Neurofuzzy Segmentation of Microarray Images

S. Battiato, G. M. Farinella, G. Gallo, G. C. Guarnera
Dipartimento di Matematica e Informatica
Università di Catania
Viale Andrea Doria 6, 95125, Catania
{*battiato, gfarinella, gallo, guarnera*}*@dmi.unict.it*

Abstract

In this paper we propose a novel microarray segmentation strategy to separate background and foreground signals in microarray images making use of a neurofuzzy processing pipeline. In particular a Kohonen Self Organizing Map followed by a Fuzzy K-Mean classifier are employed to properly manage critical cases like saturated spot and spike noise. To speed up the overall process a Hilbert sampling is performed together with an ad-hoc analysis of statistical distribution of signals. Experiments confirm the validity of the proposed technique both in terms of measured and visual inspection quality.

1. Introduction and Motivations

DNA microarray is rapidly evolving as fundamental biotechnology for genes expression profiling and biomedical studies. In a typical microarray experiment, two 16-bit TIFF images are obtained by using a microarray scanner. The corresponding pixel intensities in the images are proportional to the expressions of genes.

To extract the biological meaning of the results in a microarray experiment the images have to be properly processed. The main tasks involved in microarray image processing are to assign coordinates to the spot locations (Gridding) and to discriminate foreground pixels vs. background pixels (Segmentation). After these phases the intensity of the segmented spots is extracted end evaluated through suitable quality measures [1]. The final purely quantitative data are eventually analyzed by biomedical researchers.

Reliable interpretation of the extracted data crucially depends on the robustness and the effectiveness of the involved image processing algorithms. In particular a good segmentation of the spots is fundamental for having no corrupted data in terms of false foreground (background classified as gene expression) or false background (gene signals classified as background). Segmentation in this case is difficult because of the high variability in shape and luminance of spots. Random noise (biological and instrumental) moreover is introduced in each phase of the experiment. Finally, each microarray experiment is influenced by the specific system used to perform it [6].

For all the reasons above, microarray experiments present a high variability. This makes difficult the task of building appropriate training sets for the learning phase of supervised segmentation algorithms.

Existing methods for microarray spots segmentation can be classified mainly in four groups [6]:

1. *Fixed circle*: all image spots are enclosed in circles of constant diameter;
2. *Adaptive circle*: circle diameter is estimated separately for each spot;
3. *Adaptive shape*: no restriction on the spots shape;
4. *Histogram based*: shape is based on statistical signal distribution without using spatial information.

A comparison of the main existing methods and software for microarray image segmentation can be found in [1].

Recently, Principal Component Analysis (PCA) was employed to de-correlate the signal from the noise by projecting the microarray spots on the "eigenspots" space [5]. Despite to the straightforward natural application of PCA to images containing just one spot, this seems do not to be a real case: microarray can contain thousand of spots, and the process to locate each spot within the microarray (gridding phase) should be solved before to apply the method. The method proposed in [2] uses snake model to capture the boundary of a spot and Fisher discrimination between signal and background.

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

The method is applied only on manual extracted sub-microarray containing a single spot.

The assumption that the spot centers and manufacturer parameters are known is too strong: a useful microarray segmentation algorithm should be able to locate spots automatically and should be independent from such parameters. Moreover a robust quality measure able to take into account the irregularities of spot size, the signal-to-noise-ratio, the variability in local background, level of local background and the percentage of saturated pixel should be used to evaluate the real performances of a segmentation technique.

In this paper we propose a new unsupervised segmentation algorithm that takes into account the spatial relationship between pixels together with information pixel-wise. First, a Kohonen Self-Organizing Map (SOM) [3] is used to discriminate between foreground and background pixels. A Hilbert curve based sampling [4], on microarray images pixels is employed to speed up the self organization phase of SOM, preserving the underlying spatial relationship between pixels. The initial segmentation obtained through SOM is then refined using a fuzzy K-means clustering. In this way, local as well as global information are taken into account in the overall segmentation process. The proposed approach is applied to the whole microarray image without manual preprocessing and with no assumption on parameters of manufacturer. To assess the performance of the proposed method we use the q_{index} quality measure as defined in [1]. Experimental results confirm the effectiveness and robustness of the proposed approach in terms of visual and quality measures.

The paper is organized as follows: the proposed segmentation algorithm is described in Section 2. In Section 3 the dataset and the experimental setup are detailed while results are reported in Section 4. Finally summary and conclusion are given in Section 5.

2. Proposed Model

The overall schema of the proposed approach is summarized in Figure 1. The technique is applied to each channel separately and results are then combined together to obtain a binary map (signal vs. background) of the microarray.

To take into account the main characteristics of the input image, such as noise, spot intensities and so on, we use as training-set for the SOM, the image itself. Hence, a new training phase is performed for every new input image. In order to reduce the number of samples and the learning time, a sub-sampling technique based on the Hilbert curve is used [4]. A recursive method generates a Hilbert curve which overlap the $M \times N$ image. The number of iterations is related to the dimension of the input image, in order to obtain an optimal sampling rate on different images (see Section 3 for the settings adopted in our experiments). A discretization of the sampling curve defines the sample positions. The space filling property of the Hilbert curve allows to use only a subset of the input image as training set, saving remarkably on the learning time whitout any loss in terms of overall segmentation performances. Hilbert sampling is moreover more robust than uniform spatial sampling because it guarantees a balanced sampling, beetween the two involved classes of signals (i.e. background and foreground).

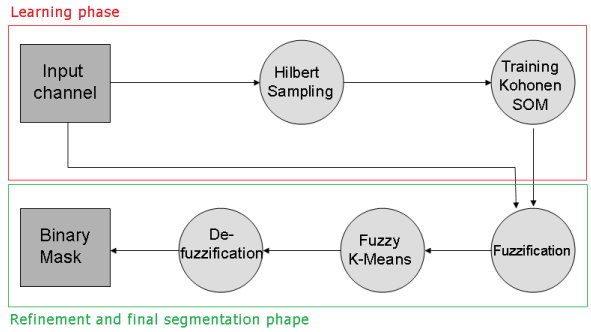

Figure 1. Schema of the proposed model.

The simple pixel intensity value is not enough to achieve a correct classification into foreground and background, because many factors affect the spatial distributions and the statistic distributions of such classes: local neighborhood information have to be considered to obtain features that are useful to a SOM segmentation process. In the following we briefly describe the rationale of the selected features.

In a typical microarray image there are more pixels belonging to background than foreground, so very likely, pixels with a value lower than the global median (G_{Median}) come from background cluster. On the other hand, foreground pixels show very high values (particularly in presence of high saturated spots). Observe that in this case the global average value ($G_{Average}$) is always greater than the global median. A pixel with a value between global median and global average values requires further investigation. In Figure 2, a typical histogram of a single channel microarray image is shown, where the global G_{Median} and $G_{Average}$ are overimposed to the plot. Three different value intervals result from this process (Figure 2). We use the pixel value in addition to a score defined as follows:

$$score(pixel) = \begin{cases} \alpha & \text{if } pixel < G_{Median} \\ \beta & \text{if } G_{Median} \leq pixel \leq G_{Average} \\ \gamma & \text{if } G_{Average} < pixel \end{cases}$$

where $\alpha, \beta, \gamma \in [0, 1]$.

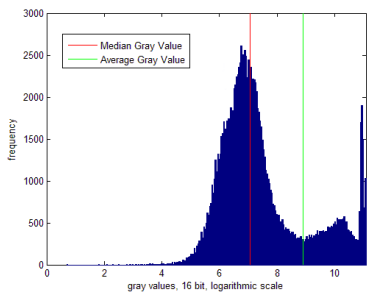

Figure 2. Typical histogram of intensities.

The difference between each pixel gray value and the average gray value in the 3×3 and 5×5 neighborhood windows complete the set of local features[1].

A SOM is used to perform binary segmentation using these features. Convergence is achieved within a fixed number of epochs through adapting the learning parameters. Because of the relevant presence of many critical cases, the classification obtained with the SOM has to be furtherly refined.

A Fuzzy K-Means algorithm is employed to deal with ambiguous pixels. Starting from the two classes derived by the output layer of the SOM, each k-dimensional feature vector ($k = 4$ in our case) is evaluated by a fuzzy membership extraction procedure, which assigns a membership value related with the distance to the centroid of foreground and background classes. The standard membership functions is iteratively evaluated and centroids are updated at each iteration.

After that fuzzy K-means convergence, pixels are defuzzified, by assigning each element to the cluster for which it shows the maximum degree of membership, with respect to the final centroids c_1 and c_2.

3. Experimental setup

To evaluate performances of the proposed technique we made both visual and numerical comparisons with respect to the segmentation results obtained using MISP [1]. MISP is an adaptive and fully automatic strategy to segment microarray images which processes each microarray image to produce five semantic regions: background, local background, red channel foreground, green channel foreground, red channel and green channel foreground. The MISP strategy is employed in MIAF framework, a complete microarray image analysis tool able to automatically perform gridding, segmentation, visualization, data extraction and spot evaluation after automatic alignment of the microarray image if it is affected by rotation problems. See [1] for more details.

Three dataset set was used for testing purpose:

- The *Whole Yeast Genome* microarrays, freely downloadable at the MAGIC Tool Website[2].
- The *Whole Genome of Saccharomyces Cerevisiae* microarrays, freely downloadable at Pat Browns lab homepage[3].
- A collection of microarray images available in the Stanford Microarray Database (SMD). In particular, we refer to the experiments *ExpID 1573934* and *ExpID 51509*. (See [1] for more details.)

Note that MISP has been compared with state-of-the-art solutions (i.e., Genepix, Scanalyze), obtaining in almost all cases the best performances [1].

The parameters α, β and γ used to obtain the *score* feature of each pixels (see Section 2) have been learned on a training set of microarray images. Specifically, a full search procedure was employed on a grid of 1331 points equispaced in the three dimensional space $\alpha \times \beta \times \gamma$=[0,1]$\times$[0,1]$\times$[0,1]. In each run the q_{index} of each spot of the microarray used as training set was obtained and the average of the segmentation quality (SQ) associated to the involved grid point was stored. The triple (α, β, γ) corresponding to the maximum SQ have been choosen to perform the proposed segmentation on the test microarray images. The result presented in next section have been obtained using ($\alpha = 0.6, \beta = 0.9, \gamma = 1$). Other involved parameters are heuristically set as follows: fuzziness factor m = 2 (the most common choise), number of epochs is fixed to 100. Seven iterations were employed for Hilbert Sampling.

4. Results overview

To assess the quality of the proposed method three different experiments have been employed: quality measures, scatter plot, visual comparison. Various quality measures have been proposed in the literature. Here

[1] All the local features are normalized in [0,1].

[2] http://www.bio.davidson.edu/projects/magic/magic.html

[3] http://brownlab.stanford.edu/

we use the quality index (q_{index}) we have introduced in [1]. The quality index is computed on each spot and encloses information about the signal-to-noise-ratio, the local background variability and the excessively high local background (see [1] for more details).
The plot of the q_{index} relative to each spot in a microarray with 288 spots is reported in Figure 3. In Figure 4 we show a magnified detail of a small region in a microarray, which consists of 25 spots. Visual results of the segmentation methods are reported, in addition to the q_{index} plot. The new solution is able to outperform in almost all cases MISP, both considering measured and visual inspection quality. To further evaluate the segmentation strategy we compared the extracted data using a scatter plot method [1]. Scatter plots usually consist of a large body of data, and the closer the data points come when plotted to making a straight line, the higher the correlation between the two variables, or the stronger the relationship. The microarray used for this test was obtained by duplicating a single channel image and applying in the copy a small random variation for each spot center location, to simulate typical acquisition problems. Since there are no changes in pixel value, the ratio for each spot is equal to 1 (perfect correlation). The original single-channel image shows some typical problems with spot size and shape, grid rotation, and nonhomogeneous background. As shown in Figure 5 also in this case the proposed approach is slightly better than MISP results.

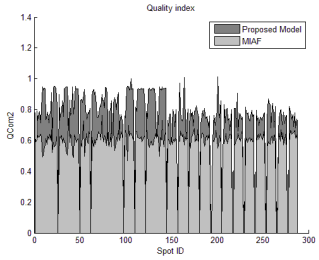

Figure 3. Quality index results.

5. Summary and conclusion

In this paper, we have proposed a segmentation strategy for microarray images based on SOM and fuzzy K-means. To speed up the self organization phase of SOM, an Hilbert curves based sampling have been employed. The overall process is fully unsupervised. Experimental results confirm that the proposed segmentation pipeline outperform our previous solution [1] both on numerical and visual comparison.

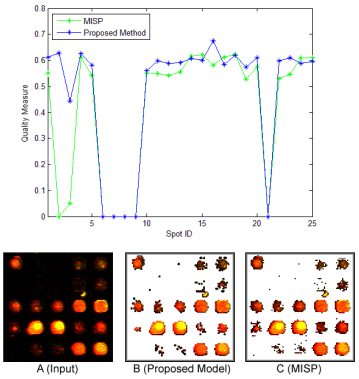

Figure 4. Proposed approach vs. MISP.

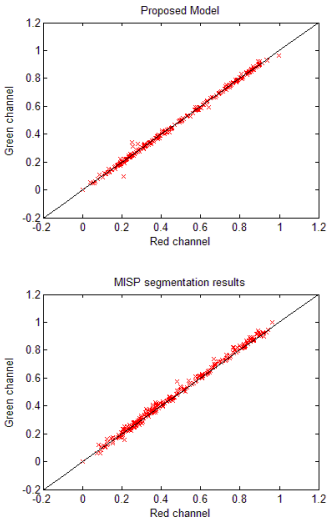

Figure 5. Scatter plots comparison.

References

[1] S. Battiato, G. D. Blasi, G. Farinella, G. Gallo, and G. Guarnera. Adaptive techniques for microarray image analysis with related quality assessment. *Journal of Electronic Imaging*, 16(4), 2007.

[2] J. Ho and W. Hwang. Segmenting microarray image spots using an active contour approach. In *Int. Conf. Image Processing*, 2007.

[3] T. Kohonen. *Self-Organizing Maps*. Springer, 2001.

[4] B. Moon, H. v. Jagadish, C. Faloutsos, and J. H. Saltz. Analysis of the clustering properties of the hilbert space-filling curve. *IEEE Trans. on KDE*, 2001.

[5] S. A. Tsaftaris, R. Ahuja, D. Shiell, and A. K. Katsaggelos. Dna microarray image intensity extraction using eigenspots. In *Int. Conf. Image Processing*, 2007.

[6] J. Yang, M. Buckley, S. Dudoit, and T. Speed. Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics*, 11(1):108–136, 2002.