

# *Statistical Applications in Genetics and Molecular Biology*

---

*Volume 3, Issue 1*

2004

*Article 11*

---

## Saturation and Quantization Reduction in Microarray Experiments using Two Scans at Different Sensitivities

Jorge García de la Nava\*      Sacha van Hijum<sup>†</sup>  
Oswaldo Trelles<sup>‡</sup>

\*Dept. Computer Architecture, University of Málaga (Spain), [gdl@ac.uma.es](mailto:gdl@ac.uma.es)

<sup>†</sup>Department of Molecular Genetics, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, The Netherlands, [S.A.F.T.van.Hijum@biol.rug.nl](mailto:S.A.F.T.van.Hijum@biol.rug.nl)

<sup>‡</sup>Dept. Computer Architecture, University of Málaga (Spain), [ots@ac.uma.es](mailto:ots@ac.uma.es)

Copyright ©2004 by the authors. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher, bepress. *Statistical Applications in Genetics and Molecular Biology* is produced by The Berkeley Electronic Press (bepress). <http://www.bepress.com/sagmb>

# Saturation and Quantization Reduction in Microarray Experiments using Two Scans at Different Sensitivities \*

Jorge García de la Nava, Sacha van Hijum, and Oswaldo Trelles

## Abstract

We present a mathematical model to extend the dynamic range of gene expression data measured by laser scanners. The strategy is based on the rather simple but novel idea of producing two images with different scanner sensitivities, obtaining two different sets of expression values: the first is a low-sensitivity measure to obtain high expression values which would be saturated in a high-sensitivity measure; the second, by the converse strategy, obtains additional information about the low-expression levels. Two mathematical models based on linear and gamma curves are presented for relating the two measurements to each other and producing a coherent and extended range of values. The procedure minimizes the quantization relative error and avoids the collateral effects of saturation. Since most of the current scanner devices are able to adjust the saturation level, the strategy can be considered as a universal solution, and not dependent on the image processing software used for reading the DNA chip. Various tests have been performed, on both proprietary and public domain data sets, showing a reduction of the saturation and quantization effects, not achievable by other methods, with a more complete description of gene-expression data and with a reasonable computational complexity.

**KEYWORDS:** microarray, preprocessing, saturation, quantization

---

\*This work has been partially supported by grant QLK3-2001-01473 under the EU sub-programme area “Quality of Life and Management of Living Resources” - Key Action “The Cell factory”. The authors would like to thank Carlos Óscar Sánchez Sorzano for carefully reading the manuscript. We specially thank all people who provided data to us: Nathalie Goupil Feuillerat and Celine Gobert (Danone Vitapole, France), Charlotte Barrière and Eric Guédon (Institut National de la Recherche Agronomique, France) and Ana Dopazo from the Spanish National Center for Oncologic Research (CNIO).

## 1.- Introduction

Gene-expression levels are used to determine the response of an organism to a particular environmental condition. The DNA microarray technology (see Schena *et al.*, 1995) has enabled the simultaneous analysis of thousands of gene transcriptions in different developmental stages, tissue types, clinical conditions, organisms, etc. The availability of such expression data affords insight into the functions of genes as well as their interactions.

However, measuring these expression levels, like any other analytical methodology, has its propensity to error. Much effort has gone into data pre-processing research for countering the many sources of systematic and random variation (for instance, see Dudoit *et al.*, 2001, Finkelstein *et al.*, 2002, Yang *et al.*, 2002, and Ideker *et al.*, 2000). In this work we have centred our attention on two sources of systematic errors that perturb measurements: saturation and quantization, which mostly arise from limitations of the data acquisition device.

The saturation and quantization problems observed in the microarray scanning devices context have been partially addressed by several authors. Dudley *et al.* (2002) deal with increasing the dynamic range of measures through a new methodology for experiments: “First, instead of cohybridizing labeled experimental and control samples, we hybridize each sample against labeled oligos complementary to every microarray feature”. This means that experiments that follow the standard methodology of Schena *et al.* (1995) cannot take direct advantage of their work.

Romualdi *et al.* (2003) propose multi-scanning at image level seeking a higher contrast but not oriented to error reduction. A drawback of this method is the need of aligned images which actually are not easy to produce. Lyng *et al.* (2004) focus in the non-linear aspects of saturation in PMTs and corrections for them, but not on quantization.

Our proposed procedure covers both aspects through a general and robust model suitable for a broad collection of scanning devices presenting different saturation behavior. It is suitable to the standard methodology and it is applied after image processing, avoiding the disadvantages of previous works.

### 1.1.- Saturation and quantization

Saturation is a consequence of the finite range of the acquisition device, which renders the relation between measured intensity and real intensity non-linear. Signals whose values lie outside the dynamic range are forced inside by device lim-

its. Since the dynamic range of the gene expressions is large, saturation in signals of greater intensity (bright spots) is a recurring phenomenon.

