Jurnal Teknologi

Full Paper

COMPARATIVE ANALYSIS OF SHRINKAGE COVARIANCE MATRIX USING MICROARRAYS DATA

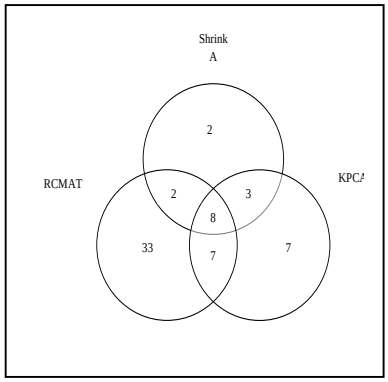
Suryaefiza Karjanto[a*], Norazan Mohamed Ramli[b], Nor Azura Md Ghani[b]

[a]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Melaka (Kampus Jasin), Melaka, Malaysia
[b]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

Article history
Received
25 October 2015
Received in revised form
14 December 2015
Accepted
9 February *2016*

*Corresponding author
suryaefiza@gmail.com

Graphical abstract



Abstract

The DNA microarray technologies permit scientists to depict the expression of genes for related samples. This relationship between genes is analysed using Hotelling's T^2 as a multivariate test statistic but the disadvantage of this test, when used in microarray studies is the number of samples is larger than the number of variables. This study discovers the potential of the shrinkage approach to estimate the covariance matrix specifically when the high dimensionality problem happened. Consequently, the sample covariance matrix in Hotelling's T^2 statistic is not positive definite and become singular thus cannot be inverted. In this research, the Hotelling's T^2 statistic is combined with a shrinkage approach as an alternative estimation to estimate the covariance matrix to detect significant gene sets. The multivariate test statistic of classical Hotelling's T^2 is used to integrate the correlation when assessing changes in activity level across biological conditions. The performances of the proposed methods were assessed using real data study. Shrinkage covariance matrix approach indicates a better result for detection of differentially expressed gene sets as compared to other methods.

Keywords: Hotelling'sT^2; gene set analysis; shrinkage covariance matrix

1.0 INTRODUCTION

DNA microarray has been a very successful tool among high-throughput techniques to examine gene and protein functions using a large amount of data. In recent times, many researchers have relentlessly continued the attempt to improve the analysis of genes by conducting in-depth studies in biological research which help to improve the interpretability and understanding of microarray data analyses. Before the development of the microarray technology, techniques for the analysis of gene expression include Northern blotting [1], differential display [2] and serial analysis of gene expression (SAGE) [3]. However, these techniques are only able to monitor a limited number of gene expressions at one time. In addition, the sensitivity and complexity of some procedures are still in question. Then, the development of microarrays at the end of the last century has improved the ability of scientists to enhance the understanding of gene expression.

Analyzing a large amount of produced data may pose a big problem that need to be overcome. Although various microarray experiments generate a lot of data, discovering the subtle knowledge is still a challenge faced by researchers in this area. Many researchers are not experienced in converting tens of thousands of noisy data points into accurate analysis and reliable interpretation of biologic information making DNA microarray analysis as a challenge. Although some researchers realize the importance of cooperating with skilled biostatisticians to analyse microarray data but the number of skilled

biostatisticians is insufficient. Therefore, the researchers are simply using available software to analyse the microarray data without knowledge of potential drawbacks.

In general, there are two types of microarrays data:

i. **One-colour spotted microarrays** provide estimations of the absolute levels of gene expression. The comparison of the two samples requires two separate one-dye hybridizations (green wavy lines). The collected data represent absolute values of gene expression because only a single dye is used. Both samples of complementary DNA (cDNA) are hybridized and scan separately. The gene is active or inactive are measured by superimposing images obtained from different chips. The advantages of one-colour spotted are lower noise levels and smaller inter-array variability. The only drawback for this type of microarray is the cost of production is more expensive which mean fewer experiments could be conducted [4].

