Detection of Artificial Contamination in E. Coli Microarray Data

DÍAZ[1], F., MALUTAN[1,2], R., GÓMEZ[1], P., RODELLAR[1], V., BORDA[2], M.
[1]Departamento de Arquitectura y Tecnología de Sistemas Informáticos, [2]Communications Department
[1]Universidad Politécnica de Madrid, [2] Technical University of Cluj-Napoca
[1]Campus de Montegancedo, s/n, 28660, Boadilla del Monte, [2]George Baritiu, nº 26-28, 400027
[1]SPAIN, [2]ROMANIA

Abstract: - Oligonucleotide Microarrays technologies offer the possibility of simultaneously monitoring thousands of hybridi-zation reactions. These arrays show high potential for many medical and scientific applications as gene expression monitoring, sequence analysis, and genotyping. This is possible because high densities of probe tests may be included in the surface of silicide compounds. Nevertheless Microarrays are exposed to errors during manufacturing, similar to silicon circuit electronics and the hybridization process may be contaminated by different reasons. Other source of errors is due to optical noise during scanning and processing, or to interactions between molecular structures and light (dispersion among others). To reduce some of these effects are used replicates of experiments with the cost of increasing expenses. In order to detect noise contamination in Microarray Data Images well-known computational techniques are proposed to help in visual analysis. The use of image transformation from the space domain to the frequency domain gives the possibility of processing it with filtering algorithms for image enhancement. Some experiments with Escherichia Coli Antisense microarray are shown to check the effectiveness of these approaches.

Key-Words: - microarray, oligonucleotide chips, probe set, low pass filtering, FFT

1 Introduction

Oligonucleotide microarrays developed by different commercial companies are becoming a most powerful technology for use in different fields related with Medicine, Biology and Pharmacology among others. These high-density arrays are designed to orderly sequence genetic information alone and are synthesized in situ using a combination of photolithography and oligonucleotide chemistry. RNAs present at frequencies of 1:300,000 are unambiguously detected, and the detection is efficient over more than three orders of magnitude. This method provides a way to use directly the growing body of sequence information for highly parallel experimental investigations. Because of the combinatorial nature of the molecular dynamics involved and the ability to synthesize small arrays containing hundreds of thousands of specifically chosen oligonucleotides, the method is easily scalable to the simultaneous monitoring of gene expression. Commercially available microarrays contain up to 500,000 unique probes corresponding to tens of thousands of gene expression measurements [8]. Probe cells are 18-50 micron square-shaped features on the chip containing millions of copies of a single 25-mer probe [6][7].

Messenger RNA (mRNA) is extracted from the cell and converted to cDNA. It then undergoes amplification and labeling before fragmentation and hybridization against 25-mer oligonucleotides on the surface of the chip. After washing-off unhybridized material, the chip is scanned in a confocal laser scanner and the resulting image is processed by computer.

2 Structure of Microarray Data

2.1 Probe Set definition

In an oligonucleotide array a gene is represented by a set of 11-25 probe cells, called *perfect match probes* (PM). The multiple oligonucleotides that represent a gene [1] are designed in such a way that they can hybridize to different regions of the RNA corresponding to the gene under test and act as a series of multiple independent detectors for that gene.

Each perfect match probe is paired with an artificially created *mismatch probe* (MM) that is tailored changing the center base of the corresponding perfect match sequence to its complementary base. The mismatch probe is intended to play the role of an internal control test for hybridization specificity to its particular hybridization site. The hybridization level by the perfect match probe represents specific hybridization and should be stronger than nonspecific hybridization level expressed by the mismatch probe. In addition, if the PM levels are consistently larger than the MM levels for a probe set, this global effect is more likely to be indicative of the actual expression of the mRNA corresponding to that gene in the sample rather than being a result of random activity.

