NORMALIZATION OF CDNA MICROARRAY DATA BASED ON LEAST ABSOLUTE DEVIATION REGRESSION

J. Ramírez, J.L. Paredes[*]

Universidad de Los Andes
Department of Electrical Engineering
Merida, 5101. Venezuela

G. Arce

University of Delaware
Department of Electrical and Computer Engineering
Newark, DE 19716

ABSTRACT

This paper proposes a method of normalization of cDNA microarray data. This approach uses all gene data to estimate the normalization parameters and it obtains these parameters using an algorithm based on Least Absolute Deviation (LAD) regression. This method normalizes iteratively each microarray set after the estimation of the normalization parameters which uses a LAD regression algorithm. The normalization method has a robust performance since it assumes that the errors between arrays follow a Laplacian distribution, leading to the mean absolute error minimization as a performance criterion to be achieved. The proposed normalization method was evaluated using three performance measures and they show that LAD based normalization method minimizes the errors and provides a more consistent replicated data spread with respect to a least square based method.

1. INTRODUCTION

Microarray technology has become in the standard experiment for large-scale analysis of cells gene expression. This technology has emerged as a powerful tool since it can be used to measure the expression levels for thousands of genes simultaneously from a single sample. However, the microarray experiments are contaminated by several sources of measurement errors which affect the quality of results. The sources of measurement errors are generally assigned to the amount of samples used in the experiment, the amount of label applied to the samples, the label emission efficiency and the scanning parameters.

To overcome these drawbacks, the generation of multiple microarrays from the same tissues has been developed in order to obtain a more reliable microarray data [1]. However, the same microarray experiment does not generate the same gene expression data. Indeed, each replicated microarray may have substantial variations with respect to another microarray. Since the microarrays have several experimental limitations which affect the gene expression levels, the normalization of the replicated microarray data has emerged as a necessary preprocessing stage before any further gene expression data analysis.

The goal in a microarray data normalization algorithm is to remove or minimize the measurement errors between microarray data sets. More precisely, all the microarray replicated data are mapped to a new data set such that a performance criterium that compares each replicated set and a reference set is achieved. The replicated data normalization methods commonly assumes that the gene expression data sets are related by a scale factor and an offset correction [2]. This leads naturally to linear regression techniques as the evident approach to estimate the normalization coefficients. Some previous normalization methods are based on Least Square (LS) algorithms that normalize the replicated microarray data assuming that the errors between replicated data sets follow a Gaussian distribution. However, it has been widely reported that microarray data is inherently noisy and, therefore, several nonlinear operations have been recently introduced which try to minimize the effects of outliers on the estimation of normalization coefficients [3, 4].

This paper proposes a new method of normalization based on Least Absolute Deviation (LAD) regression. This method uses all gene set to compute the normalization parameters under the assumption that the errors between replicated data sets follow a Laplacian model. The proposed approach normalizes iteratively each microarray set after the estimation of normalization parameters that are found through the LAD regression iterative algorithm. The performance of the proposed iterative normalization method is evaluated using three different performance measures: Mean Square Error (MSE), Mean Absoloute Error (MAE) and Boxplots. These performance measures show that LAD based algorithm minimizes the error values and provides a more consistent replicated data spread respect to a LS based algorithm [4].

2. PRELIMINARY BACKGROUND

Without lost of generality, let us assume that we have two gene expression data sets obtained from replicated microarrays. Although, in general, more than two replicated microarrays are available, the normalization is performed on pairs of microarray expression data sets, where one data set is considered as the reference set and the other is the floating point set that has to be normalized respect to the reference set. Consider that $\mathbf{y} = [y_1, y_2, \dots, y_N]$ are the expression levels outputted from the reference set. Furthermore, let $\mathbf{x} = [x_1, x_2 \dots, x_N]$ be the expression levels of the replicated microarray to be normalized with respect to the reference set.

