Biomedical Statistics and Informatics
2017; 2(4): 166-171
http://www.sciencepublishinggroup.com/j/bsi
doi: 10.11648/j.bsi.20170204.16

Review Article

Statistical Tests for Identification of Differentially Expressed Genes in Microarray Data

Harun or Rashid, Arefin Mowla, Siddikur Rahman, Siraj-Ud-Doulah, Bipul Hossen*

Department of Statistics, Faculty of Science, Begum Rokeya University, Rangpur, Bangladesh

Email address:

hb0910009@gmail.com (H. or Rashid), arefinmowla.milu@gmail.com (A. Mowla), siddikur_brur@yahoo.com (S. Rahman), sdoulah_brur@yahoo.com (Siraj-Ud-Doulah), mbipu.ru@gmail.com (B. Hossen)

*Corresponding author

To cite this article:

Harun or Rashid, Arefin Mowla, Siddikur Rahman, Siraj-Ud-Doulah, Bipul Hossen. Statistical Tests for Identification of Differentially Expressed Genes in Microarray Data. *Biomedical Statistics and Informatics.* Vol. 2, No. 4, 2017, pp. 166-171. doi: 10.11648/j.bsi.20170204.16

Received: July 29, 2017; **Accepted**: August 30, 2017; **Published**: October 20, 2017

Abstract: Gene expression assay provide a fast and organic way to identity disease markers relevant to clinical trial in modern age. In microarray experiments, differentially expressed genes, or discriminator genes, are the genes with considerably different expression patterns in two user-defined groups. Typically microarray data consists of huge amount of genes, and which genes are responsible or differentiable for a particular disease. Identification of differentially expressed genes across multiple conditions has become a vigorous statistical problem in analyzing large-scale microarray data. In this perspective, we considered a simulated data and real data sets (Head and Neck cancer). This paper uses some statistical methods: t-test, Wilcoxon signed-rank sum test and renewed approach to detect the differential expression of genes between conditions and finding the required number of differentially expressed genes. Additionally Principal Component Analysis (PCA) and largest difference from mean and data methods are used for visualizing outliers and finding numerical outliers respectively. If introducing some artificial outliers to simulated and real data sets and these outliers are not affected or not related to the differentially expressed genes. Results reveal that 25, 126 and 385 differentially expressed genes are identified by using t-test, Wilcoxon Rank sum test and Renewed Approach respectively. Among the three methods 23 common genes those are may be responsible for cancer disease. This paper shows that the two samples mean test (t-test) is perfectly used to identify the differentially expressed genes in microarray data.

Keywords: Microarray Gene Expression Data, T-Test, Renewed Approach, Wilcoxon Signed Rank Test, Differentially Expressed Genes, Outlier

1. Introduction

Microarray technology has been introduced as an important tool in genome research during the last two decades and it is used to simultaneously evaluate the expression levels of thousands of genes [1]. Microarray studies produce very large-scale data sets, requiring statistical methods planned for their analysis. An important goal of microarray studies is to find genes with different level of expression across various types of tissue samples. Gene expression in cells is of concerning because it allows a way to pinpoint disease markers that are related to medical treatments [2]. A term that many researchers may want to perform that used to identify which genes in a cell are differentially expressed in microarray data. For example, a researcher may need to prosecute an idea to discover differentially expressed genes between two different conditions. For explanation purposes it is used to be between healthy patients and cancer patients. Microarray analysis is to allow the researcher to find which genes are expressed differently between these two different groups of patients. Then researchers will be able to develop a treatment to the targeted differentially expressed genes and create a more effective type of therapy.

A microarray is used to analyze gene expressions data analysis. Mostly the gene expression microarray technology is available in two types of stages, single channel microarrays (Affymetrix) and double channel microarrays (cDNA) [3, 10, 13]. In this technology, easily used these methods to identify those genes that are differentially expressed. This study conducted under both of the simulated

and real data sets.

