Gene Selection using Multidimensional False Discovery Rate

A. Moussa[1], **M. Maouene**[1], **and B. Vannier**[2]
[1]LTI Laboratory, National School of Applied Sciences Abdelmalek Essaadi University Tangier, Morocco
[2]IPBC, University of Poitiers, Poitiers, France
Contact Author: *Ahmed Moussa ; amoussa@ensat.ac.ma*

Abstract - *This paper proposes our algorithm for gene selection in microarray data analysis comparing conditions with replicates. Based on background noise computation in replicate array, this algorithm uses the global False Discovery Rate based on 'Between' group and 'Within' group comparisons of replicates to select the set of differential expressed genes. This method uses two types of statistics that lead to improve the selection procedure when confronted to very high background noise. Using simulated datasets and the well knows Latin square data, the behavior of the proposed method is compared to results of some algorithms.*

Keywords: Gene Selection; Replicates; False Discovery Rate; Local and global FDR.

1 Introduction

The most basic question one can ask in a transcriptional profiling experiment is which genes' expression levels changed significantly [1]. Answering this question involves many considerations. There may be two experimental conditions or many, the conditions may be independent or related to each other in some way, or there may be many different combinations of experimental variables. In each of these situations, the main goal is to identify genes expressed above background levels (absolute analysis), and/or that are differentially expressed (DE) between conditions of interest. In this work we are interested to genes that are DE between replicated conditions.

A standard statistical test to detect significant changes between repeated measurements of a variable in two groups is the t-test; It can be generalized to multiple groups via the ANOVA F-statistic [2]. Variations on the t-test statistic for microarray analysis are abundant [3, 4, and 5].

For microarray studies focusing on finding sets of predictive genes, a simple method proposed by [6] computes the probability that a given gene identified as differentially expressed is a false positive by means of 'false discovery rate' (FDR). A permutation-based approximation of this method, assuming that each gene is an independent test, is implemented in the Significant Analysis of Microarray (SAM) program [3].

The variation present in microarray data poses the challenge of determining whether differences between expression measurements are caused by biological difference, or by technical variations. The best way to address this question is to use replicates for each condition studied. There are two primary types of replicates: technical and biological. Technical replicates involve taking one sample from the same source tube and analyzing it across multiple conditions (multiple microarrays). Biological replicates are different samples measured across multiple conditions (multiple samples). The use of replicates offers three major advantages:

- Replicates can be used to measure variations in the experiment so that statistical tests can be applied to evaluate differences. This property will be more explored in this paper.

- Averaging across replicates increases the precision of gene expression measurements and allows the detection of smaller changes to be detected. As the number of replicates increases, both the detectable difference from background and the detectable fold change decrease [7].

- Replicates can be compared to detect outlier results (that may occur) due to aberrations within the arrays, the samples, or the experimental procedures. The presence of outlier sample can have a severe impact on the interpretation of data. Most array platforms have internal controls to detect various problems in an experiment. However, internal controls can not identify all issues.

Multiple studies have shown that fold change on its own is an unreliable indicator [7]. If multiple measurements (i.e. replicates) exist for each gene within each condition, the measurement of variations can be estimated [8].

2 Local and Global FDR

Noting V the random variable representing the number of false discoveries and R the number of significant results obtained from a particular multiple testing procedure, [6] defined the FDR by :

$$FDR = E(V/R)\ if\ R>0, and\ 0\ otherwise \quad (1)$$

The positive FDR (pFDR) defined by [9] (for R>0), is:

$$pFDR = Pr(H = 0/T \in \Gamma) = \frac{\pi_0 P_r(T \in \Gamma / H = 0)}{P_r(T \in \Gamma)} \quad (2)$$

where H is the variable such as H = 0 if the null hypothesis H_0 is true, H = 1 if the alternative hypothesis H_1 is true, π_0= Pr(H = 0) is the probability of not being modified and T is the test statistic used for all tested hypotheses. pFDR and FDR are asymptotically equivalent and, in the following, we will note FDR for both of them.

