Genomics & Informatics Vol. 2(4) 180-183, December 2004

Comparative Statistic Module (CSM) for Significant Gene Selection

Young-Jin Kim*, Hyo-Mi Kim*, Sang-Bae Kim, Chan Park, Kuchan Kimm and InSong Koh

Division of Epidemiology and Bioinformatics, National Genome Research Institute, National Institute of Health, 5 Nokbun-dong, Eunpyeong-gu, Seoul 122-701, Korea

*These authors contributed equally to this work.

Abstract

Comparative Statistic Module(CSM) provides more reliable list of significant genes to genomics researchers by offering the commonly selected genes and a method of choice by calculating the rank of each statistical test based on the average ranking of common genes across the five statistical methods, i.e. t-test, Kruskal-Wallis (Wilcoxon signed rank) test, SAM, two sample multiple test, and Empirical Bayesian test. This statistical analysis module is implemented in Perl, and R languages.

Availability: CSM is freely available from http://cmams.ngri.re.kr.

Keywords: data analysis, gene expression, microarray, significant gene selection, statistic module

Introduction

Microarray technology is a powerful approach for revealing the patterns of coordinately regulated genes. Due to the considerable amount and intrinsic variation of microarray experiment data, statistical approaches have been used as a way to obtain useful biological information (Lobenhofer *et al.*, 2001). Most common task in analysis of microarray data is to infer differentially expressed genes among two or more samples. Various statistical methods for identifying differential expression have been suggested to determine significant genes out of the whole list of genes. According to the various modeling assumptions underlying specific tests, statistical methods are divided into three classes; parametric, nonparametric, and semiparametric (Cui and Churchill, 2003). Different statistical methods using a single set of data may produce slightly or considerably different results due to specific statistical assumptions and their data dependency. A significant gene by one statistical method could be insignificant by another (Pan, 2002). No single statistic is universally optimal and there seldom exists any basis or guidance for a particular statistical method of choice. Therefore, it is important not only to select as many differentially expressed genes as possible but also to identify statistically meaningful genes. We implemented Comparative Statistic Module (CSM) as a module for previously developed cDNA Microarray data Analysis and Management System (cMAMS) (Kim *et al.*, 2004). CSM provides genomics researchers with robust significant gene selection and suggests more reliable one among the five statistical methods by comparing the statistic ranking order of the common genes from different statistic results. The used statistical analysis methods are t-test (Lowry, 2004), Kruskal-Wallis test (also known as Wilcoxon signed rank test in case of two sample analysis) (Lowry, 2004), SAM (Tusher *et al.*, 2002), two sample multiple test (Ge *et al.*, 2003; Dudoit and Ge, 2004), and Empirical Bayesian model (Efron *et al.*, 2001).

Statistical Methods

A statistical test consists of two steps. The first step is to build a summary test statistic. The next step is to determine the significance level with the given test statistic. Usually, the significance level is constructed under the specified or estimated modeling assumption. In statistical analysis of microarray data, a relatively less sufficient sample size than other types of statistical data could mislead model assumption.

The simplest method called 'fold change determination' for differential expression is to evaluate the log ratio by more than a certain arbitrary cut-off value. It is known to be unreliable for various reasons. Recently many sophisticated statistical methods have been proposed, but it still remains to be controversial how accurate these methods are. We considered the five well-known methods to identify the statistically significant genes and to choose a more reliable method of choice if needed.

T-test is a widely used statistical method. However, its major problem is the strong assumption on the null

*Corresponding author: E-mail insong@nih.go.kr, Tel +82-31-436-8627, Fax +82-31-436-8636
Accepted 9 December 2004

Table 1. An example of comparative analysis result of Comparative Statistic Module (CSM) with the ALL/AML data set (Golub et al., 1999).

Test Method	Cut-off Value (P value)	Number of Significant Genes	Rank SUM	Rank SUM/ Number of Common Genes*	Method of Choice Rank
T-test	0.05	1,045	227,680	342.8916	2
Kruskal-Wallis	0.05	1,055	233,335	351.4081	4
2-sample multiple	0.05	883	297,503	448.0467	5
E-Bayes	0.05	714	228,453	344.0557	3
SAM	10% FDR**	1,121	227,672	342.8795	1***

*the number of commonly selected genes = 663
/*Strictly speaking, the SAM rank is not comparable with other test ranks, although estimated corresponding FDR cut-off value for p-value of 0.05 is around 0.1

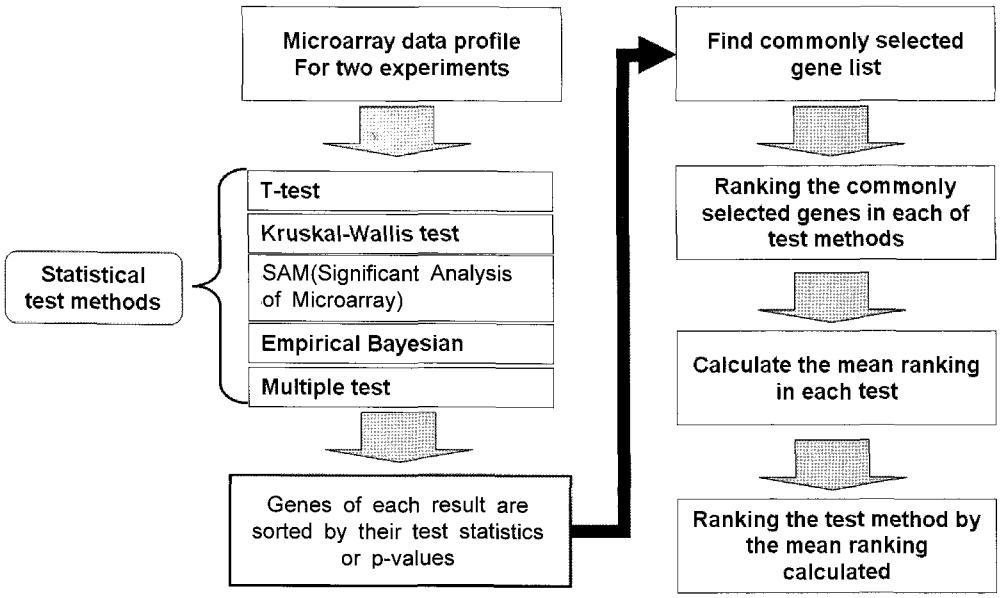

Fig. 1. Architecture of Comparative Statistic Module (CSM). Each statistical test is performed based on a specified cut-off value and sample group settings set by users. Commonly selected gene sets are provided with the brief comparative result summary of the five statistical methods.

distributions of the test statistics. The Wilcoxon rank-sum test is a nonparametric alternative to the two-sample t-test which is based on the order in which the observations from the two samples fall. When the assumption of the two-sample t-test holds, the Wilcoxon test is somewhat less likely to detect a location shift than the two-sample t-test is (Lowry, 2004). Two sample multiple test uses an adjusted p-value for complementing large multiplicity problems in which thousands of hypotheses are tested simultaneously (Dudoit and Ge, 2004). There are many criteria for calculating the adjusted p-value in two sample multiple test. But it is hard to know which criterion is the best. Efron and colleagues (2001) developed a simple Empirical Bayes test. This approach produces reliable *a posteriori* probabilities of activity differences for each gene, starting with a minimum of *a priori* assumptions. Significance Analysis of Microarray (SAM) assigns a score to each gene on the basis of change in gene expression relative to the standard deviation of repeated measurements. SAM uses permutations of the repeated measurements to estimate the percentage of genes identified by chance, i.e. the false discovery rate (Tusher *et al.*, 2002).

Comparative Statistic Module

CSM for differentially expressed gene is implemented using the aforementioned five statistical methods. These methods determine significant genes by two experimental groups. This module presents commonly selected genes first and proposes a method of choice among the five statistical methods by assigning ranks by calculating the average rank of the commonly selected genes (Table 1 and Fig. 1).

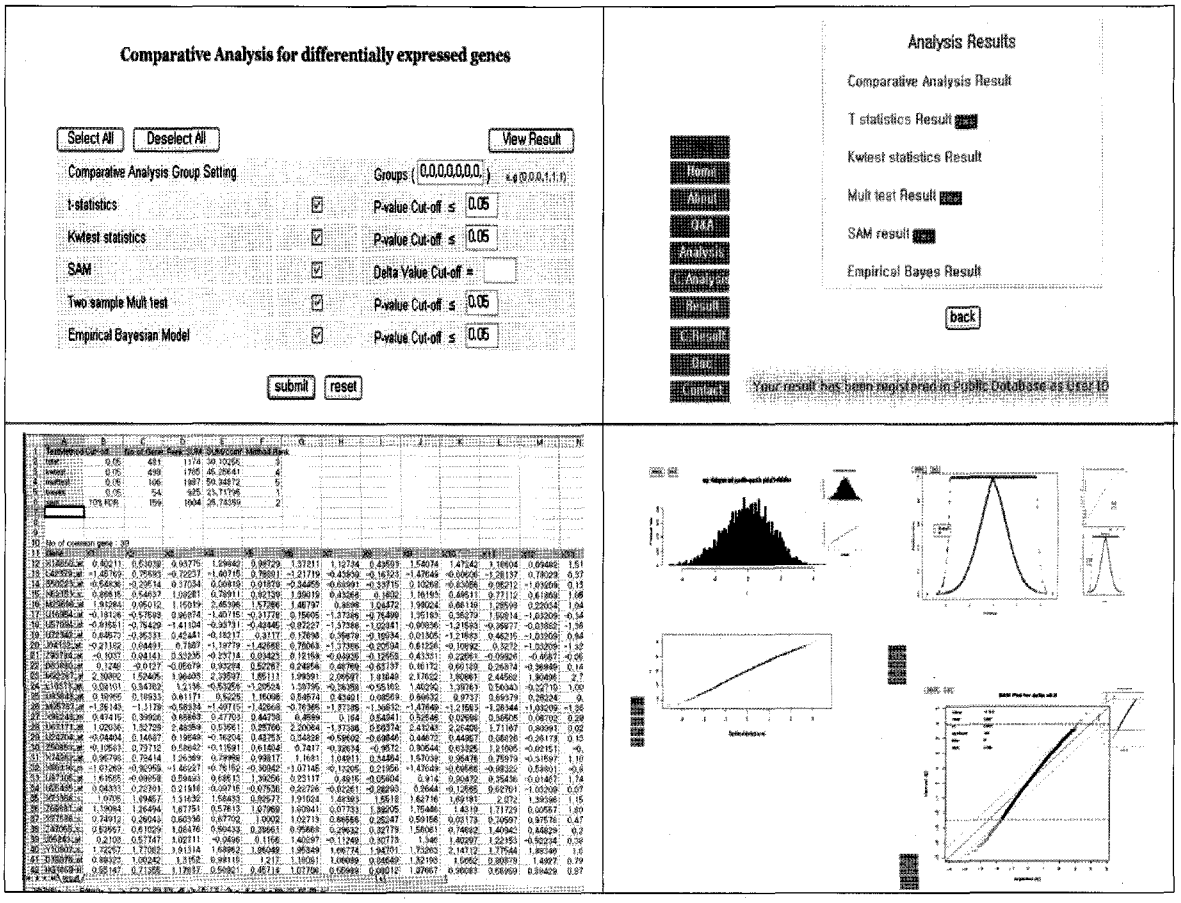

Fig. 2. General interface of Comparative Statistic Module (CSM). Data input with comma-separated values file(.csv). Upper left screenshot: Users select methods and adjust cut-off values of the statistical methods. Upper right screenshot: the analysis result menu for the five statistical methods. Lower left screenshot: the comparative analysis result screen shows two result tables. The top one is a brief comparative result summary of the five statistical methods and the bottom one is the list of the commonly selected genes with a specified significance level for each method. Lower right screenshot: Several charts show the results of t-test, two sample multiple test and SAM test.

Each statistical method lists up the significant genes and their ranks at a specified significance level. The specific rank of the same gene could be very different across the five statistical methods due to difference in specific statistical assumptions. Some significant genes by one statistical method could be insignificant by another, but some genes could be commonly listed in every test. We suggest that these common genes listed in every test would be more reliable than those listed in a single statistical test result. Experimental confirmation is desired to be performed first with these robustly listed common genes.

A user inputs a data set containing gene expression profiles. The CSM module computes the rank of each gene in each test satisfying a certain significance level of p-value (e.g. p=0.05) which is given by a user. It outputs the commonly selected gene set including all significant genes in each test, average rank of each test, and method of choice rank among the five tests (Fig. 2). We provide the ranks of the five tests to decide a method of choice for given input data. We performed a sample run for the CSM module with ALL/AML data set (Golub *et al.*, 1999). The result is shown in Table 1. The number of common genes is 663 and SAM is suggested as the method of choice with the given data. CSM helps to choose an optimal statistical method for significant gene selection by providing a brief summary of comparative result for the five statistical methods.

The current version aims to increase the statistical significance of differentially expressed genes that might become interesting targets for further studies. The implemented CSM is added as a module to cMAMS with R sources (*http://www.r-project.org*). Despite the fact that p-value is not suitable for small sample size data as

in microarray experiments, in which the sample size is less enough to ensure normal distribution, the four tests except SAM in this module use p-value as significance level measures. Future version will attempt to convert each p-value into corresponding False Positive Rate (FDR) to improve the statistical robustness for identifying common genes and to decide a method of choice in the five test methods.

Acknowledgements

This study was supported by the intramural fund of the National Institute of Health, Korea. Authors would like to thank Mr. Joshua Yang for his proof reading of the manuscript.

References

Cui, X. and Churchill, G.A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.* 4(4):210.

Dudoit, S. and Ge, Y. (2004). Bioconductor's multtest package. *http://www.mssm.edu/faculty/yongchao-ge/multtest/multtest.pdf.*

Efron, B., Tishirani, R., Storey, J.D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 96, 1151-1160.

Ge, Y., Dudoit, S., and Speed T.P. (2003). Resampling-based multiple testing for microarray data analysis. *TEST* 12, 1-44 (plus discussion p. 44-77). (Technical Report)

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomeld, C.D., and Lander, E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531-537.

Kim, S.B., Kim, Y.J., Kim, H.M. Jung, H.Y., Lee, E.J., Park, J.S., Park, Y.J., and Koh, I.S. (2004). cMAMS: cDNA Microarray data Analysis and Management System. Proceedings of the 31st Korea information *Science Society Spring Conference* 2, 247-249.

Lobenhofer, E.K., Bushel, P.R., Afshari, C.A., and Hamadeh, H.K. (2001). Progress in the application of DNA microarrays. *Environ. Health Perspect* 109, 881-891.

Lowry, R. (2004). Concept and Applications of inferential statistics. *http://faculty.vassar.edu/lowry/webtext.html.*

Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* 18, 546-554.

The R Project for Statistical Computing. (2004). *http://www.r-project.org.*

Tusher, V.G., Narsimhan, B., Tibshirani, R., and Chu, G. (2002). Significance analysis of microarrays. User Guide and Technical Documentation. *http://www-statstanford.edu/~tibs/SAM/.*