A comparative study of individual and ensemble majority vote cDNA microarray image segmentation schemes, originating from a spot-adjustable based restoration framework

Antonis Daskalakis\textsuperscript{a,}\textsuperscript{*}, Dimitris Glotsos\textsuperscript{b}, Spiros Kostopoulos\textsuperscript{a}, Dionisis Cavouras\textsuperscript{b}, George Nikiforidis\textsuperscript{a}

\textsuperscript{a} Medical Image Processing and Analysis Laboratory (MIPA), Department of Medical Physics, School of Medicine, University of Patras, Rio, Patras 265 04, Greece
\textsuperscript{b} Department of Medical Instruments Technology, Technological Educational Institute of Athens, Ag. Spyridonos Street, Aigaleo, 122 10 Athens, Greece

\textbf{A R T I C L E  I N F O}

Article history:
Received 19 June 2008
Received in revised form 23 September 2008
Accepted 12 January 2009

Keywords:
cDNA microarray
Segmentation
Restoration
Majority vote segmentation

\textbf{A B S T R A C T}

The aim of this study was to comparatively evaluate the performances of various segmentation algorithms, in conjunction with a noise reduction step, for gene expression levels intensity extraction in cDNA microarray images. Different segmentation algorithms, based on histogram and unsupervised classification methods, which have never been previously employed in microarray image analysis, were employed either individually or in ensemble majority vote structures for separating spot-images from background pixels. The performances of segmentation algorithms or ensemble structures were evaluated by assessing the validity and reproducibility of gene expression levels extraction in simulated and real cDNA microarray images. By processing high quality simulated images, the highest segmentation accuracy was achieved by an ensemble structure (Histogram Concavity, Gaussian Kernelized Fuzzy-C-Means, Seeded Region Growing). Optimum performance in terms of processing time and segmentation precision for low quality simulated and replicated real cDNA microarray images was attained by the Histogram Concavity algorithm.

© 2009 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Microarray technology has become one of the most powerful tools in the field of genetic research \cite{1,2}. Microarrays enable the high throughput profiling of thousands of distinct gene expression levels in a single hybridization experiment and, thus, provide to molecular biologists and bioinformaticians a global, simultaneous view, on the transcription levels of an organism’s genes under a range of conditions or processes \cite{3,4}.

In a typical microarray experiment, thousands of DNA sequences, called probes, are obtained from the genes of interest (targets) and are printed on a glass microscope slide by a robotic arrayer, thus, forming circular spots \cite{5,6}. In order to create a genome expression profile of a biological system with microarrays, one has to compare the relative abundance
Fig. 1 – Flowchart illustrating the basic stages of the proposed methodology.

of each one of these gene sequences in two samples. Accordingly, from each sample, the messenger RNA (Ribonucleic acid) is extracted, isolated, and labelled using different fluorescent dyes (the most usually used are the red-fluorescent dye Cy5 and the green-fluorescent dye Cy3 with emissions in the 630–660 nm and 510–550 nm respectively). Following, the two samples are mixed and hybridized within the arrayed DNA spots [7]. After hybridization, the arrays are scanned using lasers that excite each dye at the appropriate wavelength. The relative fluorescence between each dye, on each spot, is then recorded using methods contingent upon the nature of the labelling reaction, i.e. confocal laser scanner, and charged couple devices [8,9]. The output of such systems is two 16-bit TIFF images, one for each fluorescent channel. The relative intensities of the spots at each channel represent the relative abundance of the DNA product in each of the two samples.

In order to extract those relative intensities from the microarray images, a well-known algorithmic pipeline is utilized [10]. The latter employs a series of image processing techniques namely gridding, segmentation, and intensity extraction [11]. Subsequently, the extracted transcription levels are further analyzed, through data mining techniques that (i) assist biologists in understanding function of genes and (ii) facilitate the extraction of meaningful biological conclusions [12]. Notwithstanding, microarrays suffer from an inherent lack of precision due to multiple sources of noise (biological and experimental), which are introduced in a microarray experiment. Noise complicates the task of microarray image analysis and, thus, it directly affects subsequent data analysis [14]. As a consequence, improper treatment of noise might preclude the extraction of meaningful biological conclusions [13].