ii. **Two-colour spotted microarrays** are typically hybridized with cDNA prepared from control (cyanine 3-Cy3) and treatment (cyanine 5-Cy5) samples. Both samples are labeled with two different fluorophores: Cy3 with the green colour and Cy5 with the red colour. The samples are mixed and hybridized or bind together to a single microarray. Then the fluorescence intensity of the two fluorophores are measured using microarray scanner for each spot on the microarray slide. If a certain gene is very active. it produces more labeled cDNAs which hybridize to the DNA on the microarray slide thus generate a very bright fluorescent spot and if a gene is less active, it will give results in dimmer fluorescent spot. Then, if a gene is inactive then there is no fluorescence will be produced at all. If a particular gene is more expressed in treatment sample than in control sample then the spot is red, (up-regulated in treatment sample) and if the gene is more expressed in the control sample then the spot is green (down-regulated in treatment sample). If a particular gene is equally expressed in treatment and control samples then the spot is yellow. In addition, black represents the specific gene not express neither in the treatment nor control sample conducted [4].

Our study provides an alternative to estimate covariance matrix for identifying differential gene sets. The shrinkage estimators are expected to estimate covariance matrix when maximum likelihood estimator no longer provides an unbiased estimation. Our objective in this paper is to comprehensively test the proposed new shrinkage covariance matrix from our previous extension works [5] for detecting significant gene sets between different samples using real microarray data sets. We stated in Section 2 about the impact of high dimensionality problem or when the number of genes is larger than the number of samples in sample covariance matrix. We also described in Section 3 that the real microarray data sets is to evaluate the performance of our proposed methods in detecting significant gene sets. Then, Section 4 will describe the results and discussion and finally the Section 5 will summarise the findings.

2.0 PROPOSED SHRINKAGE COVARIANCE MATRIX

This study provides an alternative to estimate covariance matrix using shrinkage method based on the definition of [6, 7, 8, 9]. There were three proposed methods and we referred them as ShrinkA, ShrinkB and ShrinkC for the rest of this study. Our methodology is extended from our previous study [5] but this study validated the proposed shrinkage conariance matrix using real microarray data sets.

The following notations are used to describe experimental data generated in the form of two-colour spotted microarrays. Let n represent the number of slides/samples, and p is the total number of genes in a gene set. Let X_{ki} be the expression level for gene i=1, . . . , p of sample k=1, . . . , n from red colour spotted or the treatment group and X_{kj} be the expression level for gene j=1, . . . , p of sample k=1, . . . , n from green colour spotted or the control group. The expression level vectors for samples k from the treatment and control groups can be expressed as $X_i = (X_{k1}, .., X_{ki}, .., X_{kp})^T$ and $X_j = (X_{k1}, .., X_{kj}, .., X_{kp})^T$, respectively.

The proposed methods provide an alternative to estimate covariance matrix using shrinkage method based on the definition of Ledoit and Wolf [7, 8, 9] and Schafer and Strimmer [6]. The approach is adapted to Hotelling's T^2 and is extended to gene set analysis in the microarray study. Throughout this study, three different methods are proposed and they will be termed as ShrinkA, ShrinkB and ShrinkC for the rest of this thesis. Generally, the algorithm for the three proposed methods is outlined below:

Step 1: Prepare the data sets with the preprocessing procedure using suitable and transformation method and normalization method (if necessary). The most common transformation in microarray data analysis is using logarithmic base two for all expression of genes:

$$X^*_{ki} = \log_2(X_{ki}) \quad (1)$$

Each of the expression level of the gene for each group is normalized which every extreme value are replaced by the winsorize median absolute deviation. The upper limit of extreme value is replaced by:

$$X_{ki}^{+}=\begin{cases} median\,(X_{ki})+a\cdot MAD & ,X_{ki}>median+a\cdot MAD \\ X_{ki} & ,otherwise \end{cases} \quad (2)$$

while the lower limit of extreme value is replaced by:

$$X_{ki}^{-}=\begin{cases} median\,(X_{ki})-a\cdot MAD & ,X_{ki}<median-a\cdot MAD \\ X_{ki} & ,otherwise \end{cases} \quad (3)$$

where three is used in this study as the chosen multiplier, *a* according to Yates and Reimers [10]. The *MAD* is median absolute deviation which is formulated as below:

$$MAD=\underset{i}{median}\left\{\left|I_i-\underset{j}{median}\{I_j\}\right|\right\} \quad (4)$$

for a univariate data set $I_1, I_2, \ldots, I_n$.