Quantization occurs when dealing with finite precision (i.e. when digitizing). All the possible and unlimited physical values have to be encoded by a reduced set of discrete values. These discrete values are called symbols and the process of mapping each interval of continuous values into its corresponding symbol is called quantization (see Gray and Neuhoff, 1998). For instance, when a slide is scanned at low sensitivity, the signals of lower intensity (dark spots) may occur in the interval of the first symbol. Usually, this symbol is zero which means utterly black. The information that these dark spots could provide would subsequently be lost when reading them as black.

## 1.2.- Proposal

We propose a novel idea of producing two different observations (i.e., two images) from a DNA chip using different scanning sensitivities, thus obtaining two different sets of read intensity values. A scheme of this technique is depicted in Figure 1. The first observation is a low sensitivity measure ( $L'$ ) useful for obtaining a non-saturated measure of bright spots. The second one, a high sensitivity measure ( $H'$ ), yields a better definition of the dark spots avoiding quantization to zero.

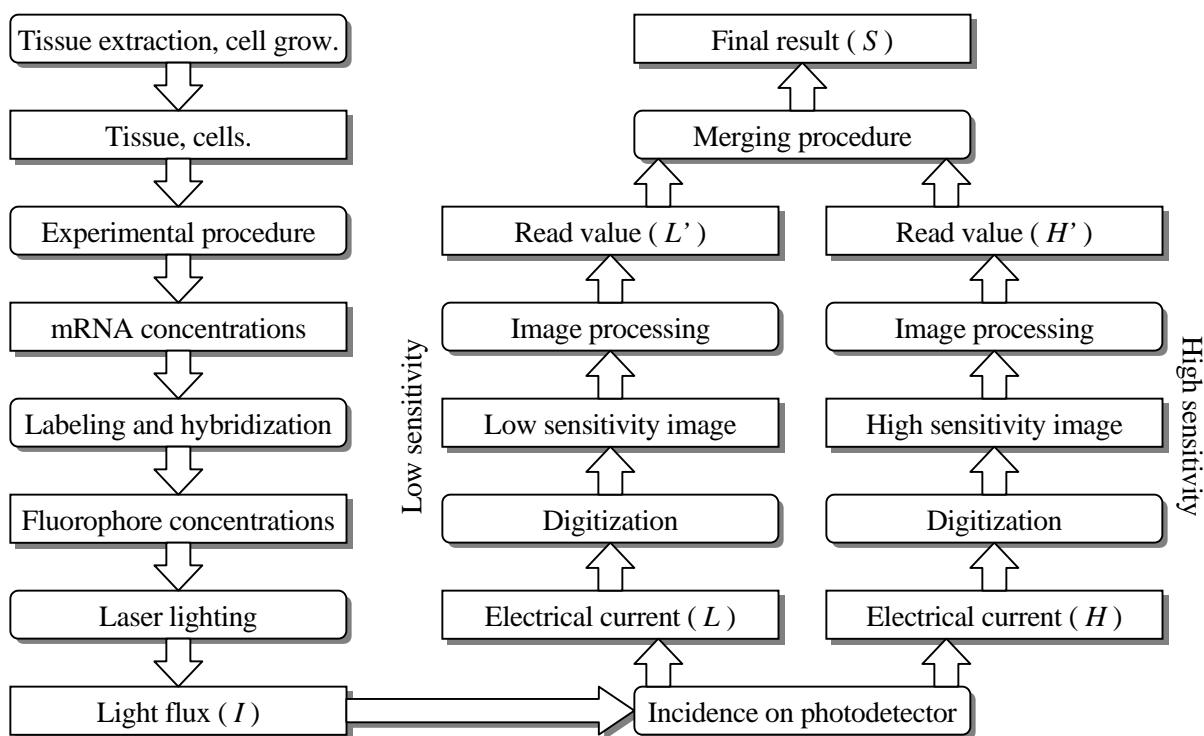
An important question arises now regarding the method for producing a single result from both images ( $L'$  and  $H'$ ). To solve this and attain a result value ( $S$ ) as proportional as possible to the actual light flux ( $I$ ), a robust mathematical model is presented for relating the two scans to each other and producing a less saturated, less affected by quantization, coherent and extended range of values.

A user friendly implementation of this proposal have been included in our PreP application (García de la Nava *et al.*, 2003). The 2Scan algorithms here described were used for processing data from several sources, and results showed an increase of quality in these data by a reduction of the saturation and quantization effects.

## 2.- System and Methods

### 2.1.- Gene-expression basis

DNA microarray slides or chips typically consist of thousands of spots containing immobilized DNA molecules. Each spot is placed at a defined location on the slide or chip surface and has the (or a part of the) DNA sequence corresponding to a specific gene. The mRNA molecules of a certain experimental condition are reverse transcribed into cDNA and labeled either directly or indirectly with a fluorescent dye (red and green fluorophores). The labeled cDNA is subsequently hybridized to the DNA microarray. The signals of the red and green fluorophore intensities are scanned by a chip scanner, producing a two channel image. After image analysis the measures of the red and green intensities for each spot in the array serves as tabulated input for the data processing pipeline (see Duggan *et al.*, 1999).



