

## Use of Splines for Normalization of Microarray Gene Expression Data

**Júlia P Soler**<sup>1\*</sup>, **Fernando F Rosa**<sup>1</sup>, **Silvana Chiavegatto**<sup>2</sup>, **Ivy Aneas**<sup>3</sup>, **José E Krieger**<sup>3</sup>

<sup>1</sup> Institute of Mathematics and Statistics, Department of Statistics, University of Sao Paulo, Brazil

<sup>2</sup> Institute of Psiquiatry, Department of Psiquiatry, University of Sao Paulo Medical School, SP, Brazil

<sup>3</sup> Laboratory of Genetics and Molecular Cardiology, Heart Institute, University of Sao Paulo Medical School, Brazil

\*Send correspondence to Júlia M. P. Soler

Titles: Doctor in Statistics by University of Sao Paulo

Post-doctor in Statistical Methods in Genetics by Southwest Foundation for Biomedical Research (San Antonio, TX, USA)

Address: Instituto de Matemática e Estatística

Rua do Matão 1010

05508-900 São Paulo, Brazil

Email: pavan@ime.usp.br

## Use of Splines for Normalization of Microarray Gene Expression Data

### Abstract

Microarray experiments are part of a new class of biotechnologies which allow the monitoring of expression levels for thousands of genes simultaneously. The measurements generated by these studies are subject to multiple sources of experimental variation. Some of these variations are considered systematic and may be explicitly corrected through data normalizations with the objective of cleaning and improving the quality of the measures of gene expression. Usual approaches consider a limited class of non-parametric techniques such as the lowess adjustments. However, there are more adaptive non-parametric methodologies that may be adopted for the problem. In this article we apply splines smoothing for normalization of gene expression data from cDNA microarray experiments. The union/intersection metric is adopted for comparison of the different methods, which explores the measures under dye-swap slides. The analysis uses a small study, consisting of RNA samples, from four rat groups generated from two rat strains, hypertensive SHR and congenic SHR-BN rat strains, evaluated under controlled condition and after 2 week NaCl treatment. Collectively, these results indicate that the smoothing splines approach is more efficient in normalizing microarray data than the global and lowess adjustments, controlling the dye bias and allowing a clearer identification of differently expressed genes in microarray experiments.

**Keywords:** cDNA microarray, normalization, lowess, splines, union/intersection metric

## **1. Introduction**

With the recent availability of the human and several model systems genomes, DNA array technology has become a powerful tool for researchers to assess global patterns of gene expression enabling a better knowledge of a variety of complex biological problems (DARVASI, 2003). For cDNA microarrays two (or more) different fluorescent dyes, such as the phycoerythrin Cy3 and Cy5, are labeled on different mRNA samples of interest, monitoring the gene expressions on the same array. The different stages of a microarray experiment include experimental design, normalization, exploratory analysis (typically, multivariate approaches) and the identification of differently expressed genes. Major challenges are present in each step of the analysis requiring intensive combined use of computational and statistical tools to minimize limitations of this method.

Typically, biological and technical variations occur on microarray experiments. The biological variations are random and may be controlled through replicates and, when appropriated, by pooling mRNA samples from some set of individuals (CHURCHILL, 2002; KERR; CHURCHILL, 2001). Known sources of technical variation include low signal intensity, bias due to choice and incorporation of dyes, variable spot shape or position on the array, amount of cDNA material and others (CHURCHILL, 2002). These variations are systematic and some of them may be avoided through a rigorous experimental control, otherwise compromising the identification of differently expressed genes. Data normalization methodologies have been commonly used for cleaning and improving the quality of the measures of gene expression. Usual approaches include a limited class of non-parametric techniques, as lowess adjustments, applied before the identification of differently expressed genes. However, there are more adaptive non-parametric methodologies that may be adopted. In this article, we

apply splines smoothing for normalization of gene expression data. This methodology has been successfully used for smoothing curves with circadian patterns (IRIZARRY, 2002) and we introduce its application for image analysis in cDNA microarray experiments.

In Section 2.1 we describe the data set used in this article. Section 2.2 presents normalization techniques and, in Section 2.3, we propose a criterion for comparison among them considering the union/intersection metric. Results for the normalization adjustments, using our data set, are showed in Section 3. Section 4 presents discussions and conclusions.

