Comparison of Statistical Tools for Microarray Data Analysis

O. Kaissi[1], A. Moussa[1], A. Ghacham[1], B. Vannier[2]

[1]LTI Laboratory,ENSA , University Adbelmalek Essaadi , Tangier, Morocco

[2]IPBC,University of Poitiers,France

Contact :amoussa@uae.ac.ma

Abstract - *The present paper proposes a comparative study of two statistical tools integrated in R-Bioconductor Project, Expander, and Bioinformatics ToolBox of Mathworks, for gene selection in microarray data analysis. The main objective is to show the impact of results on selected genes when using statistical algorithms under different environments. This study compares results related to two data sets, the first one is the well knows Latin Square Affymetrix data, and the second one is provided from a public data base.*

Keywords: Gene Selection, Statistical Algorithm, Soft Tools Comparison

1 Introduction

The technology of DNA microarrays currently experiencing an exceptional growth and has attracted tremendous interest in the scientific community. This interest lies in its efficiency; speed of obtaining results; and in its ability to study the expression of thousands of genes simultaneously [1].

The use of microarray in various fields including biology and health, allows development of several technologies grafting and in situ [2, 3]. Therefore several computational and statistical tools were developed to store, analyze and organize data [4].

A DNA chip consists of a DNA fragment immobilized on a solid support according to an ordered arrangement. The principle is based on the chip hybridization using a probe carrying the radioactive labeling [5]. Intensity of the signal generated is measured using a scanner. Image obtained, is analyzed to quantify the level of gene expression. Given the volume of data generated by this technology, several statistical methods based on the statistical t-test [6] were developed under some soft- tools for analyzing and selecting genes. But, the literature remains very poor in comparative studies showing the impact of the used algorithm and used materials in gene selection procedure. For this, the study proposed in this paper comes to show the performance of the statistical algorithm when using different soft tools.

This paper is organized as follows: an overview on Affymetrix technology and description of the three soft tools and statistical methods used in gene selection are given in section 2. In section 3, we present our comparative study of the data sets with some explanatory plots. We concluded this paper by discussing the results of this study.

2. Technologies and Tools

Affymetrix Gene Chip represents a very reliable and standardized technology for genome-wide gene expression screening [7]. In this technology; probe sets of 11–20 pairs with 25-mer oligonucleotides are used to detect a single transcript. Each oligonucleotide pair consists of a probe with perfect match to the target (PM probe) and another probe with a single base mismatch in the 13th position (MM probe) [8].

In the absolute analysis the goal is to answer the question: if the transcript of a particular gene is present or absent? The advantage to answer this question is that we can easily evaluate the expression and interpretation of results, by comparing the p-values expression levels off all genes to threshold α1 and α2. Affymetrix technology offers two levels by default of α1 and α2 significances (α1=0, 04 and α2=0, 06). Genes with expression p-values under α1 are called Present, genes with expression p-valueshigher then α2 are called Absent,the genes with p values between α1 and α2are calledMarginal (Fig.1).

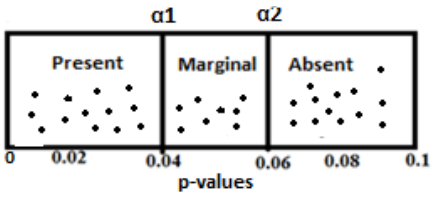

Fig.1: Significance levels in absolute analysis study

When the experiments concerned comparison of two conditions (treated # baseline) the objective of the comparative analysis is to answer the question: does the expression of a transcript on a chip (treated) change significantly with respect to the other chip (baseline)? In this context, five possible distinct answers are: Increase, Decrease, Marginal Decrease, Marginal Increase and No Change. These detections calls are giving by comparing change p-values of each gene thefour thresholds chosen by the analysis for Affymetrix technology. Those thresholds are given in the Fig.2 [9].

Based on absolute and comparative analysis results, several methods have been developed to select the genes of interest. Many of these methods would be quite appropriate if genes would be analyzed one at a time. Some methods like T-test, ANOVA and F-test can easily be carriedout for many genes simultaneously [10],

In the case of a lot of experiments, statistical test for selection is difficult to apply and multiple corrections need to be made. The most common multiple comparisons correction is the Bonferoni correction [11]: Rather than adjusting p-values for individual genes, he suggests to control the False-Discovery Rate (FDR) which is the fraction of false positives among the genes that are called, changed [12].