Over the years many methods have been used to perform the analysis of microarray data. To analyze the large number of genes in the complicated biological systems, researchers usually group the genes with similar expression profiles [3]. That means, one group is from normal cells and another is from cancer cells. Several statistical methods for such an analysis have been proposed. A straightforward method is to use the traditional two-sample t-test [4]. Thomas [5] Proposed a regression modeling approach. Pan [6] suggested a mixture model allows, which follows the basic idea of A Renewed Approach [9] to the Nonparametric Analysis and Willcoxon Rank sum test [8]. These methods are used to nursing expression levels of thousands of genes simultaneously and detecting those genes that are differentially expressed.

There are a small number of analyses in recent literature for identifying the differentially expressed genes the performance of different method applied to microarray gene expression data. Recently, some methods to identify and analyze the differentially expressed genes based on microarray glioma data [11, 12] were worked for identifying differentially expressed genes and signature pathways of human prostate cancer. Most recently, the identification of robust clustering methods in gene expression data analysis are discussed in [13]. However all of the author's demonstrated different methods in different studies are better measure to analyze the differentially expressed genes in microarray data in their analysis.

Motivated by this problem it is important to consider all of the methods for identifying differentially expressed genes by which method are comparatively best. This paper considered two-sample t-test, Wilcoxon signed-rank sum test and renewed approach for identification of differentially expressed genes with replicated measurements of expression levels of each gene under each condition. Additionally PCA and largest difference from mean and data methods are used for visualizing outliers and finding numerical outliers respectively that are implemented to both simulated and real data set.

2. Materials and Methods

2.1. Data Description

Suppose that Y_{ji} is the expression level of gene j in array i (j=1,… n; j=1,…, n, n_1+1, ... n_1+n_2). Suppose that the first n_1 and last $n_2$arrays are obtained under two conditions respectively. A general statistical model can be written as

$$Y_{ji} = a_j + b_j X_i + \varepsilon_{ji}$$

Where X_i=1 for 1≤ i ≤ n_1 and X_i = 0 for $n_1+1 \leq$ i ≤ n_1+n_2), and ε_j are random errors with mean 0. Hence a_j+ b_j and a_j are the mean expression levels of gene j under two conditions respectively. Determining how many a gene has differential expression is equivalent to testing for the null hypothesis

H_0: b_j =0 against H_1: b_j ≠0.

A statistical test consists of two parts. The first is to construct a summary test statistic. The second is to determine the significant level or p-value associated with test statistic. The p-value is usually calculated based on null distribution of the test statistic (i.e the distribution of the test statistic under H_0), which may be specified or estimated via modeling assumptions.

This study considered a simulated data and real data sets. To create a simulated data of a 4400 (100×44) gene expression microarray (including 20 differentially expressed genes) are created by adding a single simulated gene to the current data set. More about the real data set (Head & Neck Cancer), Data platform GPL8300: [HG_U95Av2] Affymetrix Human Genome U95 Version 2 Array. Dataset Record GDS2520. Sample count: 44. Data set consist of 12,625 genes. There are two distinct part of this data set. One part consists of 22 samples from normal mucosa. Another part consists of 22 samples from cancer. Both samples are taken from same patients. That is each gene contains 22 normal samples and 22 cancer samples.

2.2. Two Sample T-Test

Two samples mean t-test is a widely recommended approach in modern days for finding differentially expressed genes. It should use two samples mean t-test when sample sizes and variances are unequal between groups, and gives the same result when sample sizes and variances are equal.

The test statistic is,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S.\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}}$$

Where, the pooled standard deviation is-

$$S = \sqrt{\frac{(n_1-1){S_1}^2 + (n_2-1){S_2}^2}{n_1+n_2-2}}$$

This t-test is also known as student t-test. It is the method of calculating the degrees of freedom is related to the idea of the satterthwaite approximation. The good performance of this test, compared with many other alternatives (e.g., approximating by n_1+n_2-2), has been well documented [14]. Ignoring the part of multiple comparison adjustment [15], adopt the same t-statistic, but calculate the p-value by permutation.