Data provided from microarray in gene expression analysis can be considered as composed of two subpopulations of genes, those for which the null hypothesis is true (unmodified genes or non DE genes), and those for which the alternative hypothesis is true (modified genes or DE genes). Let $p_i, i = 1,\ldots,m$ be the *P-values* calculated for the m tested hypotheses. Let *P* be the random variable for which the *P-values* are the observations and let *f* be the marginal probability density function (pdf) of *P*. Denote f_0 the conditional *pdf* of *P* under the null hypothesis and f_1 the conditional *pdf* of *P* under the alternative hypothesis. Then:

$$f(p) = \pi_0 f_0(p) + (1-\pi_0) f_1(p) \quad (3)$$

In this setting, the local false discovery rate is:

$$fdr(p) = \pi_0 \frac{f_0(p)}{f(p)} \quad (4)$$

The local fdr can be interpreted as the expected proportion of false positives if genes with observed statistic are declared DE. Alternatively, it can be seen as the posterior probability of a gene being non-DE.

The main problem is the π_0 estimation. One solution assumes that the marginal distribution of the P-values arises from a beta-uniform mixture distribution. The model parameters are estimated using the maximum-likelihood method [10]. However, the widely estimator for π_0 is the one proposed by [11]. Using a tuning parameter $\lambda \in [0,1]$, π_0 is estimated by:

$$\pi_0^e = \frac{\neq\neq \{p_i > \lambda; i = 1,\ldots,m\}}{m(1-\lambda)} \quad (5)$$

In [12], the local fdr is generalized to multidimensional fdr for more one P statistic. For example in the two dimensional case, we can use two different statistics P_1 *and* P_2 that capture different aspect of the information contained in the data. The obtained fdr-2D can be expressed as:

$$fdr2D(p_1, p_2) = \pi_0 \frac{f_0(p_1, p_2)}{f(p_1, p_2)} \quad (6)$$

An already established graphical display for studying the trade-off between effect size and significance is the volcano plot of log10-P-values versus fold changes [13], corresponding to:

$$p_{1i} = mean(x_{i1}) - mean(x_{i2}) \text{ and } p_{2i} = -log_{10} P - value_i \quad (7)$$

where $mean(x_{i1})$ is gene-wise group mean.

In multidimensional case, the global FDR is the average of the local fdr for all used statistics. This FDR is a useful relationship for characterizing a collection of genes declared DE by local methods. Suppose R is a rejection region such that all genes with multidimensional statistics $p \in R$ are called DE. The global FDR associated with genes in R is [12]:

$$FDR(R) = E(fdr(p)/R) \quad (8)$$

This means that the global FDR of gene lists found by fdr2D can be computed by simple averaging of the reported local fdr values, and consequently, fdr2D can be compared easily with other procedures in terms of its implied global FDR.

Please use the styles contained in this document for: Title, Abstract, Keywords, Heading 1, Heading 2, Body Text, Equations, References, Figures, and Captions.
Do not add any page numbers and do not use footers and headers (it is ok to have footnotes).

3 Method Description

3.1 Between and Within Group Comparisons

Consider the example where we have to compare two experiments (Traited # Control) with three replicates. For the available microarrays, we can process in term of statistics, to two types of comparisons: 'Between' group comparisons that concern chips providing from the two samples "Fig.1". And 'Within' group comparison that concern chips inside biological or technical replicates "Fig.3".

For each set of comparison, a multidimensional fdr2D, based on statistics of equation 7 may be computed. These statistics can be summarized in two volcano plots where the first one represents results of 'Between' group comparison "Fig.3": in this plot the significance correspond to the average (-log10 P-value) across all the 'Between' groups comparison and the average Signal Log-Ratio (SLR) obtained from average fold change across all the 'Between' group comparison. And the second one show the same statistics related to the 'Within' group comparison "Fig.4". This latter informs about the experiments background noise [14]. In fact, gene stimulated in 'within' group comparisons inform about amplitude and act of experimental background noise. When this noise is very low, all genes SLR are falling around 0 in this plot.