Noise confounds all microarray image processing steps (i.e. gridding, segmentation, intensity extraction), and mostly the segmentation step. Inaccurate determination of spots boundaries, during the segmentation process, evokes wrong estimation of the relative mean spots’ intensities. Therefore, the reproducibility and validity of the gene expression levels, derived from microarrays, is degraded. To this end, several studies have been proposed [15–18], in which image enhancement is suggested as a solution to the microarray image noise problem. Results of these studies have indicated a superior quality of the enhanced images, without however examining whether enhancement leads to more accurate spot segmentation or to reduction of the variability of the extracted gene expression levels. Nonetheless, as the findings of our previous study [19] have revealed, individual restoration of spot-images, when combined with spots structural information, facilitates spot's segmentation and, consequently, improves genes quantification in terms of validity and reproducibility.

Notwithstanding, the fairly common low quality of microarrays [20] renders the task of microarray segmentation challenging and, thus, the need arises to investigate the performance of various segmentation approaches, especially in cases where microarray image quality is low. To this end, several approaches for the demanding task of microarray segmentation have been proposed [21–26]. Among them, histogram based segmentation techniques and unsupervised classification techniques have been proven to provide higher reproducibility in the extracted intensities for images characterized as highly noisy [14,20]. Those techniques, however, have not examined whether, if preceded by a noise reduction step (adjustable spot-image restoration [19]), might lead to more accurate spot segmentation and to a further improvement of the reproducibility of the extracted gene expression levels.

In this study a thorough investigation of the performance of various cDNA microarray image segmentation techniques, originating from an adjustable spot-image restoration framework, is presented. Explicitly, the present study comprises three cascade stages: (i) a spot-images restoration stage, presented by our group elsewhere [19], in which a robust framework of microarray image processing techniques, designed to take into account the effect of local spot-image noise in microarray images, was utilized, (ii) a spot-images segmentation stage, in which 5 different segmentation techniques were employed individually and in ensemble majority
vote schemes in order to separate spot-images pixels as belonging either to spot-images signal region or spot-images background region, and (iii) a spot-images intensity extraction stage, in which spots’ relative intensities were extracted from the spot-image signal regions (as they were defined in the previous stage (ii)). Regarding the spot-images segmentation stage, the techniques employed were: (1) histogram based techniques (thresholding based on: (a) histogram’s moment [27], (b) histogram’s concavity [27], (c) histogram’s entropy [27]); (2) Region Growing techniques (Seeded Region Growing (SRG) [28]); and (3) unsupervised classification techniques (Gaussian-Kernelized Fuzzy C-Means (GK-FCM) [29]). To the best of the authors’ knowledge, the employed segmentation techniques, except for the SRG, have never been used in the demanding task of microarray segmentation. The performances of individual and ensemble segmentation schemes were evaluated on both simulated and real cDNA microarray images, in terms of valid (simulated images) and reproducible (real images) extraction of gene expression levels, which is the case of concern in microarray image processing.

2. Methods

Prior to spots segmentation stage, a common noise reduction step, through the spots restoration stage, was employed (see Section 2.1). Subsequently, restored images were segmented utilizing various segmentation techniques (see Section 2.2) in order to provide genes expression levels through a spot-image intensity extraction stage (see Section 2.3). A flowchart of the steps followed, in order to extract spots intensities using the proposed methodology, is presented in Fig. 1.

2.1. Spot-images restoration stage

In order to take into account the effect of noise in cDNA microarray images and, thus, facilitate segmentation, a robust framework of image processing and analysis techniques [19] was utilized. Accordingly, the framework comprised the following stages: (1) gridding, for facilitating spot identification, (2) clustering (unsupervised discrimination between spot and background pixels), applied to spot-image for automatic local noise assessment, and (3) modeling of local image restoration process, for spot-image conditioning. Details of the employed framework are provided elsewhere [19].

2.2. Spot-images segmentation stage

Individual restored cDNA microarray spot-images were segmented utilizing various segmentation techniques either individually or in a majority vote ensemble scheme.

