Gene expression reliability estimation through cluster-based analysis

Luca Sterpone, Alfredo Benso, Stefano Di Carlo, Gianfranco Politano
Dipartimento di Automatica e Informatica (DAUIN)
Politecnico di Torino
Torino, Italy
contact mail: luca.sterpone@polito.it

Abstract—**Gene expression is the fundamental control of the structure and functions of the cellular versatility and adaptability of any organisms. The measurement of gene expressions is performed on images generated by optical inspection of microarray devices which allow the simultaneous analysis of thousands of genes. The images produced by these devices are used to calculate the expression levels of mRNA in order to draw diagnostic information related to human disease. The quality measures are mandatory in genes classification and in the decision-making diagnostic. However, microarrays are characterized by imperfections due to sample contaminations, scratches, precipitation or imperfect gridding and spot detection. The automatic and efficient quality measurement of microarray is needed in order to discriminate faulty gene expression levels. In this paper we present a new method for estimate the quality degree and the data's reliability of a microarray analysis. The efficiency of the proposed approach in terms of genes expression classification has been demonstrated through a clustering supervised analysis performed on a set of three different histological samples related to the *Lymphoma*'s cancer disease.**

Keywords-DNA microarray, profiling, gene expression, reliability classification

I. INTRODUCTION

Nowadays, the study of organisms are based essentially on genomic. Numerous genome related projects create several thousands of biological meaningful information and enable the exploration of gene functions belonging to Deoxyribonucleic Acid (DNA) sequences. Analyze the gene functionalities allow to determine the cellular process that have been disrupted or compromised thus providing a window of the gene's biological role. Several methods have been used in the past to report the gene's expression. Many of these methods are based on fairly labor-intensive operations. Recently, the gene expression analysis have been revolutionized by the introduction of DNA microarrays. These devices allow to analyze the gene expression by the visual inspection of Ribonucleic Acid (RNA) produced by thousands of genes to be monitored at once. By examining the expression of several genes simultaneously, it is possible to identify and study the gene expression patterns of a type of cellular's tissue under a certain physiological condition [1]. Physically, a DNA microarray consists of a solid glass or silicon surface, studded with a large number of DNA fragments, each containing a nucleotide sequence that serves as a probe for a specific gene. DNA microarrays have been used to examine in particular the gene expression signature of different types of human cancer cells, providing the study of additional layer of information useful for predicting gene functions in relation to cancer diseases.

Once the hybridization process is completed, the DNA microarray is visual inspected by an automated scanning-laser microscope that scans a microarray slide with several blocks of two dimensional (2-D) arrays where the DNA fragments are localized. The result is recorded in the form of an image, where the most expressed genes are indicated by an higher intensity with different color channels ranging from the green cyanine dyes, Cy3, and the red cyanine Cy5. The measurement of the gene expression levels is obtained analyzing the resulting image. Nevertheless the DNA fragments contained in the microarray have prior known positions due to their regular structure, several issues during the biological process influence the quality of the measurement.

Imperfections in microarray are due to sample contamination, scratches, precipitation, imperfect gridding or segmentation. These imperfections affect the extractions of the gene expression levels compromising the microarray classification. Since the microarray analysis are performed on recorded images, quality measurements is retrospective and thus need automatic tools to detect, censor or flagging specific genes that are not correctly expressed and if considered will contribute erroneously to the microarray classification [2]. Automatic tools exist to estimate the DNA segment shape in cDNA array using a metric specialized in the automatic gridding and in the local qualifications of the spot [3] [4]. However, the major challenge remain in the identification of the appearance irregularity of a grid and on the measurement of the illumination noise that corrupts the expected characteristics of the genetic markers where DNA fragments are placed.

A previous work proposed a method for determining individual DNA fragments and their borders in order to maximize the detected DNA fragments and to compensate the errors introduced by artifacts [5], however that algorithm does not provide any evidence of the quality measures on the microarray images. A recent work proposes an algebraic framework for count faulty DNA fragments, however this method is able to produce only an average probability of failure detected fragments, while no reliability information are given about the level of confidence and the accuracy of the gene expression levels computation [6].