2.3. Renewed Approach

A renewed approach is most widely used popular method to identify the differentially expressed genes. [9] Proposed a nonparametric analysis of replicated microarray studies. This method is based on the mixture model proximity given in [6] and uses nonparametric kernel estimation to calculate the distributions of the test statistics. The idea of the mixture model proximity is to combine two test statistics, the common t-test statistic with a second null statistic. Under the null (H_0) hypothesis both test statistics should have the same

distribution. However this was not the case for the null statistic given in [6] and proposed a new version of the null statistic [6, 20]. Combine the method of [9] with this corrected test statistic [9]. In simulation data and also for real data it can use this renewed approach method to find exact number of differentially expressed genes.

2.4. Wilcoxon Signed-Rank Sum Test

The Wilcoxon signed-rank sum test is a non-parametric statistical hypothesis test used when comparing two related samples, matched samples, or frequent measurements on a single sample to assess whether their population mean ranks differ (i.e. it is a paired difference test). Today's the Wilcoxon signed-rank sum test is most powerful test to find out the differentially expressed genes in microarray data sets under the test statistic,

$$W = \sum_{i=0}^{N_r} [sgn(x_{2,i} - x_{1,i}).R_i]$$

For large sample this test statistic follows the normal distribution under the given test statistics,

$$z = \frac{W}{\sigma_W}, \sigma_W = \sqrt{(N_r(N_r+1)(2N_r+1))/6)}$$

If $|z| \geq z_{critical}$ then reject H_0 (two-sided test)

Alternatively, one-sided tests can be realized with either the exact or the approximate distribution. P-value can also be calculated.

2.5. PCA Method for Visualization of Outliers

Principal component analysis (PCA) is one of the most vital techniques for detecting outliers in various applications such as microarray technology. Most PCA-based models for outlier detection handle in batch mode [16], where the model is first informed using training data and is then used to test the others data for outliers. PCA is a statistical multivariate analysis technique which shows the correlation among variables and represents the data into a new set of few variables showing the maximum variance. These variables are identified as principal components (PCs) and each PC is a linear combination of main variables [17]. The coefficient vectors of this linear combination define the corresponding principal components direction. PCA can be formulated as an optimization problem which reduces the reconstruction error as

$$\min_{p \varepsilon R^{n \times k}, \|p\|=I} \sum_{i=1}^{t} \|(x_i - \mu) - pp^T(x_i - \mu)\|^2$$

This is the mathematical view of PCA and need to find the various PCA's with the concepts of eigen value and eigen vector.

2.6. Largest Difference from Mean and Data

In microarray data, finds value with largest difference between it and sample mean, which can be an outlier. Outliers should be investigated cautiously. Often they include valuable information about the process under investigation or the data gathering and recording process. Before assuming the possible elimination of these points from the data, one should try to understand why they revealed and whether it is likely similar values will continue to appear. Of course, outliers are often bad data points. [18] Proposed an outliers detection procedure, which is popularly known as largest difference from mean and data for finding numerical outliers.

3. Experiments and Results

3.1. Data Analysis

The Anderson-Darling normality test [19] is techniques to detect the normality pattern of the identification of differentially gene expression microarray data. Our simulated data are scaled by their normality pattern. By applying Anderson darling normality test in real data (Head & Neck Cancer) we get 7382 genes whose both conditions (normal & cancer) come from normal distribution, 1807 genes whose both condition come from out of normal distribution and 3436 genes whose one condition comes from normal distribution only.

3.2. Outliers Analysis

Mainly the microarray gene expression data is so much noisy, mixture with expression pattern, down regulated and up regulated i.e., outliers may be presented. Here it use two methods for monitoring the outliers one is PCA method for visualizing outliers and another is largest difference from mean and data for finding numerical outliers. Adding 11 artificial unusual observations in 4400 (100×44) gene expression microarray simulated data sets also gives the same result.