2.2.1. Histogram moments (HM)

Given a spot-image \( f \) with \( n \) pixels, the histogram moments technique, for a 2 class problem, searches for an appropriate
Table 1 – Mean values (in terms of intensity) of the PE segmentation metric for both channels of high quality simulated images as obtained by the boxplots of Figs. 1 and 2.

<table>
<thead>
<tr>
<th>High quality</th>
<th>HM</th>
<th>HC</th>
<th>HE</th>
<th>SRG</th>
<th>GKFCM</th>
<th>HM</th>
<th>GC</th>
<th>CM</th>
<th>SRG</th>
<th>GKFCM</th>
<th>HM</th>
<th>HE</th>
<th>HE</th>
<th>SRG</th>
<th>GKFCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red channel</td>
<td>0.076</td>
<td>0.067</td>
<td>0.277</td>
<td>0.044</td>
<td>0.246</td>
<td>0.031</td>
<td>0.030</td>
<td>0.032</td>
<td>0.080</td>
<td>0.053</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Green channel</td>
<td>0.072</td>
<td>0.064</td>
<td>0.275</td>
<td>0.040</td>
<td>0.245</td>
<td>0.027</td>
<td>0.026</td>
<td>0.028</td>
<td>0.076</td>
<td>0.050</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

threshold value, which separates $f$ into two pixel classes: i/the spot-image signal class and ii/the spot-image background class.

Initially, the spot-image $i$th moment is calculated from the spot-image $f$ histogram as

$$m_i = \frac{1}{n} \sum_j n_j (h_j)^i, \quad i = 1, 2, 3, \ldots$$  \hspace{1cm} (1)

where $n_j$ is the total number of pixels in $f$ with gray value $h_j$.

Considering that image $f$ is a blurred version of an ideal bilevel image $g$, which consists of pixels with only 2 gray values $h_0$ and $h_1$ ($h_0 < h_1$), then the histogram moment thresholding technique searches for a threshold value $t$ such that if all pixel gray values of the spot-image signal class are replaced by gray value $h_0$ and all pixel gray values of the spot-image background class are replaced by gray value $h_1$, then the first three moments (taking as $m_0 = 1$) of image $f$ are preserved in the resulting bilevel image $g$ [27].

Accordingly, if $p_0$ and $p_1$ are the fractions of the pixels belonging to the spot-image signal and background classes respectively, the threshold value $t$ is calculated as the closest value to the $p_0$-tile of the histogram of $f$, defined as:

$$p_0 = \frac{1}{n} \sum_{h_j \leq t} n_j,$$  \hspace{1cm} (2)

where the values of $p_0$ and $p_1$ are obtained as the solution of equations:

$$p_0 + p_1 = 1$$  \hspace{1cm} (3)

and

$$m_i' = m_i, \quad i = 1, 2, 3$$  \hspace{1cm} (4)

Fig. 3 – PE segmentation metric boxplots for the green channel of the high quality simulated images as obtained by individual and ensemble majority vote segmentation schemes. Obelisks are PE values characterized as outliers.
Table 2 – Mean values (in terms of intensity) of the PE segmentation metric for both channels of low quality simulated images as obtained by the boxplots of Figs. 3 and 4.

<table>
<thead>
<tr>
<th>Low quality</th>
<th>HM</th>
<th>HC</th>
<th>HE</th>
<th>SRG</th>
<th>GKFCM SRG</th>
<th>HM</th>
<th>GKFCM SRG</th>
<th>HE</th>
<th>GKFCM SRG</th>
<th>HM</th>
<th>HC</th>
<th>HE</th>
<th>SRG</th>
<th>GKFCM SRG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red channel</td>
<td>0.302</td>
<td>0.110</td>
<td>0.287</td>
<td>0.208</td>
<td>0.283</td>
<td>0.153</td>
<td>0.151</td>
<td>0.155</td>
<td>0.243</td>
<td>0.193</td>
<td>0.149</td>
<td>0.152</td>
<td>0.241</td>
<td>0.139</td>
</tr>
<tr>
<td>Green channel</td>
<td>0.300</td>
<td>0.107</td>
<td>0.285</td>
<td>0.207</td>
<td>0.282</td>
<td>0.150</td>
<td>0.149</td>
<td>0.152</td>
<td>0.241</td>
<td>0.193</td>
<td>0.149</td>
<td>0.152</td>
<td>0.241</td>
<td>0.135</td>
</tr>
</tbody>
</table>

