IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 13, 2020, accepted August 6, 2020, date of publication August 27, 2020, date of current version September 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3019844

Microarray Image Analysis: From Image Processing Methods to Gene Expression Levels Estimation

BOGDAN BELEAN[1], ROBERT GUTT[1], CARMEN COSTEA[2], AND OVIDIU BALACESCU[3]

[1]Center of Advanced Research and Technologies for Alternative Energies, National Institute for Research and Development of Isotopic and Molecular Technologies, 400293 Cluj-Napoca, Romania
[2]Department of Mathematics, Faculty of Automation and Computer Science, Technical University of Cluj-Napoca, 400114 Cluj-Napoca, Romania
[3]Department of Genetics, Genomics and Experimental Pathology, The Oncology Institute, Prof. Dr. Ion Chiricuta, 400015 Cluj-Napoca, Romania

Corresponding author: Bogdan Belean (bogdan.belean@itim-cj.ro)

This work was supported in part by the National Authority for Scientific Research and Innovation (ANCSI) through the Core Program under Project PN19 35 01 01, in part by the Advanced Researches and Applications for Alternative Energies and Hi-Tech Engineering under Grant 36N/2019, and in part by the Ministry of Education and Research, Romania, under Project 32PFE/19.10.2018.

ABSTRACT Microarray image processing leads to the characterization of gene expression levels simultaneously, for all cellular transcripts (mRNAs) in a single experiment. The calculation of expression levels for each microarray spot/gene is a crucial step to extract valuable information. By measuring the mRNA levels for the whole genome, the microarray experiments are capable to study functionality, pathological phenotype, and response of cells to a pharmaceutical treatment. The processing of the extensive number of non-homogeneous data contained in microarray images is still a challenge. We propose a density based spatial clustering procedure driven by a level-set approach for microarray spot segmentation together with a complete set of quality measures used to evaluate the proposed method compared with existing approaches for gene expression levels estimation. The set of quality measures used for evaluation include: regression ratios, intensity ratios, mean absolute error, coefficient of variation and fold change factor. We applied the proposed image processing pipeline to a set of microarray images and compared our results with the ones delivered by Genepix, using the aforementioned quality measures. The advantage of our proposed method is highlighted by a selection of up-regulated genes that had been identified exclusively by our approach. These genes prove to add valuable information regarding the biological mechanism activated as a response of Arabidopsis T to pathogen infection.

INDEX TERMS Gene expression, level-set segmentation, clustering, haustorium formation.

I. INTRODUCTION

Deciphering the whole human genome following the completion of the Human Genome Project in April 2003, has led to a fundamental transformation of molecular biology assessments, including a change in the concepts of research and the requirement of high fidelity and increased capabilities of the supporting technologies. The new concept of functional genomics attempts to describe the dynamic aspects of cellular functionality from a holistic perspective. Nowadays, the interrogation of genomic functionality relies on microarray technology to assess the gene expression levels simultaneously for all cellular transcripts (mRNAs) in a single experiment. By measuring the mRNA levels for the whole genome, the microarray experiments are capable to study functionality, pathological phenotype, and response of cells to a pharmaceutical treatment [1]. The workflow of a microarray experiment includes, besides the procedure of measurement, a step of extensive data analysis. Standardized protocols and design methods exist for measurements, but the processing of the extensive number of non-homogeneous data is often still a challenge. Further on, both the classical flow of microarray image processing and the quality measures used for evaluation are described.

In microarray experiments, RNA extracted from biological sample is synthesized to microarray targets. The targets are either single-stranded DNAs or RNAs, representing specific genes, labeled with fluorescent markers. One or two labels

The associate editor coordinating the review of this manuscript and approving it for publication was Zhigao Zheng.

(e.g. the dyes Cy3, and Cy5) can be utilized in the same hybridization measurement. The microarray targets hybridize on a microarray slide with sub-sequences (probes) of the genes within the whole genome, each gene being associated with a fluorescent spot. A laser scanning with appropriate wavelengths produces a TIFF images for each fluorescent label. Typically, in a two color microarray experiment, a microarray image called probe, associated with one fluorescent label, is compared with a reference image, recorded with the other fluorescent label. The expression levels of genes are calculated based on intensities of the fluorescent light, using specific microarray image processing techniques [2]. The image processing techniques are classified as: (1) preprocessing, to correct image rotation and to enhance weakly expressed spots; (2) grid alignment, to determine the location for each microarray spot; (3) segmentation, to perform pixel classification (i.e. the determination of the pixels belonging either to the microarry spot or to its local background) and spot intensity features extraction (4) data normalization for gene expression levels estimation.

A great deal of research has been conducted for developing image processing techniques for microarray spot intensity extraction. Contrast enhancement techniques were proposed for enhancing the visibility of microarray spots [3]. As reported in [4], [5], automatic grid alignment is performed using a SVM based approach, whereas based on a spot selection step, a set of grid lines are placed over the image in order to separate each pair of consecutive rows and columns. Unsupervised grid alignment methods for microarray images have been also proposed [6], [7] based on the use of optimal multilevel thresholding followed by a refinement procedure to find the positions of the sub-grids in the image and the positions of the spots in each detected sub-grid. For the segmentation of microarray spots, adaptive pixel clustering techniques were used in [8]–[10]. Alternate spatial methods, such as the snake fisher model or 3D spot modeling were used for spot segmentation in [11] and [12], respectively. Combining observed intensity and spatial information, spot segmentation is performed in [13] and [14] using Markov random field modeling. Background and foreground pixel classification is also achieved by using the growing concentric hexagon algorithm [15]. Considering the increased size of microarray images together with the large variety of image processing technique for spot intensity features extraction, state of the art research proposes also microarray image compression approaches [16] and parallel implementations for spots segmentation [17].

Image processing achievements within the state of the art approaches lack the interpretation of the results accuracy improvements from the point of view of biological significance. In a quantitative comparison study, the biologists track the relative changes in intensities for the same spot from the sample and reference image. Significant relative changes represent the differentially expressed genes. Consequently, the paper proposes a density based spatial clustering procedure driven by a level-set procedure for spot segmentation and intensity extraction, together with a complete set of specific quality measures (e.g. coefficient of determination, regression ratios, intensity ratios, mean absolute error, coefficient of variation) for evaluating the proposed image processing workflow. Once the intensity extraction is performed and validated, a normalization procedure is applied to correct for intensity-depended patterns in the spot intensities distribution. Further on, the log odd ratios for each microarray spot (i.e. the fold change factor Fc) are computed and the differentially expressed genes are estimated and compared to the ones detected by existing software platforms.

II. MICROARRAY IMAGE PROCESSING METHODS

Typically, the microarray images are stored in the Tagged Image File Format (TIFF) as a two-dimensional array of intensities. In a two colour microarray experiment, two microarray images are available, each image being recorded from a specific cyanine dye. The images are denoted by I_{Cy3} and I_{Cy5}, corresponding to Cy3 and Cy5 dyes, respectively. Figure 2 shows an example of a microarray image, corresponding to the microarray experiment sample identified as GSM333341 from the Gene Expression Omnibus data repository [18] (https://www.ncbi.nlm.nih.gov/geo). Microarray images represent a collection of microarray spots arranged in one or more sub-grids, each grid representing a two dimensional array of spots. Image processing technique are used further on in order to determine spot location within each sub-grid, spot sizes, spot intensities and background intensities values which are typically delivered as raw data parameters for microarray image analysis and interpretation. Our proposed image processing techniques for automatic microarray image processing were implemented in Matlab and they are presented further on.

