform Edge detection using Bi-dimensional Empirical Mode Decomposition

Perform morphological filling on the edge image. Perform mean intensity of the pixels on each sub-grid from the morphologically filled sub-gridded image. Consider the sub-grid with maximum mean intensity. That sub-grid is called optimal sub-grid. Perform spot detection on this optimal sub-grid.

Calculation of Horizontal and Vertical Intensity profiles
Horizontal and vertical intensity projection profiles of binary image (Morphological Filled Image) are the sum of pixel intensities along each row and column respectively. Let $M_{j}$ indicates the filled image of size $M \times N$. Then the intensity projection profile along $j^{th}$ row and $j^{th}$ column are computed.

$$S_{j} = M_{ji} = \sum_{i=1}^{N} M_{j}(i, j) \text{ for } i=1,2,...,M$$

$$S_{j} = M_{ji} = \sum_{i=1}^{N} M_{j}(i, j) \text{ for } j=1,2,...,N$$

Calculation of row width (RW) and column width (CW)
The values in $S_{i}$ are used for identification of row width and values in $S_{j}$ are used for identification of column width. The values in $S_{i}$ and $S_{j}$ looks like zeros followed by nonzero again zeros and so on.

The procedure for row width calculation is as follows:
for $i = 1$ to $M$,

- Count the number of zeroes in $S_{i}$ (between the elements in $S_{i}$ having nonzero) = $W_{yp}$, Where $p = 1, 2, 3, ..., k$.
- Count the number of non-zeros in $S_{i}$ (between the elements in $S_{i}$ having zero) = $W_{sp}$, Where $p = 1, 2, 3, ..., k$.

for $p = 1, 2, ..., k$

$$W_{yp} = W_{yp} + (W_{yp} + W_{yp+1})/2,$$

Row width (RW) = median ($W_{p}$).
Using the value RW draw horizontal grid lines at positions of $i$, where $i=1, 2, ..., M$ with step increment RW.
for $i=1$ to $M$ step RW
for $j=1$ to $N$

$G[i,j] = 1$;

The procedure for column width calculation is as follows:
for $j = 1$ to $N$,

- Count the number of zeroes in $S_{j}$ (between the elements in $S_{j}$ having nonzero) = $W_{yp}$, Where $p = 1, 2, 3, ..., k$.
- Count the number of non-zeros in $S_{j}$ (between the elements in $S_{j}$ having zero) = $W_{sp}$, Where $p = 1, 2, 3, ..., k$.

for $p = 1, 2, ..., k$

$$W_{yp} = W_{yp} + (W_{yp} + W_{yp+1})/2,$$

Column width (CW) = median ($W_{p}$).
Using the value CW draw horizontal grid lines at positions of $j$, where $j=1, 2, ..., N$ with step increment CW.
for $j=1$ to $N$ step RW
for $i=1$ to $M$

$G[i,j] = 1$;

Map these grid matrix $G$ onto the sub-grid image.

Figure 3: Spot Detection Algorithm
Algorithm for Estimation of Centroids and Number of Clusters using EMD:

Let $h(k)$ be the histogram for the input image $I$ with $k=0, ..., G$ and $G$ being the maximum intensity value in the image.

Divide the histogram $h(k)$ into IMFs using empirical mode decomposition. The first IMF carries the histogram noise, irregularities and sharp details of the histogram, while the last IMF and residue describe the trend of the histogram.

Consider the summation of intermediate IMFs as follows:

$$I_{INT} = \sum_{i=2}^{n} IMF_i$$

where $n$ is the number of IMFs.

Determine all local minima in $I_{INT}$.

$$I^* = \{ \min_{i \in I_{INT}} I_i \}$$

$I^*$ is the vector carrying all local minima.

The truncation step is expressed as follows:

$$thr = \frac{1}{2N_{I^*}} \sum_{i \in I^*} I_i$$

where $N_{I^*}$ is the number of local minima belonging to $I^*$ and $I_i$ is the local minima belonging to vector $I^*$.

$$I' = \{ I_i' \}$$

where $I_i'$ consists of all local minima which are less than the estimated threshold $thr$.

In the second truncation step, the algorithm applies an iterative procedure that calculates the number of image pixels belonging to each candidate image cluster and prunes the cluster with smallest number of image pixels (less than 2 percent of total number of image pixels). The pruned candidate clusters are merged with their closest image clusters.

The number of elements in final vector $I'$ represents the number of clusters denoted by $NC$.

Determine the local maxima in $I_{INT}$.

$$I_{M}^* = \{ \min_{i \in I_{INT}} I_i \}$$

$IM^*$ is the vector carrying all local maxima.

Thresholding: Find the peaks (local maxima) whose value is higher than one percent of the maximum peak in $h(k)$.

Remove the peaks which are very close. This is done by checking the difference between the grey levels of the two individual peaks. If the difference is less than 20, then the peak with lowest value is removed.

Neighbouring pixels that lead to the same peak are grouped together. The values of the identified peaks represent the initial centroids of the input image.

Figure 4: Algorithm for estimating clustering parameters
4. CLUSTERING ALGORITHMS

In this article, k-mean [9], moving k-mean [10], and FCM algorithm [11] are used to segment microarray images. The two parameters required by these algorithms are estimated using the algorithm described in Section 3.