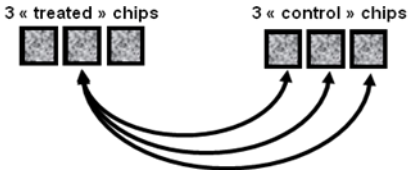

Figure 1: 'Between' group comparisons

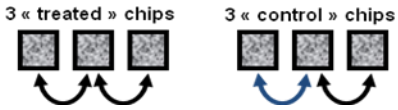

Figure 2: 'Within' group comparisons

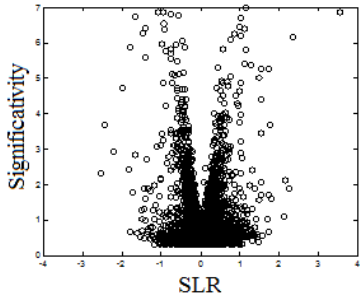

Figure 3: volcano plot of 'between' group comparison

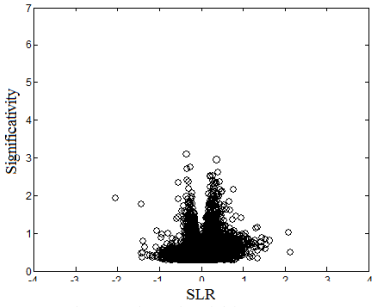

Figure 4: volcano plot of 'within' group comparison

3.2 Local fdr and Replicates

To illustrate our procedure, we use first the local fdr as described in section I. For the two sets of comparison we use the same statistics and the same null hypothesis H = 0. In this context the local fdr for 'Between' group comparison and 'within' group comparison are :

$$fdr^b(p) = \pi_0^b \frac{f_0^b(p)}{f(p)} \text{ and } fdr^w(p) = \pi_0^w \frac{f_0^w(p)}{f(p)}$$

Without loss of generality the expression:

$$FDR = \frac{fdr^w(p)}{fdr^b(p)} = \frac{\pi_0^w f_0^w(p)}{\pi_0^b f_0^b(p)} \quad (9)$$

interpreted as the expected proportion of false positives if genes with observed statistic are declared DE, is the common local FDR with the same null hypothesis[15].

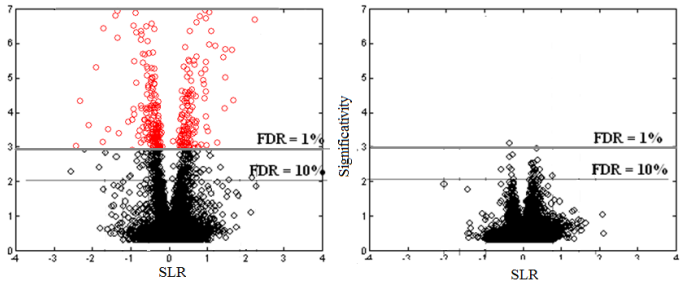

Figure 5 : volcano plot for 'between' and 'within' group comparison with the same null hypothesis.

The FDR of equation 9 changes from 0 to 1 according to the cutoff fixed by the analyst. Each FDR-cutoff value correspond to one value of significativity (-log10 (P-value-cutoff)). But in certain case, especially when the 'Within' group comparison presents a high degree of noise, this curve may not be straight monotonous and two FDR-cutoff values can corresponds to the same significativity "Fig.6". This not advisable behavior is corrected by a curve smoothing (FDR versus Significativity) with a monotonic quadratic function, where the smoothed curve guarantees the FDR uniqueness versus significance correspondence "Fig.6".

The proposed method works well when the noise observed in 'within groups' comparison is moderate. But when the background noise is high, the FDR is not well informative, and it is very difficult to find the appropriate function to extrapolate the curve FDR versus Significance. Thus, to improve the method we used two statistics (-log10 Pvalue and SLR) to generalyze this concept to the global FDR.

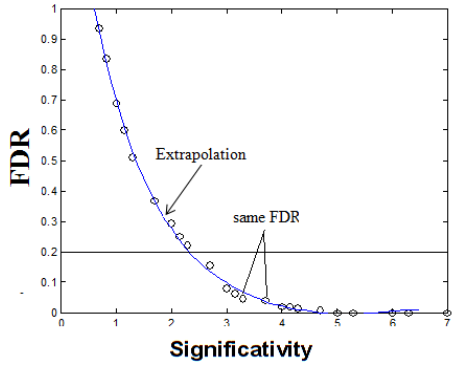

Figure 6 : smoothing the FDR vs Significativity plot

3.3 Global fdr-2D and Replicates

This solution introduces the SLR information in the selection method [16].As explained in the last section we use the local FDR for both the significance and fold change statistics. The use of two different statistics that test the same null hypothesis, but have different power against t-statistics and fold changes, comparable with the proposal made by [16], is another possibility. Thus, this method takes into account the information provided by both signals and replicates and gives a best estimate of background noise in microarray.