## **2. Materials and Methods**

### **2.1. Microarrays Data**

Using QTL interval mapping we identified 5 quantitative trait loci (QTLs) that, collectively, explain 43% of the total systolic blood pressure variation in rats (SCHORK et al., 1995). We, then, developed congenic rat strains using the hypertensive strain background in which each of the mapped chromosomal regions was replaced with the normotensive strain counterpart. All these congenic strains show alteration in the basal or NaCl load blood pressure phenotypes (data not shown).

Briefly, the study examined mRNA samples from a pool of kidney tissues extracted from a set of 3 animals, randomly selected from four rat groups: SHR hypertensive and congenic SHR-BN rat strain for chromosome 2 evaluated under controlled and after 2 weeks NaCl load. Experimental design obeyed the scheme showed in Table 1. It was used dye-swap slides and only one replicate of each experimental condition in this small trial. Congenic animals were obtained by molecular marker assisted selection after several backcrosses between the BN and SHR rat strains,

considering the finding of animals genetically equivalent to SHR hypertensive except to the region in chromosome 2. Thus, SHR-BN congenic animals are supposed to have lower blood pressure values than the SHR. For this experiment, 14K rat cDNA arrays obtained from a Norwegian Core Facility were used consisting of 13,799 sequence verified cDNA probes from Research Genetics and 10 in-house rat cDNAs printed in duplicates on Corning CMT Gaps II slides. Also included in the same array were 10 cDNA from plants. The total number of spots on the array is 26,912.

Table 1. Microarray design.

<b>Condition</b>	<b>Group</b>	<b>Slide</b>
Controlled	SHR-Cy5 x SHR-BN-Cy3	Rat 85
	SHR-Cy3 x SHR-BN-Cy5	Rat 89
NaCl exposition	SHR-Cy5 x SHR-BN-Cy3	Rat 88
	SHR-Cy3 x SHR-BN-Cy5	Rat 86

## 2.2. Normalization Methods

In cDNA microarrays the two-color system measurements (green and red) represent the abundance of two mRNA samples cohybridized (with Cy3 and Cy5, respectively) onto each arrayed gene. Such type of array is supposed to reduce part of experimental variation, because the pair  $(R,G)$  of responses is evaluated at the same time in each spot. In design of experiments, when two responses are paired into a block factor, it is usual to resume the data through the difference between the responses, like in the paired “t”-test. In microarray data, more informative displays of the gene expression intensity values are obtained from the transformation to the logarithm scale of the original values. Thus, ratio-based decisions of the gene expression levels are frequently adopted, which represent differences between the responses under log scale. Therefore, it is useful to transform  $(R,G)$ , the original pair of responses, to  $(\log_2 R, \log_2 G)$  and, further, to

$(M,A)$ , where  $M = \log_2 R/G$  and  $A = \log_2 \sqrt{RG}$  (DUDOIT et al., 2002). This new axis system,  $(M,A)$ , corresponds to a clockwise coordinate system rotation by  $45^\circ$ .

It is used to display the gene expression intensities for a pair of samples through an  $MxA$  plot, aiming a spatial identification of the points. Here, the horizontal coordinate  $A$  is a measure of the average transcription level, while the  $M$  is a measure of differential transcription, or of random variation. In a larger context, variance by average scatter-plots are traditionally used in exploratory analysis for diagnoses of data distribution, characterizing heteroscedasticity and dependences.

In cDNA array technology, several authors have addressed the problem of normalization of the intensity responses to reduce the technical variation and to obtain comparable intensities values from different arrayed genes (THOMAS et al., 2001; DUDOIT et al., 2002; YANG et al., 2002). The normalization methodologies are applied to the  $MxA$  plots, transforming the  $M$  coordinates to  $M^*$ , giving origin to an  $M^*xA$  plot, commonly used for identification of differently expressed genes.

In global normalization, proposed by Chen et al. (1997), it is assumed that red and green intensities are related by a constant factor ( $R = k G$ ). For the  $j$ -th spot we take the transformation

$$M_j^* = M_j - k ,$$

where  $k$  is obtained as the median of  $M_j$  values. This was the initially proposed and most used approach. However, in many cases, systematic variations may be observed in the  $MxA$  plot, like a dependence of the deviation of the values  $M$  on the mean intensity  $A$ , and this normalization technique does not take that into account. A more robust approach, commonly found in the literature (YANG et al., 2002) is to use the lowest

robust scatter plot smoother (CLEVELAND, 1979). Let  $c(A_j)$  denote the lowess fit to the  $Mx_A$  plot. We then take the transformation

$$M_j^* = M_j - c(A_j), \quad (1)$$

for every spot  $j$ .

