Fully automatic procedure for accurate spot addressing in microarray images

Mónica G. Larese[1,2,3⋆] and Juan Carlos Gómez[1,3]

[1] Centro Intern. Franco Argentino de Cs. de la Inform. y de Sistemas, CONICET
[2] Lab. de Inv. en Señales e Intel. Comput., FICH, Univ. Nac. del Litoral, Argentina
[3] Lab. for System Dyn. and Signal Proces., FCEIA, Univ. Nac. de Rosario, Argentina
larese@cifasis-conicet.gov.ar,jcgomez@fceia.unr.edu.ar

Abstract. *In this paper a novel procedure based on texture spatial characterization techniques is proposed aimed at automatically addressing spots in microarray images. The algorithm relies on the regular and pseudo-periodic patterns of spots (texture primitives). An automatic procedure is proposed to segment the autocorrelation functions of subgrid images and accurately determine the locations of the peaks. These candidate peaks (vectors) are next used to compute the displacement vectors that fully characterize the spatial arrangement of spots, describing the spot spacing and angle of rotation of the pattern. A refinement procedure is then applied to improve the accuracy of the norms and angles of the displacement vectors based on the frequency distribution of the regions of dominance of the candidate vectors. Experiments based on artificial and real images are encouraging and show improvements regarding robustness against image rotations, and accuracy, over results provided by state-of-the-art methods.*

1 Introduction

Complementary DNA (cDNA) microarrays are a powerful high throughput technology developed in the last decade which allows researchers to analyze the behavior and interaction of thousands of genes simultaneously. Each spot in a microarray image represents the hybridization level of a single gene under study.

A fundamental step in microarray image analysis is the addressing (*i.e.*, detection of the spatial location) of spots within image subgrids, in order to measure the hybridization levels. Even though spots are regularly located, this task is difficult due to the low quality of the images: non-uniform illumination, non-homogeneous background, missing and/or faulty spots, low contrast, presence of noise and artifacts, subgrids' rotation and/or misalignment, among other common problems arising from a physical experiment. Current methods aimed at addressing spots include semiautomatic [1,2,3] and automatic [4,5,6] procedures. Automatic methods are desirable due to the large amount of information to be processed (thousands of spots *per* image, dozens or hundreds of images *per* experiment).

⋆ Author to whom all correspondence should be addressed.

XXXIV Conferencia Latinoamericana de Informática (CLEI 2008), pp. 1050-1059, 2008.

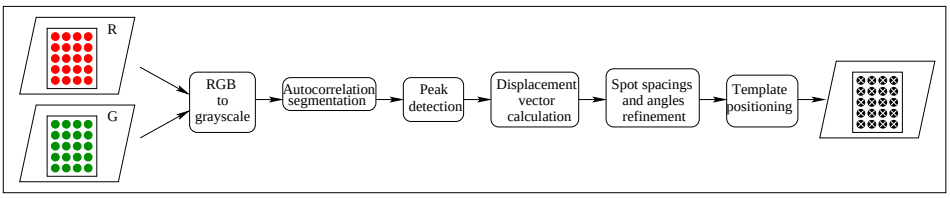


Fig. 1. Block diagram for the spot addressing algorithm.

In this paper, a novel algorithm is proposed to automatically locate the spots in microarray images. The approach relies on the pseudo-periodic patterns and regularity of spots. A new approach based on techniques from texture spatial characterization is proposed, where spots are considered as the texture primitives. An autocorrelation segmentation procedure is introduced in order to accurately estimate the two displacement vectors *i.e.*, spanning vectors, which completely characterize the pseudo-regular pattern of primitives (row and column spot spacings and angle of rotation of the pattern). The norms and angles of the two spanning vectors are refined to improve accuracy based on the frequency distribution of the norms and angles of the regions of dominance corresponding to all the most prominent peaks extracted from the segmented autocorrelation function. These two spanning vectors define an ideal regular template which is finally slightly deformed in a local way to adjust it to the real image. The whole process is depicted in the block diagram in Fig. 1.

The rest of the paper is organized as follows. Section 2 presents the procedure carried out to segment the 2D autocorrelation function and detect the candidate peaks. The displacement vectors calculation is detailed in Section 3 whereas the spacings and angles refinement procedure is explained in Section 4. In Section 5 the template positioning steps are detailed. Experimental results on synthetic and real images are discussed in Section 6. Finally, some conclusions and future work are presented in Section 7.