A. PREPROCESSING

The microarray image preprocessing techniques were used for image enhancement, rotation correction and sub-grid detection. Let the bi-dimensional array of intensities $I = (p_{i,j})$ represent the input microarray image, stored as a 16 x bits gray-scale TIFF image. A novel approach to enhance the weakly expressed spot is proposed, based on a point-wise hyperbolic tangent transformation $r_{i,j} = tanh(p_{i,j})$. According to tangent hyperbolic function representation, the luminance information $p_{i,j}$ is mapped into $r_{i,j}$ values, with respect to a threshold k_{thr}. The threshold is estimated based on the microarray background image $B = (b_{i,j})$, constructed using morphological opening on the input microarray image I.

Consequently, the intensity values $p_{i,j}$ bellow the k_{thr} threshold are mapped according to the *tanh* function representation on its negative domain values, whereas the intensity values $p_{i,j}$ over k_{thr} are mapped according to the representation on its positive domain values. The resulted $r_{i,j}$ values are normalized in such manner that their histogram fits the n - bits dynamic range of the original microarray image. In Figure 1 the effect of the enhancement procedure for weekly expressed spots is shown in case of the microarray

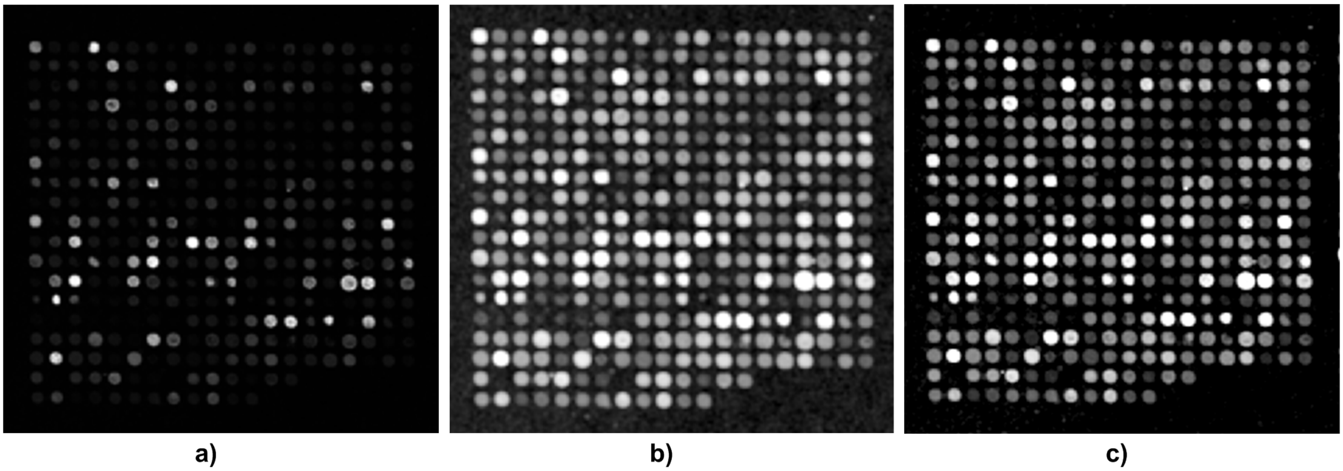

FIGURE 1. a) Group of spot from the original ExpID GSM333341 microarray image, b) the resulted image after applying the logarithm based enhancement, c) the resulted image after applying the proposed tangent hyperbolic based enhancement.

image sample GSM 333341 recorded using Cy3 fluorescent label. The microarray spots enhanced by classic logarithmic transformation are presented in Fig. 1a, whereas Fig. 1b shows the resulted image after tangent hyperbolic transform is applied. It can be seen that the background information, which may contain artifacts, is selectively not enhanced by our proposed point-wise approach for microarray spot enhancement (see Figure 1.b). Accidental microarray image rotation is detected and corrected using Radon transform, as reported in [6]. As described in [7], spot group detection is performed using an approach based on mathematical morphology, which identifies the groups of spots using an image closing procedure. The aforementioned preprocessing techniques transformed the original image I into several I_p images, each containing one group of spots.

B. GRID ALIGNMENT

For automatic grid alignment, we applied independently the support vector machine (SVM) based approach proposed by Bariamis *et al.* [4] to each preprocessed I_p image. The method consists of estimating the distance between consecutive rows and columns followed by a spot detection step, which is used to generate x_i vectors and their respective class, with i from 1 to the total number of selected spots S. The x_i vector of class -1 and x_i vector of class 1 are used to mark the area between two consecutive rows or columns, (see Fig. 2). Further on, SVM classifiers are used; the basic principle of an SVM classifier is that it produces the normal vectors w_i, which maximize the margin between x_i vectors of different classes. In this manner, SVM classifiers determine two sets of horizontal h_i and vertical v_j normal vectors, denoted by $H = \{h_i : i = \overline{1 \ldots M+1}\}$ and $V = \{v_j : j = \overline{1 \ldots N+1}\}$ respectively, with M and N representing the number of microarray spot per rows and columns, respectively. The vectors magnitudes $|h_i|$ and $|v_j|$ represent the positions of each horizontal and vertical grid lines, on the abscissa or ordinate axes, respectively.

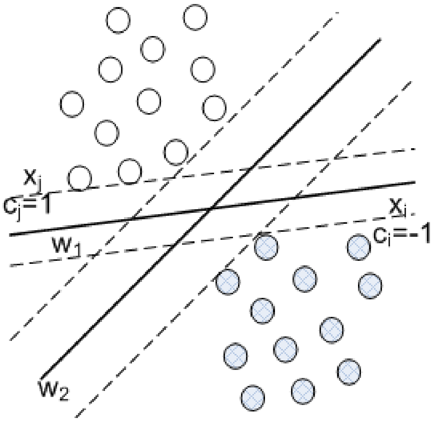

FIGURE 2. The determination of the normal vectors w_i which maximize the spatial separation between the vectors x_i of two different classes; the determined vectors correspond to the grid lines positions used for the separation of consecutive lines and columns of the microarray spots.

Consequently, the two sets determine the overall grid within the microarray image.

C. SEGMENTATION

The segmentation is performed in three steps: (1) selection of the rectangular area for each spot, (2) initial pixels classification into foreground and background based on level-set approach, and (3) a refinement of the initial classification using density based spatial clustering.

In the first step, based on the grid alignment procedure, a rectangular area R_{spot} is associated to each microarray spot from both I_{Cy3} and I_{Cy5} images. For each rectangular area R_{spot}, the foreground represents pixels corresponding to the microarray spot, the background corresponds to the local background, whereas the exclusion zone represents a selection of pixels which correspond neither to foreground or background.