Step 2: Compute the shrinkage target according to the proposed approach.

Step 3: Search the optimal shrinkage intensity using the related definition of the proposed method.

Step 4: Substitute the sample covariance matrix in Hotelling's T^2 using the results in **Step 2** and **Step 3**.

Step 5: Compute Hotelling's T^2 for each of all the gene sets that are measured in data sets as explained in:

$$T^2=\frac{n_1 n_2}{n}\left(\bar{X}_i-\bar{X}_j\right)'\left(S_{shrink}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right)^{-1}\left(\bar{X}_i-\bar{X}_j\right) \quad (5)$$

where the mean, $\bar{X}_i$ was defined as:

$$\bar{X}_i=\frac{1}{n}\sum_{k=1}^{n}X_{ki} \quad (6)$$

and the $\bar{X}_j$ is the mean for X_{kj} and S_{shrink} is shrinkage estimator as modelled in:

$$S_{shrink}=\alpha T_{ij}+(1-\alpha)S_{ij} \quad (7)$$

The sample covariance matrix, S_{ij} was defined as:

$$S_{ij}=\frac{1}{n-1}\sum_{k=1}^{n}\left(X_{ki}-\bar{X}_i\right)\left(X_{kj}-\bar{X}_j\right) \quad (8)$$

where shrinkage target, T_{ij} and shrinkage intensity, α was defined as:

$$\alpha=\max\left\{0,\min\left\{\frac{\kappa}{n},1\right\}\right\} \quad (9)$$

where κ was a constant and *n* is the number of samples. The constant κ could be written as:

$$\kappa=\frac{\pi-\rho}{\gamma} \quad (10)$$

where π was the sum of asymptotic variances of the entries of the sample covariance matrix scaled by $\sqrt{n}$. ρ was the sum of asymptotic covariances of the entries of the shrinkage target with the entries of the sample covariance matrix scaled by $\sqrt{n}$. γ was the measurement of the misspecification of the (population) shrinkage target. If κ were known, we could use κ/n as the shrinkage intensity in practice. Unfortunately, κ is unknown, so we searched for a consistent estimator for κ by $\hat{\kappa}$. This is done by finding consistent estimators for the three estimators π, ρ and γ that is $\hat{\pi}$, $\hat{\rho}$ and $\hat{\gamma}$. The proposed methods ensured the covariance matrix was always a positive definite and well defined. Table 1 showed the shrinkage target and shrinkage intensity for ShrinkA, ShrinkB and ShrinkC.

Step 6: Permute samples for each gene set thus claim the significance of gene sets according to permutation testing. The discussion of permutation testing elaborated in:

$$\hat{p}=\frac{\sum_{i=1}^{M}I\left(t_i\geq t^*\right)}{M} \quad (11)$$

where *M* is the permutation test be used, where $t_i, i=1,\ldots,M$ is Hotelling's T^2 statistic that compute from the permutation. Generally, the algorithm for the permutation testing is summarized as below:

Step 1: Permute the number of samples differently for each test.

Step 2: Compute the sum for all Hotelling's T^2 statistic from permutation testing that exceed the original Hotelling's T^2 statistic.

Step 3: Divide the summation with the number of permutations to determine the significance of *p*-values from permutation testing.

3.0 A REAL DATA STUDY

For examining the similarity across methods (in terms of gene sets detected as having differential expression), three gene set analysis methods were applied to diabetes data set originally from [11] who initiated the gene set analysis method. Currently Gene Set Enrichment Analysis (GSEA) is the most well-known and widely used approach to gene set analysis. For this study, the gene expression from 17 persons with normal glucose tolerance and 17 persons with Type 2 diabetes mellitus samples are used as a comparison.

