MAQ – A Bioinformatics Tool for Automatic Macroarray Analysis

António dos Anjos
[1]Dpt. of Electronic and Informatics Engineering
University of Algarve
8005-139 Faro, Portugal
aanjos@ualg.pt

Ricardo Leite
[2]CCMAR - Center of Marine Sciences
Molecular Biology Laboratory
University of Algarve
8005-139 Faro, Portugal
rleite@ualg.pt

M. L. Cancela[2]
lcancela@ualg.pt

Hamid Reza Shahbazkia[1]
hshah@ualg.pt

Abstract

This article presents a prototype of an application for the analysis of cDNA/gene expression data and contribution for genome annotation from macroarrays. The application described here allows the quantification of a set of experiments with macroarrays of the same size. It generates *html* reports, easily shared, showing the expression of genes contained in each membrane and a global report comparing films (expositions) and probes. Although initially the objective of this application was macroarray analysis, with minor adjustments, it can be used for other densitometric analysis involving circular dish plates such as petri dish plates, or multiple cell plates. It aims to be a simple, but effective, tool that does not require the user to have previous knowledge in image processing.

Keywords

Genomics, Macroarray, Segmentation.

1 Introduction

Gene expression analysis, by cDNA macroarray construction and hybridization[10], has been very useful for use in numerous assays[2, 6]. Disease prevention and diagnosis[7] based on gene expression patterns are some of the applications for gene expression analysis. Some companies like Imagenes[1] are still offering services based on macroarray analysis and, despite all the recent developments in gene expression quantification, macroarrays are still used as a entry-level technique. Experiments based on macroarray hybridization often produce a great amount of data, specially if the number of conditions/genes is high. The final data usually consists of autoradiography films with impressed signal, representing relative gene expressions. This kind of experiments, very frequently, require the densitometric quantification of each spot

---
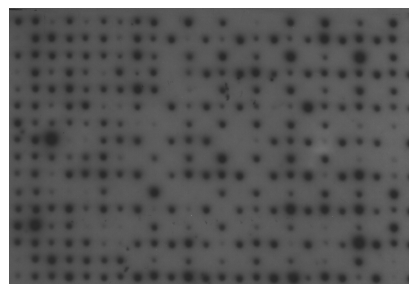
[1]http://www.imagenes-bio.de/services/c_macroarrays/

Figure 1: Macroarray film image.

in the macroarray and, also, that patterns are found among the films/probes allowing simple comparisons between normalized scenarios. In most cases, this work is done using expensive software, or free but not very friendly software, that also relies on pattern interpretation by the user. Work has shown that, for a correct quantification, it is important to be able to subtract the background noise from the spots' intensity [3]. For all these reasons, it is very difficult, time consuming and inaccurate for the users to extract conclusions from such data, specially if the analysis was done manually by visual inspection. There are applications that can help to accomplish these tasks but, they require a lot of user input, are complex to utilize and very expensive, like PDQuest and Quantity One from Biorad, or Imagene from Biodiscovery. In addition, most of the existing applications are not specific for this problem. An important aspect is also that this application is distributed freely in line with the relatively inexpensive technology of macroarrays.

The software application presented in this paper is simple and can be used by someone that does not have previous knowledge in image processing; requiring only a few input parameters. It also aims to be effective in suppressing the background of the macroarray film images and to automatically perform the normalization of the film results. The final aim of this application is, not to introduce another way
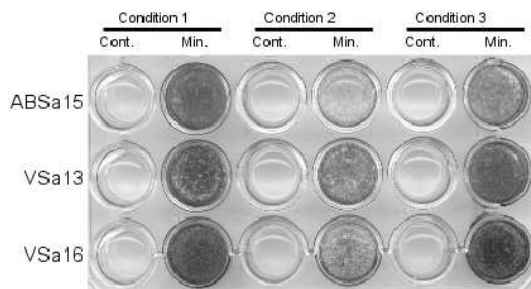
Figure 2: Mineralization from different fish cell lines.