**Figure 1.**-The proposed procedure is represented in this U-shaped diagram. Rounded boxes are processes and squared ones are measures. The downward part is the common microarray procedure leading to a lighted spot that emits photons in a light flux which is somehow related to the original mRNA concentrations. The upward part is the variation proposed. Two images are scanned using different sensitivities in the photodetector device. After digitization and image processing a merging procedure is applied which gives as result an estimated proportional value to the input light flux, avoiding saturation and quantization effects of photodetector and digitalization.

## 2.2.- Saturation model

When a device is measuring a signal that is too intense, it becomes saturated and reports the maximum value it is able, even though the actual signal is higher. This is the ideal behaviour of saturation which is called ‘clipping’. Clipping appears in some devices such as analog-to-digital converters (digitizers). Commonly the transition to saturation is more gradual and depending on its characteristics and the mathematical curve it follows, the saturation is said to be sigmoid, logarithmic, gamma, etc.

This proposal focuses on clipping and gamma saturation, the latter being a good approximation for the more complex saturation of Photodetector devices (see Lyng *et al.*, 2004, for a deeper study) and other optoelectrical devices (see Poynton, 1993). For all the scanner devices we analysed, either clipping or gamma saturation was found.

The saturation model we propose for developing the merging algorithms is described in the following equations. In them,  $I_i$  is the light flux intensity of spot  $i$ .  $L_i$  stands for the low sensitivity electrical current at photodetector and  $H_i$  for the high sensitivity one. The  $i$  subscript refers to the  $i$ -th spot. No channel distinction is done since both are to be treated independently.

Saturation is assumed to be negligible in the low sensitivity scan, so the value of  $L_i$  is given in linear terms:

$$L_i = kI_i \quad (1)$$

On the other hand,  $H_i$  is described by either a clipped linear curve or a gamma curve equation.

### 2.2.1.- Clipping saturation

In the first case, the clipped linear curve is described by:

$$H_i = \begin{cases} pI_i & \text{if } I_i < H_M / p \\ H_M & \text{otherwise} \end{cases} \quad (2)$$

$p$  being the proportionality constant between read value (high sensitivity) and spot intensity. The previous equation can be reduced to:

$$H_i = \begin{cases} mL_i & \text{if } L_i < H_M / m \\ H_M & \text{otherwise} \end{cases} \quad (3)$$

Where  $H_M$  is the saturation level (clipping level) and:

$$m = p / k \quad (4)$$

the proportionality constant between the low sensitivity and high sensitivity scans.

### 2.2.2.- Gamma saturation

In this second case, the gamma curve is defined by:

$$H_i = cI_i^\gamma \quad (5)$$

The relation between both  $H_i$  and  $L_i$  can be specified by the following equation:

$$H_i = bL_i^\gamma \quad (6)$$

Where:

$$b = c / k^\gamma \quad (7)$$

### 2.3.- Quantization model

Classically, quantization introduces an error –termed quantization noise– to the original signal, as Widrow *et al.* (1996) and Gray and Neuhoff (1998) explain. Quantization noise is centred, uniform and highly dependant on the signal to digitize but, if that signal has an underlying noise large enough, the quantization noise for each pixel becomes almost independent of it. This approximation is developed by Widrow (1961) and applied by Vanderkooy and Lipshitz (1987) on dithering. Also, since a high number of pixels per spot is present, it can then be assumed that the central limit theorem holds (see Walpole *et al.* 2002 or any other introductory text in statistics) and that their average intensity follows a normal distribution. Since the digitizer (i.e. the scanner) is the same in both measures, it can be supposed that the quantization noise is identically characterized in high and low sensitivity images too.

Then, if averaging is used in the image processing software, the quantization noise can be expressed by the following equations:

$$L'_i = L_i + \varepsilon_{L_i} \quad (8)$$

$$H'_i = H_i + \varepsilon_{H_i} \quad (9)$$

$$\sigma_{\varepsilon_{L_i}} = \sigma_{\varepsilon_{H_i}} = \sigma \quad (10)$$

$$\mu_{\varepsilon_L} = \mu_{\varepsilon_H} = 0 \quad (11)$$

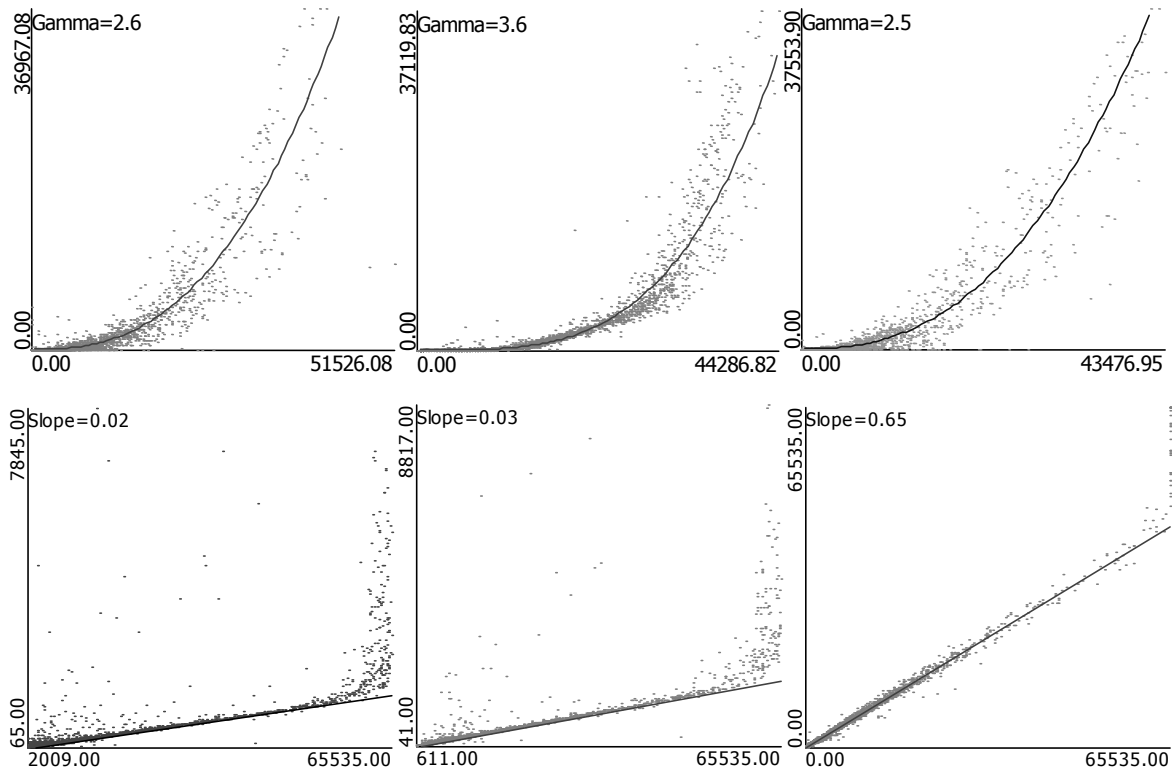
Where  $\varepsilon$  is the quantization noise for each measure,  $\mu$  its mean and  $\sigma$  its standard deviation. The variables with prime are the digitally read ones while values without prime are taken before quantization. A single scan contains thousands of spots and the above equations are applied to every spot in both the low and high sensitivity scans.

## 2.4.- Parameter Estimation Method

The model described above is only useful in practical terms if the parameters can be determined or, at least, estimated. This can be achieved by a statistical regression but, in view of the fact that outliers should be discarded, a robust regression is recommended (see Rousseeuw. and Leroy ,1987). For clipping saturation, the regression should fit the curve of equation (3) with  $m$  being the parameter to estimate. In the case of gamma saturation, the regression should fit the curve of equation (6) with  $\gamma$  and  $b$  the parameters to be estimated. See Figure 2 for regression examples on test data sets. Furthermore, the quantization noise deviation could also be estimated, but it is not necessary for the designed algorithms, as will be shown later.

## 2.5.- Practical considerations

Commercially available devices are able to provide the user with a wide range of configuration options e.g. laser power and PMT voltage are easily found. It has been described that powerful laser light (from some scanner devices) destroys the fluorescent molecules. This effect is named photobleaching (see Song *et al.*, 1995) and it is commonly stronger for the red dye than for the green dye. The models presented here do not take into account the effects of photobleaching. However, these photobleaching effects can be reduced by first scanning the red and then the green channel. Furthermore, the high sensitivity scan should be performed before the low sensitivity scan.



**Figure 2.-** Regression plots from data of six DNA microarray slides, measured at two different sensitivities. The vertical axis being the low-sensitivity measure and the horizontal axis the high-sensitivity measure. The dependency between both measures follows closely a gamma curve in the RL and RC data sets, and a clipped linear curve in the other data sets. The top graphs are from the RL data set, green channel, three different subsets. The bottom graphs are from D data set (low versus high sensitivities) red and green channels, the last graph being from the IR data set (low versus high sensitivities).

On the other hand, although the image acquisition time (i.e. scanning) can vary depending on the scanner device, the image resolution and the DNA microarray size, all our experimental partners (Danone-Vitapole; INRA-France; Molecular Genetics of the University of Groningen, the Netherlands; CNIO-Spain; CNB-Spain) have reported us that the acquisition time is not longer than two to five minutes per channel. This period is short enough for discarding slide degradation during it.

## 2.6.- Estimation

Since  $L_i$  is thought to be proportional to  $I_i$  and it lacks of saturation, it will be the value to estimate. In this estimation quantization must also be avoided. This way both saturation and quantization are removed or, at least, minimized. The estima-

tion of  $L_i$  will be written as  $S_i$  and it is calculated depending on the saturation type and the minimization approach.