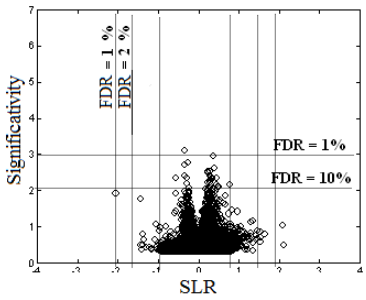

Figure 7 : FDR corresponding to the null hypothsis applied to SLR

In the selection step, the method uses conjointly FDR for significativity and FDR for SLR. This Global FDR, which uses replicates as a background adjustment is called in the next "global FDR-2D", and is expressed exactly by the equation 8.

The gene selection procedure proposed here run as follows:

1- Establish a curve, as in "Fig.6" for the studied example using a global FDR-2D values set.
2- Curve "FDR-2D versus significance" smoothing
3- Assignment of the cutoff value and search a corresponding FDR-2D in the curve (FDR-2D) cutoff

Selection of DE Genes with FDR-2D < FDR-2D cutoff

4 Results and Discussions

4.1 Simulated Dataset

We assume 10 000 genes per array with a proportion of truly non-DE genes $\pi_0 = 0.95$ throughout, and compare two independent groups with n=4 arrays per group. We further assume that the log expression values are also normally distributed in each group.

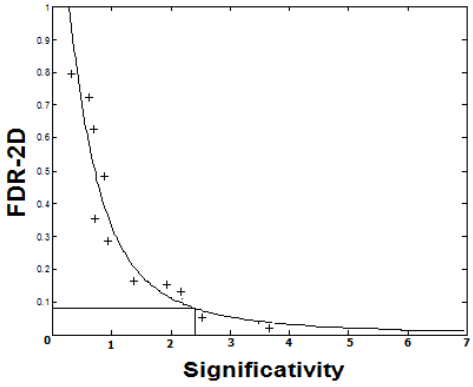

Figure 8 : smoothing the FDR vs Significativity for the simultaed dataset

We have compared results of this gene selection method to :Significance Analysis of Microarray (SAM)[3], Controlling the fdr (Benjamini method) [6], and Multidimensional local fdr [12]

In the comparison, we use three values of FDR-2D^{cutoff} e.g. 1%, 5% and 8% "Table I".

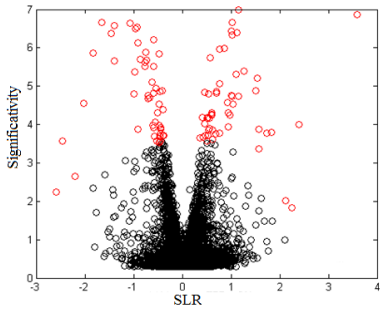

Figure 9 : Gene selected by the global FDR-2D^{cutoff}=5%

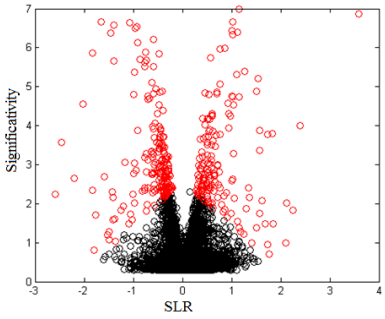

Figure 10 : FDR-2D^{cutoff}=8%

TABLE I. RESULT OF SIMULATED DATASET

	TDR			***Percentage of spike detected***		
FDR Value	1%	5%	8%	1%	5%	8%
Method 1	58.20	52.30	44.50	72.36	76.45	66.33
Method 2	68.56	45.50	35.56	66.15	70.83	67.98
Method 3	96.17	95.26	97.11	88.32	80.64	73.37
Proposed Method	**95.23**	**93.45**	**91.48**	**89.21**	**81.70**	**75.77**

4.2 Real Dataset

The proposed method was used to analyze spiked-in genes arrayed in a Latin square. In this publicly available set, 112 yeast genes and 14 human genes are cloned. Each of the labeled genes were pooled into groups and diluted to concentrations of 0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512 and 1024 pM. In every microarray experiment, 14 groups of genes in 14 different concentrations were hybridized to the microarray in the presence of a complex background of expressed human genome (30 Mb) and several control genes. For this Latin square design, 14 groups of experiments with 3 replicates for each experiment, giving a total of 42 experiments. The concentrations of the 14 in vitro transcript (IVT) groups in the first experiments are 0, 0.25, 0.5, . . . , 1024 pM, their concentrations in the second experiments are 0.25, 0.5, . . . , 1024, 0 pM, and so on [17].

The selection method proposed in this work has been applied to the Latin Square dataset. The main objective is to select a set of genes according to pre-defined P-value and compare the result with the 42 spiked-in genes. Result

summarized in "Table I" compare the results of this new selection gene method to those used in the last section for evaluating the performance of this algorithm thought simulated dataset.