In this paper we propose a new method for measure and estimate the reliability of the gene expression levels through the analysis of the DNA microarray fragments. The method is based on an analysis flow consisting of spot finding and image segmentation of DNA microarray images using the embedded dual core platform developed in [7] and integrated with two novel software modules. A quality assurance module, able to individuate, through a set of image rules, the imperfections in DNA microarray images, and a hierarchical clustering

978-1-4244-3599-9/09/$25.00 ©2009 IEEE

algorithm able to create a reliability metric on the set of analyzed microarray samples and to compute the level of accuracy (in terms of quality and reliability) of the marker genes.

The paper presents also an experimental analysis using a set of DNA microarray images related to three different histological classes: *Normal Tissue*, *Follicular Lymphoma* and *Diffuse Large B-cell Lymphoma*. The results demonstrated the capability of the proposed method to provide an estimation of the reliability degree of the gene expression levels.

II. DNA MICROARRAY SEGMENTATION ERRORS

As illustrated in figure 1 a DNA microarray image is characterized by three main objects: the DNA fragments (or spots), the Sub-grids and the Background.

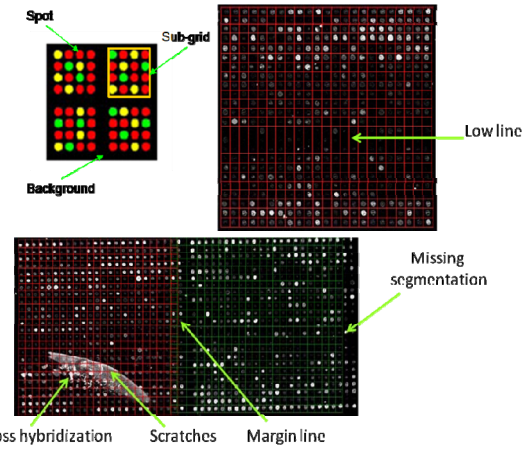

Figure 1. An example of ideal slice of a DNA microarray image on the top-left. An example of Low line error in the top-right. Some examples related to missing segmentation, cross hybridization, scratches and margin lines phenomena that can happen during the segmentation phase.

Digital DNA microarray images are characterized by two main problems: the noise level and the low pixel intensity. While in the first case, the integrity of the DNA fragment is affected by the neighborhood high intensity pixels, in the second case the DNA fragment is difficulty recognizable since, having low intensity level, is not discriminated with respect to the background of the image.

Considering the different type of errors, we classified the effects considering their regions, noise and genetic characteristics. When the region characteristics are considered, the classification is:

- *Local*: the errors are related only to a single grid, while the other grids are not affected. An example of this error is the missing segmentation, as reported in figure 1. These errors are generally correctable by moving the interested grid in the right position.
- *Expanded*: the errors are related to more than a single grid, i.e. the errors overlap two neighboring grids.

When the signal/noise ratio is considered, the errors are classified as:

- *Low line*: the missing segmentation is placed to a central region of the grid, as illustrated in figure 1. This is provoked by signal of low intensity internally to a grid.
- *Margin line*: the missing segmentation is located at the margin of a grid. This happens since the intensity of the DNA fragments progressively decrease along the x axis. Besides, in some cases the geometry of the DNA microarray is characterized by a minor number of DNA fragments in the last line of each grid. In this case the average intensity decreases and thus the autocorrelation coefficient. An example of this phenomena is illustrated in figure 1.c.

Finally, considering the genetic characteristics of each DNA fragments, two further effects are considered:

- *Cross hybridization*: the error is provoked for the annealing of a single-stranded DNA fragments to a single-stranded target DNA to which it is only partially complementary. The results of this effects are two (or more) equally wrong expressed neighborhood DNA fragments.
- *Scratches*: this type of error is due for the inclusion in the DNA fragment segmentations of pixels with artifacts (i.e. dust particles, scratches or spot contaminants). The consequence is a poor signal separation between the DNA fragments and the background, thus incrementing and distorting the correct signal/background ratio.