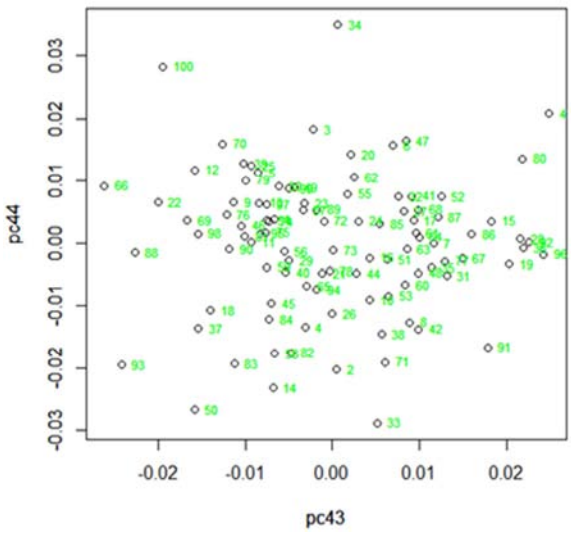

Figure 1. *PCA method for visualizing outliers in simulated data.*

That is 20 differentially expressed genes in simulated data with outliers by t-test at 5% level of significance. Again, PCA

method visualized outliers in real microarray gene expression data and 44 outliers are detected by using largest difference from mean and data. Although real data provided some outliers but these outliers are not affected or not related to the differentially expressed genes.

That means, if introduced some artificial outliers to the both simulated and real data sets and these outliers are not affected or not related to the differentially expressed genes.