Figure 3: Ready.

to quantify gene expression, but to save time and assist on results interpretation before passing to more sensitive techniques like RT-qPCR. Another important aspect is the flexibility of the software, that allows any quantification based on densitometric determination of round forms like, for example, colorimetric assays in round plates/wells. A good example of another application for this software is the analysis of Von Kossa staining[8], a calcium quantification technique that relies on a densitometric analysis, i.e. measurement of cell mineralization (See Figure 2).

In Section 2 a brief explanation about the characteristics of the images involved in this kind of experiments will be presented, then, in Section 3 it will be presented an overview of the user interface. In Section 4, the technical details of the internals of the software will be exposed. In the final section, some results will be presented and discussed.

2 Macroarray Experiments

The typical results of a macroarray based experiment are impressed on autoradiography films. Each film may have one or more macroarrays represented, and each macroarray may have hundreds or even thousands of spots. To be possible to perform an automatic analysis, the films have to be scanned and digital images produced. Figure 1 shows a typical scan of a macroarray image.

The number of rows and columns may vary but, usually, each experiment will have macroarrays with a specific number of rows and columns. Each of the cells of the macroarray will contain a spot if a given gene is being expressed. The intensity of the spot will represent the level of expression of a given gene in that sample. In the case of absence of a spot in a given cell, it means that the cDNA/gene in study is not expressed in that condition.

As it is noticeable in Figure 1, the scanned image includes noise. That may be the result of the presence of artifacts related to probe cleaning, membrane reuse, hybridization process, film processing and - even - scanning process. The darkness of the background will vary according to the film properties and exposure. Also, although not so noticeable, there is
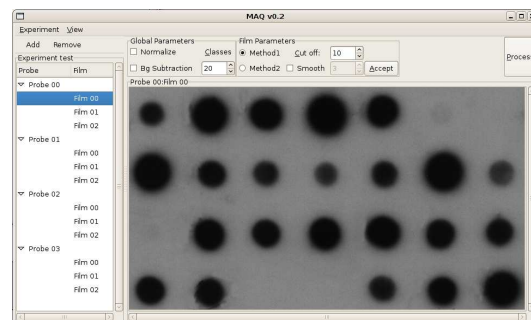
a gradient present in the image. These are problems that have to be dealt with before quantification.

These images, produced for automatic analysis, should be scanned at, at least, 300ppi. Lower resolution images will not cope very well with smoothing filters because important details of the image may be destroyed by the smoothing operations. On the other hand, images with much higher resolution will slow down the processing without many benefits.

3 Interface

The interface for the application is meant to be simple. The first dialog that pops up at the beginning of a new experiment, asks the user for the basic definitions of the experiment. It asks for the name of the experiment, the number of probes, the number of membranes/films and the size of the macroarrays. After those basic parameters are defined, the user has to assign images of each macroarray to each of the leafs of the tree generated by the application (see Fig. 3). Then, it is possible to define the film and global parameters for the experiment. As the names show, film parameters define the parameters for each film, and global parameters define the parameters for the whole experiment. The process button is activated and the user may start the automatic processing when every leaf of the tree has been assigned a macroarray image, and all the parameters are defined.
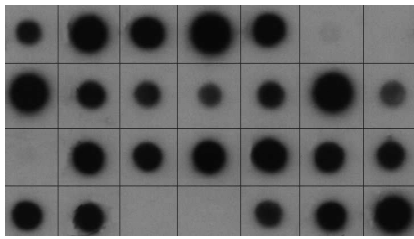
4 Technical Details

Some detailed technical information about the implementation of the several aspects of the presented application are provided in the following subsections.

4.1 Array Detection

Most of the existing applications for macroarray processing require the user to input the size of the macroarray and, only then, a grid will be generated to be manually resized and placed over the area of the macroarray. The objective is to
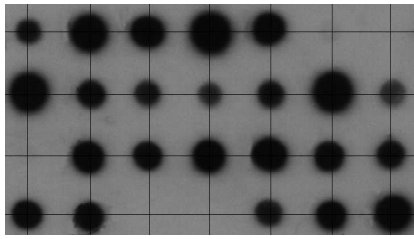
(a) Division by minima sum profile.