Considering the nature of the problem, it is expected that a small percentage of differently expressed genes will appear as outliers in the  $Mx_A$  plot. Because of this, scatter plot smoothers are needed to perform robust fits, without great influence from the outliers ( $M, A$ ). In spite of lowess being a robust non-parametric regression method, more adaptive alternatives are mentioned in the literature. In particular, it is worth noting that the lowess method has a set of parameters that must be chosen in order to perform the fit. While there might exist some empirical methods for obtaining the parameters for a given problem, there is not to date a procedure that yields optimized parameters, which in turn would guarantee an optimum fit. An alternative method is the use of smoothing splines, which is a powerful tool applied for modeling patterns in non-parametric regression problems (WAHBA; GRAVEN, 1979). When using smoothing splines it is possible to choose the location of knots and the smoothness parameter via optimum criteria. Cross validation (CV) and generalized cross validation (GCV) are popular approaches for finding an appropriate criterion, and Irizarry (2002) has proposed the minimal estimate of risk as a new criterion. In this work, we use the GCV to obtain the  $c(A_j)$  spline adjustment, implemented on the statistical package R (R Development Core Team, 2004). The smoothing splines normalization is conducted in the same fashion as the lowess one: we transform  $M$  into  $M^*$  as in equation (1), but now  $c(A_j)$  is the adjusted value for the smoothing splines fit.

The normalization methods described before can also be done by sub-arrays. For the  $j$ -th spot in the  $i$ -th sub-array we obtain the adjustment  $M_{ij}^*$  according to a given normalization technique. Those values are obtained considering each sub-array as a complete array, and then proceeding with the normalization techniques discussed, for every sub-array.

### 2.3. Comparison Criterion for Different Normalizations

One of the possible alternatives to compare different normalization methods is calculating the number of spots showing concordance between the intensities values evaluated in dye-swapped slides. However, notice that normalization of microarray data is used to reduce the variation due technical irregularities. Thus, after normalization, the intensities of two arrays, with dye assignment reversed in the second one, must be concordant only if dye bias is the most important technical error component and there is no relevant biological variation.

For  $j$ -th spot, let  $(R_j, G_j)$  and  $(M_j, A_j)$  define pairs of intensities values, and its log transformation, for mRNA samples of the treatment group labeled with red-fluorescent Cy5 dye and the reference mRNA samples labeled with green-fluorescent Cy3 dye. In addition, let  $(R'_j, A'_j)$  and  $(M'_j, A'_j)$  be defined, in the same context, for dye-swap array. And also, consider  $M_j^*$  and  $M'_j{}^*$ , the normalized intensities for the  $j$ -th spot of two arrays from dye-swap design. The following result may be assessed

$$(M_j^*, A_j) \text{ and } (M'_j{}^*, A'_j) \text{ are concordant} \Leftrightarrow M_j^* \cdot M'_j{}^* < 0. \quad (2)$$

Deviation of this inequality may occur as a result of variations due to biological tools or technical error (except dye bias component).

For our application, the number of spots showing concordance, like imposed by (2), was calculated into a restricted region, defined to the  $MxA$  plot by

$$C = \{(A_j > 10) \cap (|M_j| > 1)\}. \quad (3)$$

To define a more elegant criterion for comparison of normalization methods, and further, for evaluation of the dye bias effect, consider the union/intersection metric, used by Wilkinson and Huberman (2003) for matching gene communities, and given by

$$d(A, B) = \frac{\#\{A \cap B\}}{\#\{A \cup B\}} = \frac{\sum_{ij} \delta_{\alpha_i \beta_j}}{n + m - \sum_{ij} \delta_{\alpha_i \beta_j}}, \quad (4)$$

where  $A = (\alpha_1, \alpha_2, \dots, \alpha_n)$  and  $B = (\beta_1, \beta_2, \dots, \beta_m)$  are two sets of genes and  $\delta_{\alpha_i \beta_j}$  is a identity function for  $\alpha_i = \beta_j$ . The metric  $d(A, B)$  ranges in the interval  $[0, 1]$  and will be larger for closer the matches (or more concordant sets).