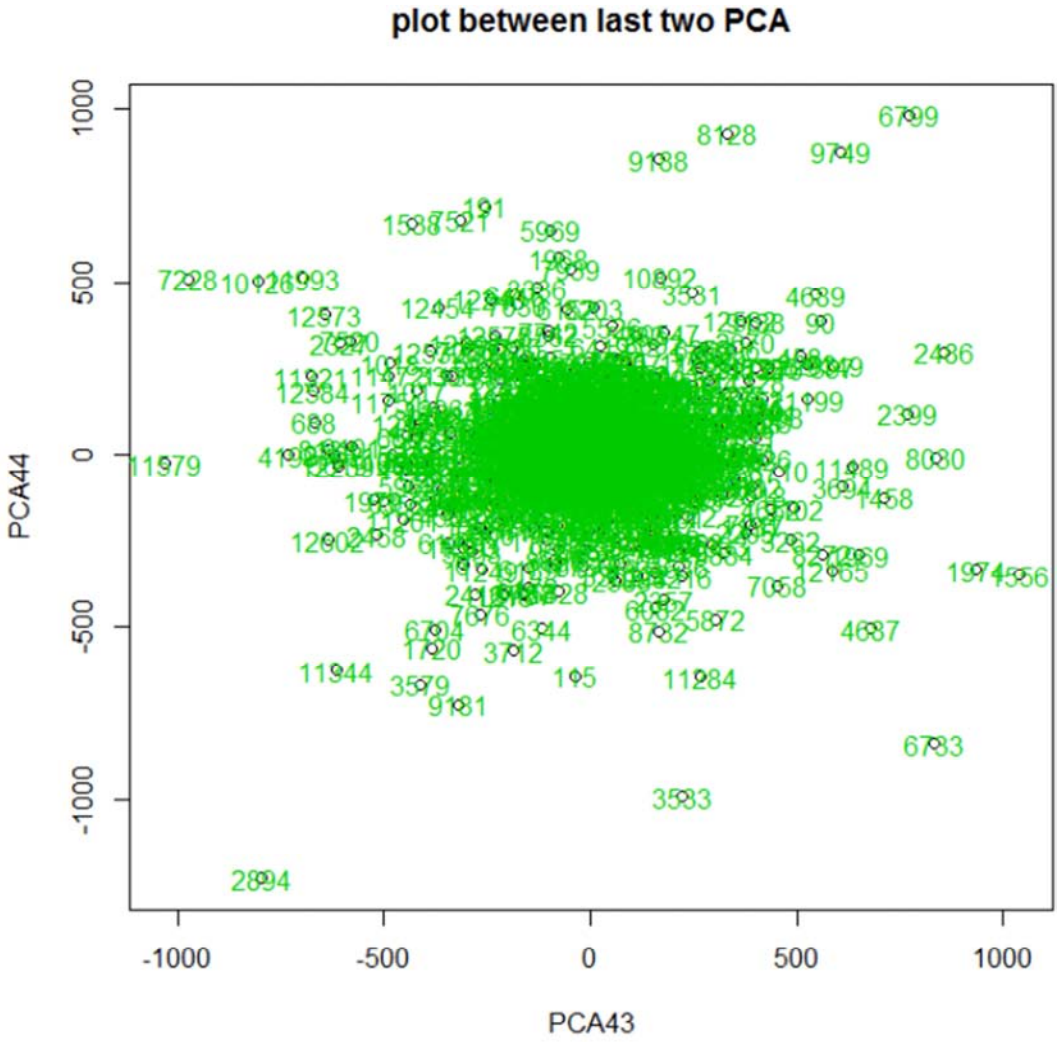

Figure 2. *PCA method for visualizing outliers in real data.*

3.3. Identification of Differentially Expressed Genes

Mainly in microarray data uses some statistical methods: t-test, Wilcoxon signed-rank sum test and renewed approach are identified the mean differential expression of genes between conditions are conducted applying correction for multiple testing and finding the required number of differentially expressed genes. To identify differentially expressed genes most times R 3.3.0 has used with extra package like degenes, tests, pca, limma, fpc, gplots and outliers. Several times Microsoft-Excel and Microsoft-Word are used as calculation and typing software.

In formulated microarray gene expression simulated data, we get 20 deferentially expressed genes by using t-test at 5% level of significance. But in renewed approach method and Wilcoxon signed-rank sum test gives more than 20 differentially expressed genes from 100 microarray gene expression simulated data at 5% level of significance.

In real microarray gene expression data, we get 25 differentially expressed genes at 5% level of significance. By applying the Wilcoxon signed-rank sum test we get 126 differentially expressed genes from the given microarray gene expression real data sets with Bonferroni Correction and FDR (False Discovery Rate) at 5% level of significance. By renewed approach method we get 385 differential expressed genes with Bonferroni Correction and FDR at 5% level of significance.

3.4. Some Common Identification of Differentially Expressed Genes in Real Data

In microarray gene expression data we have some genes these are commonly predominated at all of three methods (t-test, renewed approach and Wilcoxon signed-rank sum test). Here, 23 common genes those are detected all methods and some of the genes are responsible for cancer disease. The list of these genes with related disease is given in Table 1.

Table 1. *List of 23 Common Genes.*

S.N.	Gene name	Related disease
1	KRT13	Leukokeratosis (http://www.informatics.jax.org/disease/193900)
2	KRT4	basaloidn squamous cell carcinoma(http://www.genecards.org/cgi-bin/carddisp.pl?gene=KRT)
3	PLOD3	lysyl hydroxylase 3 deficiency, and ullrich congenital muscular dystrophy (http://www.genecards.org/cgi-bin/carddisp.pl?gene=PLOD3)
4	COL4A2	porencephaly 2, and porencephaly(http://www.genecards.org/cgibin/carddisp.pl?gene=COL4A2)
5	ITPR2	Carcinoma, Lung Diseases, Hepatitis B, Myeloid Leukemia, Kidney Neoplasm (http://www.novusbio.com/itpr2.html)
6	MAL	Breast, esophageal, ovarian, and cervical cancers(http://mcr.aacrjournals.org/content/7/2/199.abstract)
7	PPL	
8	SASH1	colon cancer, breast cancer (http://www.novusbio.com/SASH1-Antibody_NBP1-26650.html)
9	LRP10	Alzheimer‘s disease (http://www.molecularneurodegeneration.com/content/7/1/31)
10	RAF1	leopard syndrome, and murray valley encephalitis(http://www.genecards.org/cgi-bin/carddisp.pl?gene=RAF1)
11	CSTB	Alzheimer's disease (www.wikigenes.org/.../1476.ht)
12	TRIO	toxic encephalopathy, and soft tissue sarcoma (www.genecards.org/.../carddisp..)
13	TYR	oculocutaneous albinism type 1, and amelanotic melanoma(www.genecards.org/.../carddisp.....)
14	CEACAM1	Prostate carcinomas, hepatic tumors, breast cancer (http://www.wikigenes.org/e/gene/e/634.html)
15	PITX1	acute diarrhea (http://www.genecards.org/cgi-bin/carddisp.pl?gene=PITX1)
16	EPHX2	Hypercholesterolemia(http://www.ncbi.nlm.nih.gov/gene/203)
17	EXT1	hereditary multiple exostoses.... Bessel-Hagen disease; diaphyseal aclasis; exostoses, multiple hereditary; familial exostoses (ghr.nlm.nih.gov/.../hereditary)
18	TGM3	Celiac Disease (http://en.wikipedia.org/wiki/Anti-transglutaminase_antibodies)
19	11-Sep	
20	IL1RN	Tourette syndrome (http://www.ncbi.nlm.nih.gov/pubmed/20399384)
21	FSCN1	cell sarcoma, and juvenile xanthogranuloma (www.genecards.org/.../carddisp)
22	UBE2S	Carcinoma, Squamous Cell; Esophageal Neoplasms; Pemphigoi, Bullous (http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/av.cgi?db=human&c=Gene&l=UBE2S)
23	PGF	placental growth factor(https://www.ncbi.nlm.nih.gov/gene/5228)

