University of Delaware
Department of Electrical and Computer Engineering
Computer Architecture and Parallel Systems Laboratory

New Normalization Method and Error Analysis for Gene Expression Microarray Data

Stanley D. Luck†
Francisco Jose Useche G.
Wellington S. Martins
Guang R. Gao

CAPSL

September 13, 2000

†Dupont BioInformatics, Newark, Delaware, USA

Abstract

The recent development of complementary DNA microarray technology provides a powerful analytical tool for genetic research. This tool allows one to study expression levels in parallel which represents an enormous gain in terms of experimental time invested. But always while carrying out comparisons of expression data using measurements from a single array or multiple arrays, the question of normalizing data arises. An essential answer is needed to be established to the following question: Does the variation in the intensity data represent true variation in the expression level of the genes present in the analysis or is this variation the result of experimental variability? If the problem addressed by this question is not resolved, any further type of data mining process on the data is worthless. The role of normalization is to separate true variation in expression values from differences due to experimental variability. As mentioned before the microarray technology is a powerful tool that permits to study the expression level of thousands of genes at the same time, but usually in an experiment involving thousands of genes there are only few genes that are really of interest, those genes that overexpress as a "response"to the experiment practiced. Therefore it is useful to develop a method that will provide statistical information about these genes avoiding the processing of the data on the whole set of genes. In this technical memo it is proposed a new normalization method and error analysis that ultimately will provide the scientist with a statistical tool that wil allow to focus on a considerably reduced subset of genes from an originally much larger dataset. The usefullness of this type of reduction is justified under the scope of further analysis as for instance subsequent clustering techniques applied over the reduced data set.

Contents

List of Figures

1 Why normalization?

When performing several experiments over the same set of genes, which is the case of microarray gene expression experiments, it is desirable to maintain an environment where all the conditions remain constant through all the experiments, except the condition that is purposely varied in order to record change in expression level with respect to this newly changed parameter. Unfortunately the experimental process of working with gene microarrays does not fulfill this requirement. During a typical microarray experiment, many different variables and parameters different than the one being surveyed can affect the measured expression levels. Among these are the amount of applied target, extent of target labelling, efficiencies of fluor excitation and emission, detector efficiency, slide quality, dye characteristics, scanner quality, and quantification software characteristics, just to name a few. The various methods of normalization aim at removing or at least minimizing expression differences due to variability in these parameters.

2 Introduction

Normalization is the process that helps us to separate true variation in expression values from differences due to experimental variability and it is a challenge to devise a normalization method that will remove experimental variability without modifying the information of gene expression level. A very popular normalization method is the one that uses two-color fluorescence. For instance, a red labeled probe (from a healthy tissue) can be used as control to examine expression profiles in a green labeled probe prepared from an "ill" tissue. The normalized expression values for every gene are calculated as the ratio of the control and ill expression levels. Even more sophisticated methods have been devised using three colors where one color serves as control for the amount of spotted DNA and the other two colors are used to compare the samples. Another type of normalization method consists of using a set of control spots (genes) on the array. With a set of control spots it is possible to control for global variation in overall slide quality or scanning differences. The procedures mentioned above are array based methods to normalize data. However, even with multiple color fluorescence and control spots, undesired experimenttal variation can contaminate expression data. In some cases it is possible that this physical measurements for normalization are missing, making the need for additional ways of normalization an important issue. Fortunately there are other statistically based normalization methods. There are esentially two strategies that are used while performing the statistical normalization process. The first is based on a consideration of all the genes in the sample (global normalization), and the second one, on a designated subset of the genes expected to be unchanging over most circumstances. In

those cases where the expression level change is expected to be small and thus the levels are going to be close to each other the choice of global normalization is a wise decision. As the level samples become more divergent, the fraction of genes showing changed expression levels incresases, and global normalization yields a poorer estimate of normalization than would be achieved using a subset of constantly expressed genes. On this technical memo we propose a new global normalization method and error analysis so that intensity expression levels may be used for deciding significant differences in sample expressions across the gene population discernible on a microarray, by first normalizing the data and afterwards developing an error curve which identifies outliers. The new normalization method normalizes each intensity value for a specific gene at agiven experiment by taking into account all the values of that particular gene through all the experiments performed.

3 Microarray Data Normalization: A Problem Formulation

The necessity of normalization arises whenever handling gene expression data from microarrays. Every measurement as every experiment has a random component of experiment variability which does not allow to compare directly (raw intensity measurements) after obtaining the measurements between experiments for instance. This random component which is different for every experiment has to be removed. This is what we know as the normalization process. The key intuition of the normalization method is that we normalize each experiment vector independently with respect to the same variable (the norm of the expression vector), thus achieving the normalization of every experiment vector with respect to all the others.

