Comparison of Gene Set Analysis with Various Score Transformations to Test the Significance of Sets of Genes

Ivan Arreola and Dr. David Han

Department of Management of Science and Statistics, University of Texas at San Antonio, TX

ABSTRACT

Microarray analysis can help identify changes in gene expression which are characteristic to human diseases. Although genomewide RNA expression analysis has become a common tool in biomedical research, it still remains a major challenge to gain biological insight from such information. Gene Set Analysis (GSA) is an analytical method to understand the gene expression data and extract biological insight by focusing on sets of genes that share biological function, chromosomal regulation or location. This systematic mining of different gene-set collections could be useful for discovering potential interesting gene-sets for further investigation. Here, we seek to improve previously proposed GSA methods for detecting statistically significant gene sets via various score transformations.

***Keywords*:** Gene Expression, Gene Set Analysis, Gene Set Enrichment Analysis, Genomics, Micro-array Analysis

---

INTRODUCTION

Gene expression analysis, also known as pathway analysis, has become a pillar in genomics research; see Figure 1 below. Although the field has been around for more than a decade and is continually evolving, the problems still arise in identifying differentially expressed groups of genes from a set of microarray experiments. In the usual case, we have *N* genes measured on *n* microarrays under 2 distinct experimental conditions. Let n_1 and n_2 denote the sizes of microarray samples from the control and treatment groups, respectfully. Typically, *N* is large, say a few thousands while *n* is small, say a hundred or fewer [1]. The issue with this is multiple hypothesis testing, which is common in proteomics and genomics. Previously proposed methods compute a two-sample *t*-test score for each gene. Genes that have a *t*-statistic significantly larger than the pre-defined cutoff value are considered significant. The family-wise error rate (FWER) and false discovery rate (FDR) of the resulting genes are evaluated using the null distribution of the statistic.

A widely applied method called *Gene Set Enrichment Analysis* (GSEA), which is based on the signed version of Kolmogorov-Smirnov statistic, assesses the significance of predefined gene-sets, rather than individual genes. GSEA determines if members of a given gene-set are enriched using a normalized Kolmogorov-Smirnov statistic. A robust method known as *Gene Set Analysis* (GSA) proposes an alternative summary statistic for a given gene-set, called the *maxmean* statistic. It computes the average of positive (and negative) test scores for a given gene-set, and picks the larger of statistics in the absolute scale.

In studying GSEA and GSA, we found shortcomings and proposes a new way they could be improved. In our proposed methods GSA.p, we operate under the framework of GSEA and GSA to create a new summary statistic. We take the *mean* and

the *maxmean* of GSA and raise the test statistics to the power p. This increases (or decreases) the magnitudes of test scores of GSA to improve sensitivity of picking up significant gene-sets. In addition to raising test statistics to the power p, we also suggest an exponentiated version of the test statistics in order to transform the test scores of each gene and amplify the difference between two or more groups of the expression samples. Here, we provide the theoretical framework that allows us to gain biological insight in gene-set inference.

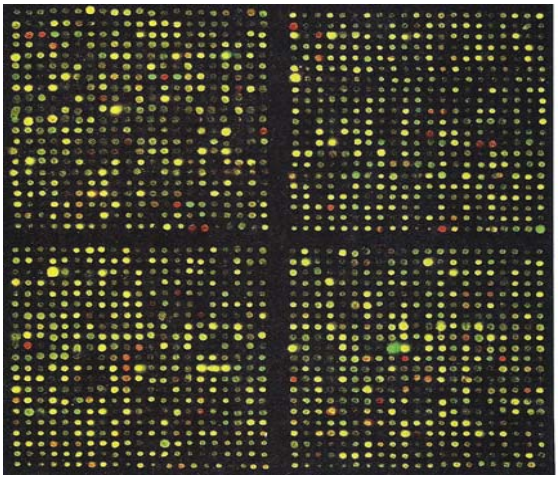

Figure 1. A sample image of a microarray experiment result; Green and red spots show differences in gene expression between two samples. Yellow spots show similar expression in both samples [5].

