A comparative study of the efficiency of four differentially expressed gene selection programs for microarray data analysis

S Rajendran, J Natarajan

Citation

S Rajendran, J Natarajan. *A comparative study of the efficiency of four differentially expressed gene selection programs for microarray data analysis*. The Internet Journal of Genomics and Proteomics. 2009 Volume 5 Number 2.

Abstract

Microarrays are part of a new class of biotechnologies which allow the monitoring of expression levels for thousands of genes in a single experiment. It is important to consider finding differentially expressed genes in a dataset of microarray experiments as differentially expressed genes are often referred as clinical markers. A number of statistical methods have been suggested for the identification of differentially expressed genes using different statistical methods and algorithms. In the present paper, an experimental investigation of four different algorithms for tracking differentially expressed genes using four publicly and freely available programs namely MeV (t-test), SAM (Significance Analysis of Microarray), EDGE (Optimal Discovery method) and iArray (Mann-Whitney test) is reported. To assess the performance of each program, 50 artificial microarray datasets with known differentially expressed genes were used for comparative study. Performance and evaluation of these programs from a biologist's perspective has been studied and reported in this paper.

INTRODUCTION

Every cell of the body contains a full set of chromosomes and identical genes. Only a fraction of these genes are turned on. "Gene expression" is the term used to describe the transcription of the information contained within the Deoxyribonucleic Acid (DNA), the repository of genetic information, into messenger RNA (mRNA) molecules that are then translated into the proteins that perform most of the critical functions of cells [1]. Disruptions or changes in gene expression are responsible for many diseases [2].

Microarray is an array of DNA molecules that permit many hybridization experiments to be performed in parallel [3, 4]. It can monitor expression levels of thousands of genes simultaneously. By analyzing microarray expression profiles one can deduce information that can provide significant understanding of the mechanism of the disease under study. However, the gene selection can be a challenging issue as the microarray data is skewed with a large number of genes in one dimension and a few samples in the other dimension. Sophisticated statistical techniques are used to extract relevant genes in a given enormous amount of microarray data [5-8].

Differentially expressed genes (DEGs) are genes whose expression levels are significantly different between two groups of experiments [9]. The genes are relevant for discovering potential pharmaceutical targets and diagnostic or prognostic markers. Identification of differential gene expression is the first task of an in depth microarray analysis. Various software packages are available to do the task. However, there are no studies to date that evaluate the performance of these methods and usage of these softwares from a biologist's perspective. The purpose of this study is to evaluate the ability of existing four softwares to correctly identify the differentially expressed genes using 50 artificial datasets.

MATERIALS AND METHODS

ARTIFICIAL MICROARRAY DATASETS

The artificial datasets used in the present study were from Shaik and Yeasin [9]. Two different models were employed to generate artificial microarray datasets viz. Lognormal model and Asymmetric Laplace distribution. The dataset contains a total number of 50 artificial microarray datasets and each artificial dataset have 4100 genes with 10 samples under one of the two conditions. The first 100 genes in each dataset were differentially expressed and the rest 4000 were non-DEGs. This process enables class labels for genes (DEGs or non-DEGs) for each generated artificially

generated microarray dataset which can be used as ground truth to quantitatively assess the performance of different softwares used in this study.

DIFFERENTIALLY EXPRESSED GENE SELECTION PROGRAMS

Various tools have been developed to measure the expression of thousands of genes to identify changes in expression between different biological states. The following four publicly and freely available programs namely MeV from Dudoit et al., [10], SAM from Tusher et al., [11], EDGE from Leek et al., [12], and iArray from Pan et al., [13] were investigated in this study.

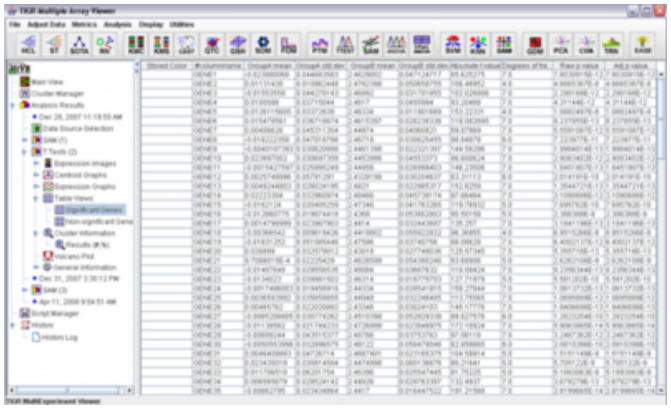
MEV

TIGR MultiExperiment Viewer (MeV) is a software for microarray data analysis. MeV is developed by a group of people at TIGR (The Institute for Genomic Research) and is freely available through TIGR web site [14]. The installation is fairly simple and running MeV requires Java Runtime Environment (JRE). A detailed instruction on installation of MeV and JRE can be found at http://compbio.utmem.edu/MSCI814/faq.php.

MeV uses a non-parametric t-test with family wise error rate corrected p- values. It allows selection of an expression pattern that has maximal difference in mean level of expression between the two groups and minimal variation of expression with each group. The output window and the list of significant genes as a result of MeV are shown in Figure 1.

Figure 1

Figure 1. Significant Genes as a Result of T-test of MeV



SAM

Significance Analysis of Microarrays (SAM) is a statistical technique for finding significant genes in a set of microarray experiments [15]. SAM is running on R platform and users were prompted to install R from its specified URL. SAM works with Microsoft Excel and hence the data should be put in an Excel spreadsheet. A detailed instruction on installation of SAM and SAM manual can be found at http://www-stat.stanford.edu/~tibs/SAM/sam.pdf

The input to SAM is gene expression measurements from a set of microarray experiments, as well as a response variable from each experiment. It uses repeated permutations of the data to determine if the expression of any genes is significantly related to the response. The cutoff for significance is determined by a tuning parameter delta, chosen by the user based on the false positive rate. In addition, one can also choose a fold change parameter, to ensure that called genes change at least a pre-specified amount. The output window for SAM Plot sheet and the list of significant genes is shown in Figure 2a and 2b.

Figure 2

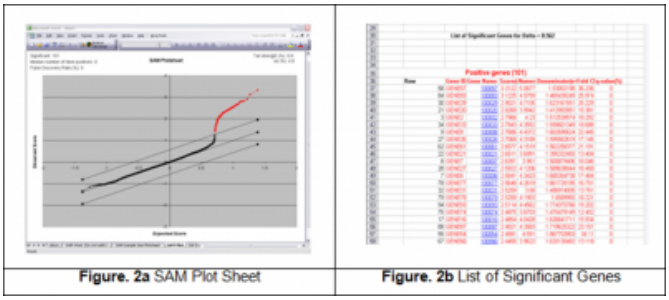


Figure. 2a SAM Plot Sheet

Figure. 2b List of Significant Genes

EDGE

EDGE is an open-source software package for the analysis of expression data [16]. The main purpose of the software is to perform significance analyses on comparative microarray experiments. EDGE is cross-platform compatible (Windows, Mac, Linux and UNIX), also running on top of the R statistical software package. EDGE includes functions for data visualization, transformation, exploratory analysis, and NCBI queries.

EDGE is based on the Optimal Discovery Procedure (ODP), which estimates the optimal rule for identifying differentially expressed genes. The user can press RECALCULATE button after changing the significance parameters to show the new list of genes that meet the redefined threshold or q-value estimation settings. Any gene in the genes called significant window can be queried on PubMed by its gene name. EDGE GUI interface with the list of significant genes is shown in Figure 3.

Figure 3

Figure 3. GUI Interface with Significantly Expressed Genes in EDGE

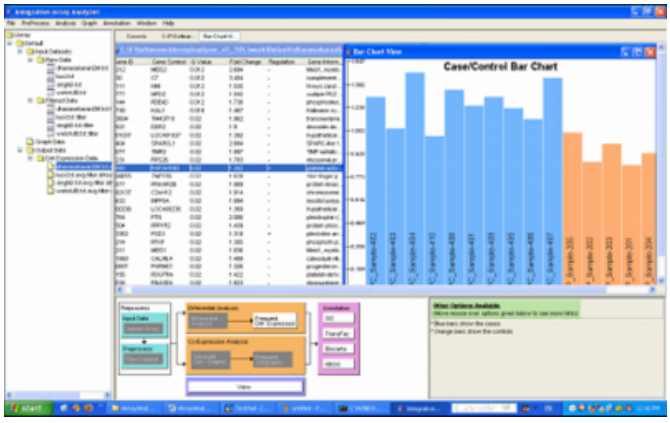
IARRAY

IArray Analyzer software (iArray for short) is a software package for analysis of cross-platform and cross-species microarray data [17]. iArray user guide which has detailed instructions on system requirements, installations procedures, and various data analysis methods can be found at http://www-stat.stanford.edu/~tibs/SAM/sam.pdf

iArray analysis module includes two steps: (1) performing differential analysis for each individual dataset (with Bonferroni or false discovery rate adjustment for multiple comparisons) and (2) identify sets of genes frequently differentially expressed in multiple datasets from results obtained in Step (1) For Step (1), two statistical methods to identify differentially expressed genes are implemented: Student's t-test and Mann– Whitney test. In the present work, we have used Mann– Whitney test. The output window and the list of significant genes as a result of iArray is shown in Figure 4.

Figure 4

Figure 4. Result of Differential Expression Analysis using iArray



RESULTS AND DISCUSSION

In the 50 artificial data sets employed in this study, the labels for the genes (DEGs/ non-DEGs) were already available. Using the class labels the performance of the each program is measured as a binary classifier where the gene is either differentially expressed (DEG) or not differentially expressed (non-DEG). Program performance was accessed by commonly used criteria such as true positives (TPs), false positives (FPs), true negatives (TNs) and false negatives (FNs). The definition of these performance measures is given as follows.

TP -> the number of true DEGs correctly identified the program

FP -> the number of true non-DEGs incorrectly identified the program

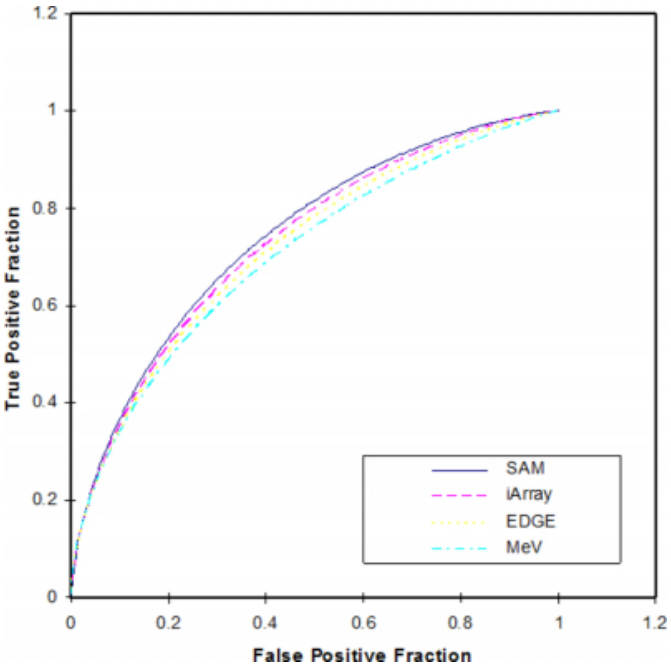
TN -> total number of true non-DEGs correctly identified by the program

FN -> the number of true DEGs incorrectly identified by the program

For a binary classifier, it is possible to accurately find TPs, FPs, TNs and FNs. Based on TPs, FPs, TNs and FNs, the final measures true positive fraction (TPF) is obtained by using the formula TPF = TP/(TP+FN) and false positive fraction (FPF) by using the formula FPF = FP/(FP+TN). These TPFs and FPFs are plotted to build the performance analysis curves. The plot of TPF vs FPF enables the comparison of performance of various programs employed in the study and shown in Figure 5.

Figure 5

Figure 5. The Performance curves for various Gene Selection Programs using Artificial Microarray Datasets



In Figure 5., blue curve indicate the output of the program SAM, rose curve indicate output of the program iArray, yellow curve indicate the output of the program EDGE and aqua curve indicates the output of MEV. Our results suggest that the all the four programs do a reasonable job in finding differentially expressed genes. However, SAM outperforms the other three programs in results. It has been found that SAM not only gives greater accuracy in results but also faster than other programs. In addition, after installation SAM available as Excel plug-in. It seems to be most advantages for biologists to run SAM easily from an Excel worksheet. EDGE gives maximum number of false negatives while detecting differentially expressed genes and iArray provides maximum number of false positives.

As a final result, the study revealed that, the installation and data analysis using SAM seems to be very simple and straight forward and also produces more accurate results when compare to other programs.

CONCLUSION

The purpose of this study is to find the most reliable method and program for differentially expressed gene selection. We presented an empirical study in which we compared four most commonly used programs MeV, SAM, iArray and EDGE. We apply these methods to 50 artificial microarray datasets, and compare, how these methods performed in DEG prediction of these test datasets. Our study has elucidated that SAM is a superlative method currently available for the detection of DEG. Apart from these the algorithm behind SAM is a robust permutation-based straightforward method that can be adapted to a broad range of situations in microarray datasets.

ACKNOWLEDGEMENTS

The authors acknowledge J. S Shaik and M. Yeasin, Department of Electrical and Computer Engineering, CVPIA Lab, University of Memphis, Memphis, USA for sharing their artificial microarray datasets. The authors would also like to thank the anonymous reviewers for their helpful suggestions and comments in improving the quality of the paper

References

1. Consortium EP: The ENCODE (ENCyclopedia Of DNA Elements) Project. Science, 306(5696):636-640, (2004).
2. Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., et al.: Multiclass cancer diagnosis using tumor gene expression signatures, PNAS, 98(26): 15149 – 15154 (2001).
3. Schulze, A., and Downward, J.: Navigating gene expression using microarrays – a technology review, Nat. Cell. Biol., E190-E1-94, (2001).
4. Schena, M., Shalon, D., Davis, R.W. and Brown, P. Q.: Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science, 270, 467-470, (1995).
5. Guyon, I: An Introduction of Variable and Feature Selection, Journal of Machine Learning Research, 3:1157–1182, (2003)
6. Getz, G., Levine, E., and Domany, E.: Coupled two-way clustering of gene microarray data, Proc of Nat Aca of Sci, USA., 97:12079–12084, (2000)
7. Mukherjee, S., Roberts, S. J., and Laan, M. J.: Data-adaptive Test Statistics for Microarray Data, Bioinformatics, 21:108–114, (2005)
8. Shaik, J., and Yeasin, M.: Adaptive Ranking and Selection of Differentially Expressed Genes from Microarray Data, WSEAS transactions on Biology and Biomedicine, 3:125–133, (2006).
9. Shaik, J. S. and Yeasin, M.: A Unified Framework for Finding Differentially Expressed Genes from Microarray Experiments, BMC Bioinformatics, 8: 347, (2007).
10. Dudoit, S., Yang, Y. H., Callow, M. J. and Speed T.: Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, Technical report 2000 Statistics Department, University of California, Berkeley (2000).
11. Tusher, V. G., Tibshirani, R. and Chu, G.: Significance analysis of microarrays applied to the ionizing radiation response, Proc. of Nat. Aca. of Sci. USA, 98: 5116-5121, (2001).
12. Leek, J. T., Monsen, E., Dabney, A. R. and Storey, J. D: EDGE: extraction and analysis of differential gene expression, Bioinformatics, 22(4):507-508, (2006).
13. Pan, F., Kamath, K., Zhang, K., Pulapura, S., et. al.: Integrative Array Analyzer: a software package for analysis

of cross-platform and cross-species microarray data, Bioinformatics, 22(13):1665-1667, (2006).
14. MEV available at http://www.tm4.org/mev.html
15. SAM available at http://www-stat.stanford.edu/~tibs/SAM/
16. EDGE available at http://www.biostat.washington.edu/software/jstorey/edge/
17. iArray available at http://zhoulab.usc.edu/iArrayAnalyzer.htm

Author Information

Sasikala Rajendran
Dr.G.R.Damodaran College of Science

Jeyakumar Natarajan
Centre of Excellence in Bioinformatics, Madurai Kamaraj University