TABLE II. RESULT OF REAL DATASET

	TDR			***Percentage of spike detected***		
FDR Value	1%	5%	8%	1%	5%	8%
Method 1	50.39	45.36	29.87	58.26	66.45	67.35
Method 2	65.44	66.21	69.52	67.6	68.84	75.38
Method 3	58.59	60.49	68.11	74.32	78.26	80.36
Proposed Method	**60.58**	**62.47**	**67.21**	**75.65**	**80.25**	**85.46**

Table 1 regroup results of four gene selection methods applied on statistical parameter of simulated dataset. The best percentage of spike detected was found by the global fdr-2D algorithm. Method3 and FDR-2D have the best percentage of spike detected. These results confirm the good behavior of the two methods in the case of simulated data. This conclusion is confirmed where the proposed algorithm have been confronted to complex data like Latin Square. In fact, in table II, the proposed method and the method 3 gives a good result of detected spike.

All of these results confirm on the one hand the good behavior of the proposed algorithm in the gene selection problem. on the other hand, it proof that when taking into account replicates of arrays by mean of the 'within' group comparison, the method allows good detection of modulations for weakly expressed genes and eliminates false positives.

References

[1] W. Liu, R. Mei, X. Di, T. Ryder, E. Hubbel, S. Dee, T. Webster, C. Harrington, M. Ho, J. Baid, S. Smeekens, "Analysis of high density expression microarray with signed-rank calls algorithms", Bioinformatics, vol. 18, N°. 12, 2002.

[2] J.-H. Zar, "Biostatistical Analysis" Prentice-Hall , Upper Saddle River, NJ, 663, 1999.

[3] V. G. Tusher, R. Tibshirani, G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response", Proc. Nat. Acad. Sci. USA 98, pp. 5116–5121, 2001

[4] T. R. Golub, et al, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", Science vol. 286, pp.531–537, 1999.

[5] F. Model, P. Adorjan, A. Olek, C. Piepenbrock, "Feature selection for DNA methylation based cancer classification" Bioinformatics vol. 17, N°. 1, pp. 157–164, 2001.

[6] Y. Benjamini, Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing", Journal of the Roy. Stat. Soc. vol. B 57, pp.289–300,1995.

[7] M. Newton, C. Kendziorski, C. Richmond, F. Blattner, K. Tsui, "On differential variability of expression ratios: improving statistical inference about gene expression changes from microarraydata", Journal of Comparative, Biol., vol. 8, pp.37-52, 2001.

[8] W. Pan, J. Lin, C. Le, "How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach", Genome Biolo. Vol. 3, N°.5, 2002.

[9] J. –D. Storey, "A direct approach to false discovery rates". Journal of the Roy. Stat. Soc. Serie B, vol. 64,pp.479-498, 2001.

[10] S. Pounds, W. Morris, "Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values", Bioinformatics, vol.19, pp.1236-1242, 2003.

[11] J.-D. Storey,R. Tibshirani, "Statistical significance for genome-wide studies", Proc. Natl Acad. Sci. USA, 100, pp. 9440–9445, 2003.

[12] A. Ploner, S. Calza, A. Gusnanto, Y. Pawitain "Multidimensional local false discovery rate for microarray studies" , Bioinformatics vol. 22 N°5, pp.556–565, 2006.

[13] R.-D.Wolfinger, et al. "Assessing gene significance from cDNA microarray expression data via mixed models", Journal of Comput. Biol., vol. 8, pp.625–637, 2001.

[14] G.-A. Churchill, "Fundamental of experimental design fo cDNA microarray", Nature genetics supplement 32, pp.490- 490, 2002.

[15] A. Moussa, M. Maouene, B. Vannier, "Multidimensional Method For Gene Selection" Proceeding of the 5th World Congres on Celular and Molecular Biology, Indor, India, 6-9 November 2009

[16] Y. H. Yang, et al., "Identifying differentially expressed genes from microarray experiments via statistic synthesis", Bioinformatics, vol. 21, pp. 1084–1093, 2005.

[17] W. Liu, and all. , "Analysis of high density expression microarray with signed-rank calls algorithms" Bioinformatics, vol. 18, N°.12, pp.1593-1599, 2002.

Knowledgments

This work was supported in part by the Moroccan National Center of Scientific and Technical Research of Grant ScVie 03/10.