### 2.6.1.- Gamma saturation, threshold approach

Once the model parameters are found, it is possible to estimate the low sensitivity value, represented by  $S_i$ , applying equation (6) on the measured and quantized high sensitivity signal, symbolized by  $H'_i$ :

$$S_i = (H'_i / b)^{1/\gamma} = \left( \frac{H_i + \varepsilon_{H_i}}{b} \right)^{1/\gamma} \quad (12)$$

Which can be developed, assuming that the quantization noise amplitude is small, in Taylor series up to the first power of  $H_i$ , as shown in the following equation:

$$S_i \approx (H_i / b)^{1/\gamma} + \varepsilon_{H_i} \left[ \frac{1}{b\gamma} (H_i / b)^{\frac{1}{\gamma}-1} \right] \quad (13)$$

Which can be rewritten as:

$$S_i \approx L_i + \varepsilon_{H_i} F(H_i) \quad (14)$$

The multiplicative factor  $F(H_i)$  is due to the quantization noise and is dependent on  $H_i$ . When this factor is less than one, the effect of the quantization noise in  $L_i$  is less than in  $L'_i$ . The method chooses the value with least noise for each spot, using  $H_i$  for deciding this.

To find the threshold of  $H_i$  that will lower  $F(H_i)$  under one, the following calculations are made:

$$F(H_i) < 1 \quad (15)$$

$$\frac{1}{b\gamma} (H_i / b)^{\frac{1}{\gamma}-1} < 1 \quad (16)$$

Because  $\gamma$  lies between 0 and 1 in the saturation zone and  $\frac{1}{\gamma}-1$  is a positive value,  $H_i$  can be found:

$$H_i < b(b\gamma)^{\frac{\gamma}{1-\gamma}} = T \quad (17)$$

The threshold  $T$  is obtained as follows. Since the real  $H_i$  value cannot be used, we can only compare the threshold against  $H_i'$ . One should consider that this expression is independent of the noise deviation.

The algorithm that this approach suggests consists of the following steps:

(i) Determine the parameters of equation  $H_i = bL_i^\gamma$  by robust regression. For large data collections (the normal case) this will result in an accurate estimation of parameters.

(ii) Compute the Threshold value  $T = b(b\gamma)^{\frac{\gamma}{1-\gamma}}$ .

(iii) For each measure pair  $(H_i', L_i')$ , take the value given by

$$\begin{cases} S_i = (H_i' / b)^{1/\gamma} & \text{if } H_i' < T \\ S_i = L_i' & \text{otherwise} \end{cases}$$

This value will minimize the quantization noise in the way shown above. However, choosing between estimates, depending on a threshold, is likely to give a higher variance which, in turn, leads to a poorer performance, as demonstrated in section 4.-.

### 2.6.2.- Gamma saturation, maximum likelihood approach

In the previous described model other information that can be used to refine the algorithm has not been used. Under the assumption that the quantization noise is characterized by a normal distribution, it is possible to find the maximum likelihood estimator.

In order to determine this, the same regression procedure is used for obtaining the parameters,  $\gamma$  and  $b$ . However, the expression of the noise will now be described by a probability density function. If the noise is seen as a perturbation of the real signal, and normal distributions are assumed, the equation to describe the probability of the values of both measures given the real values can be written as:

$$p(H_i', L_i' | H_i, L_i) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{d(H_i', L_i', H_i, L_i)}{2\sigma^2}\right) \quad (18)$$

Where  $d$  is the squared distance defined as:

$$d(H'_i, L'_i, H_i, L_i) = (H'_i - H_i)^2 + (L'_i - L_i)^2 \quad (19)$$

In seeking the maximum likelihood estimation, the equation (18) has to be maximized. This is equivalent to minimizing equation (19). If the constraints of the saturation model of equation (6) are forced, it reduces to the following equation:

$$(H'_i - bL_i^\gamma)^2 + (L'_i - L_i)^2 \quad (20)$$

where  $L_i$  is the only variable for solving, its solution ( $S_i$ ) being the maximum likelihood estimation of the real signal. This procedure is also independent of noise deviation. Equation (20) should be minimized by an optimization method. There are plenty of methods for this at Singiresu (2002).

The appropriate algorithm in case of gamma saturation is composed of the following steps:

- (i) Determine the parameters of equation  $H_i = bL_i^\gamma$  by robust regression. For large data collections (which is the normal case) this will result in an accurate estimation of parameters.
- (ii) For each pair of measures ( $H'_i, L'_i$ ) calculate the value of  $L_i$  that minimizes expression  $(H'_i - bL_i^\gamma)^2 + (L'_i - L_i)^2$ . That value is the solution  $S_i$ .
- (iii) This algorithm should iterate along the spots of the slide.

If  $n$  is the number of spots, the total complexity is  $O(n)$ .

### 2.6.3.- Clipping saturation, maximum likelihood approach

We follow the steps and arguments in the previous section up to equation (20). When using the clipping saturation model, this equation becomes:

$$\begin{cases} (H'_i - mL_i)^2 + (L'_i - L_i)^2 & \text{if } H_i < H_M \\ (H'_i - H_M)^2 + (L'_i - L_i)^2 & \text{otherwise} \end{cases} \quad (21)$$

Although both equations must be minimized at the same time, the minimization can be analytically performed.