A core element of array design, the PM-MM probe strategy, is universally applied to the production of GeneChip arrays. These probe pairs, a pair of PM probe and its corresponding MM probe, allows the quantization and subtraction of signals caused by non-specific cross-hybridization. The difference in hybridization signals between the pairs, as well as their intensity ratios, serve as presence indicators of specific target sequences. Differential estimation algorithms as MAS 5.0, MBEI or RMA [2][4][5] will evaluate the so called *expression signal* for the probe set.

The largest part of the microarray surface contains Probe Sets for the detection of gene expression levels, but other Probe Cells are also included with known values. These probes are called *quality control* (QC) probes. For example, the border around the array and the corner region are used for easily reading and the control region in the center is necessary to grant successful hybridization.

2.2 Expression Data Files

The microarrays studied are scanned with an argon-ion laser scanner. As the surface of the array is scanned, a photomultiplier tube collects and converts the fluorescent emission into electrical currents. These electrical current are converted into numeric values through an analog to digital converter to create multi-pixelated raw images (.DAT files) The .DAT files are image files, with ~10^7 pixels and a size of ~50MB. The Quality Control of the chip can provide information about the appearance of scratches or spots that represent possible chip contamination or imperfect processing.

The first step in microarray processing prior to data analysis consists in converting each multi-pixilated probe cell to a single intensity value thus transforming the raw image file (.DAT file) into a feature by feature flat file (.CEL file). The probe cell feature is scanned at a resolution of 3μm per pixel resulting in 7 pixels by 7 pixels for every probe cell for a total of approximately 49 pixels per probe cell. Taking the 75th percentile of the signal distribution for these 49 pixels creates a single intensity value for every probe cell. The single intensity value is representative of the number of targets (messenger RNA) hybridizing to multiple copies of a particular probe. The feature by feature flat file (.CEL file) is now composed of X and Y coordinates and a single intensity value for each probe cell.

2.3 Relationships between .DAT and .CEL files

Once the raw image .DAT file, has been converted into a feature by feature flat file .CEL file, assigning a single intensity value for each probe cell, it is now possible to process the file and determine the qualitative and quantitative information with a statistical algorithm. The .CEL file summarizes all the intensities from a Probe Cell. In theory all pixels from a Probe Cell must show the same intensity value but experimental results show different values. Some Probe Cells exhibit an even distribution, but others present uneven distributions of values.

The uneven distributions of the Probe Cells intensities suggest that the hybridization process contain random factors based on molecular dynamical interactions. However, the distributions of the fluorescence values in an image .DAT file are similar to the corresponding file .CEL.

Figs.1 and 2 show the log intensity histograms of .DAT and .CEL files from E.Coli array used as experimental data in this work. These plots are very similar and the mean value of the intensities is around 2. The distribution of the values to the left and right of the mean is not symmetrical. The similarity between both files can be probed with the visual analysis of the images. In order to reduce the costs of processing it is possible to analyze the .CEL file instead of the .DAT file.

In this article we assume that the information in the .CEL file summarize the contribution of hybridization and the contributions of other sources, for example, optical noise, transduction artifacts, and probe interaction.

It is known that the use of Digital Image Processing in the frequency domain provides an improvement of the image [3].