The second step, inspired by Agilent "cookie cuter" approach, consists of the classification of pixels within each rectangular area R_{spot} into foreground and background. For each rectangular area R_{spot} determined by grid alignment, a level set method is used to represent the contour of the microarray spot C_S, as the zero level set of the level set function (LSF) denoted by $\phi(x, y, t)$, as expressed by equation:

$$C_S = \{(x, y) : \phi(x, y, t) = 0\}. \tag{1}$$

The determination of the contour C_S is converted into finding solution of the partial differential equation (PDE) from (2), which is referred to as the level set evolution equation [19], were F is the speed function that controls the motion of the curve:

$$\frac{\partial \phi}{\partial t} = F|\nabla \phi|. \tag{2}$$

In image segmentation applications, is required that the LSF is smooth and accurate especially in the vicinity of its zero level set, where it describes the contour of the object to be determined. For maintaining this property of the LSF without the need of re-initialization, Li *et al.* proposed an energy formulation with distance regularization for the level set evolution in [20]. The energy functional is defined by $\varepsilon(\phi) = \mu R(\phi) + \varepsilon_{ext}(\phi)$, where $\mu > 0$ is a constant, $R(\phi) = \int_{R_{spot}} \frac{1}{2}(|\nabla\phi| - 1)^2|\nabla\phi|dxdy$ is the distance regularization term and $\varepsilon_{ext}(\phi)$ is an external energy. The minimization of $\varepsilon(\phi)$ energy is achieved by solving the following PDE equation:

$$\frac{\partial \phi}{\partial t} = \mu div\left(\frac{|\nabla\phi| - 1}{|\nabla\phi|}\nabla\phi\right) - \frac{\partial \varepsilon_{ext}}{\partial \phi}. \tag{3}$$

In order for the level set to be applied for edge detection in image processing, the ε_{ext} energy was chosen to describe edge-based information according to [20]. Let I_S be an image on a domain R_{spot}, corresponding to the rectangular area which confines and let $\phi : R_{spot} \times [0, \infty) \to \mathbb{R}$ be a LSF defined on the domain R_{spot}. The edge indicator function is defined as $g = \frac{1}{1+|\nabla G_\sigma * I_S|^2}$, where G_σ is a Gaussian kernel with standard deviation σ. Consequently, the external energy is given by $\varepsilon_{ext} = \lambda L(\phi) + \alpha A(\phi)$. The L and A energy functional are defined as $L(\phi) = \int_{R_{spot}} g\delta(\phi)|\nabla\phi|dxdy$ and $A(\phi) = \int_{R_{spot}} gH(-\phi)dxdy$, where δ and H are smooth approximations of the Dirac delta function and the Heaviside function, respectively. Considering the definition of the energy functional, the PDE equation (3) becomes:

$$\frac{\partial \phi}{\partial t} = \mu div\left(\frac{|\nabla\phi| - 1}{|\nabla\phi|}\nabla\phi\right) + \lambda\delta(\phi) div\left(g\frac{\nabla\phi}{|\nabla\phi|}\right) + \alpha g\delta(\phi) \tag{4}$$

The initial condition for the LSF is set up so the $\phi(x, y, 0)$ function represents the rectangular contour determined by the grid alignment which confines each microarray spot. Solving the equation (4), the resulted LSF function $\phi(x, y, t)$, where t is the number of iterations, represents the edge which separate the microarray spot from its local background. For each microarray R_{spot}, two curves C_{Cy3} and C_{Cy5} are determined by the zero level set of ϕ_{Cy3} and ϕ_{Cy5}, which are recorded by Cy3 and Cy5 respectively. The union of the two sets inside the curves ϕ_{Cy3} and ϕ_{Cy5} is denoted by F and yields the pixels positions considered as foreground in both Cy3 and Cy5 images. For the entire group of spots within the ExpID GSM333341, we present in Fig 3.b the zero level set of ϕ_{Cy3} determined by the resulted LSF function for each microarray spot. A selection of microarray spots which contains spots with irregular shape, spots with inner wholes and spots of regular round shape is presented in Fig. 3.c. The resulted LSF is also represented for each microarray spot, showing its convergence towards the spot edge. In can be seen that the proposed approach accounts for spots with various shapes and sizes.

In the third step, a refinement procedure is introduced, which aims to perform accurate segmentation in case of microarray spots with inner wholes. The pixels included in the rectangular area R_{spot} of the microarray images I_{Cy3} and I_{Cy5}, together with the set of foreground pixels F constitute the premises of the current procedure. Considering the closed curves determined by the zero levels of ϕ_{Cy3} and ϕ_{Cy5}, the exclusion zones E_{Cy3} and E_{Cy5} are determined by the set of pixels which fall within the two-pixel exclusion zone located around the curve in the outer side (see Fig 4). The refinement is performed using density based spatial clustering [21] applied on the two sets of pixels intensity values defined as $F \cup E_{Cy3}$ and $F \cup E_{Cy5}$, corresponding to I_{Cy3} and I_{Cy5} images, respectively.

Note that, in case of weakly expressed spots or spots with inner wholes, besides the foreground and background intensity, a third category of pixel intensity values is observed. While the higher values correspond to the microarray spot and the values close to 0 correspond to the background, the inner wholes of the spot are characterized by pixel intensity values significantly lower than the spot foreground, but higher than the background. The proposed procedure assigns the intensity of a pixel to one of the following groups: spot foreground (high value), local background (low value) or exclusion zone in case of spots with inner wholes. In this manner, two set of pixels denoted by F_{Cy3} and F_{Cy5}, corresponding to I_{Cy3} and I_{Cy5} images, respectively, are obtained for the same microarray spot. Each set is defined as the pairs of pixel indexes (i, j) relative to the microarray image I, with pixel intensity value $p(i, j)$ assigned by the clustering procedure to the foreground pixels group (high pixel intensity values). The intersection of the two sets defined as $S = F_{Cy3} \cap F_{Cy5}$ contains pixels that are called foreground of the spot (i, j) in both I_{Cy3} and I_{Cy5} images. The set S represents the segmentation mask of a given spot s, applied on both Cy3 and Cy5 images in order to determine the pixels of spot s accounted for median spot intensity computation. The local background estimation in case of a microarray spot from Cy3 dye is denoted by B_{Cy3} and computed as the median intensity over pixels located at position $R_{spot} - F$. In a similar manner, $B_{Cy5} = median(R_{spot} - F)$ denotes the local background in case of spots recorded from Cy5 dye. The median intensities of each spot are computed over the set of pixels S from the Cy3 image