5. EXPERIMENTAL RESULTS

The qualitative and quantitative analysis of proposed sub-gridding and spot detection algorithms are performed on two microarray images of breast cancer a CHG tumor tissue. The qualitative analysis of proposed sub-algorithm is shown in figure 5. The qualitative analysis of proposed spot detection algorithm is shown in figure 6. Table 1 shows a quantitative analysis of the proposed mesh generation algorithms in comparison with existing methods.

The precision of the mesh generation algorithm is estimated by the formula

\[
\text{Percentage accuracy} = \frac{\text{Number of spots perfectly gridded}}{\text{Total number of spots}} \times 100
\]

After sub-gridding and spot detection algorithms, individual spots are extracted. These individual spots are segmented using clustering algorithms with initial values are estimated using proposed estimation of number of clusters and centroids algorithm. A sample spot is extracted from the image and segmented using clustering algorithm. Three different clustering algorithms are used in the presented work. Table 2 shows the number of iterations used by different clustering algorithms in segmenting the spot image with and without estimation of initial parameters. Table 3 shows the MSE values [11] of different segmentation algorithms. The segmented step on a single spot is shown in figure 7.
Figure 5: Sub-gridding

(a) Microarray grayscale image  
(b) Edges of spots  
(c) Filling  
(d) $M_{pi}$ plot (rows)  
(e) $M_{pj}$ plot (columns)  
(f) Gridding (spot area)  
(g) Centroid  
(h) FCM output
Gridding and Segmentation Method for DNA Microarray Images

![Figure 6: Spot-Detection](Image)

<table>
<thead>
<tr>
<th>Method</th>
<th>Percentage Accuracy Image 1</th>
<th>Percentage Accuracy Image 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hirata [Microarray Gridding by mathematical Morphology]</td>
<td>87</td>
<td>79</td>
</tr>
<tr>
<td>Jain [Fully Automatic Quantification of microarray image data]</td>
<td>89</td>
<td>81</td>
</tr>
<tr>
<td>Wang [Precise Gridding of microarray images by detecting and correcting rotations in subarrays]</td>
<td>91</td>
<td>82</td>
</tr>
<tr>
<td>Shuqing Zhao [Microarray Image Processing Based on Mathematical Morphology]</td>
<td>90</td>
<td>84</td>
</tr>
<tr>
<td>Deepa J [A New Gridding Technique for High Density Microarray Images Using Intensity Projection Profile of Best Sub Image]</td>
<td>92</td>
<td>88</td>
</tr>
</tbody>
</table>

Proposed

96 91

Table 3: MSE values

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE Values (Compartment No 1) In image 1</th>
<th>MSE Values (Compartment No 8) In image 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means</td>
<td>282.781</td>
<td>346.47</td>
</tr>
<tr>
<td>Moving K-means</td>
<td>226.92</td>
<td>238.79</td>
</tr>
<tr>
<td>Fuzzy C-means</td>
<td>208.327</td>
<td>196.276</td>
</tr>
</tbody>
</table>
Table 2: Comparison of iterative step numbers

<table>
<thead>
<tr>
<th>Compartment No</th>
<th>Clustering Algorithm</th>
<th>Iterative Steps (without ECNC)</th>
<th>Iterative Steps (with ECNC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>K-means</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Moving k-means</td>
<td>14</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Fuzzy C-means</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>K-means</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Moving k-means</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Fuzzy C-means</td>
<td>19</td>
<td>11</td>
</tr>
</tbody>
</table>

Image 1

Gridding

Spot area2

Histogram

IMF1

IMF2

IMF3

IMF4

IMF5

IMF6

Combined IMF

Minima Points
### Gridding and Segmentation Method for DNA Microarray Images

<table>
<thead>
<tr>
<th>Maxima Points</th>
<th>Parameters</th>
<th>FCM Image</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>clusters - two centroids: 3, 155</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Image 2</th>
<th>Gridding</th>
<th>Spot area</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="grid.png" alt="Grid" /></td>
<td><img src="spot.png" alt="Spot" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Histogram</th>
<th>IMF1</th>
<th>IMF2</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="histogram.png" alt="Histogram" /></td>
<td><img src="imf1.png" alt="IMF1" /></td>
<td><img src="imf2.png" alt="IMF2" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IMF3</th>
<th>IMF4</th>
<th>IMF5</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="imf3.png" alt="IMF3" /></td>
<td><img src="imf4.png" alt="IMF4" /></td>
<td><img src="imf5.png" alt="IMF5" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IMF6</th>
<th>Combined IMF</th>
<th>Minima points</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="imf6.png" alt="IMF6" /></td>
<td><img src="combined.png" alt="Combined IMF" /></td>
<td><img src="minima.png" alt="Minima" /></td>
</tr>
</tbody>
</table>
6. CONCLUSIONS
The microarray image analysis is a sequential process with gridding, segmentation and quantification process. Any error in the gridding and segmentation stages, the gene expression value will be affected. This paper presents an automatic gridding algorithm and estimates the parameters required for clustering in segmentation process. The experimental results show the proposed algorithm grids the image with 96% accuracy and decreases the number of iterations by estimating the required parameters for segmenting the image by clustering algorithm.

REFERENCES