(b) Division by maxima sum profile.

Figure 4: Breaking in areas.

(a) Macroarray Cell.

(b) Histogram.

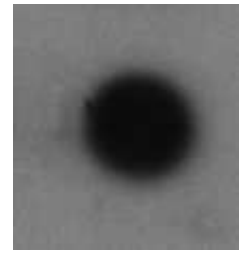Figure 5: Two classes are noticeable in the histograms.

isolate each cell of the macroarray. In this way there will be only one spot in each of the cells. Our approach tries to isolate the cells automatically. Using the information about the macroarray size $(n * m)$ provided on the first dialog, the division is performed in the following way:

- Equally divide the image in m rows and n columns;

- Calculate the intensity sum profile of the rows and of the columns, and find the minimum sum of each of the rows and each of the columns (see Fig. 4(a));

- Find the maximum intensity sum in between each of the limits determined on the previous step (see Fig. 4(b)).
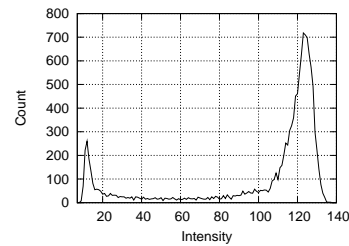
It is very important that this division is made correctly because, in this way, it will be possible to know which macroarray cell is being analyzed and, more important, it will be possible to perform regional processing of the image in terms of spot detection and quantification.

4.2 Image Thresholding

For a good quantification of the spot, one of the most important steps, if not the most important, is to decide what is background and what is foreground or, in our case, what is the background and what is the spot. That may be achieved by a technique called Thresholding. Image thresholding consists of setting an intensity level that will define which pixels belong to the foreground. Only the pixels below a given intensity level, the threshold level, will be maintained. As mentioned on Section 2, these images may present a gradient in the background. This makes it impossible to define a threshold level

for the whole image. Nevertheless, because a grid that separates each spot into a cell has been defined, it is possible to apply a regional threshold level or, in other words, apply a threshold to each of the cells. As each experiment may have many probes, each probe may have many macroarrays, and each macroarray may have hundreds of spots, it is not convenient for the scientist to be choosing one threshold level for each cell on each macroarray.

After analyzing the histograms of the image of each cell on the macroarrays, it was obvious that there were two distinct classes (hills) and, so, the optimum threshold level would be the one in the middle of those two classes (see Fig. 5(b)). For more details see [1].

The more intuitive method to find the intensity level that separates the two hills, is to find the minima between the two paramounts. That is not trivial because the gradient of the histogram is always changing its signal. To try to solve that problem, one could smooth the histogram and, luckily, if the chosen kernel was the right size and both hills had enough representativity not to be swept by the smoothing operation, it would be possible to find that minima.

A more robust technique was developed by [1]. It consists of weighting the histogram. To better understand this idea a brief explanation is given.

Assuming that an image is a 2D grayscale intensity function with N pixels with graylevels from 0 to L and the number of pixels of graylevel i denoted by f_i, lets consider an histogram with two "hills". If we imagine the histogram on a scale, we will get what is represented by Figure 6.

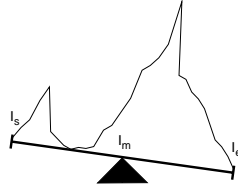It is possible to see that the scale is not balanced. There is

Figure 6: Histogram on a scale.

(a) Problem.

(b) Correction.

Figure 8: Corrected with minimum representativity.

more weight on its right side than on its left side. If weight is removed from the heavier side until the lightest side weights more and the indexes are adjusted, the scale will fall to the opposite side. If this process keeps repeating, the base of the scale (the triangle in Fig. 6) will end up at the intended position.