2.2.2. Histogram entropy (HE)

Given a spot-image $f$, the probability distribution of its gray-levels can be defined as $p_1, p_2, \ldots, p_n$. In order to separate the spot-image signal from the spot-image background, this probability distribution is appropriately modified to derive two distinct probability distributions: one for the pixels belonging to the spot-image region (defined for discrete values from 1 to $s$) and one for the pixels of spot-image background (defined for discrete values from $s+1$ to $n$) defined as:

**Signal**:

$$p_1 \frac{p_s}{p_s}, p_2 \frac{p_s}{p_s}, \ldots, p_s \frac{p_s}{p_s}$$

(5)

and

**Background**:

$$\frac{p_{s+1}}{1-p_s}, \frac{p_{s+2}}{1-p_s}, \ldots, \frac{p_n}{1-p_s}$$

(6)

Accordingly, the entropies associated with each distribution are defined as:

$$H_{\text{signal}} = -\sum_{i=1}^{s} \frac{p_i}{p_s} \ln \frac{p_i}{p_s}$$

(7)

and

$$H_{\text{background}} = -\sum_{i=s+1}^{n} \frac{p_i}{1-p_s} \ln \frac{p_i}{1-p_s}$$

(8)

The appropriate threshold value $t$, that separates the spot-image signal region from the spot-image background region, is then calculated as the gray-level value, which maximizes the quantity [27]:

$$t = \arg \max \left\{ H_{\text{signal}} + H_{\text{background}} \right\}$$

(9)

Fig. 4 – Boxplots illustrating the PE segmentation metric for the red channel of the low quality simulated images as obtained by individual and ensemble majority vote segmentation schemes. Obelisks are PE values characterized as outliers.
2.2.3. **Histogram concavity (HC)**
Given a spot-image $f$ and by defining the spot-image histogram $H$ over a set of gray-levels $g_0, g_1, \ldots, g_{t-1}$, the height of the histogram may be obtained at these gray-levels as $h(g_0), h(g_1), \ldots, h(g_{t-1})$, where $h(g_i) \neq 0$ for all $i$. Then $H$ might be regarded as a two-dimensional region from which its convex hull $\bar{H}$ might be constructed and, thus, a threshold value $t$ might be determined by the histogram’s concavities. Accordingly, possible threshold values are those gray-values at which the difference $\bar{h}(g_i) - h(g_i)$ (where $\bar{h}(g_i)$ is the height $\bar{H}$ and $h(g_i)$ is the height of $H$ at gray level $g_i$) has local maxima [27]. In order to eliminate spurious concavities, which are attributed as “noisy” histogram’s spikes, a balance measure is introduced [30]:

$$E_i = \left\{ \frac{\sum_{j=g_0}^{t-1} h(j)}{\sum_{j=g_i}^{t-1} h(j)} \right\}$$

(10)

where $E_i$ measures the balance of the histogram about the gray level $g_i$. Thus, maxima of $\bar{h} - h$, when the value of $E$ is small, are attributed as spurious maxima and, consequently, are ignored.

2.2.4. **Seeded region growing (SRG)**
Given a spot-image $f$, the SRG algorithm separates the spot-image signal region from the spot-image background region by iteratively growing contiguous signal pixel regions that differ statistically from the background region [19]. Providing a set of predefined seed points as starting points for the segmentation, the algorithm iteratively grows those pixel regions by including the most homogenous pixels from the neighborhood to the segmented regions [11,28]. In our approach, the initial seed point at each spot-image was predefined at the spots’ restoration stage through the clustering step (the clustering step resulted in 2 regions: (i) spot-image signal region and (ii) spot-image background region. As initial seed the centroid of the spot-image signal region was utilized). This iterative procedure of growing pixel regions within each spot-image continued until all pixels of the spot-image were assigned to either the spot-image signal region or its background.