3.5. Comparative Summary of Three Methods

The different explicit expression of the test statistic of these three methods are given i.e., these three methods are identified some different types of differentially expressed genes in microarray data. Among these methods we get some common genes to identify the differentially expressed genes in microarray data. Using t test we get 25 Deferentially Expressed genes, by Wilcoxon Rank sum test we get 126 Deferentially Expressed genes, by Renewed Approach we get 385 Differentially Expressed genes. We get 23 common genes those are detected all the three methods and some of the genes are responsible for cancer disease. The summary results are shown in Table 2 and Figure 3 among these different statistical methods.

Table 2. *Summary results.*

Methods	Differentially Expressed Genes (DEG)	Accuracy
t-test	25	0.92
Renewed Approach	385	0.06
Wilcoxon Signed Rank Sum Test	126	0.18

Table 2 and Figure 3 illustrates two samples mean test (t-test) is the most powerful widely used method to detect the differentially expressed genes in huge number of data set in gene expression microarray real data analysis.

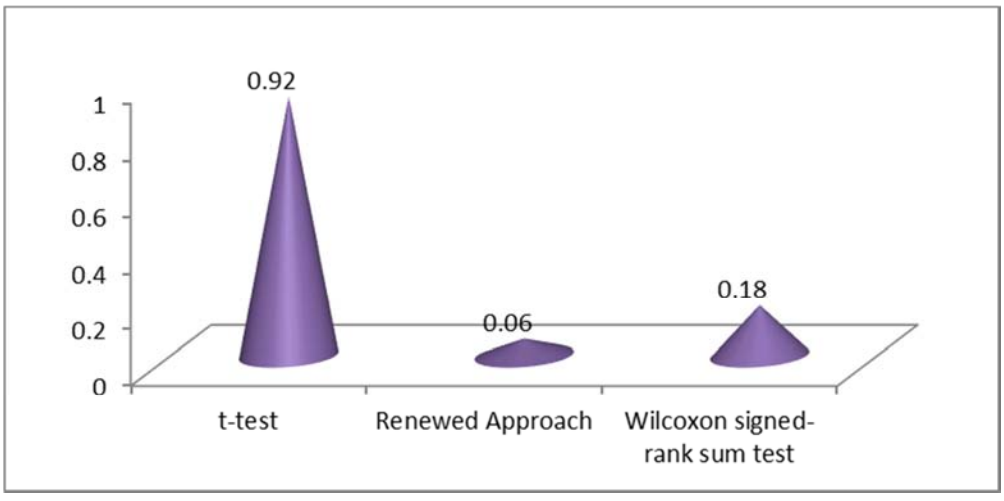

Figure 3. *Summary Results.*

4. Conclusion

Identify a small number of differentially expressed genes for accurate classification of gene samples is essential for the development of diagnostic tests. The aim of this paper is to examine the identification of differentially expressed genes in microarray data. Results reveal that the identification of differentially expressed genes by using such kind of different statistical methods - two samples mean test (t-test), Bonferroni correction and FDR, Renewed method, Wilcoxon signed-rank sum test. Additionally PCA and large sample mean test for outlier's detection with simulated data set and gene expression microarray real data set. To best of our knowledge, two samples mean test (t-test) method is perfectly used to detect the differentially expressed genes in microarray data according to their accuracy.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful comments.