We adapt the metric given in (4) to measure the concordance between the normalization procedures. Let  $C$  defined by expression 2, and  $C'$  under the terms of  $MxA$  plot. For the  $k$ -th normalization method we define  $d_k(C, C')$  as the concordance metric between subsets  $C$  and  $C'$ , with  $\delta_{c_i c'_j}$  an identity function for  $c_i = c'_j$ . We choose that normalization with maximum  $d$  value, since it promoted the closest match between  $C$  and  $C'$ , and thus was the most successful in reducing the variations due to technical effects, specially the dye bias component.

### 3. Results

The histogram of intensities from each sample for array Rat 85 is showed in Figure 1. For both, Cy3 and Cy5, a non-symmetric unimodal distribution is observed, characterized by most of its mass at small intensities. Equivalent results were found for the others slides considered in the study (Rat 89, Rat 88, Rat 86).

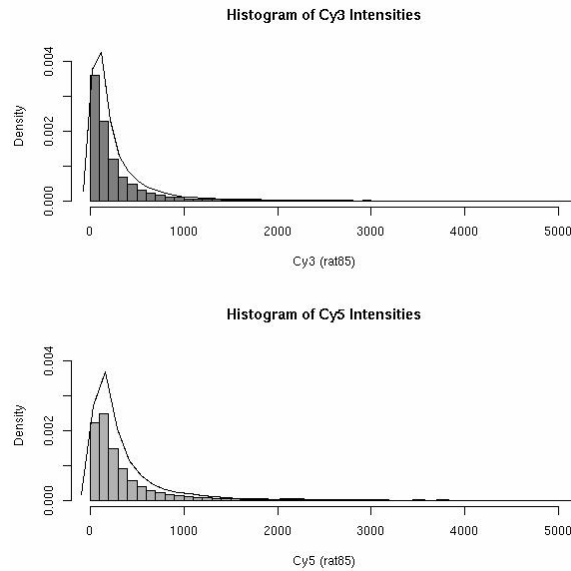


Figure 1. Histogram of probe intensities at the Cy3 and Cy5 dyes for array Rat 85.

Most of the data in Figure 1 is squeezed into a tiny corner in the bottom left of the histogram. More informative displays may be obtained from double-logarithmic scale, as showed by *MxA* plot from Rat 85 data in Figure 2. In addition, global, lowess and splines normalizations were applied for each array in our study, considering all probes at a time or making the analysis by sub-arrays. The results for array Rat 85 are presented in Figure 3. The straight-line represents the vertical axis origin ( $M=0$ ) and the other curve signalizes the adjustment obtained by the normalization method. The best result was found by the splines by sub-array technique, since the systematic variation was quasi-completely eliminated. Figure 4 shows the results of the splines by sub-array normalization for slides Rat 85, Rat 86, Rat 88 and Rat 89.

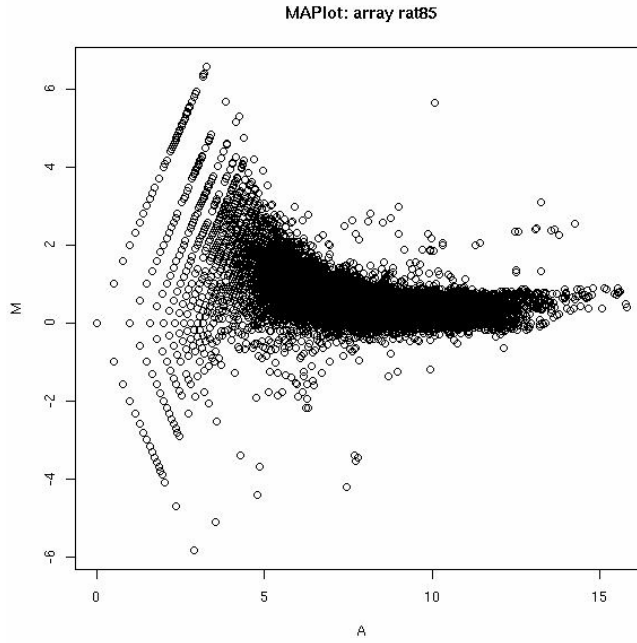


Figure 2.  $MxA$  plot for Rat 85 data.