For the normalization analysis it is assumed that there are N experiments which correspond to N microarrays. Each microarray chip contains K different genes or species.

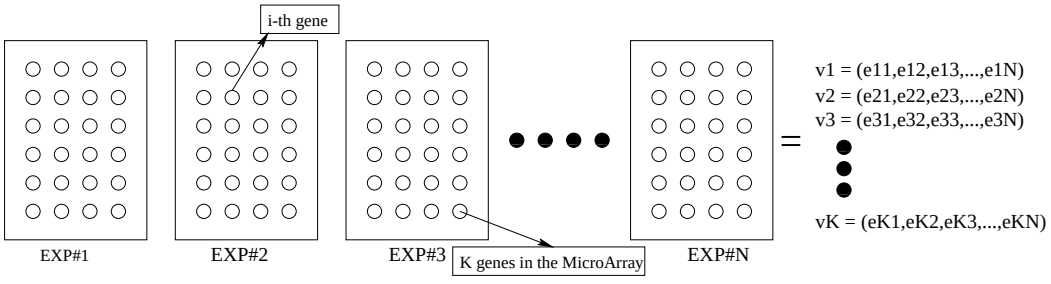


Figure 1: Basic Experiment Setup

Figure 1 exemplifies the definition of the expression vectors. An expression vector is a vector in the N dimensional space where each component (spot) represents a physical

intensity value (this measurement is achieved by processing digitized microarray images) for a given gene in a given experiment. Let us give a more complete definition of the terms involved in the analysis:

We define an expression vector as:

$$\upsilon_i = \{e_{i1}, e_{i2}, e_{i3}, ..., e_{iN}\} \quad for\ i = 1, ..., K. \tag{1}$$

where e_{ij} with j=1..N represents the intensity value obtained for the i-th gene in the j-th experiment. An expression vector contains all the expression levels related to a particular gene.

We define an experiment vector as:

$$\varepsilon_j = \{e_{1j}, e_{2j}, ..., e_{Kj}\} \quad for\ j = 1, ..., N. \tag{2}$$

where e_{ij} with i=1..K represents the intensity value obtained for the i-th gene in the j-th experiment. An experiment vector contains all the intensities belonging to a particular experiment.

The whole data for the N experiments are thought as a matrix as shown below in (3) and (4).

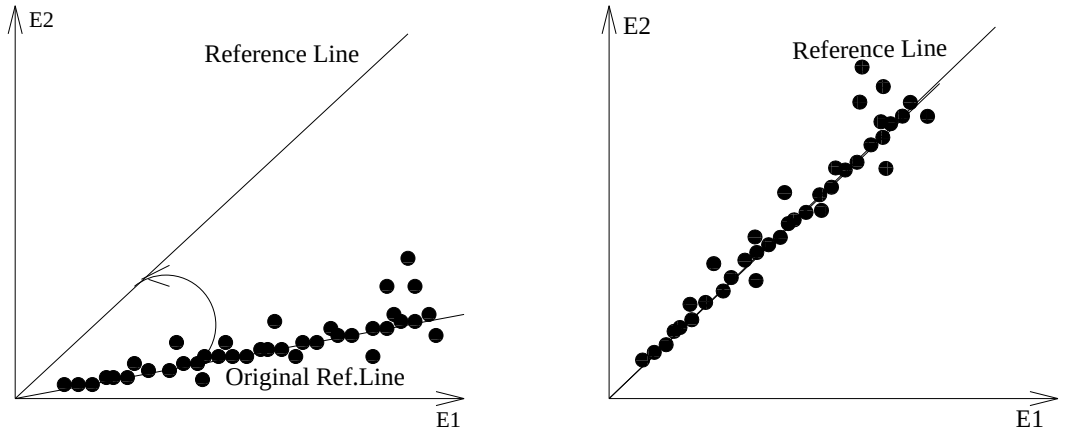
	E1	**E2**	**E3**	...	**EN**
v1	$e11$	$e12$	$e13$	...	$e1N$
v2	$e21$	$e22$	$e23$	...	$e2N$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
vK	$eK1$	$eK2$	$eK3$	...	eKN

(3)

$$m = \begin{pmatrix} e11 & e12 & e13 & ... & e1N \\ e21 & e22 & e23 & ... & e2N \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ eK1 & eK2 & eK3 & ... & eKN \end{pmatrix} \tag{4}$$

Where the rows are expression vectors and the columns are experiment vectors. We have K expression vectors of dimension N. Therefore we can consider we are working in the N dimensional intensity space.