2 Autocorrelation segmentation and peak detection

The original microarray images considered in this paper are color images. In order to develop further procedures they are first converted to grayscale images in a preprocessing step.

Autocorrelation was proposed for regular texture structure characterization, *i.e.*, extraction of texture primitives and displacement vectors describing the spatial arrangement of the primitives, on general purpose images by Lin *et al.* [7] and then applied with slight variations by Liu *et al.* [8]. As it is well known, the autocorrelation function has the same periodicity as the image, showing equidistant peaks separated by the fundamental period. In addition, autocorrelation is more robust than Fourier transform in the presence of non-correlated noise,

since peaks are stronger in the autocorrelation function and therefore easier to detect [7].

In order to detect maximal points in the autocorrelation, Lin *et al.* [7] proposed to apply a smoothing filter to the autocorrelation surface in order to eliminate irregularities and easily detect the local maxima. Liu *et al.* [8] detect preliminary candidate peaks by means of non-maximal suppression. However, in both cases the positions of the peaks correspond to discrete coordinates, since the autocorrelation is a discrete 2D function. This procedure leads to discrete displacement vectors, propagating the errors to the estimated lattice.

In the present paper, a new approach is presented to extract the spanning vectors using subpixel precision. In order to do this, the centers of the segmented components are detected in the autocorrelation image through the following steps:

1. **Edge detection**: the edges in the 2D autocorrelation function are extracted by first applying a LoG (Laplacian of Gaussian) edge detector and then looking for the zero-crossings [9]. After this step, one of the two cases shown in Figures 2(b) and (e) may occur, *i.e.*, either separated or non-separated components, for the autocorrelation functions depicted in Figures 2(a) and (d), respectively.
2. **Morphological reconstruction**: holes inside object boundaries are filled by means of morphological binary reconstruction.

   In the case of already separated components (Fig. 2(b)), skip step 3, otherwise:

3. **Morphological binary opening**: a morphological binary opening is developed trying circular structuring elements with incremental radius, until all the components are separated.
4. **Connected components labelling**.
5. **Boundary components deletion**: the components on the image border are incomplete, so their centroids are biased.
6. **Centroid calculation**: the centroid coordinates are regarded as the location of the candidate peaks of the autocorrelation function, as shown in Figures 2(c) and (f) for each one of the two possible cases.

At the end of this procedure, if the components are not appropriately separated yet, or too many components are erased due to the opening stage, the contrast of the microarray image is enhanced via an intensity logarithmic transformation, and the new autocorrelation function is segmented using the steps detailed above. The whole procedure is done fully automatically.

3 Displacement vectors calculation

In order to find the two vectors that generate the lattice, special care must be taken. The goal is to find the two shortest displacement vectors satisfying linear

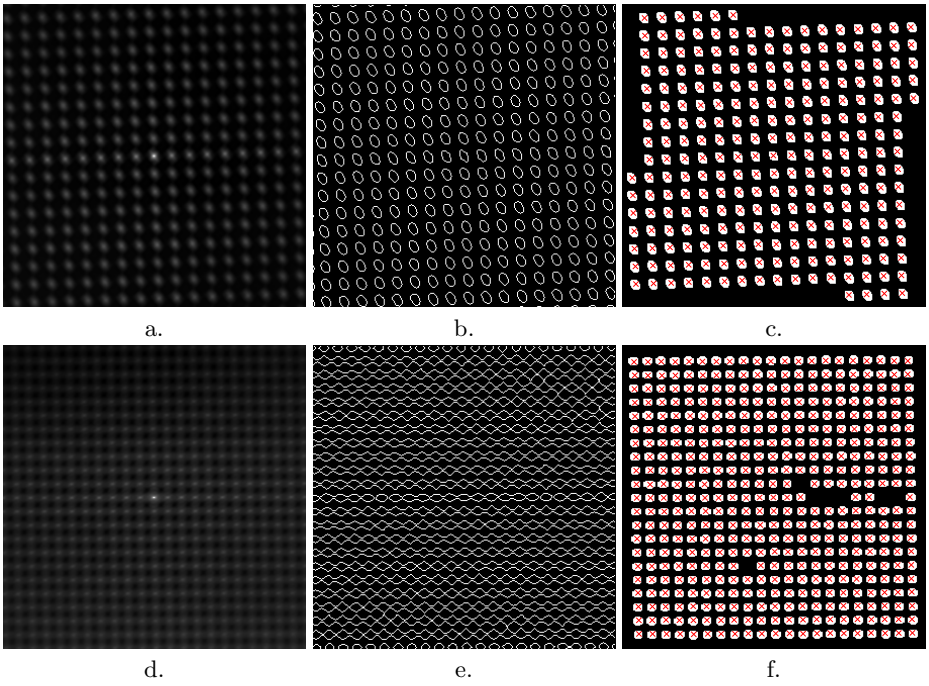


Fig. 2. Typical autocorrelation functions, with the corresponding border and segmented images (crosses in (c) and (f) stand for estimated peaks). **a**. Autocorrelation image (case 1). **b**. Border detection of (a). **c**. Segmentation of (a). **d**. Autocorrelation image (case 2). **e**. Border detection of (d). **f**. Segmentation of (d).

independence which are able to generate the whole lattice. They are neither the shortest vectors (due to spurious peaks) nor the largest ones (due to scaled versions of the targets). The approach based on regions of dominance proposed in [8] is applied to the centroids computed in Section 2 in order to determine the most prominent candidate peaks (regarded as vectors) to be considered in the displacement vectors computation. Next, the procedure described in Lin *et al.* [7], which is based on the generalized Hough transform, is implemented in order to find the two vectors that generate the spot lattice.

4 Spot spacing and angle of rotation refinement

The norms of the two spanning vectors describe the spot row and column spacings. Their angle describe the deviation in each axis direction. In order to improve accuracy even more, a histogram with the sizes of the regions of dominance for all the candidate vectors is constructed, as well as a histogram of the corresponding angles. The norms and angles of the two displacement vectors computed at the end of Section 3 are used as entries to each one of these histograms, and the

weighted mean of the corresponding isolated region in each histogram is regarded as the corrected norm and angle for each spanning vector.

The procedure is illustrated in Fig. 3, where the sizes (angles) of the regions of dominance histogram is depicted. The corrected norm (angle) Δ_c for each one of the two spanning vectors is computed as

$$\Delta_c = \frac{1}{\sum_{i=1}^{n} f_i} \sum_{i=1}^{n} f_i \Delta_i \tag{1}$$

where f_i and $\Delta_i, i = 1, \ldots, n$ stand for the frequencies, and sizes (angles) of the regions of dominance, respectively. In Fig. 3, Δ' represents the norm (angle) computed in Section 3 for each one of the displacement vectors ($n = 5$ in the example).

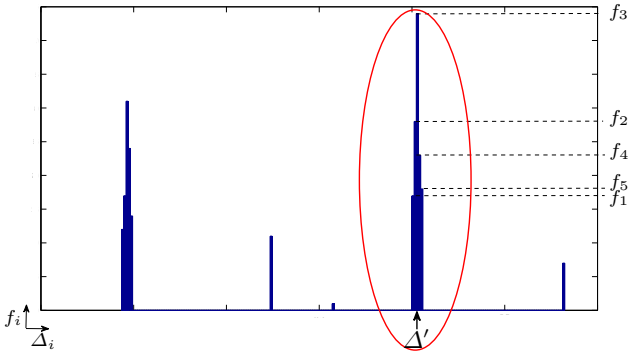
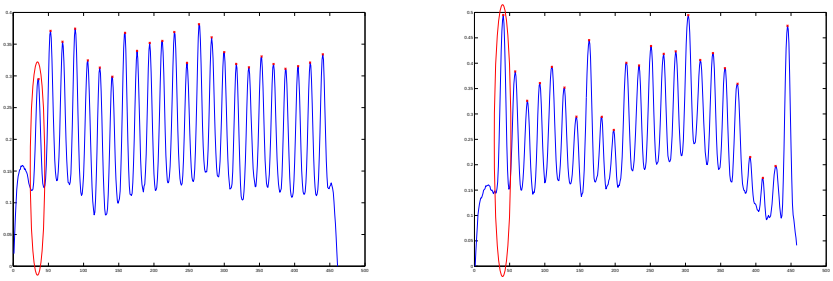


Fig. 3. Illustration of the procedure for spots spacing and angle of rotation refinement.

5 Template positioning

A template grid can be constructed using the two previously estimated displacement vectors. The number of rows and columns are known *a priori* from the microarrayer configuration. It is desirable that the top-leftmost spot in the template coincides with the top-leftmost spot of the real subgrid. With this purpose, the original image is temporarily corrected for rotation and background uniformity. The horizontal and vertical smoothed and detrended profiles are computed, and the most prominent peaks are detected. Figures 4(a) and (b) show typical profiles for a real microarray subgrid. The coordinates of the first peak are then back-rotated to the original position of the image, yielding the location from which the template can be spanned.

On the other hand, the centroids of all the connected components of the original microarray subgrid are computed. This procedure allows potential spot centers detection. However, many spurious centers arise due to the presence of artifacts and noise. Overlapped spots cause only one centroid to be detected. Moreover, the centers of missing spots or spots with low contrast cannot be detected at all. These potential spot centers are next used to adjust the position of each spot in the template using local search of observed spot centers in the neighborhood around the template centers.



a. b.

Fig. 4. Smoothed and detrended intensity profiles and local maxima (red crosses). The first local maximum from each profile is surrounded in red. **a.** Horizontal profile. **b.** Vertical profile.

6 Experimental Results

In order to validate the proposed algorithm, experiments on synthetic and real microarray images were performed. Accuracy was assessed by means of the RMSE (Root-Mean-Square-Error) between the estimated row and column spot locations, (x', y'), and the real ones, (x, y). The RMSE is computed as

$$RMSE = \sqrt{\frac{1}{G \times M \times N} \sum_{i=1}^{G} \sum_{j=1}^{M \times N} [(x_{i,j} - x'_{i,j})^2 + (y_{i,j} - y'_{i,j})^2]} \qquad (2)$$

where M and N stand for the total number of spots in each row and column, respectively, for a single subgrid i, and G is the total number of subgrids in the microarray image. Following this notation, the estimated coordinates $(x'_{i,j}, y'_{i,j})$ refer to the j-th spot (in lexicographic row-by-row arrangement) located at coordinates (x, y) in the i-th subgrid.

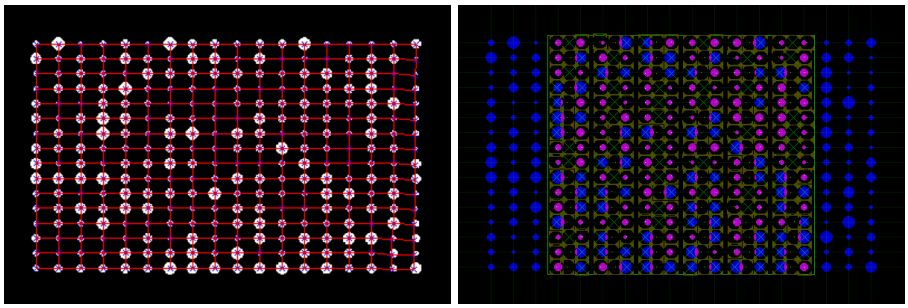
6.1 Computer generated images

Two type of experiments were performed using synthetic images. In the first case, subgrids with spot row spacing different from column spacing, and random

spot sizes, were analyzed with the proposed method and UCSF-Spot automatic algorithm [6]. The results are shown in Figures 5(a) and (b), respectively. As can be observed from Fig. 5(b) the UCSF-Spot algorithm fails to address the spots, trying to unify the row and column spacings. On the other side, the proposed algorithm succeeds in locating the spots, as shown in Fig. 5(a), where the blue crosses indicate the estimated spot centers.

In the second set of experiments, subgrids were generated with equal row and column spacings, but the locations of the spots were randomly altered from the regular lattice following a Gaussian distribution with zero mean and variances 0, 1 and 4. Spot sizes were randomly set. The images were rotated with angles in the range $[-5, 5]$ degrees and then analyzed with the proposed method and the UCSF-Spot algorithm.

In Fig. 6 the RMSE for the proposed (solid line) and UCSF-Spot (dashed line) algorithms as a function of the image rotation angle is depicted for spot location variances equal to 0 (circles), 1 (stars) and 4 (triangles). As can be observed, the algorithm introduced in this paper outperforms the UCSF-Spot algorithm in the given range. In addition, it can be noticed that the UCSF-Spot algorithm fails to locate the spots when the absolute rotation angle is greater than 1 degree, obtaining very high RMSEs. On the other hand, the RMSE values arising from the application of the proposed method hold constant for all the rotation angles.

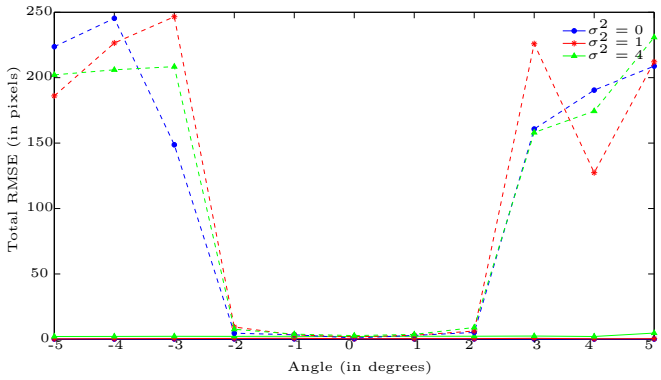


a. b.

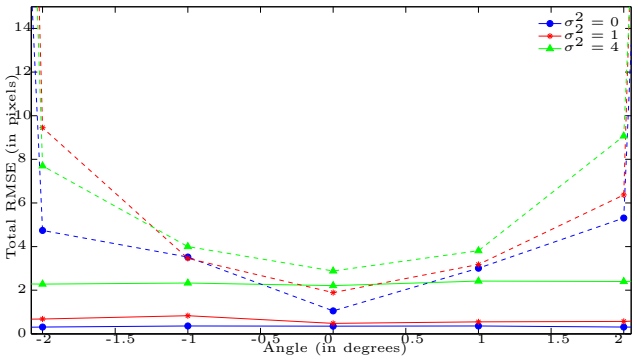
Fig. 5. Results of addressing a simulated image for the: **a.** Proposed algorithm. **b.** UCSF-Spot algorithm.

6.2 Real microarray images

The proposed method was compared to the UCSF-Spot algorithm on real images, and the accuracy of each method was measured through the RMSE calculation. Typical real microarray images were collected from the Stanford Microarray Database [10]. Details about the images, RMSEs in the x and y directions and



a.



b.

Fig. 6. RMSE for the proposed and UCSF-Spot algorithms (solid and dashed lines, respectively) as a function of the image rotation angle (bottom image is a zoomed version of the top image).

total RMSEs are reported in Table 1. In all the cases under analysis, the proposed method obtained a lower RMSE when compared to UCSF-Spot.

The high RMSE obtained when addressing image lc7b017rex2 can be attributed to the image rotation angle of approximately -2.2 degrees. However, all the other images have rotation angles below 0.5 degrees in absolute value. Specially images 21028 and 41602 are not rotated at all.

In addition, the result of applying the proposed automatic procedure to a typical microarray image is shown in Fig. 7, where the intersection of the overlapping lines define the estimated spot centers. As can be noticed from Fig. 7, the proposed algorithm succeeds in estimating the location of the spots. The whole algorithm takes approximately 12 seconds for a typical subgrid like this on a 1.6 GHz AMD-64 under Matlab and Linux, including I/O operations.

Table 1. Accuracy of spot addressing in terms of the RMSE (in pixels) for the proposed automatic algorithm and the UCSF-Spot algorithm described in [6].

Image ID	# spots	RMSE for the proposed method			RMSE for UCSF-Spot		
		$RMSE_x$	$RMSE_y$	Total $RMSE$	$RMSE_x$	$RMSE_y$	Total $RMSE$
lc7b070rex2 [11]	9216	1.88	2.21	2.90	44.21	4.97	44.49
lc7b017rex2 [11]	9216	1.01	1.87	2.13	66.89	10.80	67.75
lc7b0104rex2 [11]	9216	1.08	1.69	2.01	70.23	8.67	70.76
21028 [12]	43008	1.27	1.72	2.14	49.12	1.53	49.14
16275 [12]	45312	2.81	2.40	3.70	10.40	11.90	15.80
43957 [12]	43008	1.41	2.11	2.53	3.40	1.90	3.89
41602 [12]	43008	1.34	1.70	2.17	6.42	10.57	12.36
15739 [13]	9216	2.49	3.22	4.07	7.67	6.45	10.02

7 Concluding Remarks

In this paper an automatic approach is proposed to address the location of microarray subgrid spot centers. The method relies on the assumption that spotted microarray images can be regarded as regular texture images and consequently texture characterization techniques are suitable to be applied. This is due to the regularity and pseudo-periodicity exhibited by microarray images.