2.2.5. **Gaussian-kernelized fuzzy-C-means (GK-FCM)**
Given a spot-image $f$, the Gaussian-Kernelized Fuzzy C-Means (GK-FCM) unsupervised classification algorithm [29] was employed to discriminate the spot-image signal pixels from pixels belonging to surrounding background. The GK-FCM, based on a random initialization process, searches iteratively for cluster centers (centroids) that minimized the dissimilarity function:

$$J = \sum_{i=1}^{M} \sum_{j=1}^{N} u_{i,j} \left( 1 - \exp \left( -\frac{d_{i,j}(x_j, c_i)}{\sigma^2} \right) \right)$$

(11)

where $x_j, j=1, 2, \ldots, N$ are the pixels of the spot-image, $c_i, i=1, 2, \ldots, M$ are the cluster centers, $d_{i,j}$ is the Euclidean distance between centroid $c_i$ and data point $x_j$, $\sigma$ is the spread of the

![Fig. 5 – PE segmentation metric boxplots for the green channel of the high quality simulated images as obtained by individual and ensemble majority vote segmentation schemes. Obelisks are PE values characterized as outliers.](image)
Fig. 6 – M–A plots for a randomly selected simulated image employing the HM (a), HC (b), HE (c), SRG (d), GKFCM (e), HM-GKFCM-SRG (f), HC-GKFCM-SRG (g), HE-GKFCM-SRG (h), HM-HC-HE (i), and HM-HC-HE GKFCM-SRG (j) segmentation techniques.
Gaussian Kernel ($\sigma = 0.5$), $u_{ij}$ is the element of a fuzzy membership function matrix $U = [u_{ij}]$ with values $0 \leq u_{ij} \leq 1$, and $m$ is a weighting component ($m = 2$). The output of the iterative procedure is two clusters containing the pixels belonging to spot-image signal region and spot-image background-region.

2.2.6. Majority vote ensemble segmentation scheme

In order to explore whether the performance of individual segmentation techniques is increased when combined in an ensemble segmentation scheme, the majority vote rule was utilized [31]. The majority vote scheme goes with the decision when there is a consensus for it or at least more than half of the individual segmentation techniques agree on it. Consequently, segmentation outputs from each technique were appropriately modified (pixels belonging to the spot-image signal region were labeled 1 while pixels belonging to the spot-image background region were labeled 0) and were combined to form the final segmentation decision that satisfies:

$$G_r(X) = \sum_{i=1}^{R} d_{r,i}(X)$$  \hspace{1cm} (12)
where \( r \) is the class, \( X \) is the unknown spot-image pixel, \( i=1, 2, \ldots \), \( R \) is the odd number of segmentation techniques involved in the majority vote scheme, \( d_{ri} \) is the binary decision value \( \{0, 1\} \), 0 corresponds to spot-image background class and 1 to spot-image signal class. For a two class problem, if \( G_1(X) > G_2(X) \), the unknown pixel \( X \) is classified to class 1 (signal), otherwise \( X \) is classified to class 2 (background).

2.3. Spot-images intensity extraction stage

In order to extract spots relative intensities, spot-image signal regions, as characterized either by individual segmentation techniques or by the ensemble majority vote segmentation scheme, were referred to the original spot-images, since intensities in the processed spots-images were altered by the restoration process. Subsequently, all
3. Material

Individual and ensemble segmentation schemes performances were evaluated employing both simulated (in terms of validity) and real cDNA (in terms of reproducibility) microarray images.

3.1. Simulated cDNA images

A total of 100 publicly available [33] simulated cDNA microarrays images (50 high quality images and 50 low quality images) were evaluated. Each image contained 1000 spots, with realistic characteristics, produced by Nykter et al.’s [13] microarray simulator. High quality images were characterized by low level noise and low variability in spot sizes and shapes while low quality images

extracted intensities were normalized using global normalization [32].
Fig. 8 – Boxplots depict the pairwise MAE between all replicates (totally 10 MAE values from which the mean value for each spot is illustrated here). Obelisks are MAE values characterized as outliers.

were characterized by higher noise levels, varying spot diameters, and more irregularities in the spots’ shapes [20].

In each simulated image, pixels were pre-assigned as belonging to spot-image signal region or background. Thus, the pixel-based segmentation accuracy of individual and ensemble segmentation schemes, in terms of validity, was assessed employing the Probability of Error (PE) metric defined as [34]:

$$PE = P(O) \times P(B|O) + P(B) \times P(O|B)$$  \hspace{1cm} (13)

where $P(B|O)$ is the probability of error in classifying objects as background, $P(O|B)$ is the probability of error in classifying background as objects, $P(O)$ and $P(B)$ are a priori probabilities of objects and background in spot-images. For our case, the signal region of the spot-image is considered to be the object that must be discriminated from its background.

3.2. Real cDNA images

A publicly available dataset [33] was employed, comprising 5 real cDNA microarrays images (each one containing 12,288 spots). Microarray images were produced according to a replicated experiment, in which five replicated hybridizations from one experiment were performed. Due to the replication process, each spot should have the same gene expression ratio and, thus, the performances of individual and ensemble segmentation schemes, in terms of reproducibility, were accessed through the pairwise Mean Absolute Error (MAE) between the replicated images (altogether 10 pairwise MAE values) [20].

4. Results

Regarding the majority vote ensemble classification scheme, all possible combinations for the available segmentation techniques were evaluated. Boxplots of Figs. 2 and 3 illustrate the segmentation accuracy for the high quality simulated images, employing the PE segmentation metric (as described in Section 3.1) for the red and green channels. Table 1 provides the mean values of the PE boxplots of Figs. 2 and 3, in terms of intensities, for the evaluated 50 high quality simulated images for both channels (red and green). Boxplots of Figs. 4 and 5 depict the segmentation accuracy for the low quality simulated images, as calculated by the PE segmentation metric for the intensities extracted from both channels. Table 2 summarizes the mean values of the PE boxplots (Figs. 4 and 5), in terms of intensity, for the evaluated 50 low quality simulated microarray images. Fig. 6 illustrates the M–A plots for the segmentation methods employed in the present study (Fig. 6a–j), considering a randomly selected simulated microarray image. Fig. 7 shows one low quality spot-image and corresponding binary segmented images (Fig. 7a–k).
Fig. 9 – M–A plots for a randomly selected real cDNA microarray image employing the HM (a), HC (b), HE (c), SRG (d), GKFCM (e), HM-GKFCM-SRG (f), HC-GKFCM-SRG (g), HE-GKFCM-SRG (h), HM-HE (i), and HM-HE-GKFCM-SRG (j) segmentation techniques.
Regarding real cDNA microarray images, Fig. 8 shows the calculated pairwise MAE between the extracted expression ratios of all possible pairs (as described in Section 3.2) for the dataset of the 5 replicated images. Table 3 provides the mean values of the pairwise MAE (Fig. 8), as they were calculated for the extracted expression levels of the 5 replicated images.

Similarly to Figs. 6 and 7, Fig. 9 a–j and Fig. 10 a–k illustrate M–A plots and single spot-image segmentation results, but this time considering real microarray images.

5. Discussion

Microarray technology suffers from several sources of errors that affect the accuracy and validity of the extracted gene expression levels, influencing, in this way inference of accurate biological conclusions. The most affected stage in the microarray image analysis pipeline is the segmentation stage. Notwithstanding, as the results of our previous study have revealed [19], when noise reduction is applied locally on each spot-image, prior to the segmentation stage, the results of the segmentation stage are improved. Thus, in this study, the performances of various segmentation algorithms were explored, either on individual or ensemble majority vote schemes, in conjunction with a noise reduction step, which was based on an adjustable spot-image restoration framework [19].

Since performance evaluation of microarray image segmentation algorithms is not a straightforward task to consider in real microarray images [20], a dataset of 100 simulated images was employed. In each one of these simulated images, the a priori knowledge of spot-image signal and background regions allowed for the assessment of the segmentation accuracy through the Probability of Error segmentation metric. Moreover, the variability of the simulated images (50 images of high and 50 of low quality) rendered the possibility to demonstrate the sensitivity of the evaluated segmentation schemes (individual and ensemble) plausible.

As expected, the performances of the segmentation algorithms on good quality images were higher as compared to those on low quality images. This is clearly supported by the
obtained segmentation accuracies, as they are depicted on Tables 1 and 2. Accuracies for both individual and ensemble segmentation schemes were reduced as the quality of the images was degraded.