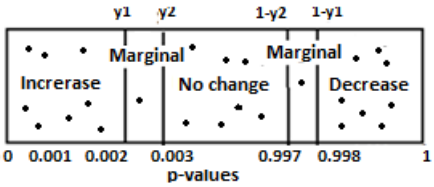

Fig.2: Significance levels in analyzer comparative study

In the comparison study of this work, we have chosen two well used methods for gene selection:

The SAM statistical algorithm [13]

The FDR controlling algorithm [11]

These algorithms, integrated in three software tools, are used as gene selection tools. Before presenting results, we recall in the two followed subsections the used data and software tools.

We used two data sets available on the public databases (NCBI and EBI) [14,15].

The first data set [16] includes 14 samples each of three replicated microarray oligonucleotides, in which multiple RNAs were added to the growing concentrations a common RNA preparation. Genes that should show variations in intensity are known (spikes genes), for this these data are generally used as references to validate developement algorithms and software.

The second data we used provide from the article [17]. In this study we have to compare healthy and affected individuals, where this last have a dysfunction of lymphocytes .Different samples were taken for each dysfunction: 10 samples with Waldenstrom Macroglobulinemia (WM), 12 with Multiple Myeloma (MM), 11 with Chronic Lymphocytic Leukemia (CLL), with normal cases, 8 of B Lymphocytes (NBL), and 5 Plasma Cells (NPC). The differentially expressed genes explain relationship between the various syndromes or dysfunction [17].

Several software's has been developed to facilitate the analysis of microarray data. In this context, the most used free softwares is Bioconductor. However, Bioinformatics ToolBox of Mathworks and Expander offer a convivial interface to analyze data provided from microarray.

Standardization of the chips is applied on all chips and assumes that the distributions of intensities must be homogeneous. Several studies have focused on the performance of different normalization methods. In this study we use the Robust Multichip Analysis algorithm (RMA). This last provides accurate estimation of inter-array variability through a robust background correction and quantile normalization computed over the whole dataset [18]. The first used software is Bioconductor that is a collaborative project using the statistical programming language R [19].It allows statistical analysis on the use of different packages grouped under the name "biocLite". Bioconductor develops between other free applications especially designed for the analysis of biological data including microarray.For the analysis of Affymetrix chips with Bioconductor, we must first ensure that the Affymetrix libraries are installed [20]. The selection of differentially expressed genesis realized by the "limma" package integrated in Bioconductor.

To assess the significance of genes, it is interesting to compare the value of 'fold change' which gives the direction of the stimulation of the gene, with the significance that quantified the importance of this direction. The volcano plot (Fig.3) arranges the genes along two axis that represent statistical significance and biological significance.

Bioinformatics ToolsBox of Mathworks offers biologists an open systems environment and stretch in which to explore ideas, prototype share new algorithms, and build applications for the analysis and simulation of biological systems [21]. It also offers interactive tools for designing and editing graphics (Fig.4).

Expander (Analyzerand Expression Displayer) is integrated software for the analysis of gene expression data. It was originally designed as a classification tool [22]. Today it has evolved to support all stages of data analysis chips, from the normalization of raw data to the inference of regulatory networks transcriptional [23].

3. Results and Discussions

We analyzed the performance of statistical tests integrated in Soft Tools cited below using Latin square and Leukemia data. Results are evaluated on the with the percentages of True Detection Rate (TDR=number of Spike detected / number of modulated genes reported). In leukemia data we consider the 69 genes cited in the work of [17] as spikes.