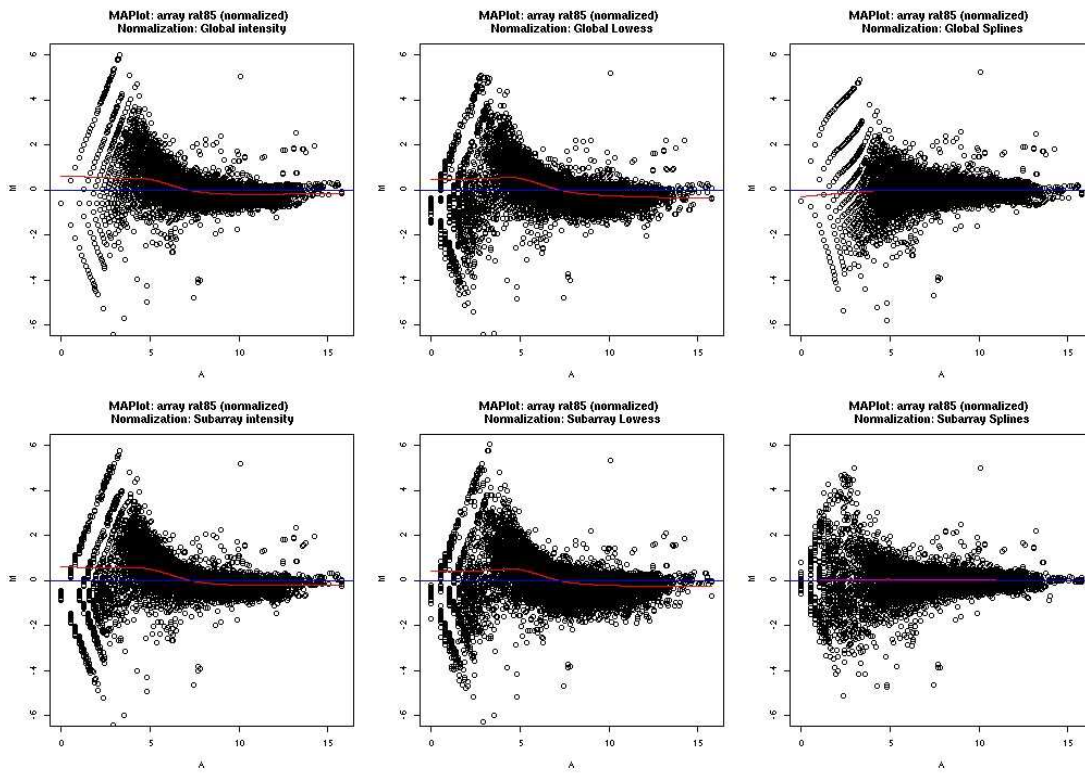


Figure 3. Different normalization methods for array Rat 85.

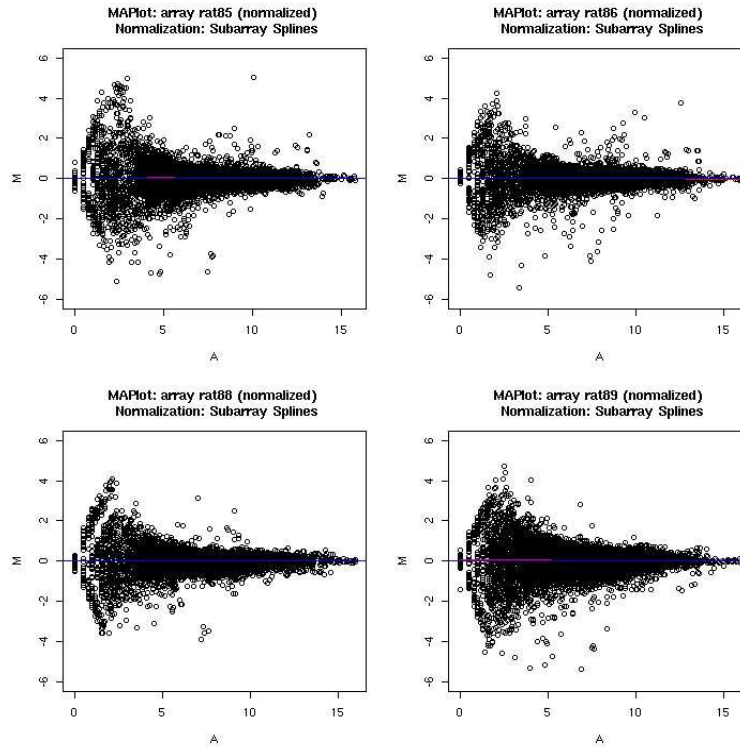


Figure 4. Sub-array splines normalizations for arrays Rat 85, Rat 86, Rat 88 and Rat 89.

Comparing lowess and splines normalization methods, we calculated the number of concordant genes (probes or spots) according to the criterion (2). The number of genes obeying restriction (3) was 109, for array Rat 85, and such set of genes was adopted in all calculation. The results are in Table 2. The values obtained for union/intersection metric are presented in Table 3. The results show the superiority of the normalization using splines methodology.

Table 2. Number of concordant spots.

Array / Reversed Dye	Lowess	Splines
<b>Rat 85 / Rat 89</b>	<b>25</b>	<b>39</b>
<b>Rat 86 / Rat 88</b>	<b>20</b>	<b>38</b>

Table 3. Union/intersection metric values.

<b>Array / Reversed Dye</b>	<b>Lowess</b>	<b>Splines</b>
<b>Rat 85 / Rat 89</b>	<b>0.0</b>	<b>0.21</b>
<b>Rat 86 / Rat 88</b>	<b>0.1</b>	<b>0.20</b>

#### 4. Discussion

Usually, in microarrays experiments the probe selections may be assumed as a quasi-random process. Thus, the intensities empirical distribution may be used to make decisions about the true distribution of gene expression intensities. The histogram showed in Figure 1 indicates that the transcription levels of a large fraction of genes are approximately the same, across the samples of hypertensive and congenic rat strains, and that only a relatively small proportion of genes will vary significantly in expression between the two mRNA samples. These findings have been described by others authors, under different treatments and for a large class of experimental unities, indicating that, in general, a small set of genes are involved in regulatory processes.

The convenience in adopting the logarithm transformation to represent microarray data is justified through the asymmetric shape found for the data distribution. In addition, the coordinate changing to  $(M,A)$  is very useful for detection of the variation components which are affecting the data. Depending whether a systematic deviation of the empirical values  $M$  across the line  $M=0$  is observed, we may take decisions about the effects of the biological and technical variation components. Figure 2 shows that while the variance of  $M$  is relatively small and approximately constant for large average intensities  $A$ , it becomes larger as  $A$  decreases, consistent with the presence of heteroscedasticity in our data. This is assumed to be due to technical error, which is

often corrected by normalization techniques. The data presented here show that splines smoothing gives better adjustment than methods currently used for this purpose. Using the union/intersection metric, we also found better results applying splines ( $d=0.21$  and  $d=0.20$ ) compared with lowess fit. In addition, others sources of technical (besides dye bias) and biological variation must be taken into account for normalization, since the concordance metric was low ( $d \ll 1$ ). Wolfinger et al. (2001) propose a parametric version to normalize microarray data applying a combined mixed model. However, the main limitation of this methodology is to normalize the data under a linear model.

In the present study, the main objective was to perform an exploratory analysis of the different normalization techniques rather than the identification of genes differently expressed. However, upon comparison of genes differently expressed in kidney samples from the two rat strains we found (data not shown): (i) a small group of genes (5 genes), that may be considered to be affected by moderated biological variation, showing high signal intensity so that they are easily detected under any normalization technique; (ii) several genes (50 genes), showing moderated expression, that are extremely affected by the normalization method, which may underlie high susceptibility to (possibly due to be highly affected by) technical variation, and (iii) a set of differently expressed genes (3 genes) uncovered only when data from paired slides, in dye-swapped design, are combined, as proposed by Yang et al. (2001).

Collectively, these results indicate that the smoothing splines approach is more efficient in normalizing microarray data than the global and lowess adjustments, controlling the dye bias and allowing a clearer identification of differently expressed genes in these experiments.

## Acknowledgments

This work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP grant 01/00009-0). The computational resources were provided by the Mathematics and Statistics Institute, University of Sao Paulo (IME-USP).