It is possible, and usually this is the case, that if we pairwise compare the experiment vectors (taken from the raw data) there will be one or more pairs that are consistently

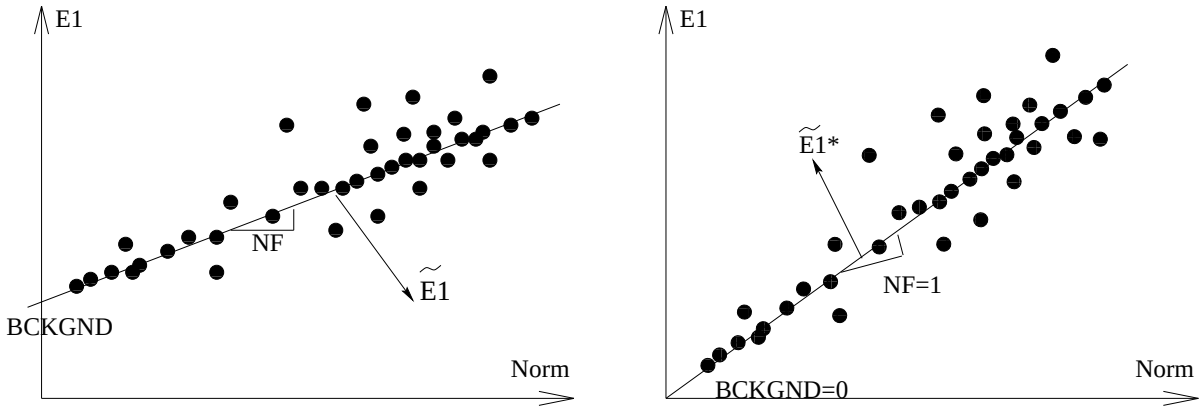


(a) (b)

Figure 2: Effect of Normalization on intensity data

and significantly different from each other. Where the differences are due to experimental variability. For example it might be very misleading to compare the intensity values of the two experiment vectors in figure 2(a). The same two experiment vectors plotted after normalization figure 2(b) show clearly that most of the genes fall in the identity line as would be expected for two different experiments performed over the same set of genes. In the general case we have in total N experiment vectors that have to be normalized.

The key intuition of the normalization method is that we normalize each experiment vector independently with respect to the same variable (the norm of the expression vector), thus achieving the normalization of every experiment vector with respect to all the others. This idea can be better understood through the following analysis:



(a) (b)

Figure 3: Normalization for experiment vector E1

We assume that for each experiment vector ε_j the experimental variability can be modeled as a linear average and therefore we perform a linear regression [1] over the set of pairs of data (Norm(v_i), e_{ij}) for i=1,...,K. As can be seen in figure 3.

The $Norm\ (v_i)$ is defined as:

$$Norm\ (v_i) = \sqrt{(e_{i1})^2 + (e_{i2})^2 + ... + (e_{iN})^2} \tag{5}$$

In other words, to each component e_{ij} of the experiment vector we associate the corresponding norm of the i-th expression vector (Norm(v_i)). This idea can be better understood under the scope of the following analysis:

The linear average of the data $\tilde{\mathbf{E_1}}$ is assumed to be given by:

$$\tilde{\mathbf{E_1}} = (NF) * \mathbf{Norm} + BCKGND$$

$\tilde{\mathbf{E_1}}$ is represented by the line in figure 3(a).

Where :

$$\text{BCKGND} = \text{Background, NF} = \text{Normalization Factor}$$

The NF (representing the slope of the line fitting the intensity values) and the BCKGND (representing the cross of this line with the intensity axis) are given by the linear regression.

If we perform the following operation over $\tilde{\mathbf{E_1}}$ to obtain $\tilde{\mathbf{E_{1*}}}$, figure 3(b), we will move the whole line, corresponding to the fit, towards the identity line shifting at the same time all the intensity values towards the identity line:

$$\tilde{\mathbf{E_{1*}}} = \frac{\tilde{\mathbf{E_1}} - BCKGND}{NF} = Norm$$

And if this procedure is repeated for all experiment vectors, we will obtain the desired normalization, given by:

$$\tilde{\mathbf{E_{1*}}} = \tilde{\mathbf{E_{2*}}} = \tilde{\mathbf{E_{3*}}} = ... = Norm$$

This means that all the expression vectors were moved towards the Reference Line (identity) as seen in figure 2(b) for two dimensions.

The actual process of normalization is implemented in each compenent of the experiment vector as shown:

$$e_{ijN} = \frac{e_{ij} - BCKGND}{NF} \tag{6}$$

Because this is a global normalization method, it is assumed that all the expression vectors will be around the Reference Line (weakly disturbed), in order for the normalization to give an accurate result. Once done the normalization we are certain of the "cleanness"of our data but still we are faced with the challenge of analyzing thousands of genes. We will show through our error analysis that many of those genes can be discarded due to their low statistical significance and therefore we will find a smaller more manageable subset of genes.

4 Error Analysis

The main objective of the error analysis is to provide information about genes that had a statistically significant expression with respect to the expression of the rest of the genes in the same set. This partition of the genes in two sets benefits any further analysis processing.

4.1 Definition of terms

First let us define the following terms in order to make the error analysis easier to understand:

Any expression vector has, in the N dimensional intensity space, the following characteristics associated with it.

- The Norm or vector length

$$Norm = \sqrt{ei1^2 + ei2^2 + ... + eiN^2} \tag{7}$$

- The angle θ which is defined as the angle between each particular expression vector and the Reference Line.

- The projection P (dot product between the expression vector and the Reference Line) which is the projection of the expression vector over the Reference Line defined as:

$$P_i = \frac{\sum_{j=1}^{N} e_{ij}}{\sqrt{N}} \tag{8}$$

- The deviation δ defined as the Euclidean distance between the expression vector and its projection vector on the Reference Line.

$$\delta \ = \ \|v_i - p_i\| \tag{9}$$

  Where in this case p_i is the vector with the same direction as the Reference Line and Norm $=$ P_i.

It is important to mension that the Reference Line is a unit vector in direction $\frac{1}{\sqrt{N}} * (1,\ 1,\ 1,\ ...,\ 1)$.

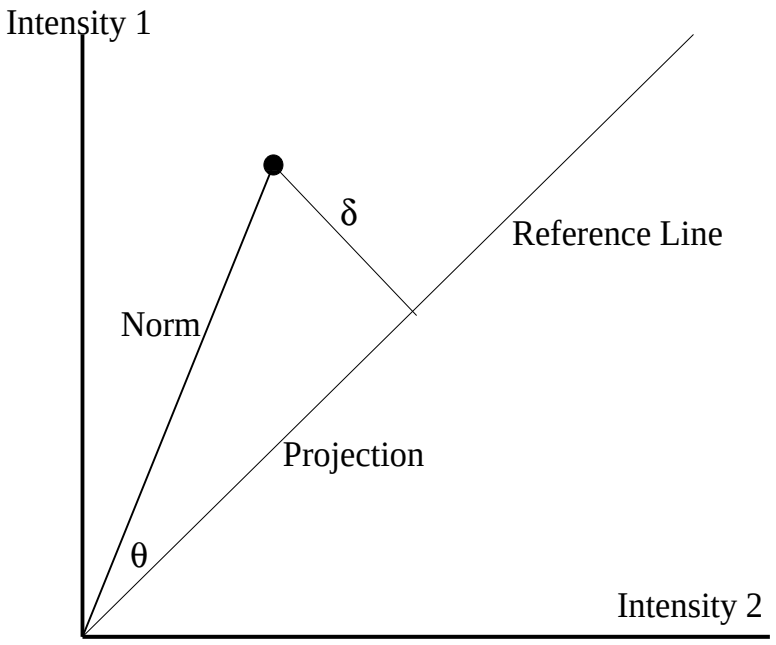


Figure 4: Characteristics of an expression vector

These characteristics of an expression vector are defined for the N dimensional intensity space. These concepts can be better grasped if we consider them in two dimensions as shown in figure 4.

4.2 Analysis

The analysis is subdivided in two main steps:

- Variance of Deviation
- Error Curve Analysis

4.2.1 Variance of Deviation

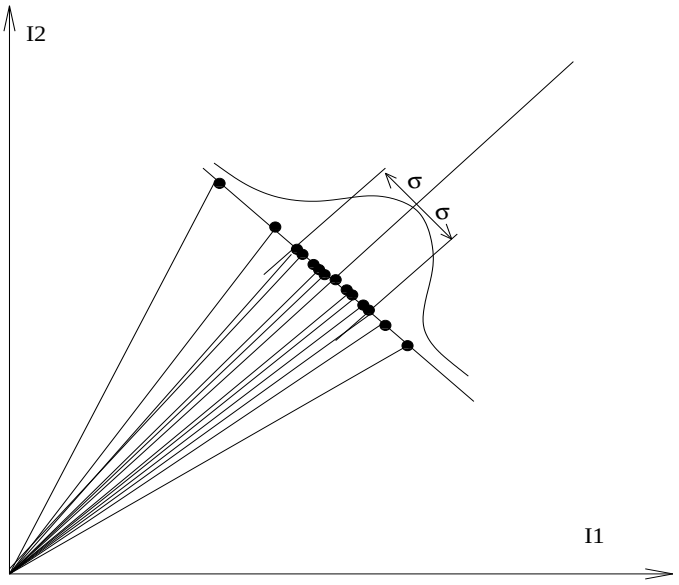


Figure 5: Representation of the distribution for the deviation at a particular projection value

As was mentioned previously it is assumed that the experiments are weakly disturbed and therefore all the expression vectors will be distributed as a type of "cloud" around the reference line. It is also assumed that for each projection value on the reference line the deviations of all expression vectors with this particular projection value will be normally distributed around the reference line. This concept is illustrated for two dimensions in figure 5. Because the expression vectors are normally distributed around the Reference Line a good statistical delimiter is the standard deviation σ of the deviation δ. In order to calculate σ we calculate first the variance of δ which is σ^2. This is done by making a nonliner regression where the independent variable is the Norm of the expression vector and the dependent variable is the square of the deviation δ^2 which will give us a fit that represents the variance of δ as explained below. We define the variance of the deviation for a particular Norm value as:

$$VAR[\delta] = E\{(\delta - \mu_\delta)^2\} \tag{10}$$

$$Since \quad \mu_\delta = 0 \tag{11}$$

$$Then \quad VAR[\delta] = E\{(\delta)^2\} \tag{12}$$

Where $E\{(\delta)^2\}$ is equivalent to the fit done by the non linear regression. Then after the regression the variance for the deviation is approximated as:

$$VAR[\delta] = C + A * Norm_i^B = \sigma^2 \tag{13}$$

Where A, B, and C are coefficients given by the non linear regression analysis [1].

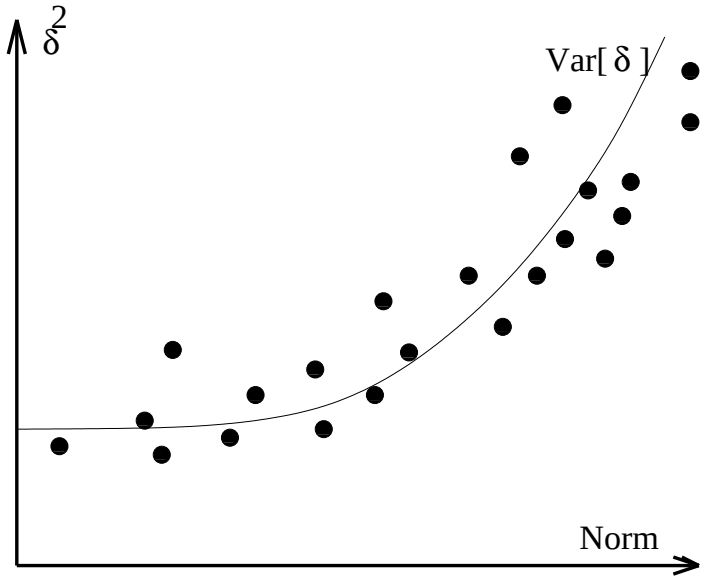


Figure 6: Curve fit to calculate Var[δ] in plot δ^2 vs Norm

This type of approximation is illustrated in figure 6.

4.2.2 Error Curve Analysis

After the previous step in the analysis we will now define the outlier genes. In other words, those genes that should really be considered for subsequent analysis. After having calculated the standard deviation σ we can now proceed to use this information to apply a statistical delimiter that will show us which genes have a significant statistical expression level. If in this case we calculate, with the coefficients already obtained through the nonlinear regression, the standard deviation σ in terms of the projection, we obtain:

$$VAR[\delta] = C + A * P_i^B = \sigma^2 \tag{14}$$

We will be able to give for each projection value a delimiter consisting in a value proportional to the calculated standard deviation σ. This type of analysis can be visualized in two dimensions in the figure 7, where the constant of proportionality in this case is 1.

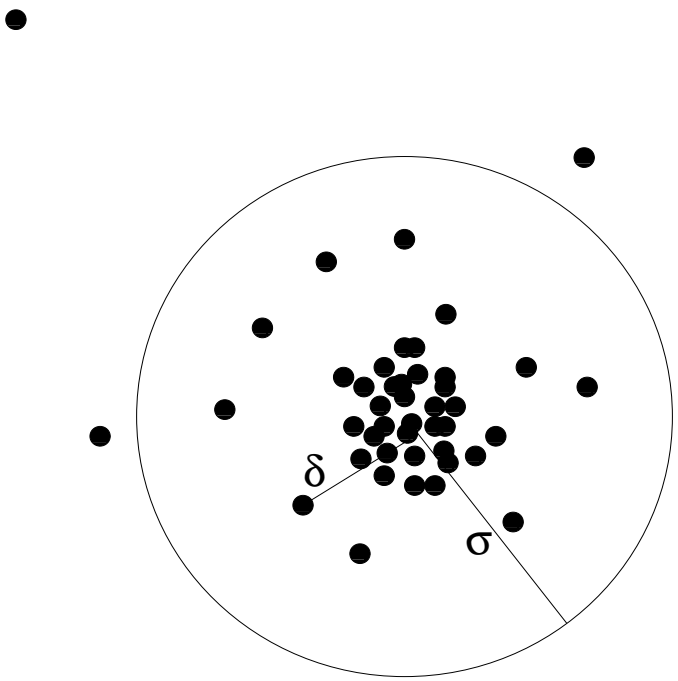


Figure 7: Distribution of expression vectors around the Reference Line in a perpendicular plane. The circle represents the delimiter imposed by the standard deviation σ

Imagine the circle is in a plane perpendicular to the reference line as shown in figure 5. Taking into account this type of analysis, we now can make a 2-dimensional plot were all this information (in the n dimensional intensity space) about genes that lie within or outside the delimiting region will be condensed. For each projection value P_i there is an associated perpendicular plane and circular delimiting region. Thus this can be plotted in two dimensions using as the independent variable the projection, and as the dependent varible the tangent of the angle between the expression vector and the Reference Line which is directly related with the deviation by:

$$tan(\theta_i) = \frac{\delta}{P_i} \tag{15}$$

The function tangent is used in order to give a better resolution for larger values of θ. The delimiting circular region (what we have called the error curve) will be represented as a single value for each projection value given by:

$$errorcurve(P_i) = \frac{\sigma}{P_i} \tag{16}$$

This is like we had folded the whole circle (see figure 7) into a single point with a distance σ from the Reference Line. The values above the error curve will be considered

outlier genes. The result of such analysis will be a plot like the one of figure 8.

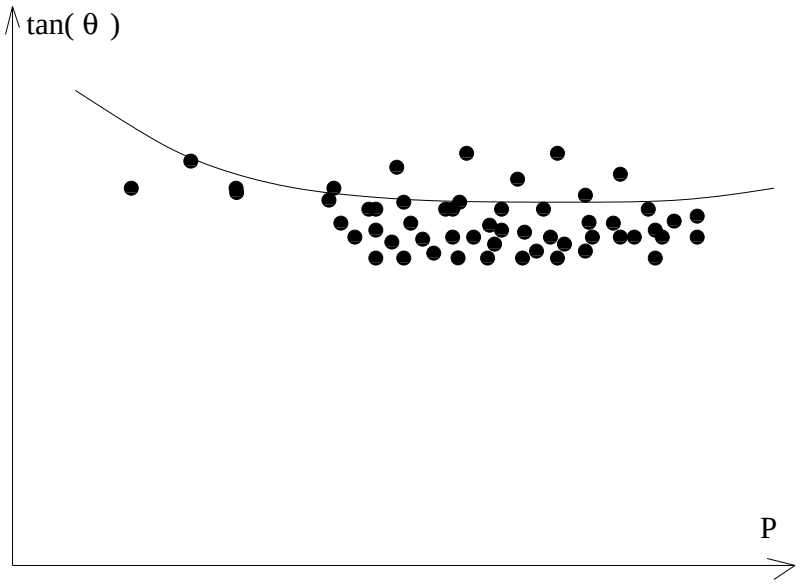


Figure 8: Error Curve analysis: tan (θ) vs P_i

5 A visualization tool for the analysis

For the purpose of implementing the analysis it was developed an application as a means for visualization of the results. This tool was developed in java (jdk 1.2) programming language. The JClass Chart API was used to produce the charts that implement the graphics previously mentioned in the analysis. The values in both axes ("x"and "y") were presented in logarithmic scale due to the large range of the data. The data used as input for the application with the information about gene identifiers and intensity values must be in a simple text format. The graphics created by the application are shown in figures 9 -14.

6 Applying the analysis to real data

The analysis previously proposed was applied to two different set of data. The first set of data belongs to Dupont and consisted in 8 experiments with a total of 4598 different genes. And the second set of data was the same set of data used in [3] with a total of 6151 genes (Pat Brown Data).

- Variance of Deviation

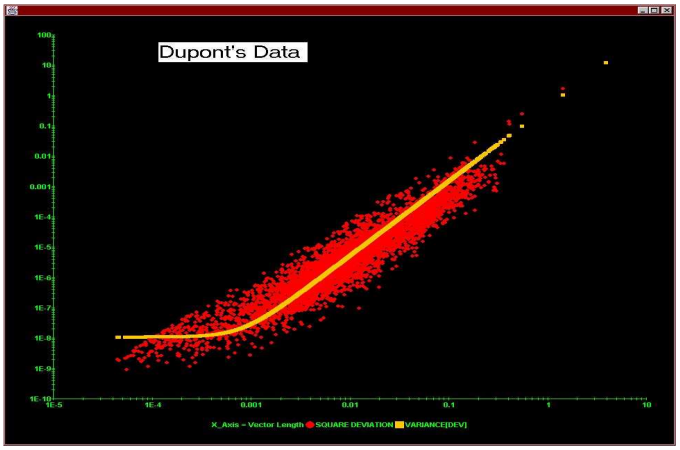


Figure 9: Dupont's Data: Square Deviation vs Vect.Length

For the Dupont set of data we present the results for the curve fit (figure 9) in order to obtain the variance of the deviation. We can observe that the fit performs well under this set of data. In this case in particular we can observe that for small intensity values (small values in the Vector Length axis) the fit is unaffected by the noise and converges to a constant value.

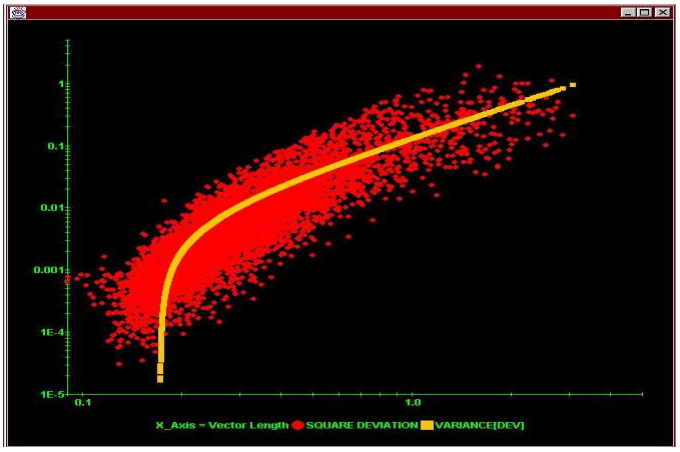


Figure 10: Brown's Data: Square Deviation vs Vect.Length (Without background correction)

For the Brown Data set (figure 10) we present the results obtained in the variance of the deviation analysis. Initially the curve fit done by the linear regression converged for small intensity values towards very small values. This was because for very small intensity values the proportional noise in the signal is too significative.

In order to perform the appropriate curve fit to the data it was necessary to background correct the values of the variance. Basically the constant expression for

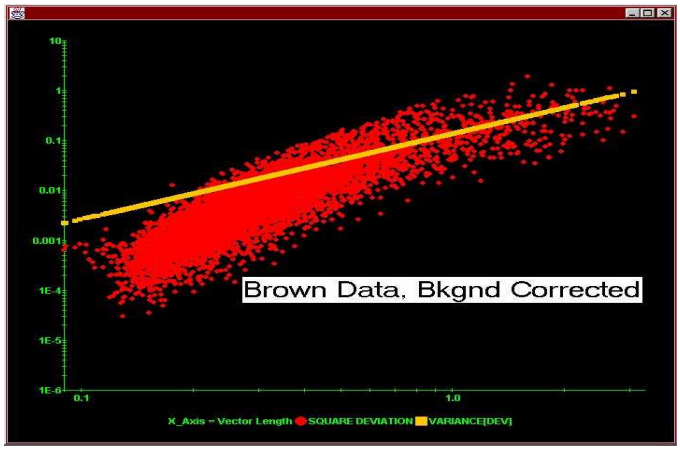


Figure 11: Brown's Data: Square Deviation vs Vect.Length (With background correction)

the variance was limited to 1xE-4 which was the average intensity value (noise) calculated from the original data (figure 11).

- Error Analysis

  For the final error analysis we present the curves obtained for both sets of data.

Figure 12: Dupont Data $\tan(\theta)$ vs. Projection (Delimiter value = 1 * σ)

In figure 12 we can see the outlier set definition for a delimiting value of one standard deviation in this case the outlier set of genes is about one half of the original gene set.