It is common to assume that the floating point set is related to the reference set by scaling and shifting operations [2, 3], i.e, $y_i = ax_i + b + \eta_i$, where a is the scale factor between the data pairs $\{x_i, y_i\}|_{i=1}^{N}$, b is the shift correction and $\eta_i|_{i=1}^{N}$ are the unobservable errors, with N as the number of genes in each microarray. The goal in a normalization method is to find the best values of a and b such that a performance criterium that compares the reference set to the floating point set is achieved. For instance, regression methods estimate the normalization coefficients assuming that the unobservable errors $\{\eta_i\}|_{i=1}^{N}$ obey a specific distribution, thus, a and b are

[*] This research was supported in part by the University of Los Andes Council of Scientific, Humanistic and Technological Development (CDCHT) under the project I-768-04-02-B.

found such that they minimize the effects caused by $\{\eta_i\}|_{i=1}^{N}$.

A first approach in the estimation of scale and shift parameters assumes that the unobservable errors η_i follow a zero-mean Gaussian Distribution. This leads, naturally, to the mean square error (MSE) between the reference and floating point sets as performance criterion to be minimized. This is achieved by the well-known least square (LS) regression. A shortcoming of this approach, however, is that it uses all the genes in the set to estimate a and b, and this may lead to unacceptable results since not all genes are reliable due to the presence of outliers.

In view of this limitation, Wang et al. in [4] developed a normalization algorithm that estimates iteratively the normalization coefficients (a, b) using a subset of reliable genes. The basic idea is to select, at each iteration, a subset of reliable genes by defining a gene selection window over a scatter plot and estimate the normalization coefficients based on this subset of control genes, adding thus robustness to the normalization approach.

Several drawbacks on this approach can be observed. First, the minimization of the MSE using only the control genes does not necessarily lead to the minimum MSE over all gene data set. Second, the choice of a good gene selection window is critical for the convergence and accuracy of the algorithm.

3. NORMALIZATION METHOD BASED ON LEAST ABSOLUTE DEVIATION REGRESSION

The normalization methods described above are based on the assumption that the unobservable errors follow a Gaussian model. However, it has been reported that outliers are likely to occur in microarray data [5]. To mitigate the effects of outliers over the estimation of scale and shift parameters, robust regression methods are needed. Furthermore, it has been shown in [6] that modelling expression data using heavier-than-Gaussian tail distributions (such as Laplacian distribution) leads to more accuracy results in gene classification.

In this paper, we propose a microarray normalization method which assumes that the unobservable errors follow a Laplacian model, leading to least absolute deviation (LAD) regression as a optimization technique to compute the normalization coefficients, that by itself, adds robustness, avoiding thus the pre-selection of reliable genes. Therefore, the natural performance criterion to be minimized is the mean absolute error (MAE) between the reference set and normalized set. That is,

$$MAE = \sum_{i=1}^{N} |y_i - (ax_i + b)| \tag{1}$$

where a and b are, as before, the scale and shift coefficients to be determined.

Unlike LS regression, in LAD regression there are not closed-form expressions for the computation of a and b, so we have to resort to an iterative approach to find the values of a and b that minimizes Eq. (1) [7, 8]. Finding the optimum values for a and b can be thought of as a two stage-procedure. First, the scale factor a is fixed to an initial value, say $a = a_0$ and the MAE function becomes a one-parameter function depending on b:

$$F(b) = \sum_{i=1}^{N} |y_i - a_0 x_i - b| \tag{2}$$

Note that Eq. (2) is the Maximum Likelihood estimator (MLE) of location for a Laplacian distribution, where $y_i - a_0 x_i$ are the observation samples and b is the location parameter to be determined. The goal, thus, is to find the location parameter b that minimizes $F(b)$. It is well-known that $F(b)$ reaches its minimum at the sample median [9], thus the shift correction b is obtained as:

$$b = \text{MED}\left(y_i - a_0 x_i \Big|_{i=1}^{N}\right) \tag{3}$$

In the second stage, the shift correction is fixed to the value computed previously and now the MAE becomes a function of the scale factor a, namely