## References

- CLEVELAND, W.S. Robust locally weighted regression and smoothing scatterplots. **Journal of the American Statistical Association** , v. 74, n. 368, p. 829-836, 1979.
- CHEN, Y.; DOUGHERTY, E.R. and BITTNER, M.L. Ratio-based decisions and the quantitative analysis of cDNA microarrays images. **Journal of Biomedical Optics**, v. 2, n.4, p. 364-374, 1997.
- CHURCHILL, G.A. Fundamentals of experimental design for cDNA microarrays. **Nature Genetics Suppl** , v.32, p. 490-495, 2002.
- DARVASI, A. Genomics: Gene expression meets genetics. **Nature**, v. 422, p. 269-270, 2003.
- DUDOIT, S.; YANG, Y.H.; SPEED, T.P. and CALLOW, M.J. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. **Statistica Sinica**, v.12, p. 111-139, 2002.
- DURBIN, B and ROCKE, DM. Exact and approximate variance-stabilizing transformations for two-color microarrays. **Bioinformatics**, v.18, p. S105-S110, 2002.
- KERR, M.K.; CHURCHILL, G.A. Statistical Design and the Analysis of Gene Expression Micro Array Data. **Genetic Research**, v.77, p. 123-128, 2001.
- IRIZARRY, R.A. (2002). Choosing smoothness parameters for smoothing splines by minimizing an estimate of risk. 2002. Available in:

[www.biostat.jhsph.edu/~ririzarr/papers/react-splines.pdf](http://www.biostat.jhsph.edu/~ririzarr/papers/react-splines.pdf). Accessed in May 10<sup>th</sup> 2004.

R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria. ISBN 3-900051-00-3. 2004. Available in: [www.R-project.org](http://www.R-project.org). Accessed in May 10<sup>th</sup> 2004.

SCHADT, E.E.; MONKS, E.A.; DRAKE, T.A.; LUSIS, A.J.; CHE, N.; COLINAYO, V.; RUFF, G.; MILLIGAN, S.B.; LAMB, J.R.; CAVET, G.; LINSLEY, M.M.; STOUGHTON, R.B. AND FRIEND, S.H. Genetics of gene expression surveyed in maize, mouse and man. **Nature**, v.422, p. 297-302, 2003.

SCHORK, N.J.; KRIEGER, J.E.; TROLLIET, M.R.; FRANCHINI, K.G.; KOIKE, G.; KRIEGER, E.M.; LANDER, E.S.; DZAU, V.J. and JACOB, H.J. (1995). A biometrical genome search in rats reveals the multigenic basis of blood pressure variation. *Genome Research* **5**: 164-172.

THOMAS, J.G.; OLSON, J.M.; TAPSCOTT, S.J. and ZHAO, L.P. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. **Genome Research**, v.11, p.1227-1236, 2001.

YANG, Y.H.; DUDOIT, S.; LUU, P. and SPEED, T.P. Normalization for cDNA Microarray Data. [S.1.]. 2001. Available in: [www.citeseer.nj.nec.com/406329](http://www.citeseer.nj.nec.com/406329). Accessed in May 10<sup>th</sup> 2004.

YANG, Y.H.; BUCKEY, M.J.; DUDOIT, S. and SPEED, T.P. Comparison of methods for image analysis on cDNA microarray data. **Journal of computational and Graphical Statistics**, v.11, p. 108-136, 2002.

WAHBA, G. and GRAVEN, P. Smoothing noisy data with spline functions. **Numerische Mathematik**, v. 31, p. 337-403, 1979.

WILKINSON, D.M. and HUBERMAN, B.A. A method for finding communities of related genes. (2003). Available in:

[www.pnas.org/cgi/doi/10.1073/pnas.0307740100](http://www.pnas.org/cgi/doi/10.1073/pnas.0307740100). Accessed in May 10th 2004.

WOLFINGER, R.D.; GIBSON, G.; WOLFINGER, E.D.; BENNETT, L.; HAMADEH, H.; Bushel, P.; AFSHARI, C.; PAULES, R.S. Assessing Gene Significance from cDNA Microarray Expression Data via Mixed Models. **J. Compu. Biol.**, v. 8, n.6, p. 625-637, 2001.