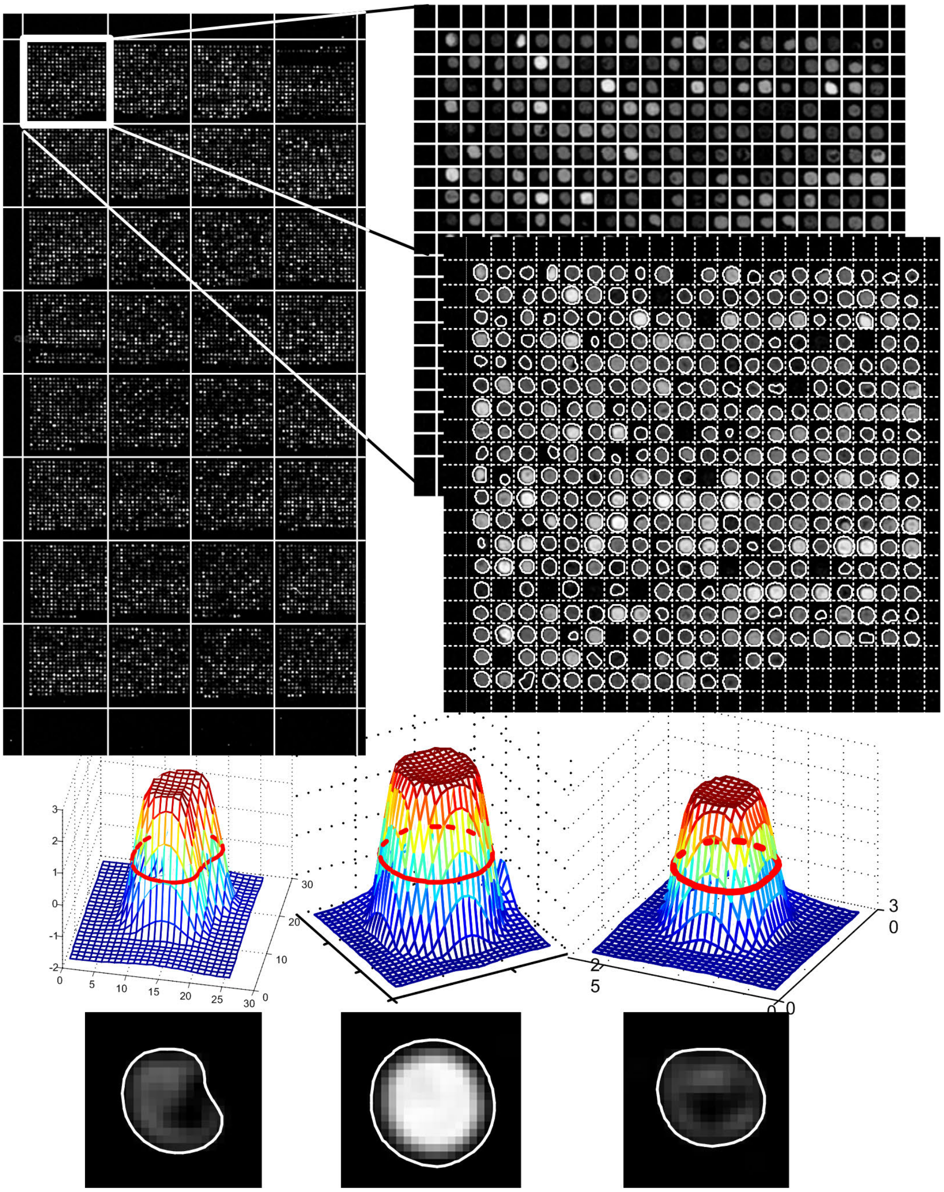

FIGURE 3. a) Localization of each spot group using mathematical morphology in case of ExpID GSM333341 microarray image, b) microarray spot segmentation using the level set approach in case of the selected spot group from a), the results of the level set segmentation on a selection of spots with the indexes 85, 25 and 118 from the spot group from figure b).

and the Cy5 image, and they are denoted by S_{Cy3} and S_{Cy5}, respectively.

D. NORMALIZATION

For the microarray spot from the I_{Cy3} image, the background corrected median intensity is given by the difference $R = S_{Cy3} - B_{Cy3}$. For a quantitative comparison study, the background corrected intensity R has to be compared with the background corrected intensity, $G = S_{Cy5} - B_{Cy5}$, of the spot at identical location in the reference image I_{Cy5}. We computed the R and G values for each of the spots in our test data set. To correct for intensity-depended patterns in the (R, G) data of a microarray, we applied the standard scatter plot smoother ''lowess'' of Cleveland and Devlin [22] with linear fit and window size of 20%, which yield the new normalized (R, G^n) values.

III. EVALUATION

The proposed image processing techniques are evaluated by means of reproducibility and biological significance. Aiming to quantify the reproducibility of the proposed segmentation techniques the following quality measures are computed: the mean absolute error (MAE) and the coefficient of variation (CV), which indicates the sameness of the spot intensities and

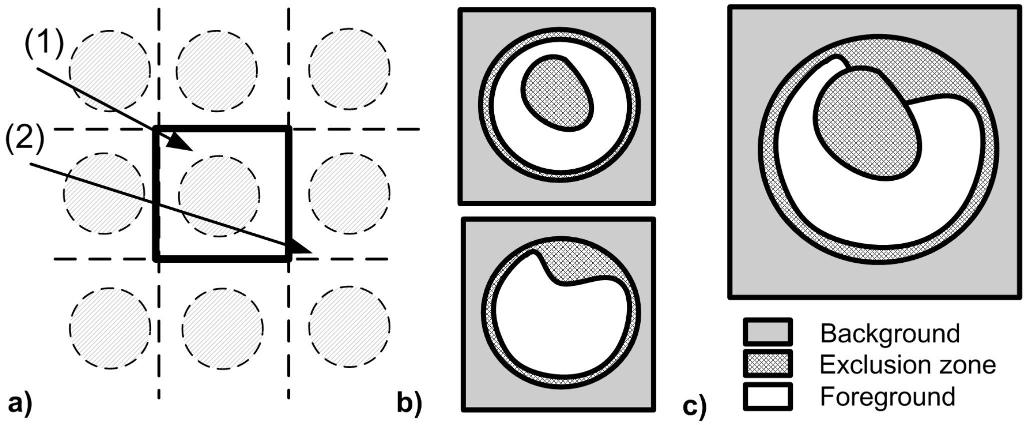

FIGURE 4. a) Schematic localization of the spot considering its neighbors, b) level set segmentation of the same spot for both the I_{Cy3} and I_{Cy5} images (upper and lower images), c) refinement of the level-set segmentation considering the background, foreground and exclusion zone.

the variation of spots intensities, respectively. The MAE is given by

$$MAE_{spot} = \frac{1}{k}\sum_{i=1}^{k}|G^n{}_i - \overline{G}|, \tag{5}$$

where k is the number of replicates, $G^n{}_i$ is the normalized mean spot intensity value and $\overline{G}$ is the spot overall mean considering the means of the corresponding spots within the n replicates. The CV is given by

$$CV_{spot} = \frac{\sigma}{\mu}, \tag{6}$$

where σ represents the standard deviation of spot intensity with subtracted background and μ denotes the mean spot intensity.

In order to validate the proposed methodology for gene expression estimation, the following comparisons are performed. The median intensity values for each spot, R and G^n, determined by our approach are compared with the ones delivered by GenePix platform by means of Pearson correlation. Conventionally, the ratio $r = \frac{R}{G^n}$ measures the relative change in gene expression expressed by a microarray spot. Additionally, the change of gene expression can be measured for a spot also by a regression ratio rR. The regression ratio rR is the slope of the linear fit through a scatter plot. The scatter plot has a point (r, g) for each pixel inside the foreground area of a spot. The values r and g are the raw intensities of the foreground pixels in the images I_{Cy3} and I_{Cy5}, respectively. Both r and rR coefficients determined by our proposed method for the whole population of spots are correlated with the values delivered by GenePix for the same parameters in case of microarray experiments under analysis. The biological significance of the results delivered by the proposed image processing workflow resides in the differentially expressed genes resulted from the microarray experiments under analysis. The selection of differentially expressed genes is done using the fold change value Fc [23], which is given by the log odd ratio off the spot intensity from the two microarray images I_{Cy3} and I_{Cy5}, sample and reference image (i.e., $log(\frac{R}{G^n})$). The differentially expressed genes are estimated using both our approach and GenePix, in order to underline the advantages of our proposed techniques.

IV. RESULTS

The data set consists of 8 images corresponding to the microarray samples having the following IDs: GSM333336, GSM333353, GSM333337 and GSM333341. Each microarray image has the size of 4000 × 1944 pixels and contains 32 spot groups with 380 spots per group. The set of images is organized in 4 pairs of (I_{Cy3}, I_{Cy5}) and represents 4 different microarray samples from the experiment presented in [24]. Features extraction is performed using our proposed approach for the entire population of spots included in the selected microarray image samples. The obtained results are compared with the ones delivered by GenePix Software in terms of reproducibility and reliability. The differences between our results and the results delivered by GenePix may reveal significant information from a biological point of view.

A. REPRODUCIBILITY AND RELIABILITY OF THE PROPOSED SPOT SEGMENTATION APPROACH

Our proposed level-set segmentation approach was applied on the previously mentioned image samples. The spot intensities R and G^n together with intensity ratios r, regression coefficients rR and fold-change factor Fc are computed for the whole population of spots included in each of the microarray images under analysis. Values determined by GenePix (e.g. in rR_{GP} the lower indices denotes that the rR parameter values are computed by GenePix) for the same sets of spots are compared with our results by means of Pearson correlation. Table 1 lists the Pearson coefficients showing the correlation between GenePix results and ours, considering the aforementioned parameters.

TABLE 1. The Table lists the Pearson coefficients resulted by comparing our computed parameters values: spot intensities *R*, *Gn*, *Fc* ratio and the regression ratio *rR* and the values delivered by GenePix for the whole population of spots (12160 counts) within the microarray experiments under analysis.

Exp. ID	(R, R_{GP})	(G^n, G^n_{GP})	(Fc, Fc_{GP})	(rR, rR_{GP})
GSM333336	0.951	0.942	0.891	0.924
GSM333353	0.959	0.946	0.873	0.912
GSM333337	0.948	0.950	0.804	0.935
GSM333341	0.963	0.938	0.922	0.940

Considering the increased Pearson correlation coefficients computed for Fc and rR values over the entire data set, the proposed approach was proven to deliver similar results as compared with GenePix. The two pairs of microarray experiments $E1 = (GSM333336, GSM333353)$ and $E2 = (GSM333337, GSM333341)$ represent each biological replicates of the same experiment. The $E1$ and $E2$ sample pairs represent gene expression profiles of infected and uninfected Arabidopsis leaves, respectively. The reproducibility of our proposed segmentation approach was evaluated by means of spot sameness (MAE) and spot variation (CV) in case of both $E1$ and $E2$ samples. The lower the MAE and CV values are the better is the performance of the proposed method. The MAE and CV values were computed both for our proposed segmentation approach and GenePix approach. In case of our level-set segmentation approach, two values were computed for each MAE and CV parameter. One corresponds to the segmentation procedure composed of steps 1 and 2 and is denoted by lower indices 1-2 (e.g. MAE_{1-2}), whereas the other values corresponds to the segmentation procedure composed of steps 1, 2 and 3 (e.g. MAE_{1-3}). It was intended to show the difference in accuracy in case the refinement procedure (i.e. segmentation step 3) is included within the segmentation or not. For the E1 experiment the following parameter values were obtained: the average value of the MAE_{1-2} coefficient was 563, the average value for MAE_{1-3} coefficient was 476, whereas the average MAE_{GP} obtained by GenePix Pro software was 551. Moreover, for the experiment E2, the average values for the MAE_{1-2}, MAE_{1-3} and MAE_{GP} were 541, 408 and 472, respectively. The level-set segmentation procedure shows similar results as compared with GenePix approach, with respect to the MAE and CV coefficient. Nevertheless, the refinement procedure, introduced as the third step of the segmentation process, showed improvement compared to the GenePix segmentation approach. The MAE_{1-3} and CV_{1-3} coefficients were, in average values, 70 and 0.242 units lower, respectively. The resulted MAE coefficients are illustrated using box plots as shown in Figure 4, whereas Table 2 shows the CV values. The CV represents a standardized measure of dispersion in case of spot pixel intensity values, independent of the unit in which the measurement has been taken. A small CV corresponds to a small variation among the pixel intensity values for a given microarray spot. Thus CV can be used as a quality measure for the spot segmentation process.

TABLE 2. Coefficient of variation obtained using both the proposed image processing workflow (CV_{1-2} and CV_{1-3}) and GenePix tools (CV_{GP}).

Experiment ID (Image channel)	Mean CV (CV_{1-2})	Mean CV (CV_{1-3})	Mean CV $(CV_{GP}.)$
GSM333336 (ICy3)	0.686	0. 517	0.663
GSM333336 (ICy5)	0.522	0.426	0.534
GSM333353 (ICy3)	0.674	0.395	0.635
GSM333353 (ICy5)	0.642	0.446	0.706
GSM333337 (ICy3)	0.667	0.398	0.684
GSM333337 (ICy5)	0.751	0.344	0.766
GSM333341 (ICy3)	0.618	0.451	0.552
GSM333341 (ICy5)	0.82	0.422	0.795

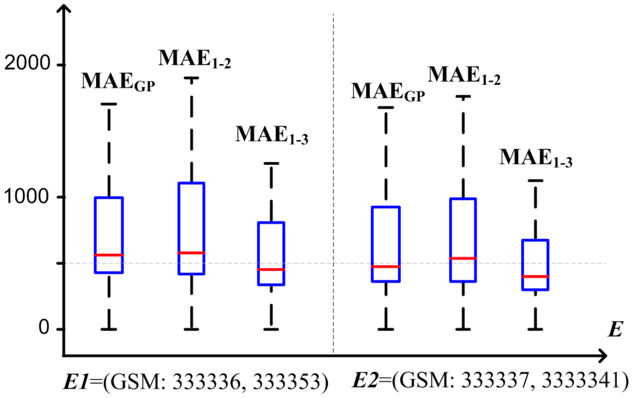

FIGURE 5. MAE plots for the whole population of spots from E1 and E2 experiments, in case of three segmentation approaches: our level-set approach steps (1) and (2), our level-set approach steps (1), (2) and (3) and the GenePix approach.

The increased correlation coefficients between our results and the ones delivered by GenePix (Table 1) together the improvement with respect the MAE and CV (Figure 5 and Table 2, respectively) of the level set segmentation showed the reliability and the reproducibility of our proposed method for expression level estimation.

B. BIOLOGICAL SIGNIFICANCE OF THE RESULTS

Once the intensity values for each microarray spot are estimated, the genes differentially expressed within a microarray experiment are determined and they are known as up/down regulated genes. The biologists and medical doctors are interested in the interpretation of the relative changes in intensities for the same spot from the sample and reference image, I_{Cy3} and I_{Cy5}, respectively. The selection of differentially expressed genes is done using the fold change value $Fc = log2\frac{R}{G^n}$, which is given by the log odd ratio off the spot intensity from the two microarray images. In order to underline the contributions of the proposed method from biological point of view, we first describe the microarray image samples and then we estimate the differentially expressed genes using both our proposed method and using GenePix approach.

This experiment describes the changes in the global gene expression profiles of susceptible Arabidopsis leaves for supporting biotrophic parasitic plants. The developmental process associated with the plant parasitism is known as haustorium formation [25]. Haustoria represent multicellular invasive organs of parasitic plants which are able to attach and to penetrate the host tissue in order to acquire water and nutrients. During this interaction, the host accumulate defense genes transcripts. The major pathogen-inducible defense pathways are regulated by salicylic acid (SA) and jasmonic acid (JA), or by complex networks interconnecting these defense pathways [24]. The experiment under analysis tries to elucidates weather these defense pathways are either not-activated or inactivated by pathogen effectors. Markers such as JAR1, involved in the JA adenylation required for the JA function, are used in order to characterize the JA defense pathway. The arabidopsis jar1-1 mutant reduces the impact of the JA-dependent pathway. This pathway becomes activated only upon interactions with pathogen effectors such as G. cichoracearum. These premises make the jar1-1 mutant useful to reveal responses associated with JA-dependent pathway. According to the experiment conducted by [24], the GSM333336 and GSM333353 samples represent biological replicates of the uninfected Arabidopsis leaves compared to a common reference, whereas GSM333337 and GSM333341 samples represent biological replicates of infected leaves compared to the same reference. The findings of the aforementioned analysis conducted by Fabro *et al.* [24] show that the JA pathway leads to enhanced resistance to G. cichoracearum and also signals effective defenses against the fungi.

Further on, we aim to determine supplementary differentially expressed genes related to the plant defense response to pathogen infection as compared to differentially expressed genes determined by Fabro *et al.* [24]. For this purpose, a visual representation of the differences in intensities between the normalized median intensity values R and G^n is given trough the MA plot in figure 6. The log odd ratio of the fold change factor Fc is represented on the ordinate, whereas the abscissa illustrates the average log intensity denoted by $log_{10}(RG^n)$. Considering our proposed feature extraction approach for microarray spots, differentially expressed genes are determined using the fold change factor $Fc > 1.75$ for estimation. Using the same Fc factor, MA plots are shown for both our proposed method and GenePix approach in upper and lower images of figure 6, respectively.

A number of 119 genes were found as up-regulted by both the proposed spot segmentation approach and GenePix software. As referred to our approach, a set of 22 supplementary genes are found to be up-regulated as compare with the GenePix analysis. A sub-set of interest from the genes determined to be up-regulated exclusively by our proposed approach is presented in Table 3. The spots corresponding to each gene from Table 3 are illustrated in Fig. 7. These genes have different roles: the At4g11320 and gene At5g43060 (index 1379 and 10565 respectively) are involved in degenerative events that decrease metabolic activities and cause the death of cells [26], [27], the At1g08650 gene (index 2151) is involved in plant stress signaling, the At5g38430 (index 2583) involved in the process of converting the carbon dioxide into energy-rich molecules, whereas the AT3G09440 gene (index 9458) are part of family of heat-shock proteins produced by plant cells in response to stressful condition [28].

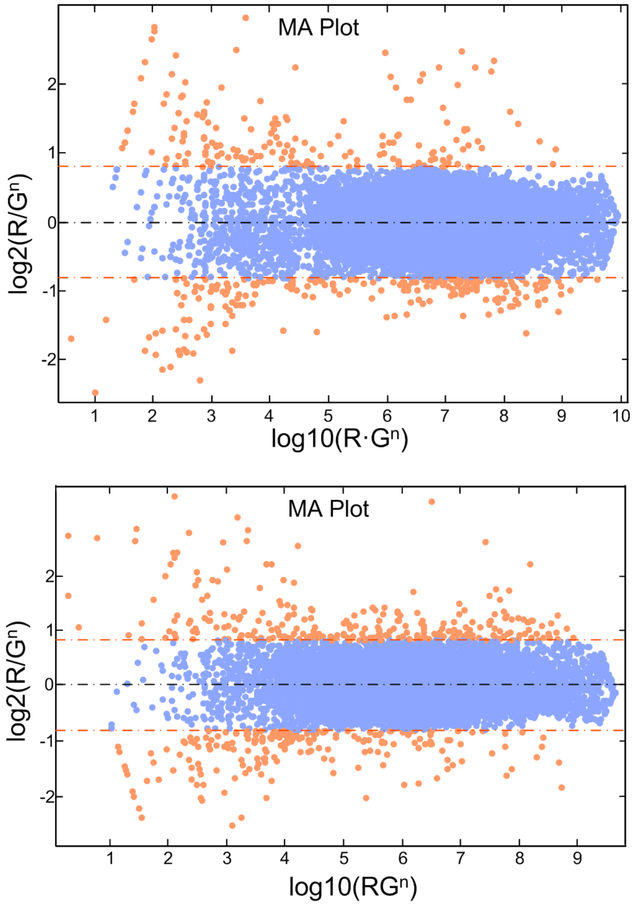

FIGURE 6. **MA plots illustrating the differences in intensities between R and G^n in case of GenePix approach (upper) and our proposed approach (lower) for gene expression levels estimation.**

The importance of these up-regulated genes in the context of the JA defense pathway is discussed next. JA signaling pathway has been extensively studied, having a key role in signaling effective resistance of Arabidopsis T to haustorium formation. A common defense feature, triggered by the JA pathway is the hypersensitive response HR, known as a form of programmed cell death occurring at the primary infection site. This local immune response limits the spread of the pathogen by reducing their access to nutrients [29]–[31]. As shown in [29], the up-regulated At4g11320 and At5g43060 genes mark the activity of the vacuolar protease RD21 which contributes to the vacuole rupture during plant hypersensitive response [29]. This rupture dramatically alters the cytoplasm by acidification and the release of enzymes acting as cell death mediators [32], [33]. Regarding the up-regulation of the AT3G09440, it is a well known plant response to various forms of stress, besides heat, leading also to the plant hypersensitive stress response and consequently to the induction of programmable cell death. According to [34], in spite of the intensive study of the JA signaling pathway, our knowledge regarding the JA signaling in plant–environment interaction is still not clear.

TABLE 3. The table list a selection of 5 supplementary expressed genes determined using our proposed approach as compared with the GenePix software, in case of Arabisopsis jar1-1 mutant supporting biotrophic parasitic plants, i.e., G. cichoracearum.

				GenePix		Determined values	
Spot index	Reporter Sequence	Description	Gene Model	I_{Cy3}	I_{Cy5}	I_{Cy3}	I_{Cy5}
1379	At4g11320	cysteine proteinase	F8L21.110	2102	881	4520	1038
2151	At1g08650	putative calcium-dependent protein kinase	F22O13.13	9189	6550	24264	14832
2583	At5g38430	ribulose bisphosphate carboxylase small chain	MXI10.15	31493	21125	44168	13970
9458	AT3G09440	heat-shock protein	F3L24.33	23614	14524	37195	18977
10565	At5g43060	cysteine proteinase RD21A	MMG4.7	17168	9622	19675	7162

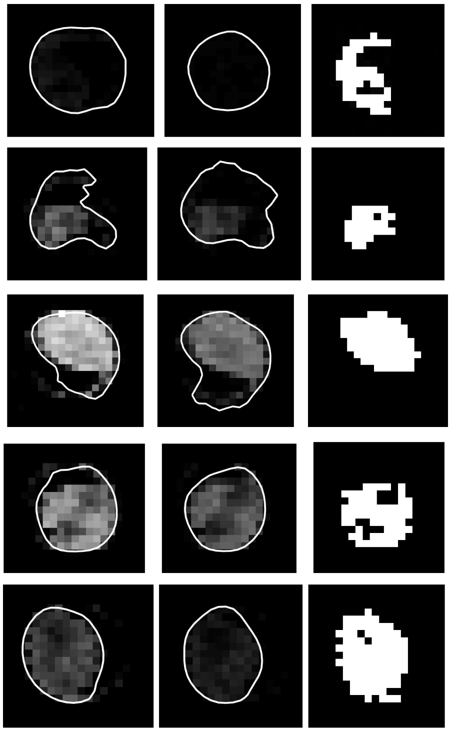

FIGURE 7. Spots taken from both I_{Cy3} (first column) and I_{Cy5} (second column) images corresponding to the up-regulated genes listed in Table 3, together with the contours marking the foreground pixels and the clustering masks (third column) used for the determination of median pixel intensity values for each spot.

Thus, the supplementary up-regulated genes underline the activation of different enzymes, showing new mechanism of action for the JA signaling.

V. CONCLUSION

We presented a complete image processing pipeline for the extraction of microarray spot intensity features. Based on a level set approach, a predefined contour is evolved leading to the separation of the foreground from background pixels for each spot. Considering spots with non-homogeneous intensity distribution and inner wholes, this initial classification of pixels yielded only a rough approximation that was insufficient to extract reliable values for all intensity features. To overcome this drawback, we introduce a density based spatial clustering procedure which determines segmentation masks for spots with non-homogeneous intensity distribution and inner wholes, underlining spot intensity values corresponding the spot on both Cy3 and Cy5 images. The proposed approach was tested on a set of microarray images. Furthermore, the results were compared with the ones delivered by GenePix for the same set of images. Similar spot intensity features were yielded for the majority of spots, as shown by the Pearson coefficients exceeding values of 0.94 and indicating strong correlation of our data (intensities) and GenePix reference data. Nevertheless, lower quality measures for spot sameness and spot variation showed an improvement of our segmentation approach as compared with GenePix solution. Moreover, based on the determined spot intensity features, up-regulated genes were determined for our data set. A supplementary number of genes were found as up-regulated by our approach that has not been reported in the reference data set. A visual inspection of the spots corresponding to the genes found as up-regulated by our proposed approach showed that our segmentation procedure fits much better to the true shape of the spot. As referred to the biological significance of the results, we identified supplementary important genes that complete the description of the plant–environment interaction mechanism activated in a defense response of Arabidopsis to pathogen infection.

AUTHORS' CONTRIBUTION

B.B, R.G., C.C. and O.B. have equally contributed to the design and implementation of the present research, to the analysis of the results and to the writing of the manuscript.

REFERENCES

[1] M. Schena, *Microarray Analysis*. New York, NY, USA: Wiley, 2003, p. 654.

[2] A. M. Campbell, W. T. Hatfield, and L. J. Heyer, “Make microarray data with known ratios,” *CBE—Life Sci. Edu.*, vol. 6, no. 3, pp. 196–197, Sep. 2007.

[3] T. Li, G. Shao, Y. Sun, and W. Shi, “Contrast enhancement for cDNA microarray image based on fourth-order moment,” *Signal, Image Video Process.*, vol. 12, no. 6, pp. 1069–1077, Sep. 2018.

[4] D. Bariamis, D. K. Iakovidis, and D. Maroulis, “M^{3}G: Maximum margin microarray gridding,” *BMC Bioinf.*, vol. 11, no. 1, p. 49, Dec. 2010.

[5] D. Bariamis, D. Maroulis, and D. K. Iakovidis, “Unsupervised SVM-based gridding for DNA microarray images,” *Computerized Med. Imag. Graph.*, vol. 34, no. 6, pp. 418–425, Sep. 2010.

[6] L. Rueda and I. Rezaeian, “A fully automatic gridding method for cDNA microarray images,” *BMC Bioinf.*, vol. 12, no. 1, pp. 1–17, Dec. 2011.

[7] Z. Gan, N. Zeng, F. Zou, J. Chen, M. Du, L. Liao, H. Li, and Y. Zhang, “Multilevel segmentation optimized by physical information for gridding of microarray images,” *IEEE Access*, vol. 7, pp. 32146–32153, 2019.

[8] J. Angulo and J. Serra, “Automatic analysis of DNA microarray images using mathematical morphology,” *Bioinformatics*, vol. 19, no. 5, pp. 553–562, Mar. 2003.

[9] G. Shao, D. Li, J. Zhang, J. Yang, and Y. Shangguan, “Automatic microarray image segmentation with clustering-based algorithms,” *PLoS ONE*, vol. 14, no. 1, Jan. 2019, Art. no. e0210075.

[10] N. Giannakeas and D. I. Fotiadis, "An automated method for gridding and clustering-based segmentation of cDNA microarray images," *Computerized Med. Imag. Graph.*, vol. 33, no. 1, pp. 40–49, Jan. 2009.

[11] J. Ho and W.-L. Hwang, "Automatic microarray spot segmentation using a snake-Fisher model," *IEEE Trans. Med. Imag.*, vol. 27, no. 6, pp. 847–857, Jun. 2008.

[12] E. Zacharia and D. E. Maroulis, "3-D spot modeling for automatic segmentation of cDNA microarray images," *IEEE Trans. Nanobiosci.*, vol. 9, no. 3, pp. 181–192, Sep. 2010.

[13] O. Demirkaya, M. H. Asyali, and M. M. Shoukri, "Segmentation of cDNA microarray spots using Markov random field modeling," *Bioinformatics*, vol. 21, no. 13, pp. 2994–3000, Jul. 2005.

[14] E. Athanasiadisa, D. Cavourasb, and S. Kostopoulos, D. Glotsos, I. Kalatzis, and G. Nikiforidis, "A wavelet-based Markov random field segmentation model in segmenting microarray experiments," *Comput. Methods Programs Biomed.*, vol. 104, pp. 307–315, 2011.

[15] N. Giannakeas, F. Kalatzis, M. G. Tsipouras, and D. I. Fotiadis, "Spot addressing for microarray images structured in hexagonal grids," *Comput. Methods Programs Biomed.*, vol. 106, no. 1, pp. 1–13, Apr. 2012.