In figure 13 (Dupont data set) we can observe the curve for the error analysis with a delimiting value for the error curve of two times the standard deviation. It can be seen that the number of outliers obviously depends on the delimiting value for the error curve.

Figure 13: Dupont's Data $\tan(\theta)$ vs. Projection (Delimiter value = 2 * σ)

For the Brown data we present the error analysis curve for a delimiting value of two standard deviations in figure 14.

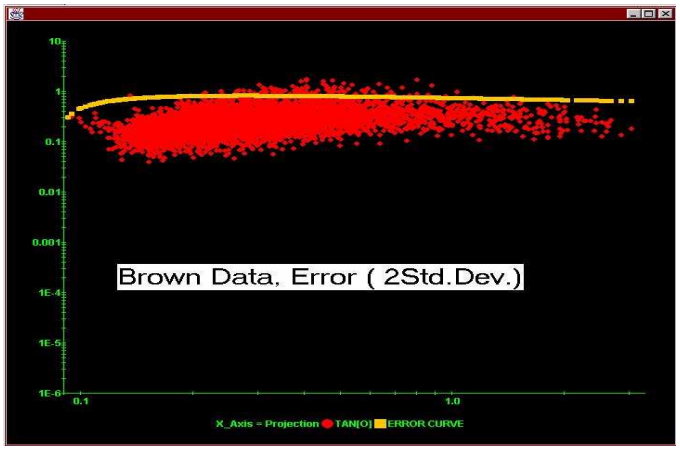


Figure 14: Brown Data $\tan(\theta)$ vs. Projection (Delimiter Value = 2 * σ)

Knowing that the statistics for each particular system are not the same, the best option would be to "fine tune" the delimiting value according to the statistics of the system under consideration and to the trade off between quality of data obtained versus not taking into account real weakly expressed genes. In the table below we show the exact number of outliers for different values of the delimiting value that were obtained for both data sets. After the set of outlier genes has been obtained in this manner the application of a clustering algorithm on this data set would deliver better results.

Data Set	Delimiting Value		Total number of genes
	σ	$2 * \sigma$	
Dupont	2002	475	4598
Brown	854	112	6151

7 Conclusions and Future Work

- In this new field of bioinformatics researchers all over the world are trying their own methods for the different processes necessary to analyze the information coming from the genetic analysis experiments. The standard procedures for all these processes are being defined by the scientific community right now. Therefore any new method that proves to work succesfully with the problem it addresses will be of great contribution for the field. Even those methods that prove not to work will have their own value as a learning experience.

- This particular method of normalization is proposed as a tool to analyze gene expression data. The main characteristics of the normalization algorithm proposed are that it facilitates the error analysis performed afterwards and the calculation of the Pearson Correlation which is a correlation used extensively by biologists. It is based on the assumption that it is very important to separate noise from real change in expression.

- The data obtained by the analysis provide valuable information about genes that have a statistically significant change in expression level.

- The error curve used in the analysis has to be defined according to the statistics of the particular system in order to be meaningful.

- For some data sets it would be necessary to perform background correction, where the background constant would be calculated as an average of the low intensity values. This would lower bound the calculated variance curve and therefore give meaningful values for the low intensity range of intensities.

- Additionaly it would be interesting to define an error threshold instead of an error curve. This threshold would be implemented using some error tolerances given to the intensity value measurements of the data instead of using one single value.

- It would be interesting to continue this work with a clustering analysis. The original main objective of this type of error estimation method is to improve the clustering analysis of error profiles. For instance Pearson correlation can be used to construct a distance matrix for expression profiles. The error analysis would indicate which of

the expression profiles are suitable for clustering and provide a score or probability for the match between a given expression profile and a cluster.

- Possible future work is to make this tool available as an application in the webpage of our group.

References

[1] Cambridge, editor. *Numerical recipes in C.* Press Syndicate of the University of Cambridge, P.O.Box 243, Cambridge, MA 02238, 1990.

[2] Yidong Chen. Ratio-based decisions and the quantitative analysis of cdna microarray images. *Journal of Biomedical Optics*, 2(4), October 1997.

[3] Joseph L. DeRisi. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Department of Biochemistry, Stanford University School of Medicine, Howard Hughes Medical Institute*, 1997.

[4] G.Getz. Coupled two-way clustering analysis of gene microarray data. May 2000.

[5] Stanley Luck. (paper on normalization and error analysis). Under Preparation, 2000.

[6] Mark Schena. *Microarray Biochip Technology.* BioTechniques Books, Natick, MA, 2000.

[7] Roland Somogyi. Making sense of gene-expression data. Technical report, Pharmainformatics, 1999.