BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 9(20)

Genes Selection Comparative Study in Microarray Data Analysis

Ouafae Kaissi[1], Eric Nimpaye[2], Tiratha Raj Singh[3], Brigitte Vannier[4], Azeddine Ibrahimi[5], Abdellatif Amrani Ghacham[1] & Ahmed Moussa[2]*

[1]LTI Laboratory, ENSA, Adbelmalek Essaadi University, Tangier, Morocco; [2]LabTIC Laboratory, ENSA, Abdelmalek Essaadi University, Tangier, Morocco; [3]Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Solan, H.P, India; [4]Research Group 2RTC, University of Poitiers, France; [5]Medical Biotechnology Laboratory, FMP, Mohammed V Suissi University, Rabat, Morocco; Ahmed Moussa Email: amoussa@uae.ac.ma; *Corresponding author

Received November 23, 2013; Accepted December 16, 2013; Published December 27, 2013

Abstract:
In response to the rapid development of DNA Microarray Technologies, many differentially expressed genes selection algorithms have been developed, and different comparison studies of these algorithms have been done. However, it is not clear how these methods compare with each other, especially when we used different developments tools. Here, we considered three commonly used differentially expressed genes selection approaches, namely: Fold Change, T-test and SAM, using Bioinformatics Matlab Toolbox and R/BioConductor. We used two datasets, issued from the affymetrix technology, to present results of used methods and software's in gene selection process. The results, in terms of sensitivity and specificity, indicate that the behavior of SAM is better compared to Fold Change and T-test using R/BioConductor. While, no practical differences were observed between the three gene selection methods when using Bioinformatics Matlab Toolbox. In face of our result, the ROC curve shows that: on the one hand R/BioConductor using SAM is favored for microarray selection compared to the other methods. And, on the other hand, results of the three studied gene selection methods using Bioinformatics Matlab Toolbox are still comparable for the two datasets used.

Keywords: Microarray data, Gene selection, R/BioConductor, Bioinformatics Matlab Toolbox, Comparative Study.

Background:
In many cases, the purpose of the microarray experiments is to compare the gene expression levels between two different conditions. Mostly, one sample is considered as reference or control and the other is considered as experiment. In all such comparative studies, the main goal is to determine the genes that are differentially expressed in the two samples being compared. In the early days the simple method of Fold Change (FC) was used, this last involves selecting genes for which the ratio of the experiment and control values is a certain distance from the experiment/control ratio **[1].** Another possible approach to gene selection is to use invariance statistical tests, like T-test. This approach essentially used the classical hypothesis testing methodology **[2].** The analysis of variance (ANOVA) is a particular interesting approach to microarray data and differentially expressed genes (DEGs) selection, the idea behind ANOVA is to build an explicit model of all sources of variance that affect the measurements and the use of the data to estimate the variance **[3]**. The following approach is the model based maximum likelihood estimation methods **[4].** This approach is very general and very powerful. Significant Analysis of Microarray (SAM) and moderated t-statistic using empirical Bayes are two tests used to address the problem of the simple T-test when the number of samples is small **[5, 6]**. More recently, Storey *et al.* **[7]** developed a new approach based on the Optimal Discovery Procedure (ODP), which aims to maximize the expected number of true positive genes for each fixed level of expected false positives. Several others

methods have been used for the DEGs selection and the literature is abundant **[8, 9, 10].** Despite the wealth of a literature concerning DEGs selection methods, there is no clear-cut guideline regarding the choice of methods and softwares. In this context, a comparative study could guide the practitioners to a suitable tool. Here, we comparatively review, three DEG selection methods: Fold change, T-test and SAM using two software's tools R/Bioconductor and Bioinformatics Matlab Toolbox. To conduct our study we used two datasets issued from the Affymetrix technology **[11]** since we showed in our last study that there is no significant difference, in term of differential analysis, between Affymetrix and other technology like Agilent **[12].** The aim is to evaluate by means of sensitivity and specificity the impact of statistical tools on the DEGs selected. In the next section, we will justify the choice of statistical tools and we will give a concise description of the DEGs selection workflow followed for all proposed algorithms and tools. In the end, we discuss the results of used algorithms and tools on well known datasets.

Methodology:

In order to assess the performance of a gene selection method we need a set of criteria able to qualify the outcome of the selection process. In our case the performance of a gene selection method has been calculated in terms of specificity and sensitivity presented in Receiver Operating Characteristics curves (ROC). In a binary decision situation, such as changed or unchanged, the results can always be divided into four categories: truly changed that are reported as changed (True Positives: TP), unchanged that are reported as changed (False Positive: FP), truly changed that are reported as changed (False Positive: FP) truly changed that are reported as such (True Negatives: TN). Based on these parameters, one can define the sensitivity and specificity that qualify the productivity of DEGs methods and tools used.