Type 2 diabetes mellitus is a disorder that interrupts the way of body uses glucose that affects over 366 million people worldwide in 2011 alone. All the cells in body need glucose to function normally and the glucose gets into the cells with producing enough insulin. If there is not enough insulin, or if or the body's cells ignore the insulin, glucose builds up in the blood. Hence, the diabetes mellitus is a metabolic disorder characterized by a high blood glucose level. There are two types of diabetes mellitus:

i. **Type 1 diabetes mellitus** results from the pancreas's failure to produce insulin and causes high blood glucose levels, which can cause dangerous if untreated.

ii. **Type 2 diabetes mellitus** results from insulin resistance, a metabolic condition in which cells fail to use insulin properly. This is the most common form of diabetes.

The samples are obtained at the time of diagnosis or before treatment with hypoglycemic medication (an abnormally low level of the sugar glucose in the blood) and under the controlled conditions of a hyperinsulinemic euglycemic clamp (the plasma insulin concentration is acutely raised up and kept at certain level by a continuous infusion of insulin).

The data provides 149 gene sets classified from a total of 22,283 genes (variables). These lists of gene sets were used to define the gene sets in the analysis presented in this study. From 149 gene sets that compiled, 113 are grouped according to involvement in metabolic pathways or gene sets and 36 consist of gene clusters that are corregulated in a mouse expression atlas [11]. The pathways or gene sets is curated from the Broad Institute of MIT and Harvard which formerly known as the Whitehead Institute/MIT Center for Genome Research or WICGR (https://www.broadinstitute.org) and NetAFFX Analysis Centre (http://www.affymetrix.com/analysis/index.affx). However the mouse expression cluster is curated from the Genomics Institute of the Novartis Research Foundation (GNF) founded in 1999 (www.gnf.org).

The performance of our approach was evaluated by comparing the results with those obtained from two other methods: (1) KPCA [12] and (2) RCMAT (Regularized Covariance Matrix Approach) [8] because their methods were also applied using Hotelling's T^2.

Table 1 The shrinkage combinations for ShrinkA, ShrinkB and ShrinkC

Type	Shrinkage Target	Shrinkage Intensity
ShrinkA	$T_{A_{ij}} = \begin{cases} s_{ii} & \text{if } i=j \\ 0 & \text{if } i \neq j \end{cases}$	$\hat{\kappa}_A = \frac{\hat{\pi} - \hat{\rho}_A}{\hat{\gamma}_A}$, $\hat{\pi} = \frac{1}{n}\sum_{k=1}^{n}\left\{(X_{ki}-\bar{X}_i)(X_{kj}-\bar{X}_j) - s_{ij}\right\}^2$ $\hat{\rho}_A = 0$ $\hat{\gamma}_A = \sum_{i=1}^{n}\sum_{j=1}^{n}(s_{ij})^2$
ShrinkB	$T_{B_{ij}} = \begin{cases} s_{ii} & \text{if } i=j \\ \sqrt{s_{ii}s_{jj}} & \text{if } i \neq j \end{cases}$	$\hat{\kappa}_B = \frac{\hat{\pi} - \hat{\rho}_B}{\hat{\gamma}_B}$, $\hat{\rho}_B = \underbrace{\sum_{i=1}^{p}\hat{\pi}_{ii}}_{on\,diagonal} + \underbrace{\sum_{i=1}^{p}\sum_{j=1,j\neq i}^{p}\frac{1}{2}\left(\sqrt{\frac{s_{jj}}{s_{ii}}}\hat{\vartheta}_{ii,ij} + \sqrt{\frac{s_{ii}}{s_{jj}}}\hat{\vartheta}_{jj,ij}\right)}_{off\,diagonal}$, $\hat{\pi}_{ii} = \frac{1}{n}\sum_{k=1}^{n}\left\{(X_{ki}-\bar{X}_i)^2 - s_{ii}\right\}^2$

		$\hat{\vartheta}_{ii,ij} = \frac{1}{n}\sum_{k=1}^{n}\left\{(X_{ki}-\bar{X}_i)^2 - S_{ii}\right\}\left\{(X_{ki}-\bar{X}_i)(X_{kj}-\bar{X}_j) - S_{ij}\right\}$ $\hat{\vartheta}_{jj,ij} = \frac{1}{n}\sum_{k=1}^{n}\left\{(X_{kj}-\bar{X}_j)^2 - S_{jj}\right\}\left\{(X_{ki}-\bar{X}_{i.})(X_{kj}-\bar{X}_j) - S_{ij}\right\}$ $\hat{\gamma}_B = \sum_{i=1}^{n}\sum_{j=1}^{n}(f_{ij}-S_{ij})^2 \quad , f_{ij} = \sqrt{S_{ii}S_{jj}}$
ShrinkC	$T_{Cij} = \begin{cases} S_{ij} & if \ \ i=j \\ \bar{r}\sqrt{S_{ii}S_{jj}} & if \ \ i \neq j \end{cases}$	$\hat{\kappa}_C = \frac{\hat{\pi}-\hat{\rho}_C}{\hat{\gamma}_C}$ $\hat{\rho}_C = \hat{\rho}_B$ $\hat{\gamma}_C = \sum_{i=1}^{n}\sum_{j=1}^{n}(f_{ij}-S_{ij})^2 \quad , f_{ij} = \bar{r}\sqrt{S_{ii}S_{jj}}$

4.0 RESULTS AND DISCUSSION

able 2 shows the RCMAT appears to detect few more gene sets, it detected 50 gene sets under the permutation *p*-value of 0.05 and KPCA method detected 25 gene sets, while our method detected 15 gene sets.

The ShrinkA method identified 15 significant gene sets out of the 148 tested, sorted by ascending nominal p-values. In this case, the overlap among the two and our method were eight gene sets. The most significant pathway in ShrinkA analysis was c25 gene set. The results also show that 4 of the 15 *p*-values of ShrinkA were less than the corresponding *p*-values produced by RCMAT while 9 were less than KPCA's *p*-values. The c25_U133_probes was most significant gene set detected by ShrinkA which corresponds to the same position for RCMAT. In the original study, the OXPHOS gene set was reported as the only significant gene set, with the *p*-value 0.034 [11]. As displayed in Table 1, ShrinkA analys was also detected OXPHOS as significant gene sets with the permutation *p*-value 0.0496 and RCMAT with the permutation *p*-value 0.0441.

The ShrinkB method identified 13 significant gene sets while ShrinkC method identified 24 significant gene sets out of the 148 gene set tested. The results also show that 9 of the 24 *p*-values of ShrinkC were less than the corresponding *p*-values produced by RCMAT while 18 were less than KPCA's *p*-values. The c25_U133_probes was most significant gene set detected by ShrinkC which corresponds to the same position for RCMAT.

Among the significant gene sets identified, there are a number of shared gene sets as shown in Figure 1. Eight gene sets shared in the three methods implying their relatedness in detecting significant gene sets. The ShrinkA shares two gene sets with RCMAT not found in KPCA and three gene sets with KPCA not found in RCMAT. Additionally, RCMAT shares seven gene sets with KPCA not found in ShrinkA. There are two gene sets that were only detected by Shrink A, while seven gene sets were only detected by KPCA. There are also 33 gene sets that had a significantly detected by RCMAT but were overall not detected by ShrinkA nor by KPCA.

Table 2 Comparison of significant of nominal (unadjusted) permutation p-values between ShrinkA, ShrinkB, ShrinkC, RCMAT and KPCA for diabetes data

Gene set	No. of gene set	Shrink A	Shrink B	Shrink C	RCMAT	KPCA
c25_U133_probes	64	0.0009	0.9400	0.0009	0.0003	0.0039
MAP00600_Sphingoglycolipid_metabolism	18	0.0147	0.0483	0.0165	0.0018	0.0036
MAP00511_N_Glycan_degradation	9	0.0191	0.0123	0.1433	0.0072	0.0066
MAP00360_Phenylalanine_metabolism	23	0.0211	0.1663	0.0207	0.0036	0.0723
MAP00252_Alanine_and_aspartate_metabolism	35	0.0238	0.9730	0.0215	0.0131	0.0472
MAP00100_Sterol_biosynthesis	21	0.0239	0.3532	0.0585	0.0986	0.5851
MAP00531_Glycosaminoglycan_degradation	18	0.0328	0.0287	0.0239	0.0043	0.0005
MAP00400_Phenylalanine_tyrosine_and_ tryptophan biosynthesis	12	0.0328	0.0513	0.0260	0.0551	0.0378
MAP00300_Lysine_biosynthesis	5	0.0363	0.2901	0.0059	0.0020	0.0089
MAP00910_Nitrogen_metabolism	31	0.0406	0.0542	0.0262	0.0094	0.0385
MAP00580_Phospholipid_degradation	10	0.0406	0.2042	0.0499	0.0798	0.0574
MAP00940_Flavonoids_stilbene_and_lignin_ Biosynthesis	4	0.0418	0.0464	0.0746	0.0583	0.0238
c23_U133_probes	109	0.0450	0.1323	0.0265	0.0035	0.1581
c34_U133_probes	452	0.0452	0.6180	0.0086	0.0311	0.2366
OXPHOS_HG-U133A_probes	114	0.0496	0.6294	0.1586	0.0441	0.1705

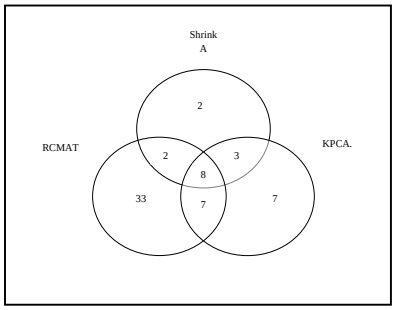


Figure 1 Significant gene sets detected by ShrinkA, RCMAT, KPCA for diabetes data.

The ShrinkB, RCMAT and KPCA share six overall common gene sets as displayed in Figure 2. However, ShrinkB has a similar agreement with RCMAT and KPCA method as shown by the number of common significantly different gene sets detected of two for both methods.

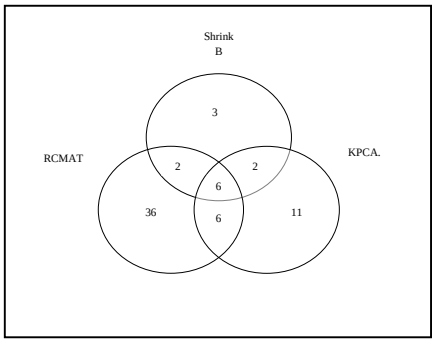


Figure 2 Significant gene sets detected by ShrinkB, RCMAT, KPCA for diabetes data.

Out of the 25 significantly different gene sets identified by ShrinkC in FIgure 3, nine are also detected by both RCMAT and KPCA method. However, ShrinkC has a stronger agreement with RCMAT, than with KPCA method as shown by the number of common significantly different gene sets detected of 11 for the former compared with only two for the latter.

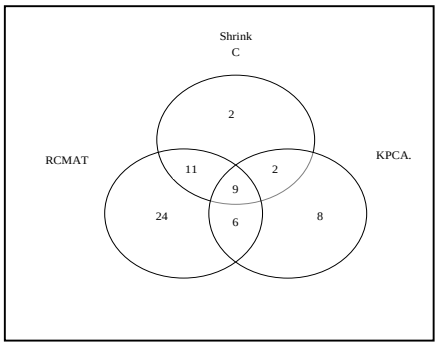


Figure 3 Significant gene sets detected by ShrinkC, RCMAT, KPCA for diabetes data.