$$F(a) = \sum_{i=1}^{N} |y_i - a x_i - b_0| \tag{4}$$

after simple algebraic manipulations, (4) becomes

$$F(a) = \sum_{i=1}^{N} |x_i| \left| \frac{y_i - b_0}{x_i} - a \right| \tag{5}$$

Upon closer examination of Eq. (5), it can be noticed that the minimization of $F(a)$ looks like the MLE of location for a Laplacian distribution, where $\{\frac{y_i - b_0}{x_i}\}$ are the observation samples, $|x_i|$ are the sample weights, and a is the location parameter to be estimated. Thus, the value of a that minimizes Eq. (5) reduces to the weighted median of the samples $\{\frac{y_i - b_0}{x_i}\}$ defined as [9]:

$$a = \text{MED}\left(|x_i| \diamond \frac{y_i - b_0}{x_i} \Bigg|_{i=1}^{N} \right) \tag{6}$$

where $\diamond$ is the replication operator defined as $k \diamond x = \overbrace{x, \ldots, x}^{k\ times}$.

Thus, the estimation of a reduces to replicating the samples $\{\frac{y_i - b_0}{x_i}\}$, $|x_i|$ times, sort the replicated samples and select the center of the "sorted-replicated" set as the scale coefficient a. Although it may be seem that the replication operator is restricted to integer value weights, a more general definition of weighted median using non-integer weights has been developed in [9] and [10].

Note that Eq. (3) uses a fixed value of a to estimate b. However, there is not a specific procedure to compute the initial value for a. An intuitive approach to estimate a is using the LS solution, and this value is used as initial value for a. After the initial scale coefficient a_0 is obtained, the algorithm starts an iterative procedure to estimate b and a. The main steps of LAD regression are summarized as follows.

LAD Regression Iterative Algorithm

Step 1) Set k=0, and compute the initial value of the scale factor $a = a_0$ using the LS solution.

Step 2) Set k=k+1, and estimate b_k using the fixed value a_{k-1} computed at the previous iteration

$$b_k = \text{MED}(y_i - a_{k-1} x_i|_{i=1}^{N}). \tag{7}$$

Step 3) Estimate a_k, using the value of b_k previously computed in step 2

$$a_k = \text{MED}\left(|x_i| \diamond \frac{y_i - b_k}{x_i} \Bigg|_{i=1}^{N} \right) \tag{8}$$

Step 4) Repeat step 2) and step 3) until a_k and b_k deviate from a_{k-1} and b_{k-1} within a tolerance band.

Once defined a procedure to estimate the regression coefficients based on MAE minimization criterium, we propose a normalization method of microarray data using the LAD regression iterative algorithm. This normalization algorithm uses all gene expression data to estimate the normalization coefficients, and avoid that the normalization effectiveness depends on the selection of gene control data.

3.1. LAD based Normalization Method

The approach to normalize the floating point set with respect to the reference data uses a double iterative procedure. The first iterative procedure estimates the normalization coefficients using the LAD regression iterative algorithm, as described above, and the second one is the normalization procedure itself.

This normalization algorithm, first selects one of the gene expression sets as the reference set and the remaining sets are considered as floating point sets. The normalization method is then applied to the reference set $\{y_i\}|_{i=1}^{N}$ and each one of the floating point sets $\{x_i\}|_{i=1}^{N}$, independently.

Initially, the proposed method defines the floating point set as the normalized set at the iteration $j = 0$, i.e. $x_i^{(j=0)} = x_i$, where j is the iteration index for the normalization loop and $x_i^{(j)}$ is the normalized data at the $j - th$ iteration. With the two data sets $\{y_i\}|_{i=1}^{N}$ and $\{x_i^{(j-1)}\}|_{i=1}^{N}$, the algorithm computes the best values of $a^{(j)}$ and $b^{(j)}$ using the LAD regression iterative algorithm; where $a^{(j)}$ and $b^{(j)}$ are, respectively, the scaling and shifting parameters computed by the LAD iterative regression algorithm, at the iteration j.