Sensitivity = TN/TN +FP Specificity = TP/TP +FN

Selected algorithm for comparison

Several researches are done on the comparison and the choice of gene selection methods. The last study in this context aims to compare the gene selection methods and classify them according to some parameters **[13].** This study shows that all DEGs selection methods may be classified in three categories: Class of methods based on deterministic FC, class of methods based on T-test, multiple T-test, P_values, and the class of methods using random permutations. Based on this classification, we selected three selection methods, each one provided from one class, and we will compare selected genes from each method under R/Bioconductor and Bioinformatics Matlab Toolbox. The first approach used is the FC, selected from the first class. This method is the most initiative approach to finding genes that differentially regulated **[1].** Typically, an arbitrary threshold is chosen and the difference is considered as significant if it is larger than the threshold. The FC method is often used because it is simple and intuitive. T-test, which is chosen from the second class, is extensively used in gene expression analysis. It demonstrates whether the difference between two groups or samples is significant **[2].** In our study P_values calculated by T-test have been used without any multiple testing corrections. The DEGs were obtained by setting two criteria of P_values <0.05 and FC >=2. The last approach provided from the third class is the SAM method **[5].** SAM uses a statistic called "relative difference" for gene. This is very similar to a t-statistic with equal variance, except that the gene "gene-specific scatter" in the denominator is offset by a fudge factor. This latter is an exchangeability factor chosen to minimize the coefficient of variation and is computed with a sliding window approach across the data. These DEGs selection methods have been already integrated in the R/BioConductor project **[14].** In Bioinformatics Matlab Toolbox the two first DEGs selection methods have been also integrated. But for the SAM algorithm under this software, it was recently implemented by **[15].**

Data sets

We applied feature selection methods to two data sets issued from the affymetrix technology. The First one is the Spike-in datasets **[16].** These data sets are designed to address questions about the correctness of microarray data and have been used extensively to compare among different analysis methods. Currently there are four major spike-in datasets available for the Affymetrix microarray platform: the Affymetrix spike-in dataset for cross platform comparisons **[17],** the Affymetrix Latin square dataset **[18]**, the Gene Logic spike-in dataset **[19]** and the Golden Spike dataset **[20].** The new wholly defined Affymetrix spike-in dataset consisting of 18 microarrays. Over 5700 RNAs are spiked in at relative concentrations ranging from 1- to 4-fold, and the arrays from each condition are balanced with respect to both total RNA amount and degree of positive versus negative fold change. The second data set is recently represented in **[21].** These cancer data sets consist of 18 breast cancer patients' usual-risk controls undergoing reduction mammoplasty (RM), and histologically normal (HN) patients. In this work, we analyze the total of 86 genes presented as differentially expressed between RM and HN samples. All Chips from the two datasets present variations due to the experiences background noise and systematic errors. For this, the standardization is applied to all chips to be sure that the distributions of intensities across chips are homogeneous. Several studies have focused on the performance of different normalization methods **[22].** In this study we use the Robust Multichip Analysis algorithm (RMA). It provides accurate estimation of inter-array variability through a robust background correction and quantile normalization computed over the whole datasets. To simplify analysis, we follow in our comparative study the workflow presented below:

Raw Affymetrix Data (.cel)→ Normalization (RMA)→ Different DEGs Selection Methods and Tools→ Results of The Comparative Study

Results & Discussions:

The aim of this paper is to understand how gene selection algorithms differ from one another and how it is influenced by using different software's. In this study, we compare results for three gene selection methods, SAM, FC, and T-test in both R/BioConductor and Bioinformatics Matlab Toolbox using two datasets. We evaluate first the influence of statistical tools on the size of selected genes using the Venn diagram then we evaluate the sensitivity and the specificity of each methods and tools using the ROC curves. The first result is represented in the **Figure 1**. A Venn diagram shows the overlap between the three gene selection methods and the genes identified as differentially expressed in each dataset (denoted by "identified genes" or "id" in **Figure 1).** The lists of genes generated as differentially expressed differ between methods, software's and datasets used. Viewpoint methods, SAM identified 80% to 89% of all genes in each dataset under R/Bioconductor relative to t-test and FC methods. While, meeting lists of each gene selection method in Bioinformatics Matlab Toolbox show a close between the three gene selection methods used with a reduced number of genes. In term of software's; the number of DEGs selected was still large in R/BioConductor , but was considered reduced in Bioinformatics Matlab toolbox (for example in spike-in data set ,1224 gene was selected by R/BioConductor face to 101 genes selected by Matlab toolbox for the same data).