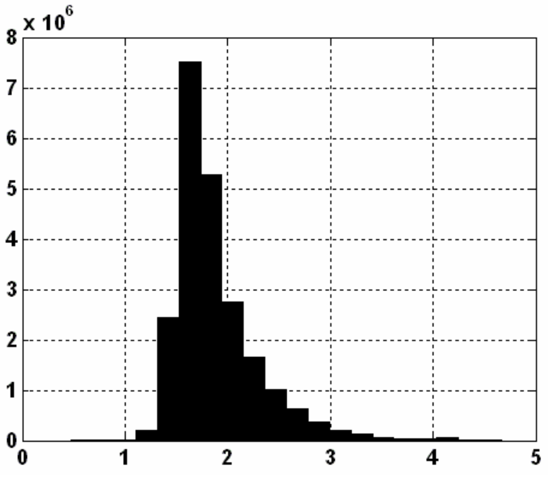

Fig. 1 Histogram of Log Data Pixel Intensity

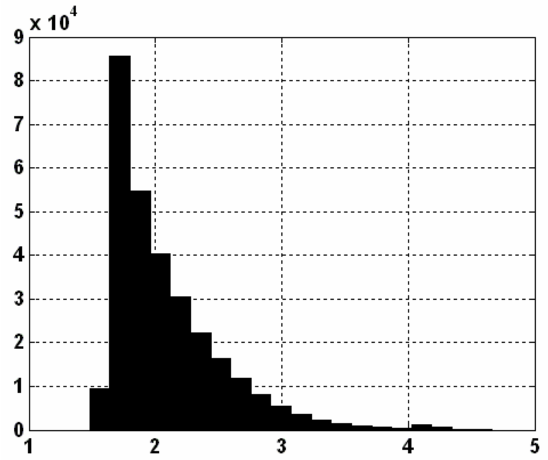

Fig.2 Histogram of Log CEL Intensity

Our goal is to prove that the processing in the frequency domain is an efficient approach to improve the images of oligonucleotide arrays and reduce the effects of corruption agents over the hybridization signal. In the following section we introduce the basic concepts of spatial domain filtering and frequency domain filtering. In the final section we present different experiments using this approach.

3 Fundamentals of Digital Image Processing

3.1 Filtering in the spatial domain

The term “spatial domain” refers to image plane itself and methods in this category are based on the direct manipulation of pixels in an image. The spatial domain processes are denoted by the expression:

$$g(x,y)=T[f(x,y)] \qquad (1)$$

where *f(x,y)* is the input image, *g(x,y)* is the output image and *T* is an operator on *f*, defined over a specified neighborhood around point *(x,y)*.

The principal approach for defining spatial neighborhoods about a point *(x,y)* is to use a square or rectangular region centered at *(x,y).* The center of the region is moved from pixel to pixel starting at the top left corner, and, as it moves, it sweeps over different neighborhoods. Operator *T* is applied at each location *(x,y)* to yield the output *g* at that location. Only the pixels in the neighborhood are used in computing the value of *g* at *(x,y).*

The neighborhood processing or spatial domain filtering requires four steps for performing the desired transformation on the image:

- defining the center point *(x,y)*;
- performing an operation that involves only the pixels in a predefined neighborhood around the center point;
- assigning the result of that operation to the response of the process at that point;
- repeating the process for every point in the image.

The concept of linear filtering has its roots in the use of the Fourier transform from signal processing in the frequency domain. The linear operations used for linear spatial filtering consist of multiplying each pixel in the neighborhood by a corresponding coefficient and adding up the results to obtain the response at each point *(x,y).*

If the neighborhood size is *m*x*n*, an equal number of coefficients are required. The coefficients are arranged as a matrix which is often called mask, kernel or filter mask. There are two closely related concepts which are used when performing linear spatial filtering, one is correlation and the other one is convolution. Correlation is a neighborhood operation in which the value of an output pixel is computed as a weighted sum of neighboring pixels. The weights are defined by the correlation kernel which represents the filter mask. The convolution is the same as correlation with the difference that the convolution kernel is obtained by rotating the correlation kernel 180 degrees.

3.2 The 2D Discreet Fourier Transform

For a *f(x,y)* function with *x=0,…,M-1* and *y=0,…,N-1* which denote an *M*x*N* image the 2D discrete Fourier transform is denoted by *F(u,v)* and is given by the equation:

$$F(u,v)=\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)e^{-j2p\left(\frac{ux}{M}+\frac{vy}{N}\right)} \qquad (2)$$

where *u = 0, …, M-1* and *v = 0, ..., N-1.*

The frequency domain is simply the coordinate system spanned by *F(u,v)* with *u* and *v* as frequency variables. This is analogous to the spatial domain which is the coordinate system spanned by *f(x,y)* with *x* and *y* as spatial variables. The *M*x*N* rectangular region defined by *u=0,…,M-1* and *v=0,…,N-1* is often referred to as the frequency rectangle (or frame) of the same size of the input image.

The inverse discrete Fourier transform is given by:

$$f(x,y)=\frac{1}{MN}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1} F(u,v)e^{j2p\left(\frac{ux}{M}+\frac{vy}{N}\right)} \qquad (3)$$

where *x = 0, …, M-1* and *y = 0, ..., N-1.*

Even if *f(x,y)* is real, its transform in general is complex. The principal method to visually analyzing a transform is to compute its spectrum, the magnitude of *F(u,v)* and display it as an image. The Fourier spectrum is defined as:

$$|F(u,v)| = \left[R^2(u,v) + I^2(u,v)\right]^{1/2} \qquad (4)$$

where *R(u,v)* and *I(u,v)* represent the real and imaginary components of *F(u,v)*.
The phase angle of transform is defined as:

$$\Phi(u,v) = \tan^{-1}\left[\frac{I(u,v)}{R(u,v)}\right] \qquad (5)$$

The preceding two functions can be used to represent *F(u,v)* in the familiar representation of a complex quantity:

$$F(u,v) = |F(u,v)| \cdot e^{-j\Phi(u,v)} \qquad (6)$$

If *f(x,y)* is real its Fourier transform is conjugate symmetric about the origin, $F(u,v)=F^{*}(-u,-v)$, which implies that the Fourier spectrum is also symmetric about the origin: |*F(u,v)*|=|*F(-u,-v)*|.

3.3 Filtering in the spatial domain

The foundation for linear filtering in both the spatial and frequency domain is the convolution theorem, which may be written as:

$$f(x,y) * h(x,y) \Leftrightarrow H(u,v) \cdot F(u,v) \qquad (7)$$

and conversely

$$f(x,y) \cdot h(x,y) \Leftrightarrow H(u,v) * F(u,v) \qquad (8)$$

Here the symbol * indicates the convolution of two functions and the expression on the sides of the double arrow constitute a Fourier transform pair. The first expression indicates that convolution of two spatial functions can be obtained by computing the inverse Fourier transform of the product of the Fourier transform of the two functions, this relation being very useful in terms of filtering. Filtering in the spatial domain consist of convolving an image *f(x,y)* with a filter mask *h(x,y)*. According to the convolution theorem the result of linear spatial convolution can be obtained in the frequency domain by multiplying the Fourier transform of the image *F(u,v)* by the Fourier transform of the filter mask H*(u,v)*. To obtain the filtered image in the spatial domain simply compute the inverse Fourier transform of the product *H(u,v)F(u,v)*. This process is similar to applying the filter mask *h(x,y)* on the image *f(x,y)* using the convolution in spatial domain.

4 Experimental Results

With the purpose of checking the low-pass filtering in the frequency domain the Escherichia Coli Antisense sample data from [8] was used. In this microarray the Probe Pairs are localized consecutively in order to localize a Probe Set in neighbour positions, being highly sensitive to local corruption, as with this arrangement the local contamination may affect entire Probe Sets.

The first step in the analysis process was the computation of the logarithm of the data and after that transforming the image to the frequency domain using the Fast Fourier Transform in two dimensions (2).

In order to filter the image we used a filter in the spatial domain. This filter is a Gaussian Lowpass filter (GLF) with the form given by the equation:

$$h_g(n_1,n_2) = e^{-\left(n_1^2+n_2^2\right)/\left(2s^2\right)} \qquad (9)$$

$$h(n_1,n_2) = \frac{h_g(n_1,n_2)}{\sum_{n_1}\sum_{n_2} h_g(n_1,n_2)} \qquad (10)$$

where n$_1$, n$_2$ specify the number of rows and columns and s is the standard deviation. This filter returns a symmetric matrix.

The Gaussian Lowpass normalized filter used is shown in Fig. 3 for a scanned image, .DAT file. It is not possible to be used directly on a .CEL image file and this is way a reduction in dimension and variance proportionality ought to be carried out. In Fig. 4 the final Gaussian lowpass filter is shown and as it may be seen its shape is similar but his size is smallest than the prior one.

Using (7) the filtering process was transposed from the spatial domain to the frequency domain; in order for this to be possible the Fourier transforms of both the .CEL image and Gaussian low-pass filter were computed accordingly to expression (2). The original .CEL image is shown Fig. 5 and after applying the GLF as described in previous section we obtain the result shown in Fig. 6. The visual inspection of the images, the original and the filtered one, indicates that this filter will increase low frequency features and reduce or eliminate high frequency ones from the image.

In the filtered image some reference point marks on the array, -the left up corner of the border and the rectangle from the center- are visible and marked with white circles. All these represent quality control probe cells and are included during the manufacture of the array, their values being known a priori before hybridization.

The filtering tries to detect the artifacts, fibers or stained spots in the scanned image with high fluorescence values.

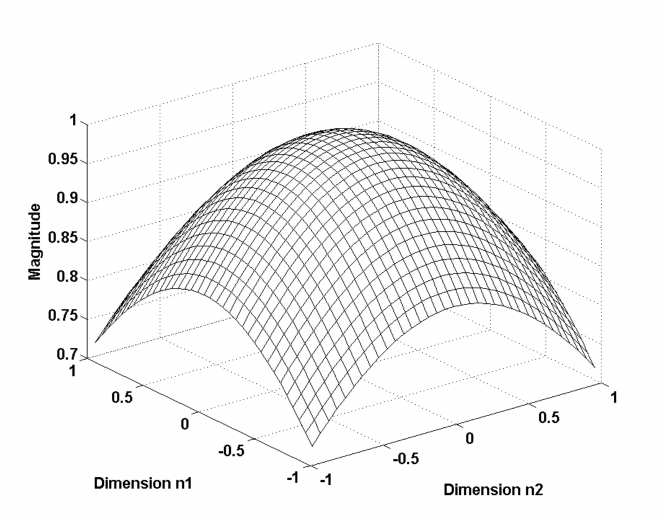

Fig. 3 Normalized GLF for .DAT image file

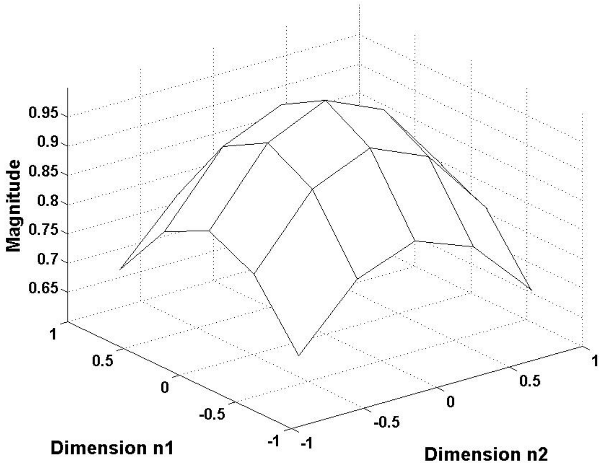

Fig. 4 Normalized GLF for .CEL imge file

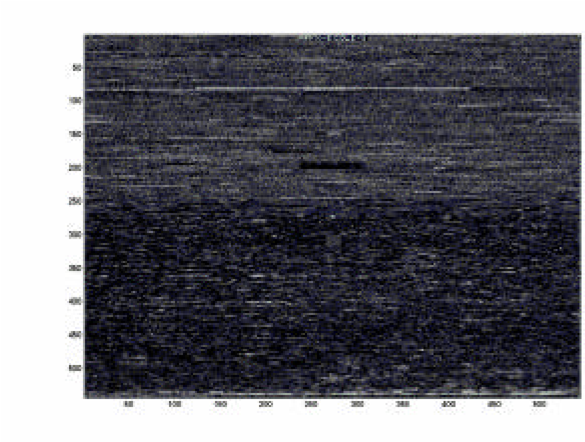

Fig. 5 Original E. Coli .CEL

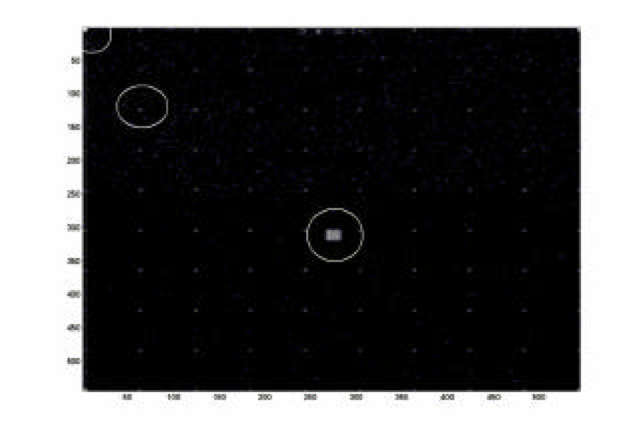

Fig. 6 Filtered E. Coli .CEL image

In order to determine the effectiveness of the Gaussian Lowpass filter in detection of artifacts and other corrupting effects other tests have been done. For such the original .CEL image was corrupted with marks in the shape of a line and an arc. Corrupting patterns presented two features: the thickness of the mark and the intensity of the pixels used to compose the mark. For the thickness two values were used to obtain a thin mark (T) and a bold mark (B) and for the intensity three qualitative ones: low (L), median (M) and high (H) according to the histogram (Fig. 2).

The results for different combinations of thickness and intensity values are being plotted in Fig. 7.

In the cases of high intensities (HT, HB) the mark is detected in the filtered image but better for low thickness. For the bold thickness the filter detects only the contour of the contamination mark. In the others cases, for medium intensity (LT, LB) the mark is invisible in the original image so the filter is not able to detect it, however in these cases the influence of the mark is insignificant. In the last cases, where the intensity is lower than the mean (LT, LB) the mark is being detected by the filter much better for a low thickness than for a bold pattern.

4 Conclusions

Through this study we re-scaled the filter design for .DAT image file to a filter for .CEL data file. This approach reduces the computation costs and proves our hypothesis about filtering in the frequency domain. This method is effective in detecting sources of spatial corruption, as artifacts, fibers or stain spots presenting large intensities values. The transformed images can be used for background estimation.

This research is being carried out under grants TIC2002-02273 and TIC2003-08756 from the *Programa Nacional de las Tecnoligía de las Información y las Comunicaciones (Spain)*