STATISTICAL METHODS

Overview of GSEA GSEA determines statistically if members of a gene-set are enriched from differentially expressed genes between two classes. First, gene expressions are ordered using signal-to-noise ratio (SNR) difference metric. The SNR is the difference of means of two classes, divided by the sum of standard deviations of the two diagnostic classes [2]. Then, for each gene-set, an enrichment measure, also known as *Enrichment Score* (ES), is calculated, which is the normalized Kolmogorov-Smirnov statistic. Let us consider ordered gene expressions $R_1, R_2, \dots, R_n$ based on the difference metric between two diagnostic classes and a gene-set S comprised of G members. Let i be the gene index and j be the sample index. Then,

$$X_i = \sqrt{\frac{G}{N-G}}$$

if R_i is not a member of S, or

$$X_i = \sqrt{\frac{N-G}{G}}$$

if R_i is a member of S.

Then, a ranking sum across all N genes is computed. We define ES to be

$$\max_{1 \le j \le N} \sum_{i=1}^{j} X_i$$

This is also known as the maximum observed deviation of the running sum, and it records the maximum enrichment score (MES). The significance of MES is computed by a permutation test of diagnostic labels from individuals. For example, consider a case where an individual is diagnosed with DM2 or NGT. DM2 is type 2 diabetes mellitus and it is a key contributor to atherosclerotic vascular disease, blindness, kidney failure, and amputation [2] while NGT stands for normal glucose tolerance. The null hypothesis is that no gene-set is associated with class distinction and the alternative is that the gene-set is associated with class distinction. To assess if a gene-set shows association with different phenotype class distinctions, class labels are permuted 1000 times and each time MES is recorded over all gene-sets. Permutation testing involves randomization of diagnostic labels and is a

dependent test on the primary diagnostic status of affected individuals.

Overview of GSEA.abs This version of GSEA also determines the significance of predefined gene-sets instead of individual genes by focusing on gene-sets, which are derived from groups of genes that share similar biological functions, chromosomal locations or regulations [3]. Similar to GSEA, GSEA.abs also considers gene expression profiles from samples that belong to two distinct classes. Then, genes are ranked based on their correlation between their expression and class distinctions by using any appropriate difference metric. To obtain the *Enrichment Score* (ES), let N represent the number of genes, k the number of samples, the exponent p for controlling the weight of each step along with a gene-set S. Then, we have the following:

$$P_{hit}(S,i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|r_j|^p}{N_R}$$

where $N_R = \sum_{g_j \in S} |r_j|^p$ and

$$P_{miss}(S,i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{1}{(N - N_H)}$$

We evaluate the genes in S (hit) weighted by their correlation and genes not in S (misses) given position i in L.

To determine if the ES of a gene-set S is significant, first we randomly assign phenotype labels and samples, reorder genes and recalculate ES. Next, we repeat the first step 1,000 times, and create a histogram of ES_{Null}. Finally, a nominal p-value of gene-set S from ES_{Null} is calculated by using the positive and negative portions of the distribution corresponding to the sign of observed ES [3]. This method does many permutations of the sample labels and recomputes the test statistic for each permuted dataset. From this information, we can compute the False Discovery Rate (FDR) of the list of significant gene-sets. Roughly speaking, FDR is equivalent to the Type-I error rate. In our situation, the FDR represents the proportion of non-significant genes-sets that were incorrectly found to be significant.

Overview of GSA In GSA, three summary statistics are calculated to determine the significance of a gene-set [1]. Given a gene expression data matrix X consisting of N genes in rows and n samples in columns, separated into two classes, n_1 control and n_2 treatment, a two-sample t-test statistic is computed for each gene in X, comparing the two classes. For convenience, let us transform the t-score t_i into the z-score z_i for the i^{th} gene in X. This is done by applying the cumulative distribution function (CDF) of the t-distribution to the t-score and then applying the quantile function of the standard normal distribution. Theoretically, we now have the following:

$$z_i \sim Normal(0,1) \quad \text{under } H_0$$

Let $\boldsymbol{z}_{\mathcal{S}} = (z_1, z_2, \dots, z_m)$ represent the set of m gene z-values in the gene-set $\mathcal{S}$ and define the gene-set enrichment test statistic to be