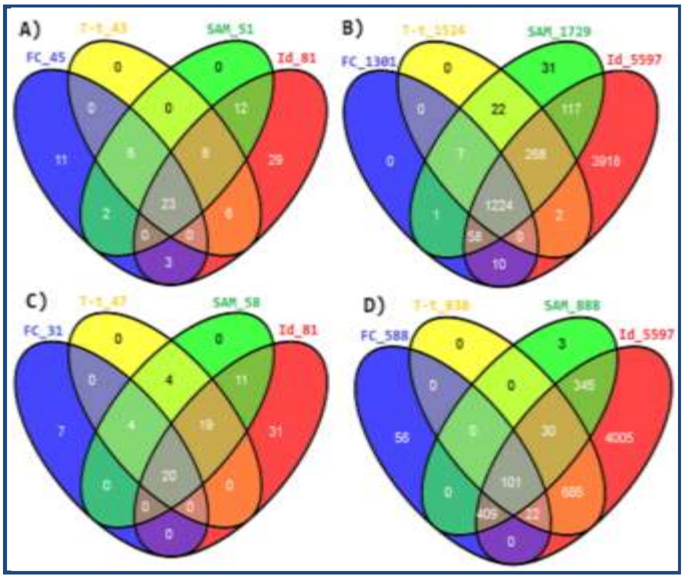

Figure 1: Overlap between the three genes selection methods and the identified genes using both softwares and datasets: **A)** R/BioConductor in Breast cancer dataset; **B)** R/Bioconductor platinum spike-in dataset; **C)** Bioinformatics Matlab Toolbox in Breast cancer dataset; **D)** Bioinformatics Matlab Toolbox in platinum spike-in dataset.

After this first study we can conclude that Bioinformatics Matlab Toolbox allows a selection of a reduced number of genes compared to R/BioConductor. But are remains unclear whether the selected genes are relevant. For this reason we used the ROC curves that compare the performance of the three methods and the software's in each dataset used in terms of sensitivity in specificity (**Figure 2). Figure 2A** represents the ROC curve using R/BioConductor for the breast cancer dataset. This figure shows that SAM and T-test report differentially expressed genes in a stricter manner than FC. This is not surprising given that the former two methods take into account the variance of an observation between different conditions, which is not the case of FC. As the figure shows, the SAM is more stringent than the T-test, the explanation coming from the "fudge factor" that reduces the bias of genes with large variances without necessarily biological significance. FC has a significant false positive rate, almost half of its results**. Figure 2B** represents the ROC curve using R/BioConductor for the platinum spike-in dataset, this figure shows that SAM is preferential than t-test and FC. **Figure 2C** represents the ROC curve using Bioinformatics Matlab Toolbox for the breast cancer dataset, this figure shows close results between the three gene selection methods used, especially between SAM and T-test. **Figure 2D** represents the ROC curve using Bioinformatics Matlab Toolbox for the Platinum spike-in dataset and shows also that the false positive rate is still comparable for the three methods.