Once obtained the scaling and shifting parameters, the normalization operation is carried out. Thus, the new normalized data set is:

$$x_i^{(j+1)} = a^{(j)} x_i^{(j)} + b^{(j)} \qquad (9)$$

where $x_i^{(j+1)}$ is the new normalized data and $x_i^{(j)}$ is either the normalized data obtained from the previous iteration or the floating point data for the first iteration. The algorithm back and forth until a convergence criterium is achieved, otherwise begin the new iteration with $\{y_i\}|_{i=1}^{N}$ and $\{x_i^{(j)}\}|_{i=1}^{N}$. This normalization method is summarized as follows:

LAD based Normalization Algorithm

Step 1) Select an expression level data set as reference data set $\{y_i\}|_{i=1}^{N}$ and one of the remaining data sets as the floating point data set.

Step 2) Start the normalization process using the reference data set and one floating point data set $\{x_i^{(j=0)}\}|_{i=1}^{N}$, where j=0 indicates that the floating point data set is considered as the normalized data at iteration j=0.

Step 3) Set $j = j + 1$, and estimate the normalization coefficients $(a^{(j)}, b^{(j)})$ using the LAD Regression Iterative Algorithm, where the reference data is y_i, and the floating point data is the normalized data at the previous iteration $x_i^{(j-1)}$.

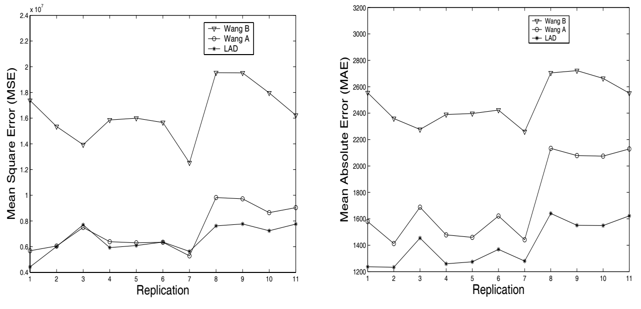
Step 4) Obtain the new normalized data set using the linear model:

$$x_i^{(j)} = a^{(j)} x_i^{(j-1)} + b^{(j)} \qquad (10)$$

Step 5) End the iteration, once $a^{(j)}$ and $b^{(j)}$ do not diverge from $a^{(j-1)}$ and $b^{(j-1)}$ within a tolerance band. Otherwise, go back to **step (3)**.

4. RESULTS AND DISCUSSION

The proposed normalization method was evaluated on public cDNA microarray data available in [11]. This microarray database is a set of 72 raw data files. The microarray data are obtained from experiments that measure gene expression in honey bee brains. Each microarray experiment consists of an array of approximately $9,000$ cDNA clones selected from a set of $20,000$ bee brain expressed sequence tags (EST). Further, each experiment uses 60 samples of honey bee brains (Apis Mellifera) to measure the gene expression levels at two different growing stages (nurse and forager).



(a) (b)

Fig. 1. (a) MSE and (b) MAE curves obtained using the various normalization methods.

We apply the normalization methods on 12 of the 72 microarray data sets, each one with $1,000$ gene expression levels obtained from intensity values of Channel-5. The gene expression levels of Channel-5 is the spot median pixel intensity obtained at scanner laser wavelength of 635 nm with the spot median background subtracted [12].