To present it in a more formal way, let I_s be the first grayscale intensity occurrence and I_e the last one. The base of the scale will be always placed at I_m, as defined by $I_m = \frac{I_s + I_e}{2}$. The weight functions of the left and right sides of the scale are defined, respectively, as $W_l = \sum_{i=I_s}^{I_m} f_i$ and $W_r = \sum_{i=I_m+1}^{I_e} f_i$, so that initially $W_l + W_r = N$.

Now we can define the following algorithm:
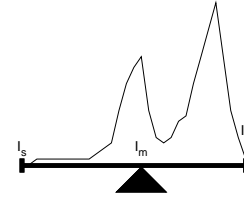
Algorithm 4.1: GET-THRESHOLD(f, I_s, I_e)

$I_m \leftarrow \frac{I_s + I_e}{2}$
$W_l \leftarrow \sum_{i=I_s}^{I_m} f_i$
$W_r \leftarrow \sum_{i=I_m+1}^{I_e} f_i$
while $I_s \neq I_e$
do
 if $W_r > W_l$
  then
   $W_r \leftarrow W_r - f_{I_e}$
   $I_e \leftarrow I_{e-1}$
   if $\frac{I_s + I_e}{2} < I_m$
    then
     $W_l \leftarrow W_l - f_{I_m}$
     $W_r \leftarrow W_r + f_{I_m}$
     $I_m \leftarrow I_{m-1}$
 else if $W_l \geq W_r$
  then
   $W_l \leftarrow W_l + f_{I_s}$
   $I_s \leftarrow I_{s+1}$
   if $\frac{I_s + I_e}{2} > I_m$
    then
     $W_l \leftarrow W_l + f_{I_m+1}$
     $W_r \leftarrow W_r - f_{I_m+1}$
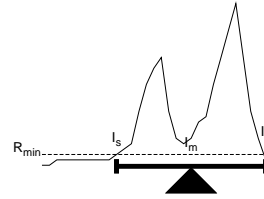 $I_m \leftarrow I_{m+1}$
return (I_m)

Figure 7 illustrates the result of applying algorithm 4.1 to the histogram presented on Figure 6.

As illustrated by Figure 8(a), in some cases there may be a problem when applying this algorithm. If the histogram, due to noise in the image, has tails at the left hand side of the first peak or at the right hand side of the second peak, the indexes may be misplaced. This will result in the index I_m "jumping" out of the middle of the two peaks, producing a bad result. This can be simply solved by passing a parameter R_{min} defining the minimum height that is to be accounted in the histogram before the first peak and after the second peak. This will allow to place I_s at the first occurrence of $f_i = R_{min}$ if moving across the histogram from intensity level 0 to L, and I_e if moving across the histogram from L to 0.

Algorithm 4.1 is very fast with $O(L)$ complexity in the worst case. Even if there are algorithms with the same order of complexity, like Otsu's, the execution time of the presented algorithm is better. Nevertheless, this algorithm may fail when the spot is very faint because in these cases the histogram may present only one peak (see Fig. 9).

For those cases, a different approach was taken. It consists of performing dynamic thresholding using the Otsu's thresholding algorithm for each pixel in the cell.

4.3 Filtering the cell

In most cases, even after the thresholding operation, there will still be some particles remaining from the background. Thus, the spot has to be distinguished from those particles. Two measurements are used to accomplish that task. One is the eccentricity and the other is the circularity. The eccentricity is based on the second order central moments and is calculated
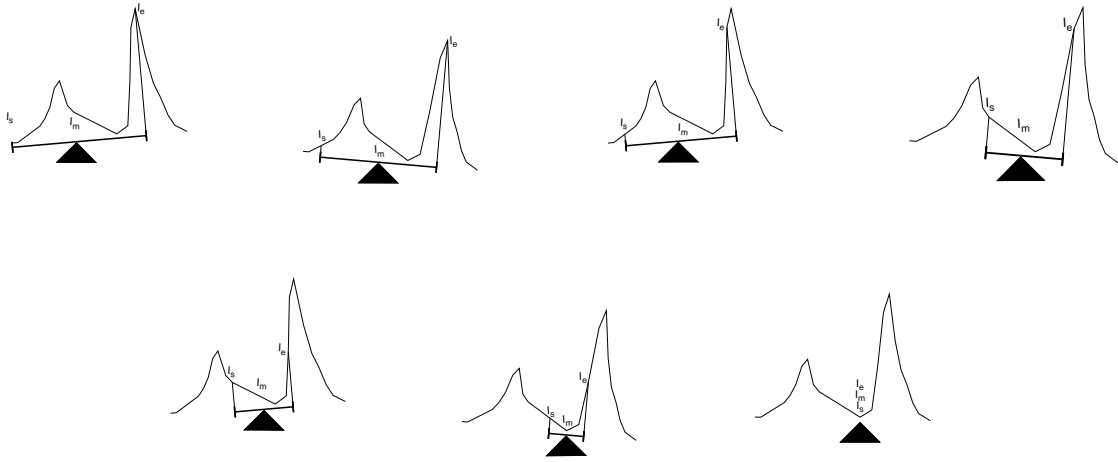
Figure 7: Applying algorithm 4.1.