References

[1] Nguyen TV, Andresen BS, Corydon TJ, Ghisla S, Abd-El Razik N, Mohsen AW, Cederbaum SD, Roe DS, Roe CR, Lench NJ, Vockley J (2002); Identification of isobutyryl-CoA dehydrogenase and its deficiency in humans. Mol Genet Metab, vol. 77, pp. 68-79.

[2] Chu G, Narasimhan B, Tibshirani R, Tusher V (2002); "SAM "Significance Analysis of Microarrays" Users Guide and technical document."

[3] Monti S, Tamayo P, Mesirov J, Golub T. (2003); Consensus clustering: a re-sampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn, vol. 52, pp. 91-118.

[4] Devore J. And Peck R (1997); "Satistics: The exploration and analysis of data", 3rd edition, Duxury Press, Pacific Grove, CA.

[5] Thomas JG, Olson JM, Tapscott SJ, Zhao (2001); An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. Genome Research, vol.11, No. 7, pp. 1227-1236.

[6] Pan W (2001); A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics, vol. 18, pp. 546-554.

[7] Efron B, Tibshirani R, Gross V, Tusher VG (2001); Empirical Bayes analysis of a microarray experiment. Journal of American Statistic Association, vol. 96, pp. 1151-1160.

[8] Tusher VG, Tibshirani R, and Chu G (2001); "Significance Analysis of Microarrays Applied to the Ionizing Radiation Response," Proceeding National Academy of Sciences USA, vol. 98, pp. 5116-5121.

[9] Jung K., Quast K., Gannoun A. and Urfer W. (2006); A renewed approach to the nonparametric analysis of replicated microarray experiments. Biometrical Journal, vol. 48, pp. 245-254.

[10] Quackenbush J (2001); Computational analysis of cDNA microarray data. Nature Reviews, vol. 6, No. 2, pp. 418-428.

[11] Chun-Ming Jiang, Xiao-Hua Wang, Jin Shu, Wei-Xia Yang, Ping Fu, Li-Li Zhuang, Guo-Ping Zhou (2015); Analysis of differentially expressed genes based on microarray data of glioma. Int J Clin Exp Med, vol. 8, pp. 17321–17332.

[12] Jennifer SM, Ariana KL, Charles JR, Qing-Xiang AS (2015); Differentially Expressed Genes and Signature Pathways of Human Prostate Cancer. PLoS One, vol. 10, No. 12, e0145322. https://doi.org/10.1371/journal.pone.0145322.

[13] Hossen Md. B. and Siraj-Ud-Doulah (2016); Identification of Robust Clustering Methods in Gene Expression Data Analysis. Current Bioinformatics, vol. 11, No. 3, pp. 01-05.

[14] Best DJ and Rayner JC (1987); Multiple Comparisons, Selection and Applications in Biometry. Vol. 30, pp. 719-724.

[15] Dudoit S, Shaffer CBJ (2003); Multiple hypothesis testing in microarray experiments. Statistical Science. vol. 18, No. 1, pp. 71–103.

[16] Alka B, Monir HS, Hassan AK (2015); Incremental principal component analysis based outlier detection methods for spatiotemporal data streams. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. II-4/W2, pp. 67-71.

[17] Jolliffe (2001); Principal Component Analysis, 2nd edition, Springer Series in Statistics.

[18] Snedecor, G. W., Cochran, W. G. (1980). Statistical Methods (seventh edition). Iowa State University, Press, Ames, Iowa.

[19] Corder, G. W., Foreman, D. I. (2009). Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach Wiley, ISBN 978-0-470-45461-9.

[20] Meiller A, Alvarez S, Drané P, Lallemand C, Blanchard B, et al. (2007); p53-dependent stimulation of redox-related genes in the lymphoid organs of gamma-irradiated mice: identification of haeme-oxygenase 1 as a direct p53 target gene. Nucleic Acids Res, vol. 20, pp. 6924–6934.

[21] Zhao LP, Prentice R and Breeden L (2001); Statistical modeling of large microarray data sets to identify stimulus-response profiles. Proc National Acedemy of Science USA, vol. 98, pp. 5631-5636.