Data Normalization and Identification of Differentially Expressed genes by Multiple Hypothesis testing procedures.

Jamal Fathima . J .I[1] and P.Venkatesan[2]

1.Research Scholar -Department of statistics National Institute For Research In Tuberculosis, Indian Council For Medical Research,Chennai,India

2. . Department of statistics National Institute For Research In Tuberculosis, Indian Council For Medical Research,Chennai,India

ABSTRACT: Normalization procedures of micro array data are presented in this work. A comparative study on identification of differentially expressed genes of hypothesis testing by regularized t-, SAM statistics is done. A discussion on non parametric approach, the permutation test based on based false discovery estimates in identifying differentially expressed genes is discussed. A study on the application of t-test and permutation tests were done

Results: Box plots using log transformations for three types of micro arrays are obtained,t-test is performed on two groups, and a histogram of the t-values are obtained there by determining the cut-off values for identifying differentially expressed genes. A comparative study of testing the significance of differentially expressed genes using different fudge factors for the standard error is done. Algorithms were implemented using the statistical programming language R

1.INTRODUCTION: In recent years with the advent of micro array technology biology has been greatly benefited in analyzing gene expression data.DNA micro array analysis allow highly parallel and simultaneous monitoring of the whole genome(Brown and Botstein,1999).Micro array technology has been increasingly used to detect differentially expressed genes (Spellman et al .,1998).The key objective of analyzing such types of experiments is comparison of different gene expression levels in varying conditions and identifying differentially expressed genes. Regular t-test and permutation tests are applied. Thus t-test can be considered as the ratio of between classes to within class variability of gene expression data. Efron et al(2001),Tusher et al (2001) and chu et al (2000) developed a different strategy where a variance component s_0 is introduced to improve the reliability of the test statistics.

2.Methedology:

2.1. Normalization: An important part of data processing is normalization. It adjusts the individual intensities such that comparisons can be made both within and between arrays in experiments. Normalization of raw data is mainly done to remove the bias which arises from variation in the micro array technology rather than from biological differences between the RNA samples or the printed probes. A difference in the data arises due to differences in print quality or from differences in ambient condition when the plates are processed. Normalization procedure differs with respect to which kind of average is used and what sources of variability are taken into account (Yang et al 2002).In micro array studies gene expression will not have the desired statistical properties such as normality or constant variance etc. Then we transform the values to get a better inference about the data. The commonly applied transformations are:

2.1.1. Logarithmic transformation:

The most commonly applied transformation is the logarithmic transformation. Logarithmic function is a monotonic function hence we apply log transformation to the micro array data. If x_{ij} represent the expression value of i^{th} gene in the j^{th} sample.

y_{ij} = log (x_{ij}). (1)

Logarithm to the base 2,10,or the natural logarithm is taken. Logarithmic transformation is applied to micro array data because it tends to provide values that are approximately normally distributed. Figure (ii),(iii),(iv) shows the box plots of log transformation of three arrays of data.

2.1.2. Box-Cox Transformation.

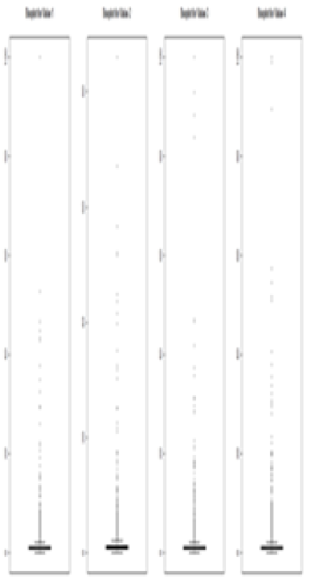
The Box-Cox transformation defines $y_{ij} = \frac{x_{ij}^{d}-1}{d} \ \forall \ i = 1,2,\ldots p, \quad j = 1,2,\ldots\ldots n.$ (2)

the square root transformation corresponds to the parameter's' being ½.