Dataset	Organism	Scanner	Software	Laser Pwr	PMT Gain	Spots	Saturation
RL	L.Lactis	GeneTac LS IV	ArrayPro 4.5		48-62	5760	Gamma
RC	Lucidea controls	GeneTac LS IV	ArrayPro 4.5		48-62	7776	Gamma
D	L.Lactis	GenePix 4000B	GenePixPro 4.1	100	310-1000	4760	Clipping
IE	B.Subtilis	Virtek ChipReader	ImaGene 5.1	5-100	600-1000	4352	Clipping
IR	L.Lactis	Virtek ChipReader	ImaGene 5.1	15-90	780-850	5760	Clipping

**Table 1.-** Description of the data sets used in this study. The dataset name, source organism or controls and technical information such as scanner model, image processing software, laser power, PMT gain and number of spots are listed. The last column describes the kind of saturation observed in the respective datasets.

$$\begin{cases} S_i = \frac{H'_i m + L'_i}{m^2 + 1} & \text{if } \frac{L'_i m - H_M Z}{1 - Z} < H'_i \\ S_i = L'_i & \text{otherwise} \end{cases} \quad (22)$$

Where  $Z$  is a distance constant:

$$Z = \sqrt{m^2 + 1} \quad (23)$$

The proposed algorithm in case of clipping saturation is formed by the following steps:

- (i) Determine the parameters of equation  $H_i = mL_i$  by robust regression. For large data collections (the normal case) this will result in a good estimation of parameters.
- (ii) For each pair of measures ( $H'_i$ ,  $L'_i$ ) estimate the value of  $L_i$  from equation (22).
- (iii) This algorithm should also iterate along the spots of the slide.

In this case, the complexity is linear, i.e.  $O(n)$ .

### 3.- Results

#### 3.1.- Data description

Data sets and variable-sensitivity images are available as supplementary material for validation purposes as well as the used data acquisition protocol. The proposed algorithms have been applied on several data sets. Table 1 summarizes the relevant parameters of each one of the data sets.

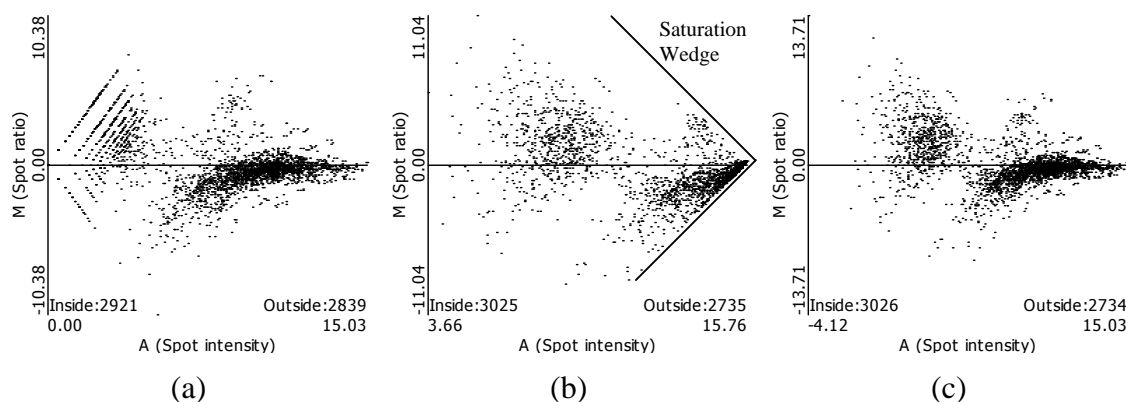
Since we are applying a repeated scanning, two data files will be written by the image processing software. Plotting this files leads us to MA graphs (see Dudoit *et al.*, 2001, for a definition) similar to those that appear in Figure 3 (a) and (b). Saturation is identified by the wedge-shaped at the right which is due to the maximum values for both channels. Quantization is recognized by the regular organization of low intensity data and its greater ratio scattering. It is easily seen that low-sensitivity data suffer from quantization and high-sensitivity data from saturation.

### 3.2.- Analysis

Once the saturation parameters are known by regression techniques (Figure 2), algorithms return the results depicted in Figure 3 (c). In this case, the clipping-saturation algorithm (2.6.3.-) was used. Comparing these results with the original data (Figure 3 (a) and (b) ) reveals that the effects of quantization and saturation were reduced.

The assumptions of gamma and clipping saturation are verified since all regressions fit data reasonably well and this is also confirmed in other data sets.

Further proofs for quantization reduction are shown by the number of infinite-valued spots in above MA graphs. These infinities appear due to a zero or negative (once the background is subtracted) intensity values. Usually this zero is a consequence of quantization and the lack of precision in the digitizer device not an actual black spot. In the graphs of Figure 3, the number infinite-valued spots are displayed by “Outside”. Comparing again, many spots that were outside in the low-sensitivity graph (a) are recovered in results (c).



