MatSAM: a Matlab implementation for Significance Analysis of Microarrays

[1]Eric Nimpaye, [2]Ouafae Kaissi, [3]Tiratha Raj Singh, [4]Brigitte Vannier, [5]Azeddine Ibrahimi, [1]Ahmed Moussa.

[1]LabTIC laboratory, ENSA, University AdbelmalekEssaadi, Tangier, Morocco.
[2]ENSA, University AdbelmalekEssaadi, Tangier, Morocco.
[3]Department of Biotechnology and Bioinformatics, Jaypee University of Information and Technology, Solan, H.P., India.
[4]Receptors, Regulation and Tumor Cells (2RCT), University of Poitiers, Poitiers, France.
[5]LTP Laboratory, FMP, Mohammed V University, Rabat, Morocco.

Article Info

Article history:

Received May 29th, 2014
Revised Jun 20th, 2014
Accepted Jul 4th, 2014

Keyword:

SAM
microarrays
matlab
expression
gene

ABSTRACT

Microarray experiments enable the simultaneous measure of expression levels of large amount of genes and have many applications. A widespread one is finding set of genes that are differentially expressed. Significance Analysis of Microarrays (SAM) helps to produce those sets using multiple testing techniques. There is unfortunately not yet a public tool enabling to do SAM using the Matlab platform. We here define MatSAM, a SAM implementation in Matlab, and show that it yields results of high confidence comparatively to those obtained by putative tools available in the R programming environment. MatSAM can be used in conjunction with Matlab Bioinformatics toolbox to perform further analysis.

Availability: MatSAM is available as source code at http://www.bioinfoindia.org/MatSAM

Corresponding Author:

Ahmed MOUSSA.
LabTIC laboratory, ENSA,
University AdbelmalekEssaadi,
Tangier, Morocco.
Email: amoussa@uae.ac.ma

How to Cite:

Nimpaye E. *et al.*MatSAM: a Matlab implementation for Significance Analysis of Microarrays. IJCB.2014; Volume 3 (Issue 2): Page 49-51.

1. INTRODUCTION

In microarray experiments, identification of variation in genes expression in different experimental conditions is a crucial step in data exploitation. Complexity of the analysis comes from the fact of testing together few conditions, many observations and hundreds or even thousands hypotheses leading to multiple testing problems [1]. The Significance Analysis of Microarrays (SAM) can help identify those expression variations [2], usually by samr [3] or siggenes [4] packages which are software written in the R programming environment. As Matlab programming environment is popular in engineering and scientific applications and to our knowledge, there is not yet a SAM implementation in the Matlab programming environment although it endows with many other tools to do microarray data analysis. We propose MatSAM to fill in that gap.

In order to run SAM using the proposed implementation, the data should be preprocessed using the already existing tools in Matlab, namely the Matlab Bioinformatics toolbox. The preprocessing is expected to lead to two data files containing, one the gene names, and the other one the gene expression values. MatSAM is run upon those files to yield outputs in the form of statistics concerning differentially expressed genes. In the next section, we give an overview of the way SAM is implemented and more details are available at the implementation web site.

2. FEATURES

The first step in order to assess the differential expressions of different genes through many experimental conditions is to choose an appropriate statistic method for testing on each gene and to compute the corresponding p-value. The resulting p-values have to pass through an adjustment process to avoid errors due to hypotheses multiplicity. Many statistical methods have been defined to adjust for multiplicity [5]. One such a method is to plot the observed test statistics against the values they would take under a combination of the null hypotheses. In the resulting Quantile-Quantile (QQ) plot, most of the data points lie approximately on the diagonal, and if genes differentially expressed exist among the tested sample conditions, they should be matching points that step aside the diagonal line in a substantial way. SAM can help measure that substantial gap.

The analyst can customize the behavior of the SAM algorithm by providing the statistical test settings such as variance (in)equality between conditions, method to use when computing gene fold changes, transformation to apply on the gene expression data. The behavior customization of MatSAM should also include the permutation procedure settings, the prior probability that a gene is differentially expressed and the method for handling missing values. All these settings have default values that the investigator could use them unchanged or adjust them to adapt to different situations upon utilization. The next step consists in mimicing the experimental data. Usually, two conditions or classes will be taken in consideration, for instance one class for healthy condition (control subjects) compared to another one for the sick condition (patients). Currently, this design is the only implemented in MatSAM, even though the SAM algorithm can support other types of designs.

Before requesting any results as MatSAM outputs, the analyst has to decide a threshold (called delta) giving approximately the desired number of differential expressed genes while maintaining a low false discovery rate. Figure 1 shows the SAM plot using the chosen delta value which is also used to generate the genes of interest (data not shown here, see the MatSAM website). MatSAM empower its users to handle all those steps and is actually a port to Matlab of the siggenes implementation as a courtesy of its author.

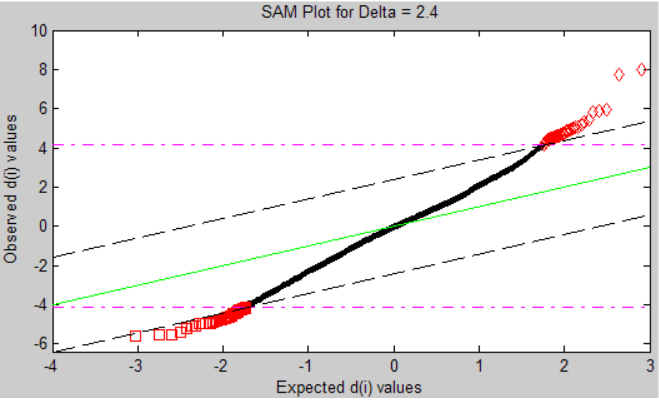

Figure 1. The SAM plot: given a delta value of 2.4, a QQ plot of expected vs. observed scores is generated. Differentially expressed genes deviate from the diagonal area: up-regulated genes are colored in red diamonds in the upright corner, and red squares in the bottom-left corner represent down-regulated ones.

To assess the MatSAM performance, we compared it to the samr and siggenes SAM implementations using receiver operating characteristics (ROC). The dataset was the GSE21344, which is also known as Platinum Spike [6]. We did not use the whole observations since the three implementations cannot run beyond very small delta values, say less than 10^{-4}, which would permit us to pull out all the observations even if the outcomes would then be of no biological interest. We rather fixed the comparison dataset size to the largest output that can samr produce given it is the SAM algorithm authors' very implementation.

Figure 2 shows that all the three implementations have poorer classification performances than random guessing[7] as their respective ROC points lay back and forth on the diagonal. It is worth noticing the values for the area under the ROC curve (AUC) for the different methods, respectively 0.455 to samr, 0.448 to siggenes, and 0.457 to MatSAM. These numbers are unsurprising since SAM is not a classification algorithm. Nevertheless, the AUCs show that all the three implementations have similar average performance (≈ 0.45) on detecting spikes in the Platinum dataset which was particularly designed, including but not limited to, run method's benchmarks. It is clear that MatSAM can be considered as reliable as the well-established samr and siggenes SAM implementations.