Regarding high quality simulated images, the highest segmentation accuracy in terms of mean Probability of Error, for both red and green channels, was achieved by the ensemble majority scheme, in which the thresholding based on histogram’s concavity, the Gaussian Kernelized Fuzzy-C-Means, and the Seeded Region Growing techniques participated (mean PE was found 0.030 and 0.026 for the red and green channel respectively) (HC-GKFCM-SRG). This

Fig. 10 – A randomly selected real cDNA microarray spot-image (Fig. 7a) and its corresponding segmentation result employing the HM (b), HC (c), HE (d), SRG (e), GKFCM (f), HM-GKFCM-SRG (g), HC-GKFCM-SRG (h), HE-GKFCM-SRG (i), HM-HC-HE (j), and HM-HC-HE-GKFCM-SRG (k) techniques.
might be attributed to the diverse theoretical framework of each segmentation algorithm. Complementary information of individual segmentation techniques was exploited by the ensemble majority vote scheme and, thus, the segmentation accuracy improved.

Regarding low quality simulated images, the highest segmentation accuracy was achieved by the thresholding technique, which is based on the histogram concavity (the mean PE was found 0.110 and 0.107 for the red and green channels respectively) (HC). This might be attributed to the fact that HC determines the threshold value, which separates the spot-image signal region from its background by analyzing the concavity of the histogram. The latter provides a good threshold value in cases where the histogram of a spot-image
is not well separated (i.e. in the ideal case the histogram of a spot-image forms a valley between the spot-image’s signal and background regions) and this is usually observed in highly degraded spot-images.

Regarding real microarray images, the actual spot-image boundaries and, thus, the spots intensity levels were not available. Therefore, an alternative method, in terms of reproducible gene expression levels, should be selected for assessing the performances of individual and ensemble majority vote schemes. Accordingly, a properly designed dataset of 5 replicated real cDNA microarray images was employed [20]. Due to the replication, each spot should have the same intensity ratio throughout the replicated experiments, and therefore the “sameness” of the replicates could be assessed by utilizing the pairwise MAE (totally 10 pairwise MAE values). Fig. 8 illustrates the boxplots of MAE as they were calculated for the spots extracted intensities ratios of the 5 replicated images and Table 3 depicts the mean values of those boxplots. The lowest MAE value, in terms of extracted intensities ratios, was achieved by the HC technique (mean MAE was found 0.153). Lower MAE values are indicative of higher segmentation performances and, thus, of more accurate extraction of gene expression levels. Since real cDNA microarrays are characterized by a fairly common low quality, those results are in line with the results obtained by the HC technique on the low quality simulated microarray images.

Additionally, it might be interesting to notice that for both simulated and real microarrays the performance of the GKF CM algorithm, in terms of validity (simulated images) and reproducibility (real images), is relative low when compared with results obtained by the rest of the evaluated techniques. This might be attributed to the GKF CM’s random initialization, which causes variations in the segmentation results each time segmentation of the same spot-image is performed.

Regarding spots intensity extraction, for a 6000 x 2160, 16-bit cDNA image containing 12,288 spots, the processing times were 490, 561, 1246, 726 and 1960 s for the individual segmentation techniques, for the HM, HC, HE, SRG and GKF CM techniques, respectively. Similarly, for each combination of the ensemble majority vote segmentation scheme, the processing times were 2882, 1435, 2449, 2039 and 3165 s for the HM-GKF CM-SRG, HC-GKF CM-SRG, HE-GKF CM-SRG, HM-HE and HM-HE-GKF CM-SRG, respectively. This might seem lengthy, however, it must be stressed that the code has not yet been optimized for speed.

6. Conclusion

As the results of this study indicate, the performances of microarray segmentation techniques are strongly dependent on the quality of the cDNA images to be processed. In cases where the quality of cDNA images is high (simulated images), the utilization of an ensemble majority vote segmentation scheme (HC-GKF CM-SRG) seems to provide the highest segmentation accuracy. However, in cases where the image quality is degraded, which is usually observed in real world experiments, the selection of a histogram based technique (such as HC) proved to provide reliable and reproducible results in reasonable processing time.

REFERENCES