[16] M. Hernandez-Cabronero, I. Blanes, A. J. Pinho, M. W. Marcellin, and J. Serra-Sagrista, "Analysis-driven lossy compression of DNA microarray images," *IEEE Trans. Med. Imag.*, vol. 35, no. 2, pp. 654–664, Feb. 2016.

[17] S. Katsigiannis, E. Zacharia, and D. Maroulis, "MIGS-GPU: Microarray image gridding and segmentation on the GPU," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 3, pp. 867–874, May 2017.

[18] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, "NCBI GEO: Archive for functional genomics data sets—Update," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D991–D995, Nov. 2012.

[19] R. Malladi, J. A. Sethian, and B. C. Vemuri, "Shape modeling with front propagation: A level set approach," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 17, no. 2, pp. 158–175, Feb. 1995.

[20] C. Li, C. Xu, C. Gui, and M. D. Fox, "Distance regularized level set evolution and its application to image segmentation," *IEEE Trans. Image Process.*, vol. 19, no. 12, pp. 3243–3254, Dec. 2010.

[21] H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 1, no. 3, pp. 231–240, May/Jun. 2011.

[22] W. S. Cleveland and S. J. Devlin, "Locally weighted regression: An approach to regression analysis by local fitting," *J. Amer. Stat. Assoc.*, vol. 83, no. 403, pp. 596–610, Sep. 1988.

[23] M. R. Dalman, A. Deeter, G. Nimishakavi, and Z.-H. Duan, "Fold change and p-value cutoffs significantly alter microarray interpretations," *BMC Bioinf.*, vol. 13, no. 2, p. S11, Dec. 2012.

[24] G. Fabro, J. A. Di Rienzo, C. A. Voigt, T. Savchenko, K. Dehesh, S. Somerville, and M. E. Alvarez, "Genome-wide expression profiling Arabidopsis at the stage of Golovinomyces cichoracearum haustorium formation," *Plant Physiol.*, vol. 146, no. 3, pp. 1421–1439, Mar. 2008.

[25] A. Kokla and C. W. Melnyk, "Developing a thief: Haustoria formation in parasitic plants developmental biology," *Developmental Biol.*, vol. 442, no. 1, pp. 53–59, 2018.

[26] T. Shindo, J. C. Misas-Villamil, A. C. Hörger, J. Song, and R. A. Van Der Hoorn, "A role in immunity for Arabidopsis cysteine protease RD21, the ortholog of the tomato immune protease C14," *PLoS ONE*, vol. 7, no. 1, 2012, Art. no. e29317.

[27] H. Liu, M. Hu, Q. Wang, L. Cheng, and Z. Zhang, "Role of papain-like cysteine proteases in plant development," *Frontiers Plant Sci.*, vol. 9, p. 1717, Dec. 2018.

[28] J. Ruan, Y. Zhou, M. Zhou, J. Yan, M. Khurshid, W. Weng, J. Cheng, and K. Zhang, "Jasmonic acid signaling pathway in plants," *Int. J. Mol. Sci.*, vol. 20, no. 10, p. 2479, May 2019.

[29] E. Thomas and R. van der Hoorn, "Ten prominent host proteases in plant-pathogen interactions," *Int. J. Mol. Sci.*, vol. 19, no. 2, p. 639, Feb. 2018.

[30] M. B. Dickman and R. Fluhr, "Centrality of host cell death in plant-microbe interactions," *Annu. Rev. Phytopathol.*, vol. 51, pp. 543–570, Aug. 2013.

[31] J. Han, H. Li, B. Yin, Y. Zhang, Y. Liu, Z. Cheng, D. Liu, and H. Lu, "The papain-like cysteine protease CEP1 is involved in programmed cell death and secondary wall thickening during xylem development in Arabidopsis," *J. Exp. Botany*, vol. 70, no. 1, pp. 205–215, Jan. 2019.

[32] A. M. Jones, "Programmed cell death in development and defense," *Plant Physiol.*, vol. 125, no. 1, pp. 94–97, Jan. 2001.

[33] N. Hatsugai, K. Yamada, S. Goto-Yamada, and I. Hara-Nishimura, "Vacuolar processing enzyme in plant programmed cell death," *Frontiers Plant Sci.*, vol. 6, p. 234, Apr. 2015.

[34] Z. Li, H. Yue, and D. Xing, "MAP kinase 6-mediated activation of vacuolar processing enzyme modulates heat shock-induced programmed cell death in Arabidopsis," *New Phytologist*, vol. 195, no. 1, pp. 85–96, Jul. 2012.

BOGDAN BELEAN received the Ph.D. degree in electronics and telecommunication engineering from the Technical University of Cluj-Napoca, in 2011. In 2012, he joined the National Institute for Research and Development of Isotopic and Molecular Technologies, Cluj-Napoca, Romania, where he is currently a Senior Researcher responsible for the development of image processing algorithms applied on bio-medical images. Since 2015, he has been also responsible for the Reconfigurable Architectures for Signal and Image Processing master course with the Technical University of Cluj-Napoca. His current research interests include bio-medical image processing, neural networks, and hardware architectures development for high performance computing.

ROBERT GUTT received the Ph.D. degree in applied mathematics from the Faculty of Mathematics and Informatics, Babes-Bolyai University, Cluj-Napoca. In 2018, he joined the Center of Advanced Research and Technologies for Alternative Energies (CETATEA), National Institute for Research and Development of Isotopic and Molecular Technologies, Cluj-Napoca. He has published over ten research articles in leading journals of mathematics and physics related to boundary value problems in fluid mechanics, optical metamaterials, and energy harvesting systems, as well as two patent applications in the fields of microwave technology and concentrating solar collectors. His current research interests include image processing algorithms, artificial intelligence for time series predication, and novel solutions for electromagnetic energy harvesting.

CARMEN COSTEA received the B.S. degree in mathematics and informatics from Babes Bolyai University, Cluj-Napoca, in 2012, and the M.S. degree in educational management from Dimitrie Cantemir University, Cluj-Napoca, in 2014. She is currently pursuing the Ph.D. degree in computer science with the Technical University of Cluj-Napoca. Her main research interests include numerical algorithms for partial differential equation and bio-medical image processing.

OVIDIU BALACESCU received the Ph.D. degree in pharmacy from the University of Medicine and Pharmacy "Iuliu Hatieganu" Cluj-Napoca, in 2006. He was a Marie Curie Fellowship of genomics with the Gustave Roussy Institute, France, from 2003 to 2004, and several international trainings in microarray (mRNA and miRNA), qRT-PCR, bioinformatics, epigenetics, methylation, and miRNA::mRNA interactor identification and validation. He is currently a Senior Researcher with the Laboratory of Functional Genomics and Experimental Pathology, The Oncology Institute "Prof. Dr. Ion Chricuta" Cluj-Napoca. He is focused on identifying relevant molecular markers in both tissue and blood, concerning cancer progression and prognosis, and studying treatment resistance mechanisms in different pathologies, including prostate, breast, and colon and cervical cancers. He has an extensive experience in fundamental and translational research, leading five national research projects in domain of cervical, prostate and breast cancer, but also international experience during an EORTC program focused on malignant melanoma.