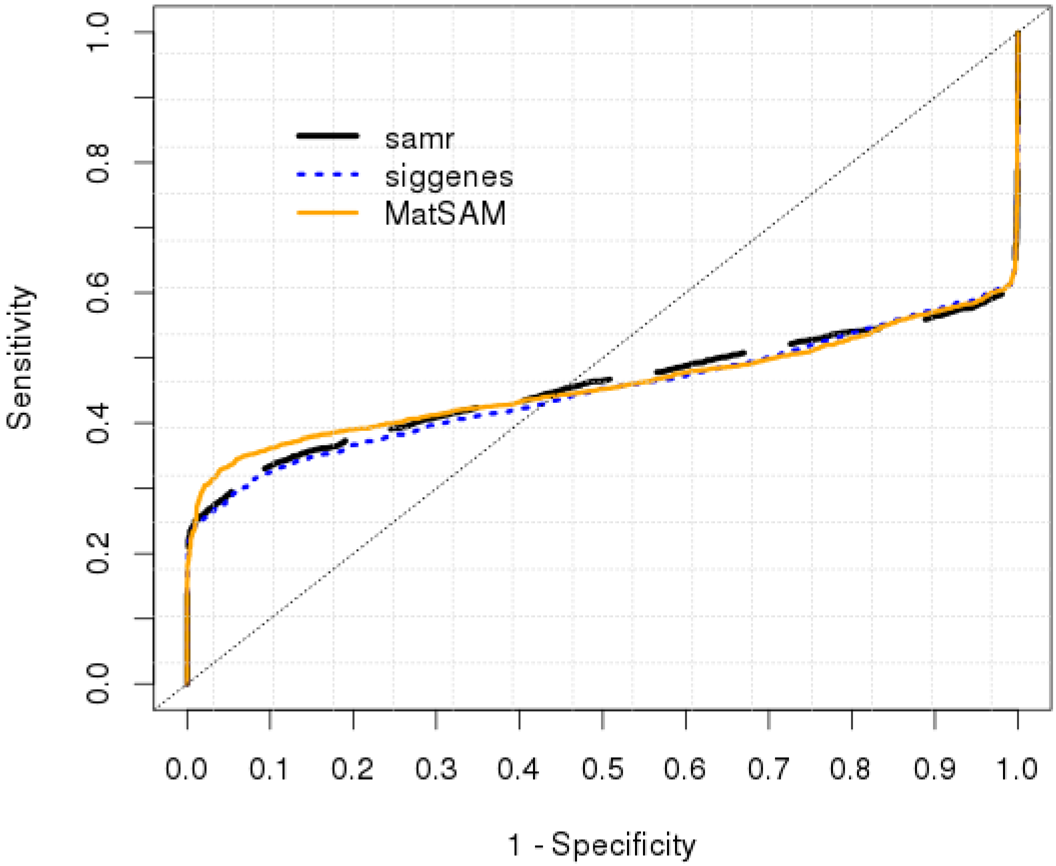

Figure 2. ROC graph showing samr, siggenes and MatSAM performances on truly and falsely classifying spikes as differentially expressed.

3. CONCLUSION

In this paper, we described MatSAM which is loaded with tools compatible with the Matlab platform and capable of performing Bioinformatics analysis using SAM which remains a very popular method for this type of analysis. We also addressed challenges of relying on the Matlab platform to run Bioinformatics analyses with no need of external tools. Obtained results are comparable with those produced by samr and siggenes R packages as tools that have proven their robustness. In a future version, MatSAM should reuse more objects provided by the Matlab Bioinformatics toolbox in order to better integrate and interact with the existing framework.

REFERENCES

[1] Kaissi O, Nimpaye E, Singh TR, Vannier B, IbrahimiA, Ghacham AA, Moussa A (2013). Gene Selection Comparative Study in Microarray Data Analysis. Bioinformation 9(20): 1019–1022.

[2] Tusher VG, Tibshirani R, Chu G (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of USA, 98(9): 5116–5121.

[3] Tibshirani R, Chu G, Narasimhan B, Li J (2011). samr: Significance Analysis of Microarrays. R package version 2.0.

[4] Schwender H (2012). siggenes: Multiple testing using SAM and Efron's empirical Bayes approaches. R package version 1.38.0.

[5] Bender R, Lange S (2001). Adjusting for multiple testing — when and how? Journal of Clinical Epidemiology 54, 343–349.

[6] Zhu Q, Miecznikowski JC, Halfon MS (2010). Preferred analysis methods for AffymetrixGeneChips. II. An expanded, balanced, wholly-defined spike-in dataset. BMC Bioinformatics, 11:285.

[7] Fawcett T (2006). An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874.