(a) Faint Spot.

(b) Histogram.

Figure 9: Algorithm 4.1 will fail in this case.

using equation 1.

$$\epsilon = \frac{(\mu_{2,0} - \mu_{0,2})^2 - 4\mu_{1,1}^2}{(\mu_{2,0} - \mu_{0,2})^2} \quad (1)$$

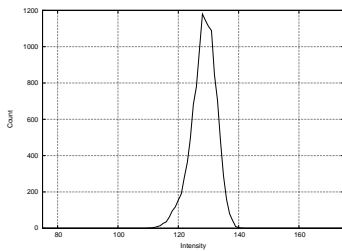The value of ϵ may vary between 0 and 1, where $\epsilon = 0$ indicates a perfectly round object and $\epsilon = 1$, a line shaped object [4]. The formula used to calculate the circularity is formula 2.

$$k = \frac{p^2}{4\pi A} \quad (2)$$

On formula 2, p represents the perimeter of the component being measured and A represents its area. For a perfect circle, k will be equal to 1 and greater than 1 for other objects. After analyzing hundreds of spots in regard to this measurements, some conclusions were taken. First, each measure by itself is not completely reliable for this problem. So, both measurements complemented each other. Figure 10(a) and Figure 10(b) show components with very similar eccentricities although the cell in Figure 10(b) does not have a valid spot. Circularity had to be used to exclude the invalid components. Figure 10(c) shows a component with a quite bad circularity but with a very good eccentricity. Given these results the approach that was taken was to accept all those with an eccentricity lower that 0.0005 and reject all the components with an eccentricity higher than 0.5. Eccentricities in between 0.0005 and 0.5 are subjected to verification of circularity. From these, the components that have circularity higher than 1.5 are rejected. This values were determined empirically.

4.4 Subtracting the background

Due to the process of exposure of the radioactive membrane to the films, the time and conditions of exposure and develop-

(a) E:0.0239; C:1.30.   (b) E:0.0253; C:5.00.   (c) E:0.0003; C:1.96.

Figure 10: Eccentricity Vs Circularity.

ment, and even the nature of the film, the resulting film will look darker than it is in reality. This will result in spots having an intensity that is the sum of the real spot intensity with the background intensity. The present application offers the possibility to correct that problem. In summary, it generates a synthetic background that will be subtracted from the image at the end of all processing. The synthetic background is generated by using the information gathered in the Section 4.1. As shown in Figure 4(b) there are areas known, for sure, to be background only. Using the lines that divide the cells, it is possible to generate a background by interpolation. To avoid pixels that are affected by long tailed noise such as salt-and-pepper noise, in each position of the grid it is calculated the median. The result will be a "medianized" grid with representative values of the background. Using the intensity values at the position of the vertical lines of the grid, the intensity values for each row in between each of the vertical lines, is calculated by interpolation. The interpolated intensity level is calculated by the linear equation $y = mx + b$, where b is the intensity value of the first grid line being used at the pixel row being generated, m is the slope between the intensity values of the grid lines at the pixel row being generated, and y is the interpolated intensity. The same process is used vertically, using the horizontal lines of the grid, to generate an interpolated vertical background. After having both backgrounds, it is calculated an average background. This will improve the estimation of the synthetic background. Figure 11(b) represents the final generated background and it can be seen that the synthetic background is quite representative of the real background.