Rocke and Lorenzate(1995),Durbin et al (2002) proposes a two component error model for gene expression data in micro array analysis . Let 'X' denote the raw expression value, μ the mean expression level and 'b' the background noise then the model of log transformation is given by

$X = b + \mu e^{\eta} + \in \quad \in \sim N(0, \sigma_{\epsilon}^{2})$ (3) , random variable η and ε are taken to be independent.

Data normalization of three groups of data applying log transformation are shown in fig(i),(iii),(iiii).



Fig(i) represents the normalized data of Group I

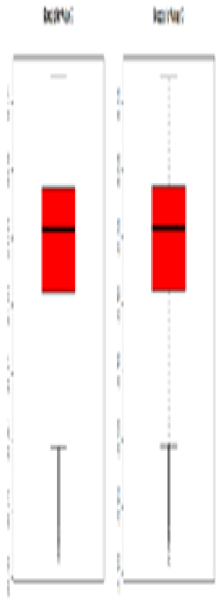
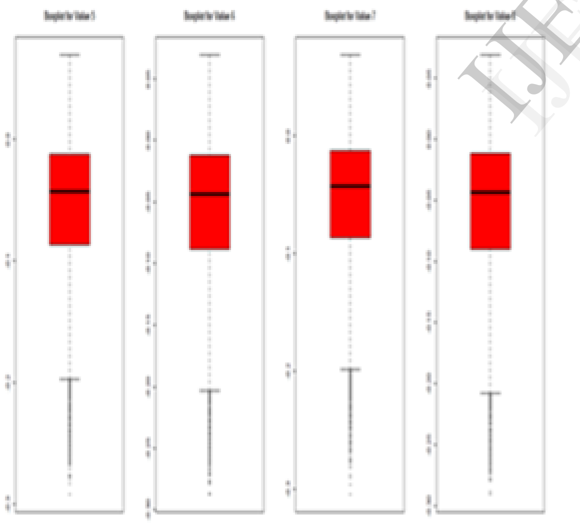


Fig (iii)represents the normalized data of Group III



Fig(ii) Represents the normalized Data of Group II

2.1.3..Square root transformation

In micro array studies the intensity readings will be proportional to the number of occurrences of fundamental molecular events such as hybridizations. The constant of proportionality will be the quantum of fluorescence, radiation produced by a single fundamental event. The fundamental event fall into two categories the true gene expression and noise. The noise contains events that are not of scientific interest. The model for such type of events is Poisson distribution. Thus the gene expression value x_{ij} is proportional to a Poisson variable.

$y_{ij} = \sqrt{x_{ij}} \quad i = 1,2, \ldots .p \; and \; j = 1,2, \ldots .n$ is the variance stabilizing function.

3.Hypothesis testing

3.1.t-test:

Comparison of two groups of samples;

The data analyzed is assumed to be normalized and divided into two subgroups. Normalization is based on the log ratios of the gene intensities. In micro array data analysis where p ≥ N the main goal is to assess the significance of individual features. The feature assessment problem leads one to multiple hypothesis testing problems. Suppose we have two samples of genes the normal and the infected group. To identify the informative or the significant genes a two-sample t-statistics is computed for each gene.

$$t_j = \frac{\bar{x}_{2j} - \bar{x}_{1j}}{se_j}, \qquad - (4)$$

$where \; \bar{x}_{kj} = \frac{\sum_i x_{ij}}{n_r} \quad r = 1,2., se_j$ is the pooled standard error for gene 'j'.

$$se_j = \hat{\sigma}_j \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad , \hat{\sigma}_j{}^2 = \frac{1}{n_1+n_2-2}\left[\Sigma\left(x_{ij} - \bar{x}_{1j}\right)^2 + \Sigma\left(x_{ij} - \bar{x}_{2j}\right)^2\right]$$

A histogram can be constructed for the t-statistics. From the histogram a cut-off value t_0,t_1 (the left and right critical values) are determined. If t_j 's are normally distributed then any value greater than the two absolute value is considered to be significant. This procedure of finding the significant genes is called the multiple testing problems. In multiple testing problem the theoretical probabilities assuming normal distribution is calculated, the p-values for each gene is calculated.