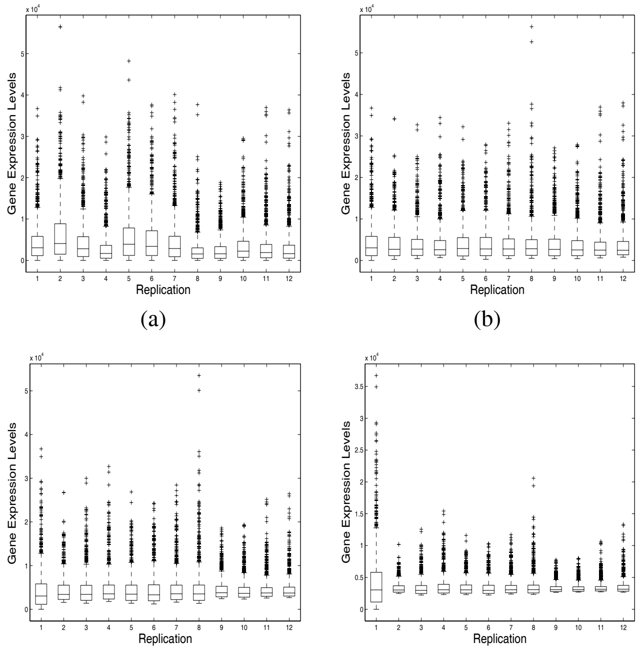
First, we test the accuracy of the normalization algorithms using the Mean Square Error (MSE) and the Mean Absolute Error (MAE) between the reference set and the normalized sets as performance measures. Next, we exploit the boxplot representation to evaluate the *between-slice* performance of the various normalization methods. In all these performance measures, the proposed normalization method are compared to Wang's normalization approach [4].

We have implemented Wang's approach and applied it to the same microarray data sets. Wang's approach selects, as initial control gene subset, those genes with standard deviation across replication less than a predefined threshold. Furthermore, since Wang's algorithm highly depends on the iterative selection of the control genes that lie inside the window function, we use two different window parameter sets. The values of the first window parameter set are fixed in $\varepsilon_{1A} = 3000$, $\varepsilon_{2A} = 16000$ and $\varepsilon_{3A} = 0.2$, and will be referred throughout the paper as Wang A. As a second selection function, we used the window parameter set defined by $\varepsilon_{1B} = 3000$, $\varepsilon_{2B} = 6000$ and $\varepsilon_{3B} = 0.2$ and they will be referred throughout the paper as Wang B. The parameters of this second window function were used by Wang et. al in [4].

Figure 1(a) shows the MSE between the reference set and normalized set for both methods. This performance measurement was used by Wang et. al in [4] to evaluate their normalization method, but the MSE was computed using only the selected control genes set. In this paper, however, since all genes are used to analyze the microarray experiment, and therefore, we compute the MSE using all gene expression levels to observe the accuracy of the normalization methods over the all data set.

As it can be seen in Fig. 1(a), LAD based method outperforms Wang's approach in almost all replicated sets. This indicates that LAD based normalization approaches are robust methods, since these approaches generate a normalized data closer to the reference data. Further, LAD based method need not a gene control subset to estimate the normalization coefficients. In contract, the difference between *Wang A* curve and *Wang B* curve evidences that the accuracy of Wang's algorithm is affected seriously by the selection of the window function.

Figure 1(b) shows the MAE between the reference set and each replicated normalized data set over all gene expression levels. The



(a)

(b)

(c)

(d)

Fig. 2. Boxplot of (a) original microarray sets, (b) LAD normalized data, (c) Wang A normalized data, (d) Wang B normalized data.

values of MAE obtained by applying LAD based normalization method are significantly smaller than the corresponding ones obtained using Wang's methods. The resultant MAE clearly shows that, under this performance criterion, the LAD based method is a better normalization procedure than Wang's method, obtaining lower error values over all replicated microarray data. These results are in concordance with the fact that LAD based normalization approach tries to minimize the MAE between the reference set and the normalized set.

A third way to evaluate our normalization methods is by looking the boxplot generated by the normalized data. The boxplot is a graphic technique widely used in the visualization of the outcomes of multiple slice microarray normalization methods [13] and allows us to display a graphical summary of the distribution of each data set.

Figure 2(a) shows the boxplot generated by the original replicated microarray data and outputted by MATLAB's *boxplot* function. Note that, there is a high variability of interquartile range across the replicated microarray sets. This behavior indicates that some replicated data sets has a larger spread of the expression levels than others. Furthermore, it can be observed that the median of expression levels varies considerably across the replicated sets, indicating that the central location of the data has different levels among all data sets.