4.5 Equalizing the films

Due to different levels of exposure, different films will often have different corresponding signal intensities. The result is that, when comparing the signal intensities among the different films, the results will be misleading. For that reason, it is mandatory that, previously to any comparison, the films are equalized. Obviously, this is a time consuming task and, for someone without any experience in image processing, this will be difficult to accomplish. To avoid manual equalization of the films, a process to automatically equalize them has been developed. It consists of calculating the average brightness of each film, excluding the areas of the previously detected spots
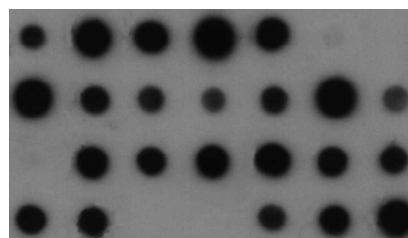
(a) Original image.
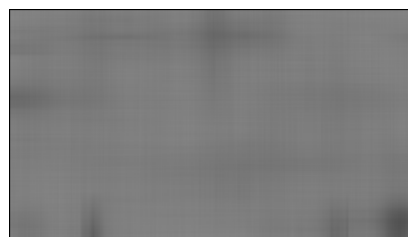
(b) Synthetic background.

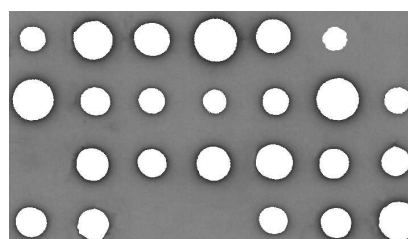Figure 11: Macroarray film and approximate background.

Figure 12: Background only.

(see Fig. 12). In this way it is calculated the average brightness of the background of each film and, thus, there can be established a correlation between level of brightness among all films. That information is used to equalize the results.

4.6 Reporting the results

Two kinds of reports are generated by this tool. The first one is produced for each of the films and, the second, is a global report. A sample of the first report mentioned can be seen in Table 1.

At the bottom of the report, an image of the processed film is also shown (see Fig. 13(a) or Fig. 13(b)), for spot identification purposes.

Although the data in the report is intended to be quite simple to understand, it is worth to discuss how some of the fields are determined. For instance, the field Area is the result of the multiplication of the number of pixels of the spot by the area of each pixel. The area of each pixel is calculated from the resolution of the image and that is done automatically without the

Table 1: Part of the report for each film.

Spot (x,y)	# of pixels	Area (cm^2)	Integrated Intensity	Relative Intensity	Class (1-20)
(1,1)	1574	5.25	351895	2.50%	7
(1,2)	3641	12.14	866699	6.15%	16
(1,3)	2902	9.67	688067	4.88%	13
(1,4)	4486	14.95	1075301	7.63%	20
(1,5)	2869	9.56	678416	4.81%	13
(1,6)	1412	4.71	187220	1.33%	3
(1,7)	0	0.00	0	0.00%	N/A
(2,1)	4081	13.60	964232	6.84%	18
(2,2)	2052	6.84	475476	3.37%	9
(2,3)	1640	5.47	363844	2.58%	7
(2,4)	1352	4.51	281412	2.00%	5
(2,5)	1737	5.79	398881	2.83%	7
(2,6)	4315	14.38	1021936	7.25%	19
(2,7)	1685	5.62	337142	2.39%	6
(3,1)	0	0.00	0	0.00%	N/A
(3,2)	2569	8.56	613734	4.35%	11
(3,3)	1976	6.59	466974	3.31%	9
(3,4)	2819	9.40	673115	4.77%	13
(3,5)	3145	10.48	746663	5.30%	14
(3,6)	2196	7.32	519574	3.69%	10
(3,7)	2003	6.68	464221	3.29%	9
(4,1)	2257	7.52	528679	3.75%	10
(4,2)	2374	7.91	551845	3.91%	10
(4,3)	0	0.00	0	0.00%	N/A
(4,4)	0	0.00	0	0.00%	N/A
(4,5)	1895	6.32	420883	2.99%	8
(4,6)	2247	7.49	533388	3.78%	10
(4,7)	3757	12.52	888709	6.30%	17

Table 2: Summary report (uncolored).