5.0 CONCLUSION

We have not only introduced a novel approach, shrinkage covariance matrix to detect significantly altered gene sets, but also investigated the performance characteristics of a subset of commonly used approaches through the analysis of real microarray data set. In the original paper of [11], they detected OXPHOS as the significant gene set while only ShrinkA approach can detected the gene set but

more work need to be done to confirm these results. According to [13], the molecular processes which contribute to skeletal muscle insulin resistance are not fully understood. In their study, they redo the experiment with other diabetes samples and discovered no alteration in OXPHOS gene expression. Furthermore, in this study, we mainly focus on the analysis of diabetes, it is advisable to use our method as well as to other diseases like cancer. It is also interesting to apply our method to before- and after-treatment data to identify the significant gene sets.

References

[1] Alwine, J. C., Kemp, D. J., and Stark, G. R. 1992. Method for Detection of Specific RNA's in Agarose Gels by Transfer to Diaobenzyloxymenthyl-Paper and Hybridization With DNA Probes. *Proc Natl Acad Sci. 74*. 5350-5353.

[2] Liang, P., and Pardee, A. B. 1992. Differential Display of Eukaryotic Messenger RNA by Means of the Polymerase Chain Reaction. *Science. 257*. 967-971.

[3] Velculescu, V. L., Zhang, B., Vogelstein., and Kinzler, K. 1995. Serial Analysis of Gene Expression. *Science, 270*, 484–487

[4] Mary-Huard, T. Daudin, J. J., Robin, S., Bitton, F., Cabannes, E., and Hilson, P. 2004. Spotting Effect in Microarray Experiments. *BMC Bioinformatics, 5*, 63. Math. Statist. And Prob. 1. 361–379.

[5] Karjanto, S., Ramli, N. M., Aripin, R. and Ghani, N. A. M. 2014. Improved Statistical Test using Shrinkage Covariance Matrix For Identifying Differential Gene Sets. *Journal Of Applied Environmental And Biological Sciences*. 1. 302-310

[6] Schäfer J. and Strimmer K., A. 2005. Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics, *Statistical Applications In Genetics And Molecular Biology*. 4(1). 32.

[7] Ledoit O. and Wolf M. 2003. Improved Estimation of the Covariance Matrix of Stock Returns with An Application to Portfolio Selection, *Journal Of Empirical Finance*. 10.5. 603-621.

[8] Ledoit O. & Wolf M. 2004. Honey, I Shrunk the Sample Covariance Matrix. *The Jurnal Of Portfolio Management*. 31(1), 110-119.

[9] Ledoit O. and Wolf M. 2003. A Well-Conditioned Estimator for Large Dimensional Covariance Matrices, *Journal Of Multivariate Analysis*. 88. 365–411.

[10] Yates P. D. and Reimers M. A. 2009. RCMAT: A Regularized Covariance Matrix Approach to Testing Gene Sets, *BMC Bioinformatics.* 10. 300.

[11] Mootha V. K., Lindgren C. M., Eriksson K.F., Subramanian A., Sihag S., Lehar L., Puigserver P., Carlsson E., Ridderstraale M., Laurila E. *et al*. 2003. Pgc-1 A-Responsive Genes Involved in Oxidative Phosphorylation Are Coordinately Downregulated in Human Diabetes. *Nature Genetics* 34(3). 267–273.

[12] Kong S. W., Pu W. T. and Park P.J. 2006. A Multivariate Approach for Integrating Genome-Wide Expression Data and Biological Knowledge. *Bioinformatics*. 22(19), 2373-2380.

[13] Gallagher, I. J., Scheele, C., Keller, P., Nielsen, A. R., Remenyi, J., Fischer, C. P., *et al.* 2010. Integration of Microrna Changes in Vivo Identifies Novel Molecular Features of Muscle Insulin Resistance in Type 2 Diabetes. *Genome Med*. 2(2). 9-9
.