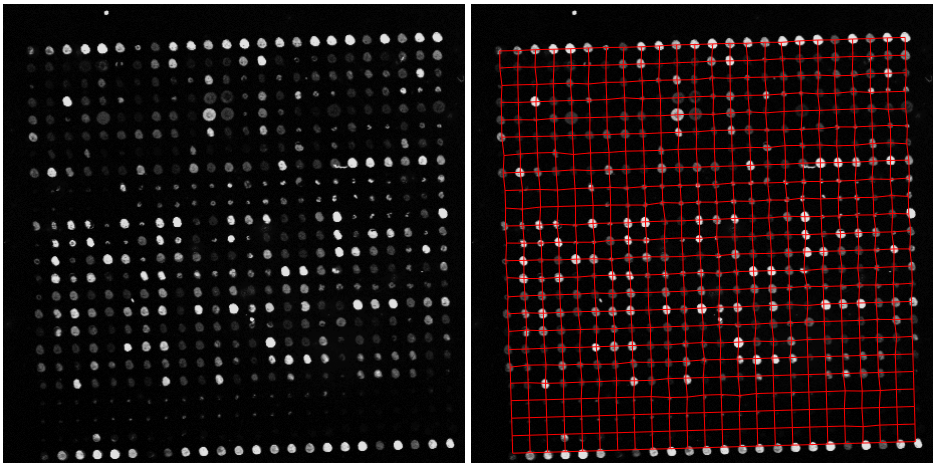
The present approach computes the displacement vectors that span the spot lattice, delivering the image angle of rotation and the row and column spot spacing. This approach is based on the computation of the generalized Hough transform. Instead of using the raw autocorrelation image, the autocorrelation is previously segmented by means of morphological binary operations and connected components detection. The centers of these components are processed to get the spanning vectors, allowing subpixel precision.

Experimental results on synthetic and real images show that the proposed method outperforms the ones provided by a state-of-the-art microarray analysis tool (namely the UCSF-Spot) specially when image rotations and unequal row and column spacings are present.

The present authors believe that the method yields promising results improving accuracy over widely used tools available in the literature. A refinement procedure based on Markov Random Fields is currently under development to further improve robustness against spot location variation, and accuracy.

References

1. L. J. Heyer, D. Z. Moskowitz *et. al.*, "Magic tool: integrated microarray data analysis," *Bioinformatics*, vol. 21, no. 9, pp. 2114–2115, 2005.



a. b.

Fig. 7. Example of real subgrid. **a.** Real subgrid taken from image lc7b017rex2 [11]. **b.** Result of the proposed addressing algorithm.

2. Y. Yang, M. Buckley, S. Dudoit, and T. Speed, "Comparison of methods for image analysis on cDNA microarray data," Tech. Rep. #584, Dep. of Stat., UCB, Nov. 2000, http://www.stat.berkeley.edu/users/terry/zarray/TechReport/584.pdf.
3. M. Eisen, "Scanalyze," 1999, http://rana.lbl.gov/EisenSoftware.html.
4. M. Katzer, F. Kummert, and G. Sagerer. Methods for automatic microarray image segmentation. *IEEE Transactions on Nano-Bioscience*, 2(4):202–214, 2003.
5. K. Hartelius and J. M. Carstensen, "Bayesian grid matching," *IEEE Trans. on P.A.M.I.*, vol. 25, no. 2, pp. 162–173, 2003.
6. A. N. Jain, T. A. Tokuyasu *et. al.*, "Fully automatic quantification of microarray image data," *Genome Research*, vol. 12, pp. 325–332, 2002.
7. H.-C. Lin, L.-L. Wang, and S.-N. Yang, "Extracting periodicity of a regular texture based on autocorrelation functions," *Pattern Recognition Letters*, vol. 18, pp. 433–443, 1997.
8. Y. Liu, R. T. Collins, and Y. Tsin, "A computational model for periodic pattern perception based on frieze and wallpaper groups," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, no. 3, pp. 354–371, 2004.
9. R. Gonzalez and R. Woods, *Digital image processing*, Prentice Hall, 2nd. edition, 2002.
10. J. Demeter, C. Beauheim, J. Gollub *et. al.*, "The stanford microarray database: implementation of new analysis tools and open source release of software," *Nucleic Acids Res.*, vol. 35(Database Issue), pp. D766–770, Jan 1 2007.
11. A. A. Alizadeh, M. B. Eisen *et. al.*, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503–511, February 2000.
12. S. Subramanian, R. B. West *et. al.*, "The gene expression profile of extraskeletal myxoid chondrosarcoma," *J. Pathol.*, vol. 206, pp. 433–444, 2005.
13. M. N. Arbeitman *et al.*, "Gene expression during the life cycle of Drosophila melanogaster," *Science*, 297:2270–2275, 2002.