Spot (x,y)	Probe 00		Probe 00	
	Film 00	Film 01	Film 00	Film 01
(1,1)	7	7	7	6
(1,2)	16	17	12	12
(1,3)	13	13	13	12
(1,4)	20	20	20	20
(1,5)	13	12	13	12
(1,6)	3	3	3	3
(1,7)	N/A	N/A	N/A	N/A
(2,1)	18	18	18	17
(2,2)	9	9	9	9
(2,3)	7	7	7	7
(2,4)	5	5	N/A	N/A
(2,5)	7	7	7	7
(2,6)	19	19	19	18
(2,7)	6	6	6	6
(3,1)	N/A	N/A	N/A	N/A
(3,2)	11	11	5	5
(3,3)	9	9	9	9
(3,4)	13	12	13	12
(3,5)	14	14	14	14
(3,6)	10	10	10	10
(3,7)	9	9	9	9
(4,1)	10	10	10	10
(4,2)	10	10	9	9
(4,3)	N/A	N/A	N/A	N/A
(4,4)	N/A	N/A	N/A	N/A
(4,5)	8	8	8	7
(4,6)	10	10	10	10
(4,7)	17	16	17	16
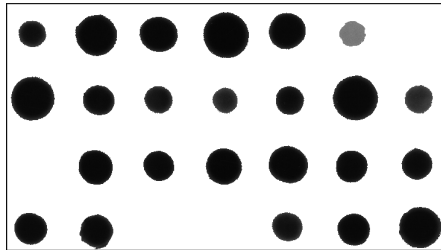
(a) Processed without background subtraction.
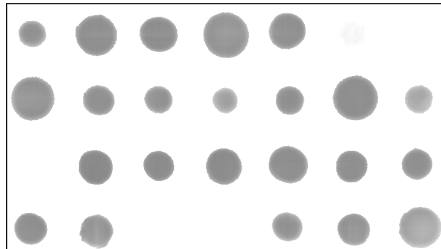
(b) Processed with background subtraction.

Figure 13: Original and processed images.

need to input the resolution of the image or the size of a pixel. The column Integrated Intensity, also called Optical Density in other publications [5], is calculated by $\sum_{x,y} I(x,y)$, with $I(x,y)$ representing the graylevel intensity at the coordinates x and y. (x,y) represents all the pixels belonging to the spot. The Relative Intensity is relative to the sum of the intensities of all the spots in the image. Also, the Class of intensity of the spot is determined according to the Integrated Intensity of the spot and the number of defined classes, which is a user defined parameter.

The global report is more straightforward to understand than the film report. It allows the user to find patterns, similarities or differences in a fast way. The class to which the spot belongs is placed side by side according to its probes and film. On top of that, each class is colored differently. The uncolored version of the report is presented in Table. 2.

5 Conclusion

In this article we have presented an algorithm for automatic spots densitometry analysis and comparison that can be used in image analysis within macroarray hybridization technology, giving a classification according to gene expression patterns. The algorithm demonstrates high level of robustness and flexibility in terms of macroarray design and size, if the analyzed images do not deviate strongly from an overall grid of regular spots. The user intervention is minimal, just providing the size of the array (number of spots in horizontal and vertical directions of the array). In the parameters section it is possible to control the background subtraction, apply two

Table 3: Comparison of the results of processing image in Figure 14.