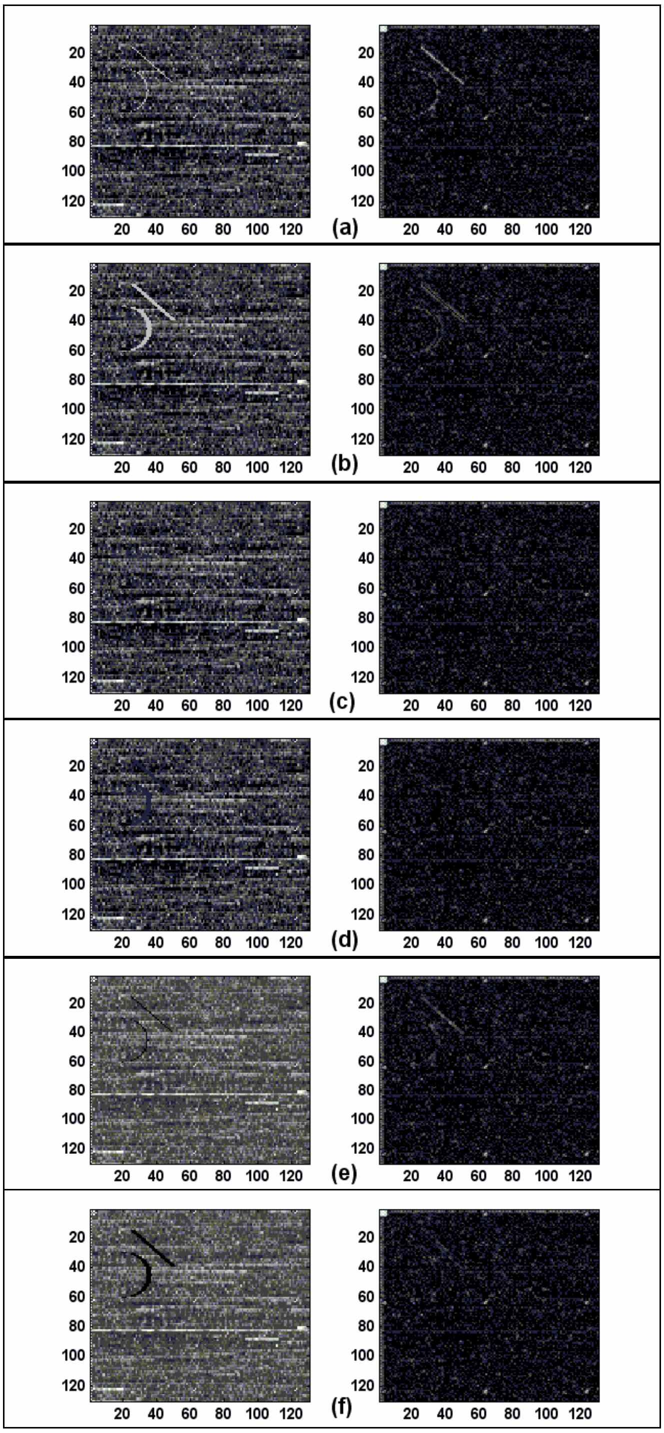

Fig. 7 Original image after contamination with (a) HT mark, (b) HB mark, (c) MT mark, (d) MB mark, (e) LT mark, (f) LB mark and the coresponding filtered image

References:

[1] Amaratunga, D. and Cabrera, J., *"Exploration and analysis of DNA microarray and protein array data"*, Ed. Wiley Interscience, Hobooken, N.J., 2004, pp. 35-36

[2] Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf , U. and Speed, T. P., "Exploration, normalization and summaries of high density oligonucleotidearray probe level data", *Biostatistics.* Vol. 4, Number 2 (2003), pp. 249-264

[3] Gonzalez, C., Woods, R. E., Eddins, S. L., *"Digital image processing using MATLAB"*, Ed. Prentice Hall, Upper Saddle River, N.J., 2004

[4] Li, C. and Wong, W. H., "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection", *Proc. Nat. Acad. Sci.*, Vol. 98 (2002), pp. 31-36

[5] Naef, F., Lim, D. A., Patil, N., and Magnasco, M., "From features to exxpresion: High denstiy oligonucleotide arrays revised", *Proc. DIMACS Workshop on Analysis of Gene Expression Data*, (2001)

[6] Robert J. Lipshutz, Stephen P. A. Fodor, Thomas R. Gingeras & David J. Lochhart, "High density synthetic oligonucleotide arrays", *Nature Genetics suppliment* vol. 21, January 1999, pp. 20-24.

[7] Washington, J. A., S. Dee, and Trulson, M., "Large-scale genomic analysis using affimetrix genechip", *Microarray Biochip technology*, Chapter 6, pp. 119-148, 2000

[8]www.affymetrix.com/support/technical/sample_data/demo_data.affx.