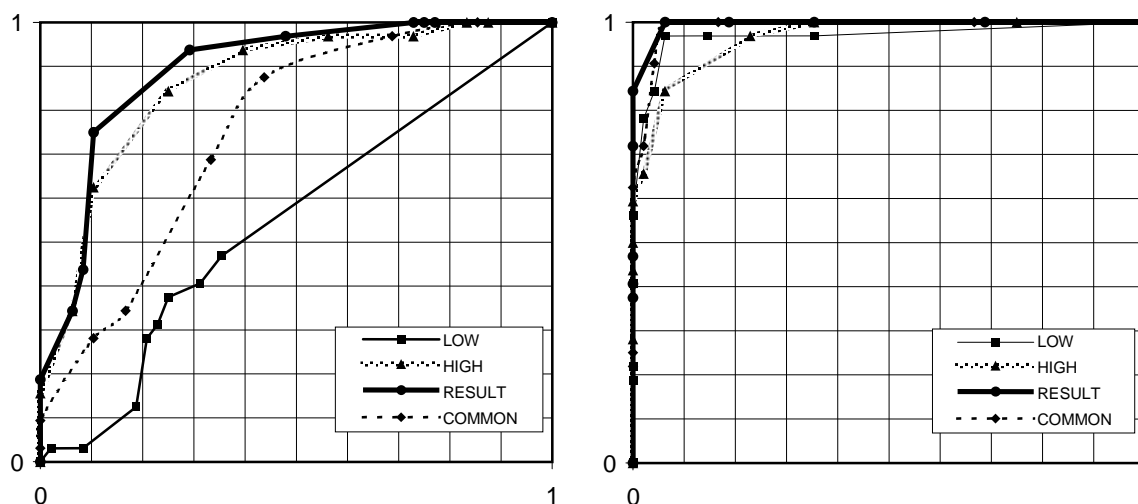
**Figure 3.-** MA graphs from D data set. (a) File “d\_m” (medium sensitivity), (b) File “d\_vh” (very high sensitivity) and (c) Result from clipping saturation, maximum likelihood algorithm. The readings “Inside” and “Outside” count the number of spots in the graph or outside it due to infinities in ratio (zero or negative signal values). All axis scales are logarithm of base 2.

### 3.3.- Differential expression

The aim of DNA-chip experiments is identifying genes that behaves differently between two conditions. Due to the noisy nature of this kind of measures, statistical methods are applied and experimenters have to check the results. Usually, a statistical test leads to the probability of the measure being caused by chance, which is called p-value. If this p-value is low enough, it can be said that the measure is not produced by noise and the gene is differentially expressed. The lower p-value, the higher significance the test has.

What is expected from the 2Scan-improved data is a better detection of differential expressed genes due to the reduction of quantization noise and saturation effects. Both quantization and saturation tends to blur mentioned genes into noise. The former via increasing global noise variance and the latter by reducing dynamic range.

The RC data set is composed of control spots whose expression ratio is known. This allows the construction of ROC curves for comparing the quality of the differential expression test before and after applying the algorithms. Figure 4 shows these curves for the RC data set in both a very low quality image (on the left) and a medium quality image (on the right). The area below the ROC curve measures the performance of the test. When the results of the algorithm are used in the dif-



**Figure 4.-** Two groups of ROC curves from RC data set once preprocessed. Thin lines with squares are the low sensitivity scan ROC curves, dotted lines with triangles the high sensitivity ones, thick lines with circles the results from the gamma maximum likelihood algorithm and dashed lines with diamonds are the commonly used sensitivity.

ferential expression test, the corresponding curve improves.

Additional information about improvement on data quality for the different data sets is presented in supplementary material, including descriptive information on the effect of adjusting the PMT gain, histogram distribution of intensities values, before and after algorithm application, etc.

#### 4.- Conclusions

Several algorithms were introduced for improving the quality of the DNA microarray data depending on the saturation produced by the scanning device. These algorithms are based on obtaining a double scan (with low and high sensitivity) of a DNA microarray slide. A model for saturation and noise is presented from which the proposed procedure arises.

The proposed algorithms use a translation curve that relates the read intensities of the two scans to each other for every spot. A gamma curve was found to be suited for PMT devices and other similar devices. A clipped linear curve was found to be suitable for analog-to-digital-conversion saturation.

For gamma saturation two adequate solutions were implemented and tested: the first algorithm is based on minimizing the error factor of first order when expanding in Taylor series, while the second algorithm is based on a maximum likelihood estimation. The former algorithm, although quicker, introduces discontinuity due to the threshold that is used. The latter algorithm assumes that the noise is Gaussian. This assumption is only valid when there are many pixels in each spot and averaging is used in the image processing software, which is the common case. For a clipped linear curve, the described maximum likelihood algorithm is proposed.

As result, data become more proportional compared to the original sets. In addition, the number of spots with intensity readings is increased, thus maximizing the amount and definition of data from a DNA microarray. The proposed methodology in replicated slides is expected to lead a reduction of the variance reduction and possibly an improved confidence. It was also shown that the complexities for the algorithms are  $O(n)$ . Although the threshold algorithm is theoretically valid, it introduces artefacts and discontinuity close to the threshold. The maximum likelihood algorithm provides a smooth transition between the low sensitivity and the high sensitivity scan at the cost of being more computationally expensive. Because in the clipping saturation case, the maximum likelihood method already supplies an analytical result, the corresponding threshold method was discarded and it is not explained here.