| Prepared Dilutions	Our Software					Quantity One v.4.6.5	
	# of Pixels	Area (cm^2)	Integrated Intensity	Relative Intensity	Class (1-20)	Volume INT*mm^2	%Adjusted Volume
1:1	27165	17.82	6417520	48.12%	20	1236	50.44075029
1:100	7760	5.09	1757682	13.18%	5	300	12.23835282
1:10000	97	0.06	8145	0.06%	N/A	37	0.01
1:10	12498	8.20	2839304	21.29%	9	472	19.2736867
1:1000	1772	1.16	268186	2.01%	1	42	1.747317103
1:5	8807	5.78	2046901	15.35%	6	399	16.2877317
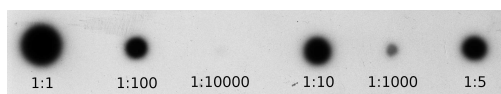
Figure 14: Dilutions used for testing.

different methods of segmentation, although these do not normally need adjusting. Practical usage of this program also revealed to be very intuitive and the entire process from film scanning to software processing is very fast, taking 5 minutes at maximum. The results are presented in a clear way and in a couple of minutes the user can evaluate different patterns between two different membranes, outperforming academic and commercial packages, with some extra functions like the class value, whereas for good quality images its efficiency is comparable (see Table 3). Other applications are possible due to the flexibility of the software allowing the usage of this application in a variety of fields in biology, specially in molecular biology and biochemistry. Even so, a final evaluation of this software should take in consideration that it is just an alternative to provide an easy approach to evaluation of patterns of gene expression on macroarrays; being at the same time a user friendly and free software.

6 Acknowledgements

References

[1] Anjos, A. and Shahbazkia, H., Bi-Level Image Thresholding - A Fast Method. BIOSIGNALS 2008, Vol:2, Pages:70-76, (2008).

[2] Miura, K. et al, Expression profiling-based identification of CO_2-responsive genes regulated by CCM1 controlling a carbon-concentrating mechanism in Chlamydomonas reinhardtii, Plant Physiology, Vol:135, 3, p:1595-1607 (2004)

[3] Kucho, K. et al, Ident. of novel clock-controlled genes by cDNA macroarray analysis in Chlamydomonas reinhardtii. Plant Molecular Biology. Vol:57, Pages:889-906 (2005).

[4] Kilian, J., Simple Image Analysis by Moments, OpenCV lib documentation (2001).

[5] Dowsey, AW, Dunn, MJ, Yang, GZ, The role of bioinformatics in two-dimensional gel electrophoresis, Proteomics, Vol:3, Pages:1567-1596, ISSN:1615-9853 (2003).

[6] Kumari, Sumita and Panjabi and Kushwaha, Hemant and Sopory, Sudhir and Singla-Pareek, Sneh and Pareek, Ashwani, Transcriptome map for seedling stage specific salinity stress response indicates a specific set of genes as candidate for saline tolerance in Oryza sativa L., Functional & Integrative Genomics, 9(1), Pages:109-23 (2009).

[7] Raymond, Josette AND Thiberge, Jean-Michel AND Kalach, Nicolas AND Bergeret, Michel AND Dupont, Christophe AND Labigne, Agnès AND Dauga, Catherine, Using Macro-Arrays to Study Routes of Infection of *Helicobacter pylori* in Three Families, PLoS ONE, Public Library of Science, Vol: 3, Pages:e2259 (2008/05).

[8] Cátia L. Marques, Marta S. Rafael, M. Leonor Cancela, and Vincent Laizé. Establishment of primary cell cultures from fish calcified tissues. Cytotechnology. 2007 September; 55(1): 9–13.

[9] N. Otsu, A Threshold Selection Method from Gray-Level Histograms, Journal of IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-9, no.1, 62-66, 1979.

[10] Diachenko L, Lau YF, Campbell AP, Chenchik A, Mogadam F, Huang B, Lukyanov S, Lukyanov K, Gurskaya N, Sverdlov ED, Siebert D. (1996) Suppression Subtractive Hybridization: A method for generating differentially regulated or tissue-specific cDNA probes and libraries. Proc. Natl. Acad. Sci. USA 93(12): 6025-6030. PMID:8650213.