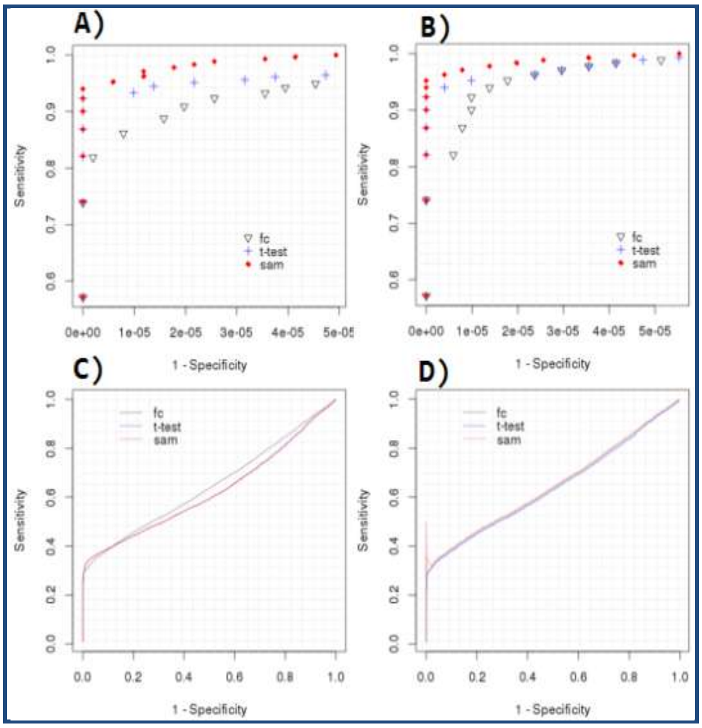

Figure 2: A) ROC curve for the three genes selection methods using both softwares and datasets: **B)** R/BioConductor in the breast cancer dataset; **C)** Bioinformatics Matlab Toolbox in breast cancer dataset; **D)** Bioinformatics Matlab Toolbox in spike-in dataset.

Conclusion:

A fundamental step of microarray studies is the identification of a small subset of DEGs from among tens of thousands of genes probed on the microarray. DEGs lists must be concordant to satisfy the scientific requirement of reproducibility, and must also be specific and sensitive for scientific relevance. A Bioinformatics Matlab Toolbox allows a selection of a reduce list of DEGs used different methods proposed relative to the R/ BioConductor. Concerning the gene selection methods, the R/BioConductor using SAM is favored for microarray selection compared to the other methods notably the FC which is uncertain. While, the ROC curve when using Bioinformatics Matlab Toolbox shows that there is a not significant difference between the three genes selection methods with reproducible results. We conclude that the SAM approach is over conservative in R/ BioConductor but the FC has an extremely low power in the two datasets used. On the other hand, the three genes selection methods display a high power in Bioinformatics Matlab toolbox. Therefore, this latter would be selected if we prefer to select a small list of DEGs. In conclusion, a DEGs list should be chosen in a manner that concurrently satisfies scientific objectives in terms of inherent tradeoffs between reproducibility, specificity, and sensitivity.

Acknowledgment:
We acknowledge support from Volubilis (French-Moroccan Grant) to AI AM. This work was supported by a grant from the NIH for H3Africa BioNet to AM & AI.

References:

[1] Chen Y *et al. J Biomed Opt*. 1997 **2**: 364 [PMID: 23014960]
[2] Baldi P & Long AD, *Bioinformatics* 2001 **17:** 509 [PMID: 11395427]
[3] Kerr MK & Churchill CA, *Biostaticstics*. 2001 **2**: 183 [PMID: 12933549]
[4] Ideker T *et al. J Comput Biol*. 2000 **7:** 805 [PMID: 11382363]
[5] Tusher VG *et al. Proc Natl Acad Sci U S A*. 2001 **98**: 5116 [PMID: 11309499]
[6] Kendziorski CM *et al. Stat Med*. 2003 **22**: 3899 [PMID: 14673946]
[7] Storey JD *et al. Biostatistics* 2007 **8**: 414 [PMID: 16928955]
[8] Pan W, *Bioinformatics* 2003 **19**: 1333 [PMID: 12874044]
[9] Park T *et al. Bioinformatics*. 2003 **19**: 694 [PMID: 12691981]
[10] Benjamini Y *et al. Behav Brain Res*. 2001 **125:** 279 [PMID: 11682119]
[11] Irizarry RA *et al. Nucleic Acids Res*. 2003 **31**: e15 [PMID: 12582260]
[12] Sabaouni I *et al. Bioinformation* 2013 **9**: 849 [PMID: 24250110]
[13] Jeffery IB *et al. BMC Bioinformatics* 2006 **26**: 359 [PMID: 16872483]
[14] http://www.bioconductor.org
[15] http://www.mathworks.com/matlabcentral/fileexchange/42346
[16] Zhu Q *et al. BMC Bioinformatics* 2010 **11**: 285 [PMID: 20507584]
[17] McCall MN & Irizarry RA, *Nucleic Acids Res*. 2008 **36**: e108 [PMID: 18676452]
[18] http://www.affymetrix.com/support/technical/sample_data/datasets.affx
[19] Irizarry RA *et al. Nucleic Acids Res*. 2003 **31**: e15 [PMID: 12582260]
[20] Choe SE *et al. Genome Biol*. 2005 **6**: R16 [PMID: 15693945]
[21] Graham K *et al. Br J Cancer*. 2010 **102**: 1284 [PMID: 20197764]
[22] Schadt EE *et al. J cell Biochem Suppl*. 2001 **37:** 120 [PMID: 11842437]

Edited by P Kangueane
Citation: Kaissi *et al.* Bioinformation 9(20): 1019-1022 (2013)
License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited