COMPARATIVE ANALYSIS BETWEEN SOFT AND TOOLS IN MICROARRAY DATA PROCESSING

Ouafae Kaissi, Ahmed Moussa, Brigitte Vannier, Abdellatif Ghacham
LTI Laboratory, ENSA, Abdelmalek Essaadi University, Tangier/ Morocco
IPBC, University of Poitiers/France
amoussa@uae.ac.ma

ABSTRACT

Motivation: In the analysis of experiments that involves the high density of oligonucleotide chips, it is important to generate list of genes or ‘targets’ from the genome wide data set that contains a lot of information. Gene selection is a process that seeks to identify the most significant genes which reveal large expressions changes between the baseline experiments and the conditions. Even though, several algorithms like T-test and other derived statistical algorithms were used for that selection process, the suitable Pvalue Cutoff remains difficult to choose. Therefore, one solution consists of using a False Discovery Rate (FDR) control. The Significance Analysis of Microarray (SAM) and the T-test Benjamini & Hochberg (BH) algorithms have been successfully used in such way. However, the reproductivity of results and their impact on the genes and/or experiments classification, while using different soft tools remain a subject of discussion.

Methods: we use two Affymetrix data sets, when we look for identifying list of genes under SAM and T-test-BH algorithms with FDR control running under R/Bioconductor project and Bioinformatics Tool Box of Math works and Expander.

Results: The list of selected genes changes significantly when using the two algorithms under R/Bioconductor project, Bioinformatics Tool Box of Math works and Expander. By means of data provided from publicly databases, we illustrate, that the permutation process of the multiple statistical T-test (SAM and BH) may affect results of selection process. Moreover, list of genes using the two Soft is affected by the choice of the Pvalue Cutoff for identifying true differential expressed genes. According to this work, we present some results clarifying sensitivity and efficiency of used soft and its influence in gene selection process. Hierarchical classification of selected genes and corresponding experiences confirm the influence of both methods and tools on the outcome of gene expression data analysis.

KEY WORDS
Microarray Data Analysis, Gene Selection, Comparative Analysis.

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.

On the other side, selection process of differentially expressed genes (DEG) across multiple conditions is one of the major goals in many microarray experiments [6]. Since one cannot analyze the raw data with thousand’s or more of genes, a variety of multiple-testing procedures for DEG selection have been developed [7]. A statistical test like t-test is the main procedure used when the goal is to detect significant level of genes expression; it can be generalized to multiple groups testing for identifying DEG [8]. In literature, the statistics t-test for microarray analysis are abundant [9, 10, and 11]. Some methods use ‘False Discovery Rate’ (FDR) control to compute the probability that a given gene is a false positive and is identified as DEG [12]. A permutation based approximation of this method, assumes that each gene is an independent test, is implemented in the Significant Analysis of Microarray (SAM) program [11].

In microarray data analysis, a comparative study seems to be a useful tool that leads the analyst to a suitable choice of methods, algorithm and analysis software. In this context, different comparisons have been implemented such as the comparison of normalization methods for high density oligonucleotides [13], comparison of selection methods [14] and comparison of statistical clustering techniques [15]. Yet, the reproducibility of result from these algorithms and their impact on the classification, when using different development tools and different
technologies of microarrays remain a stand point of debate. For that reason, this paper presents tools of comparison study that use two methods for identifying DEG and evaluate their performances on two publicly available microarray data sets. The aim is to show: on the one hand, the impact of the P-values choice on the number of detected genes. On the other hand, to discuss the performances of selection methods and their impact on the classification under both software’s. In the second section, this paper summarizes an overview of the use of Affymetrix technology in DEG analysis and describes tools and statistical methods used in genes selection. Results, discussions and conclusion are presented in the last section.

2. Technologies Tools & Data

2.1 DEG Analysis in Affymetrix

Affymetrix Gene Chip represents a very reliable and standardized technology for genome-wide gene expression screening 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).

The most widely format used for analyzing data provided from Affymetrix technology is .CEL format. This last, called “the raw data”, contains the microarray feature intensity quantification, and such data are the starting point for quality assessment and expression analysis.

Several experiences in microarray data intend to compare two conditions (treated # baseline). The objective is generally to answer the question: does the expression of a transcript on a chip (treated) change significantly with respect to the other chip (base-line)? 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 the four thresholds chosen by the analysis for Affymetrix technology. In the case of high dimensional data, for example when comparing several experiences, the detection call is a limited tool and other solution like multiple testing procedures can be used. Some of these procedures, such as the Bonferroni procedure, control the Family-Wise-Error-Rate (FWER). The other multiple testing procedures, such as the Benjamini and Hochberg (BH) procedure, control the False Discovery Rate (FDR). Another challenging aspect of microarray data analysis is to choose appropriate test statistics for different types of responses and covariates obtained from the datasets. The commonly used statistics including the t statistic and the F-statistic were originally designed for performing a single test but are not appropriate for large-scale data analysis. This motivated the development of many new statistics that borrow information across multiple genes for identifying differentially expressed genes, including a Significance Analysis of Microarrays (SAM); offering then a random testing approach which relies on relatively weak assumptions and yet are quite powerful [11].

Several of these methods have a strong and weak point, but there is no argument over the choice of a particular tool. For this we will try to argue in crossing a comparative study using two well used tool for gene selection under different algorithms:

The SAM statistical algorithm [11] and the T-test BH algorithm [12]. This choice is justified by their popularity and their availability in Expander, Bioconductor and Bioinformatics Tools Box of Math works, and software tools.

2.2 Data

We used two data sets available on the public databases (NCBI and EBI). In the Latin Square [16], set, 12 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 each microarray experiment, 14 groups of genes in 14 different concentrations were hybridized to microarray in the presence of a complex background of expressed human genes (30 Mb) and several control genes. For this Latin square design, 14 groups of experiments with 3 replicates for each one, give 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.

The second data set concern data provided from cancer projects. In fact, the variety and extent of the cancer data, gave us access to choose the second phase of data dealing a Chronic Lymphocytic Leukemia (CLL) [17]. The CLL is the most common chronic leukemia with an extremely variable clinical course. Some patients survive only a few months, whereas others have stable disease for years. The identification of novel genes mutated in CLL is important for prognostic purposes, to understand the biology of the disease and identification of targets and pathways for therapeutic intervention. In this type of data, two normal B cells isolated from peripheral blood and 5 CLL specimens have been analyzed with affymetrix (HGUA133A) microarray for expression. The main goal is to find genes that reveal a significance change between normal B cells and CLL cells.

2.3 Soft Tools

Several Soft Tools has been developed to facilitate the analysis of microarray data. In this context, the most used free software is Bioconductor. However, Bioinformatics Tool Box of Math works 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 genes is realized by the "limma" package integrated in Bioconductor.

The second used software is Bioinformatics Toolbox of Matlab works 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.

The last used software is Expander (Analyzer and 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].

Figure 1. Selected according to used algorithms and tools for Latin Square Data

Figure 2. Number of genes selected according to used algorithms and tools for Leukemia data
3. Results and Discussions

Throughout this research, we analyzed the performance of statistical tests integrated in Soft Tools cited below using Latin square that’s a 42 spikes are known and Leukemia data cited in the work of [17] as spikes. Results are evaluated with the percentages of True Detection Rate (TDR=number of Spike detected / number of modulated genes reported) and the percentage of Spike Detection based on Pvalue from 0.001 to 0.02. The graphs plotted below illustrate the variation of the Number of genes selected (TDR & spikes) according to the selection methods, depending on the software and also according to Pvalue to used algorithms and tools. Concerning the Soft Tools, the study show a large integrality between Expander and Bioconductor in the number of common selected genes compared to Matlab. In connection with the test of selection, the behavior of the SAM algorithm is not expected and some genes selected by SAM are not sorted by T-test BH algorithm under the same Pvalue Cutoff. This variation is due to a large part of the random permutation of the SAM algorithm. In fact, SAM is considered a very powerful test for selection especially in the case of a large sample size like Latin Square data set. As regards the change in number of spikes depending on the Pvalue Cutoff ; the Figueil and Figure.2 show that more the Pvalue is low, more the percentage of detection Spikes is significant.

In the same way the result deduced show the variation of TDR according to the Pvalue Cutoff. Thus, When the Pvalue increases from 0.001, the value of TDR decreases, meaning that is preferable to usually choose a small Pvalue for gene selection. (See Figure 3 and Figure 4 )

All clusters regroup control cell in the same group. But the classification of conditions show certain changes between Bioconductor and Bioinformatics Tool Box. This classification is a result of DEG selected from SAM in Matlab (see Figure 4-a), BH-T-test in Matlab (see Figure. 4-b), SAM in Bioconductor (see Figure.4-c), BH-Test in Bioconductor (see Figure. 4-d). These results confirmed
the influence of the used soft on selected genes and furthermore the classification process.

4. Conclusion

We can conclude that this analysis confirms that number of selected genes depend both on the method and soft used tools. In this sense our comparative study shows:
A large integrality between Expander and Bioconductor in the number of common selected genes compared to Matlab.

According the two selection test (SAM & t-test BH) different genes list are generated with greater variety in SAM.

Regarding the variation of the Pvalues, more the Pvalues is low; more the percentage of detection Spikes is significant for both methods and soft.

Finally, this comparative study shows that genes selected in microarray experiments may depend both on methods and tools used. In a word, we suggest the microarray data analyst to validate results and give new lists of genes by confirming the reproducibility of selected genes using various methods and tools.

Figure 5: Hierarchical Biclustering of genes and condition of the second dataset
(a): Hierarchical Biclustering of genes selected by SAM in Matlab
(b): Hierarchical Biclustering of genes selected by BH-Ttest in Matlab
(c): Hierarchical Biclustering of genes selected by SAM in R/Bioconductor
(d): Hierarchical Biclustering of genes selected by BH-Ttest in R/Bioconductor

References


[19] www.r-project.org