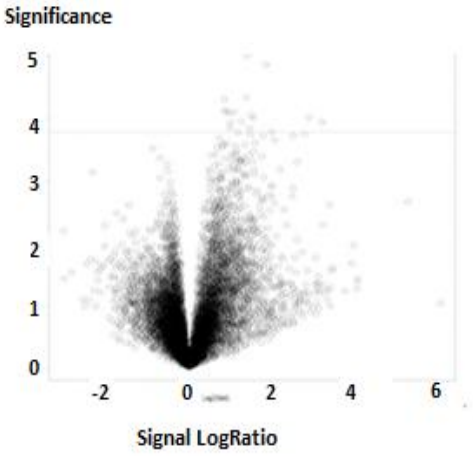

Fig.3: Volcano plot of leukemia data using Bioconductor

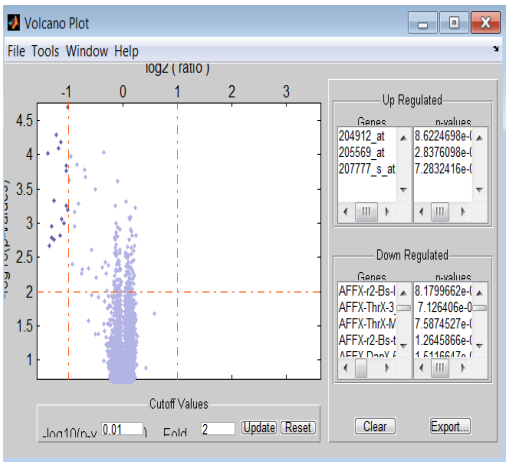

Fig.4: Volcano plot Latin Square data using Bioinformatics ToolBox of Matworks

For Both SAM and FDR controlling algorithm, we used two cutoff of pvalue for gene selection. Results are summarized in Figures 5 and 6 that represent the distributions of genes selected according to each software and each statistical algorithm.

Our comparative study allows us to define and determine that p-values 0.001is more significant than p-values of 0.01 for both SAM and FDR , and the Expander allows to select a maximum of TDR and Spike. In addition we show that this analysis confirms that selected genes depend both on the used algorithm and the used Soft Tools. This analysis gives some list of new interest genes.

Finally, we remind that this work focus the problem of used algorithm and tools in gene selection problem. In this context we have used two p-values with screening tests: FDR and SAM. To highlight the difference between these two selection methods we tested their effectiveness on three environmental developments chips Bioconductor, Bioinformatics tool box and Expander, using Latin square data and leukemia public data. We conclude that in microarray data analysis, the best way is to work with different approaches for statistical analysis at the same time for a better validation of results.

References

[1]: V. Gomases, S. Tagore and K.V. Kale, 'Microarray: an approach for current Drug targets' Current Drug Metabolism, vol. 9, pp. 221-31, 2008.

[2]: Y. F. Leung and D. Cavalieri, 'Fundamentals of cDNA microarray data Analysis', Trends Genet, vol.19, pp. 649-659,2003

[3]:.D. J Lockhart, H. Dong, M. C. Byrne, and al. 'Expression monitoring by hybridization to high-density oligonucleotid arrays' Nat. Biotechnol, vol.14,pp.1675-1680, 1996.

[4]: D. J. Duggan, M. Bittner, Y. Chen,P. Meltzer and J. M. Trent, 'Expression profiling using cDNAMicroarrays'. Nat.Genet, vol.21,pp.10-14, 1999.

[5]: V.Frouin, and X.Gidrol,.'Analyse des données d'expression issues des puces à ADN' Biofutur , vol.252, pp. 22 – 26,2005.

[6]:F. Chu and L. Wang, 'Applications of support vector machines to cancer classification with microarray data', International Journal of Neural Systems, vol. 15, pp475-484,2005

[7]: D. Lockhart, H. Dong, M. Byrne,M. Follettie, M. Gallo, M. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton and al 'High density synthetic oligonucleotide arrays'. Nat.Biotechnol, vol. 14, pp. 1675-1680, 1996.

8]: D. Choaglin, F.Mosteller, J.W. Tukey, 'Understanding Robust and Exploratory Data Analysis' Wiley, vol.79,pp.7-32, 2000.

[9]: W. mLiu, R. Mei, X. Di, T. Ryder, E. Hubbel, S. Dee, T. Webster, C.Harrington, M. Ho, J. Baid and S. Smeekens 'Analysis of High Density Expression Microarray with Signed-Rank Calls Algorithmes'. Bioinformatics, vol.12, pp.1593-1599, 2002.

[10]: D. Faller, H. U. Voss, J. Timmer and U. Hobohm. 'Normalization of DNA-microarray data by nonlinear correlation maximization' J. Comput. Biol. ,vol.10, pp. 751-762. 2003.

[11]: S.Dudoit, Y.H. Yang, M.J. Callow and T.P. Speed,'Statistical methods for identifying differentially expressed genes in replicated DNA microarray experiments'. Statist.Sinica, vol.12,pp. 111–139, 2002.

[12]: Y. Benjamani, Y. Hochberg. 'Controling the false discovery rate :a practical and powerful approach to multiple testing'. Journal of the Roy.Soc. vol. 57,pp. 289-300, 1995.

[13] : V.G. Tusher, R. Tibshirani, G. Chu, 'Significance analysis of microarrays applied to the ionizing radiation response', Proc. Nat. Acad. Sci. USA, 2001, Vol. 98, pp. 5116-5121.