In computing t-statistics for micro array data sets the standard error obtained s_i is not very reliable as the number of samples in each group is very small. This leads to underestimation of the s_i values, and the genes with small variations give rise to extreme values of t, and are therefore false positive. To overcome this problem Tusher and Tibshirani (2001) suggested a new statistics.

$$t_j = \frac{\bar{x}_{2j} - \bar{x}_{1j}}{\frac{se_j + s_0}{2}}, \qquad (5)$$

Where s_0 is the median of the se of all the genes. Efron et al(2001) suggested the fudge factor s_0 as a particular percentile.

$$t = \frac{\bar{x}_{2j} - \bar{x}_{1j}}{se_j + s_0} \sim t, - - (6)$$

In all the above cases if $s_0 = 0$ then the statistics reduces to the ordinary t-statistics (4) which is nothing but the square root of F-statistics.

3.2.Permutation tests in micro array data:

Application of t, F tests is based on normality assumption for testing the differentially expressed genes may be unreliable because the distributional assumption may not hold true. In such situations a family of tests called permutation tests or randomization tests offer an alternative testing approach. Let us consider the testing situation in which 'n' experimental units are randomly divided into two groups of n_1 units and n_2 units respectively, where n= n_1+ n_2 , n_1 units corresponds to control condition and n_2 units correspond to treatment conditions. The response measure is represented by y_{ij} for units ,j= 1,2......$n_{i,}$, i= 1,2. Let H_0 be the null hypothesis that there is no difference in the response pattern for units subject to the treatment and control conditions. When H_0 is true the random assignment of the experimental units to treatment and control conditions imply that possible assignment of 'n' observations into groups of n_1 and n2 cases are equally probable. The arrangement may be viewed as first n1 values of the permutation of n-responses may be assigned to group I and the remainder to group II. The theory of permutation tells us that there are 'A' such permutations or arrangements where

$$R = \frac{(n_1 + n_2)!}{n_1!\, n_2!} \qquad - - - - - - (7)$$

To test the hypothesis that the response pattern for the treatment and control condition share a common feature or they differ on that feature the test statistic is given by

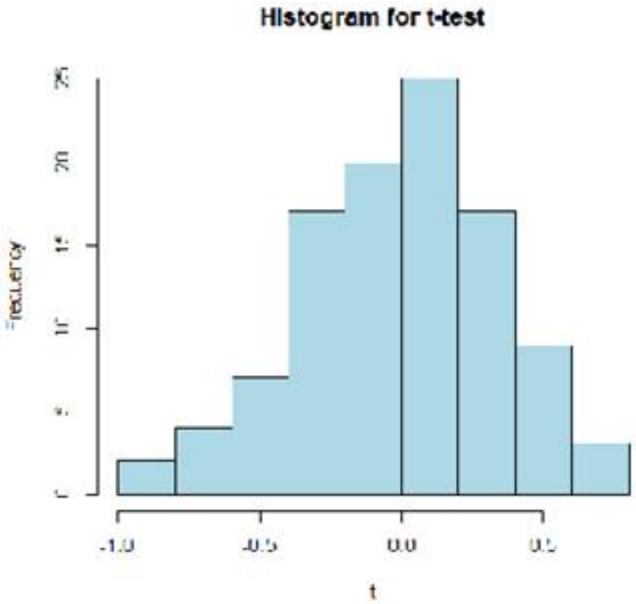
$$d^* = \frac{\sum_{j=1}^{n_2} y_{2j}}{n_2} - \frac{\sum_{j=1}^{n_1} y_{1j}}{n_1} = \bar{y}_2 - \bar{y}_1 \qquad - - - (8)$$

Between the two groups, where $\bar{y}_i$ is the mean of ith group. Accept H_0 if $|d|$ is too large otherwise reject. Under H_0 each permutation q of n responses can be considered as a realization of the experimental study .The analysis leads to a calculated difference in mean response .Let us denote the difference in mean response be denoted by d_q.The permutation procedure yields a total of R such differences d_q where R is defined in equation (1).In statistical hypothesis testing a p-value gives the consistency of the null hypothesis. In a permutation test if the null hypothesis were true the p-value is the fraction of the R calculated differences d_q that are greater or equal to the observed difference d^* in the absolute value ,that is.