Exhaustive tests have been performed on various data sets, indicating that the above-mentioned procedures are not bound to a specific device and are thus universally applicable, provided that the models are suitable.

## 5.- Supplementary material

Available at <http://chirimoyo.ac.uma.es/bitlab/suppl-2scan/>

## 6.- References

- Deguchi,T., Katoh, N. and Berns, R.S. (1999) Clarification of “Gamma” and the Accurate Characterization of CRT Monitors. *Proc. SID International Symposium*, **30**, 786-789.
- Dudley,A.M., Aach,J., Steffen,M.A. and Church,G.M. (2002) Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *PNAS*, **99**, 7554-7559
- Dudoit,S., Yang,Y.H., Luu,P. and Speed,T.P. (2001) Normalization for cDNA microarray data. *Proceedings of SPIE*, **4266**, 19.
- Duggan,D.J., Bittner,M., Chen,Y., Meltzer,P. and Trent,J.M. (1999) Expression profiling using cDNA microarrays. *Nature Genetics Supplement*, **21**, 10-14.
- Finkelstein,D.B., Gollub,J. and Cherry,J.M. (2002) Normalization and systematic measurement error in cDNA microarray data. *Joint Statistical Meeting 2000. Unpublished manuscript*.
- García de la Nava,J., van Hijum,S.A.F.T. and Trelles,O. (2003) PreP: Gene expression data pre-processing. *Bioinformatics*, **19**, 2328-2329.
- Gray,R.M. and Neuhoff,D.L. (1998) Quantization, *IEEE Transactions on Information Theory* 50th anniversary issue, **44**, 2325–2383.
- Hamamatsu Photonics (2002) Photomultiplier Tubes, Photomultiplier Tubes and Related Devices. *Catalog* June 2002 (www.hamamatsu.com) ([http://usa.hamamatsu.com/hcpdf/catsand-guides/PMTCAT\\_accessories.pdf](http://usa.hamamatsu.com/hcpdf/catsand-guides/PMTCAT_accessories.pdf))
- Ideker,T., Thorsson,V., Siegel,A.F. and Hood,L.E. (2000) Testing for Differentially-Expressed Genes by Maximum-Likelihood Analysis of Microarray Data. *Journal of Computational Biology*, **7**, 805-817.
- Lyng,H., Badiee,A., Svendsrud,D.H., Hovig,E., Myklebost,O. and Stokke,T. (2004) Profound influence of microarray scanner characteristics on gene expression ratios: analysis and procedure for correction. *BMC Genomics* **5**:10. (Provisional)
- Poynton, C. A. (1993) "Gamma" and its disguises: The Nonlinear Mappings of Intensity in Perception, CRTs, Film and Video. *The Society of Motion Picture and Television Engineers Journal*, **102**, 1099-1108.

- Romualdi,C., Trevisan,S., Celegato,B., Costa,G. and Lanfranchi,G. (2003) Improved detection of differentially expressed genes in microarray experiments through multiple scanning and image integration. *Nucleic Acids Research*, 2003, **31**:23 e149.
- Rousseeuw,R.J. and Leroy,A.M. (1987) Robust Regression and Outlier Detection. John Wiley & Sons, New York. ISBN: 0-471-85233-3
- Schena. M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* , **270**, 467-70.
- Schuchhardt, J.D., Beule, A., Malik, E., Wolski, H., Eickhoff, H., Lehrach, H. and Herzelt, H. (2000) Normalization strategies for cDNA microarrays. *Nucleic Acids Res.*, **28**, E47.
- Singiresu S.R. (2002) Applied Numerical Methods for Engineers and Scientists. *Prentice Hall*. ISBN: 0-13-089480-X.
- Song,L., Hennink,E.J., Young,I.T. and Tanke,H.J. (1995) Photobleaching kinetics of fluorescein in quantitative fluorescence microscopy. *Biophys. J.*, **68**, 2588-2600.
- Vanderkooy,J. and Lipshitz,S.P. (1987) Dither in Digital Audio. *Journal of Audio Engineering Society*, **35**, 966-975.
- Walpole,R.E., Myers,R.H., Myers,S.L. and Ye,K. (2002) Probability and Statistics for Engineers and Scientists. *Prentice Hall*. ISBN: 0-13-041529-4.
- Wang,C. and Carriedo,S. (2001) PMT Adjustment in GenePix 4000B. *Axon Instruments Inc. Unpublished manuscript*.
- Widrow,B. (1961) Statistical Analysis of Amplitude-Quantized Sampled-Data Systems. *Trans. American Institute of Electrical Eng.*, **79**, 555-568.
- Widrow,B., Kollár,I. and Liu,M.C. (1996) Statistical Theory of Quantization. *IEEE Trans. on Instrumentation and Measurement*, **45**,353-361.
- Yang,Y.H., Dudoit,S., Luu,P., Lin,D.M., Peng,V., Ngai,J. and Speed,T. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, E15.