These type of effects are the most critical ones, since they are not identifiable through gridding or segmentations algorithm and therefore require a further data analysis.

III. THE PROPOSED METHOD

The flow of the present method is illustrated in figure 2. The images of the DNA microarray samples under analysis are elaborated through the platform we developed in [7]. It performs the gridding and the segmentation of the several DNA microarray images. Two new modules have been developed: the segmentation analysis and the cluster reliability estimation.

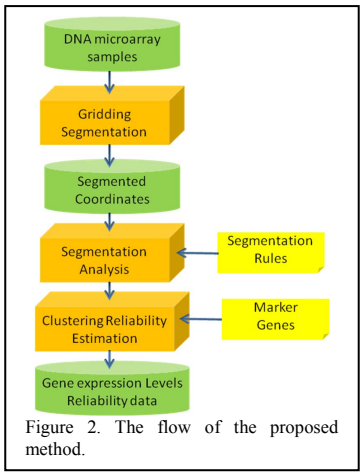

Figure 2. The flow of the proposed method.

A database containing the coordinates of the segmentation of each image is generated. The coordinates are then used in order to perform a segmentation analysis of the *region* and *signal/noise* effects. For this purpose a set of segmentation rules have been defined in order to classify each considered DNA fragment. Finally, a hierarchical clustering algorithm is performed according to the marker genes of the considered histological analysis, this algorithm is able to consider the cross-hybridization and the scratches effects. It generates the classification of the microarray samples and a list of selected gene expression levels, each one referred to a single DNA microarray, reporting the correspondent level of quality and reliability. The method generates two parameter for each marker gene: the *quality* and the *reliability* coefficients. The former indicates the regularity of the DNA fragments in terms of shape and morphology, while the second indicates the influence of that gene in the final gene expression classification.

A. *Fragments Segmentation Rules*

The Fragments Segmentation Rules are a set of typical shape used to compare each analyze spot. The set consists of:

1. *Regular spot, centered without noise*: the segmentation analysis compares the shape of the spot with a circular region.
2. *Irregular spot, centered, without noise*: the irregular morphology of the spot does not affect the results, since the analysis is performed considering only the spot's intensity value.
3. *Regular spot, not centered, without noise*: the spot segmentation is incorrect. The spot region must be moved in the corrected position in order to have the correct result.
4. *Regular spot, centered, with noise*: the existence of noise represents the major drawback of the segmentation algorithm. In this case the spot must be flagged as not correctly analyzable.

B. *Clustering Reliability Estimation Algorithm (CREA)*

The CREA algorithm works on the basis of the final segmentation and on the marker genes used for the analysis. Starting from the marker genes, it performs a hierarchical clustering with a supervised approach that uses the phenotypic information associated to each microarray sample.

The algorithm executes the following steps: segmentation selection, prior clustering and classification. The first step consists in analyzing the spots flagged by the segmentation analysis and identifying their redundant elements within the microarray. In the case the flagged spots belong to the set of marker genes, they are included in the reliability metric. The second phase executes the clustering of the samples basing on the marker genes and their redundant elements, this is considered as the *prior cluster*. The third phase creates all the possible combination of clusters and computes the reliability parameter for each gene that is expressed as the ratio between the number of clusters equal to the *prior cluster* and the total number of combination generated. By this way, the reliability parameter indicates the percentage of influence on the classification of a selected gene. Lower is the percentage of the reliability parameter, major is the probability that an error on the considered gene affects the classification.

The results of the algorithm is a dendogram diagram and a gene expression quality and reliability estimation related to the uncorrected identified genes. By exploring the generated dendogram, it is possible to identify the classification's group of the considered DNA microarray samples, while considering the gene list is possible to evaluate the influence of that genes on the classification.

IV. EXPERIMENTAL RESULTS

We validated the proposed method on a set of real data, related to the Lymphoma disease. We analyzed 16 samples of two different histological cases related to the *Follicular Lymphoma* (samples 14 – 16) and *Diffuse Large B-Cell Lymphoma* (6 – 13) and related to a sane tissue, *Normal Tissue* (1 – 5). The samples are available from the Stanford Microarray Database [8].

We performed two analysis. The first analysis has been executed on the original samples in order to identify the quality and reliability estimation. We achieved the results reported in the Figure 3.a and in the Table 1, the sample 13 results not correctly classified since it has been classified erroneously as Follicular Lymphoma. The table 1 shows the list of quality and reliability gene's characteristics computed by the CREA algorithm. Four marker genes related to the sample 13 has been identified and two of them (Oncogen ETV6 and CD22) has a lower level of reliability.

TABLE I. GENE EXPRESSIONS QUALITY AND RELIABILITY ESTIMATION

Sample #ID	Quality and Reliability Measures		
	Marker Gene Identifier	***Quality***	***Reliability***
13	Oncogen ETV6	4	12%
13	CD 22	2	5%
13	Transcriptor TFAP4	4	98%
13	PAK1	4	94%

We removed for microarray redundant gene list of the sample 13 the erroneously identified genes ETV6 and the CD22 genes and we performed a second analysis. As results we obtained the right classification of the three histological samples, as reported in figure 3.b.

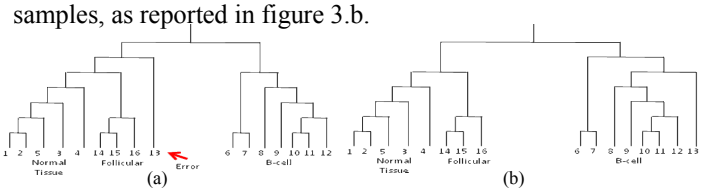

Figure 3. The figure reports the dendogram clustering diagram obtained (a) without applying the quality and reliability results, the sample #13 (B-cell Lymphoma) is erroneously classified with the Follicular Lymphoma, and (b) avoiding the false positive DNA fragments reported by the proposed method, the sample #13 is correctly classified as B-cell Lymphoma.

V. CONCLUSION

In this paper we presented a new method to estimate the quality and the reliability of gene expression analysis performed on DNA microarray. The experimental results performed on a real genomic set of DNA microarray images confirm the validity of the developed method.

REFERENCES

[1] Amos Mosseri and Eitan Hirsh, "Analysis of Gene Expression Data", Lecture 3, Tel Aviv University, 2005.

[2] D.A. Morrison and J.T. Ellis, "The design and analysis of microarray experiments: applications in parasitology", DNA Cell Biology, 22: pp 357 – 394, 2003.

[3] C.A. Glasbey and P. Ghaazl, "Combinatorial image analysis of DNA microarray features", Bioinformatics 19: 194 – 203, 2003.

[4] M. Bakay, Y. W. Chen, R. Borup, P. Zhao, K. Nagaraju, and E. P. Hoffman, "Sources of variability and effect of experimental approach on expression profiling data interpretation", BMC Bioinformatics 3: 4, 2002.

[5] K. Blekas, N. P. Galatsanos, A. Likas, and I.E. Lagaris, "Mixture Model Analysis of DNA Microarray Images", IEEE Transactions on Medical Imaging, Vol. 24, No. 7, July 2005.

[6] D. Huang, O. Milenkovic, "Superimposed coding for iterative detection of DNA microarray spot failures", IEEE International Workshop on Genomic Signal Processing and Statistics, pp. 1 – 4, 2008.

[7] L. Sterpone, M. Violante, "A new FPGA-based edge detection system for the gridding of DNA microarray images", IEEE Instrumentation amd Measurement Technology Conference, pp. 1 6, 2007.

[8] Stanford University, "Stanford Microarray Database", Available: http://smd.stanford.edu/