$$p - value = \frac{count_q(|d_q| \geq |d^*|}{R} ------(10)$$

4.Application to Micro array data:

Data studied was obtained from NCBI GENE OMNIBUS. The log intensities of 12607 genes in breast cancer data set. Data filtering is carried out by leaf and stem method. Data consists of three groups, consisting of ten samples (Group-I control group 4 samples, Group-II treatment group 4 samples Group –III normal tissues of two sample. Gene filtering has given a total of 103 outliers from the three groups. We have numbered the output from 1-103.Data normalization of the three groups are shown by Box Plot in Fig (i),(ii) and(iii)Applying t-test Between GroupI and Group-II values are computed .t-values using fudge factor s_0,taking the value of s_0 as the median of s_i Tusher and Tibshirani (2001) ,and percentiles Efron et al(2001) are calculated as t1(median),t2(45th quartile,t3(50th quartile)and t4(55th quartile). t-values shows that around 4-8 genes accounts for maximum variation or they are differentially expressed.A comparative study of the t_i's shows that the usual t-statistics and the statistics with fudge factor due to Tusher and Tibshirani(2001) shows similar results . A histogram was constructed for the t-values obtained by equation (4). A cut – off value +/-0.5. gives the differentially expressed genes.



Fig(iv) represents the histogram of t-values obtained on GroupI and GroupII

The cut-off value t_0 ,t_1 are choosen at +/-0.5, shows that lessthan ten percent of the genes are significant.

5.Summary **:Data normalization helps in removing the bias, thus helping to have a valid statistical analysis .We have applied log transformation on the intensities, one could see from the figures that the observations after transformation seems to be similar. Hypothesis testing –t-test for identifying the differentially expressed genes are performed and compared with other testing procedures for the reliability. Introducing median and percentiles shows that ordinary t-test performs well when compared to other statistics.**

References:

1..Benjamin .Y.and Hochberg.Y.(1995) Controlling the false discovery rate :a practical and powerful approach to multiple testing. Journal of Royal Statistical Society,**B57**,289-300.

2.Dudoit.S. ,Yang.W.H,Callow.J and Speed.T.P(2000)-Statistical methods for identifying differentially expressed genes in replicated cDNA micro array experiments ,Technical Report -578.

3. Dudoit.S., Shaffer.J.P. and Boldrick.J.C(2003)-Multiple Hypothesis Testing in micro array experiments ,Statistical Science,**18,**71-103.

4.Efron.B., Tibshirani .R.,Storey.J.D.,Tusher.V(2001) Empirical Bayes analysis of a micro array experiment. Journal of American Statistical Association,**96** ,1151-1160.

5. Efron.B and Tibshirani .R.J.(1993).An introduction to the Bootstrap .London Chapman and Hall.

6. Hastie.T.Tibshirani.R, andFriedman.J.(2001).-The Elements of Statistical Learning. Springer.

7. Kathleen Kerr.K, Mitchell Martin and Gary. A. Churchill. (2000)-Analysis of variance for Gene expression micro array data, journal of Computational Biology)**7** ,pages-819-837.

8. Lambert.D.(1990) Robust two-sample permutation tests. Annals of.Statistics.**13**, 606-625.

9. Storey.J.D.and Tibshirani.R.(2003).Statistical significance for genome – wide expression. Proceeding of National Academy Science, USA,**100,** 9440-9445

10.Tusher.V.G.,Tibshirani.R,Chu.G(2001)-Significance Analysis of micro arrays applied to the ionizing radiation response, .Proceeding of National Academy Science,USA,**98(9),**5116-5121

11. Y.H.Yang and T.Speed.(2002)-Design issues for cDNA micro array experiments ,Nature Review.Genet,**3**,579-588.

12.Zhao.Y.and Pan.W.(2003) Modified non-parametric approaches to detecting differentially expressed genes in replicated micro array experiments.Bio informatics,**19(9)**-1046-1054.