Figure 2(b) shows the boxplot of the normalized data sets using the LAD based normalization method. The spread of the normalized expression levels shows consistency across replicated data sets. More precisely, a much lower interquartile range variability across replicated sets and a median value close to a horizontal line is achieved using the proposed normalization method, satisfying thus the goals of a good normalization method[13].

For comparison purposes, figures 2(c) and 2(d) shows the boxplot of the normalized data for Wang A and B, respectively. These boxplots certainly show a more consistent spread of the data across replications compared to the raw data boxplot. However, the interquartile range has much more variability than the one obtained using LAD based normalization approach. Note that, the interquartile range of the last four replicated data sets is smaller than the interquartile range of the remaining sets.

5. CONCLUSIONS

In this article, we propose an iterative normalization method of cDNA microarray data based on LAD regression. This method uses all gene expression data in the estimation of the normalization coefficients and adds robustness to the estimation of the normalization parameters assuming that the errors between replicated sets obey a zero-mean Laplacian distribution. The behavior of the proposed iterative normalization method was evaluated using three performance measures, and these performance measures show that the proposed approach outperforms other methods based on linear regression. Convergence as well as computation time are ongoing research issues and will be reported elsewhere.

6. REFERENCES

[1] M. Lee, F. Kuo, and J. Sklar, "Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cdna hibridizations," *Proc. Nat. Acad. Sci.*, vol. 97, no. 18, pp. 9834–9839, Aug. 2000.

[2] J. Novak, R. Sladek, and T. Hudson, "Caracterization of variability in large-scale gene expression data: Implications for study design," *Genomics*, vol. 79, pp. 104–113, jan 2002.

[3] N. Morrison and D. Hoyle, "Normalization: Concepts and methods for normalizing microarray data," in *A Practical Approaches to Microarray Data Analysis*, D. Berrar, W. Dubitzky, and M. Granzow, Eds. Dordrecht: Kluwer Academic Publishers, 2003.

[4] Y. Wang, J. Lu, R. Lee, Z. Gu, and R. Clarke, "Iterative normalization of cDNA microarray data," *IEEE Trans. Inform. Technol. Biomed.*, vol. 6, pp. 29–37, Mar. 2002.

[5] C. Workman, L. Jensen, H. Jarmer, and R. Berka, "A new non-linear normalization method for reducing variability in dna microarray experiments," *Genome Biology*, vol. 3, no. 9, 2002. [Online]. Available: http://genomebiology.com/2002/3/9/research/0048

[6] K. Bloch and G. Arce, "Nonlinear correlation for the analysis of gene expression data," in *Proceedings of the 2002 Workshop on Genomic Signal Processing and Statistics*, Raleigh, North Carolina, Oct. 2002, pp. 11–16.

[7] Y. Li and G. Arce, "A maximum likelihood approach to least absolute deviation regression," *EURASIP Journal on Applied Signal Processing*, no. 12, pp. 1762–1769, Sept. 2004.

[8] G. Arce, *Nonlinear Signal Processing: A statistical approach*. Hoboken, NJ: Wiley Sons, 2005.

[9] L. Yin and R. Yang, "Weighted median filters: A tutorial," *IEEE Trans. Circuits Syst.*, vol. 43, pp. 157–189, Mar. 1996.

[10] G. Arce, "A general weighted median filter structure admitting negative weights," *IEEE Trans. Signal Processing*, vol. 46, pp. 3195–3205, Dec. 1998.

[11] (2005) ArrayExpress at the European Bioinformatic Institute. [Online]. Available: http://www.ebi.ac.uk/arrayexpress/

[12] Y. Chen, E. Dougherty, and M. Bittner, "Ratio-based decisions and the quantitative analisys of cDNA microarray images," *J. Biomed Opt.*, vol. 2, pp. 364–374, Oct. 1997.

[13] G. Smyth and T. Speed, "Normalization of cDNA microarray data," *Methods*, vol. 31, pp. 265–273, 2003.