[14]: www.ncbi.nlm.nih.gov/

[15]: www.ncbi.nlm.nih.gov/

[16]:www.affymetrix.com

[17]: NC. Gutiérrez and All 'Gene expression profiling of B lymphocytes and plasma cells from Waldenstrom's macroglobulinemia: comparison with expression patterns of the same cell counterpartsfrom chronic lymphocytic leukemia, multiple myeloma and normal individuals', Leukemia, vol. 21(3), pp. 541-550, 2007.

[18]: B. Bolstad, R Irizarry., MAstrand., and T .Speed, A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. Bioinformatics,vol.19, pp.185-193,2003.

[19]: www.r-project.org

[20]: G. K. Smyth 'Linear Models and Empirical Bayes Methods for Assessing Differential Expression in MicroarrayExperiments. Statistical Applications' in Genetics and Molecular Biology, vol.3 pp. 2004.

[21]:www.mathworks.com/products/bioinfo

[22]: R. Sharan, A. Maron-Katz, and R Shamir, 'CLICK and EXPANDER: a system for clustering and visualizing gene expression data. Bioinformatics vol.19,pp.,1787–1799, 2003.

[23]: R. Shamir and al., 'EXPANDER: an integrative program suite for microarray data analysis. BMC Bioinformatics vol 6 pp, 232, 2005.

Table I: Results of Latin Square Dataset

Pvalues	**0,01**				**0,001**			
Statistical Test	**T-Test (SAM)**		**FDR**		**T-Test (SAM)**		**FDR**	
	TDR	Spike	TDR	Spike	TDR	Spike	TDR	Spike
Bioconductor	40,22 %	83,33 %	53,33 %	76,19 %	54,76 %	54,76 %	55,55 %	35,71 %
Bioinformatics Tools Mathworks	35,95 %	76,19 %	42,64 %	69,04 %	49,75 %	50%	52,94 %	21,42 %
Expander	57,07 %	95,23 %	60,34 %	83,33 %	64,1%	59,52 %	65,62 %	50%

Fig.5: Number of genes selected and grouped according to the used statistical tool.

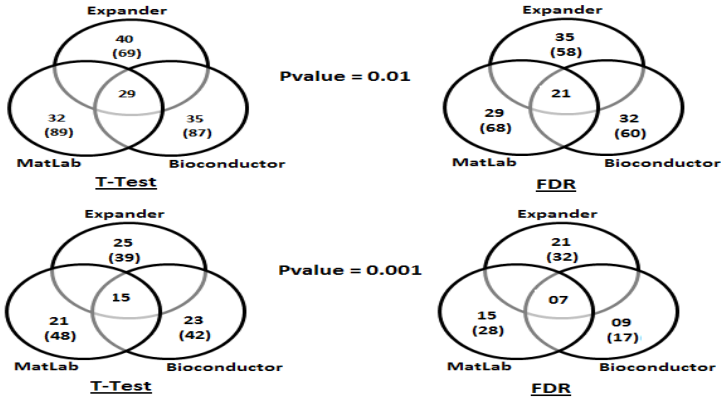

Table II: Results of Leukemia Dataset

Pvalues	**0,01**				**0,001**			
Statistical Test	**T-Test**		**FDR**		**T-Test**		**FDR**	
	TDR	Spike	TDR	Spike	TDR	Spike	TDR	Spike
Bioconductor	36,5%	86,95%	51,85%	79,71%	93,84%	56,52%	55,6%	46,37%
Bioinformatics Tools Mathworks	33,82%	79,71%	45,39%	75,36%	71,13%	50,72%	98,57%	42,02%
Expander	55,64%	89,85%	70,4%	58,5%	66,45%	65,21%	65,34%	56,52%

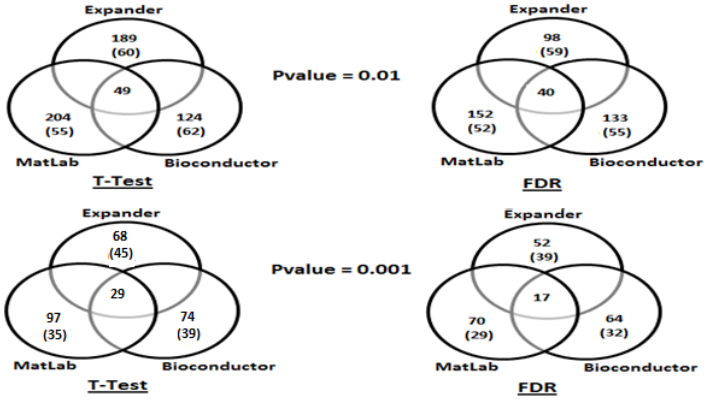

Fig.6: Number of genes selected and grouped according to the used statistical tool.