$$S = S(\boldsymbol{z}_{\mathcal{S}})$$

A large value of S indicates greater enrichment. For instance, applying a selected transformation function $s()$ to the individual z-scores, we have $s_i = s(z_i)$ and the gene-set score S can be defined as the average of s_i in $\mathcal{S}$ so that

$$S = \sum_{\mathcal{S}} \frac{s(z_i)}{m}$$

Efficient testing requires specification of the alternatives to the null selection [1]. The Poisson selection model starts with independent Poisson indicators given by

$$I_i \sim Poisson(v_i) \quad \text{where} \quad v_i = \alpha e^{\beta' s_i}/T_\beta$$

for $i = 1,2,\dots,N$. The effective choice of $S = S(\boldsymbol{z}_\mathcal{s})$ depends on the individual scoring function $s_i = s(z_i)$. Consider the following two cases.

$$s^{(1)}(z) = z \quad \text{and} \quad s^{(2)}(z) = |z|$$

$s^{(1)}$ being *mean* has power against *shift* (*location*) in $z$values while $s^{(2)}$ being *absmean* (absolute value of the mean) has power against *scale* alternatives. A two-dimensional scoring function is also suggested as follows.

$$s(z) = \left(s^{(+)}(z),\ s^{(-)}(z)\right), \begin{cases} s^{(+)}(z) = \max(z, 0) \\ s^{(-)}(z) = -\min(z, 0) \end{cases}$$

and the *maxmean* statistic is defined to be

$$S_{\max} = \max(s_\mathcal{s}^{(+)}, s_\mathcal{s}^{(-)})$$

$S_{\max}$ is able to detect large z-values in either or both directions of departure. In essence, we have the following summary statistics for GSA.

$$GSA \begin{cases} S = \sum_1^m \frac{z_i}{m} & (mean) \\ S = \sum_1^m \frac{|z_i|}{m} & (absmean) \\ S = \left\{ \left|\sum_1^m \frac{z_i^+}{m}\right|, \left|\sum_1^m \frac{z_i^-}{m}\right| \right\} & (maxmean) \end{cases}$$

Overview of GSA.p Similar to GSA, our proposed methods take the *mean* and *maxmean* summary statistics and raise the test scores to the power p. In addition to raising test statistics to the power p, we also suggest an exponentiated version of the test statistics in order to transform the test scores of each gene and amplify the difference between two classes. This increases (or decreases) the magnitudes of test scores to improve sensitivity of picking up significant gene-sets.

To define the methodology, let us consider a gene expression data matrix X consisting of N genes in rows and n samples in columns, separated into two classes, n_1 control and n_2 treatment. A two-sample *t*-test statistic is computed for each gene in X, comparing the two classes. Again, for convenience, let us transform the *t*-score t_i into the *z*-score z_i for the i^{th} gene in *X* so that

$$z_i \sim Normal(0,1) \quad \text{under } H_0$$

Let $\boldsymbol{z}_\mathcal{s} = (z_1, z_2, \dots, z_m)$ represent the set of m gene z-values in the gene-set $\mathcal{s}$ and define the gene-set enrichment test statistic to be

$$S = S(\boldsymbol{z}_\mathcal{s})$$

Applying a selected transformation function s() to the individual z-scores, we have $s_i = s(z_i)$ and the gene-set score S can be defined as the average of s_i in $\mathcal{s}$ so that

$$S = \sum_\mathcal{s} \frac{s(z_i)}{m}$$

This entails us to specify the alternative to the null distribution. The Bernoulli selection model starts with independent selection indicators given by

$$I_i \sim Bernoulli(p_i) \quad \text{with}$$

$$logit(p_i) = \alpha + \beta s_i$$

for $i = 1,2, \ldots, N$. Fundamentally, this binary indicator I_i randomly assigns the i^{th} gene to the gene-set $\mathcal{s}$ when $I_i = 1$ with the selection probability specified by p_i. Using a logistic regression model, the value of s_i influences the selection probability as desired. Under this framework, the gene-set $\mathcal{s}$ can be represented by

$$\mathcal{s} = \{i: I_i = 1\}$$

with the number of selected genes specified by

$$m = \sum_1^N I_i$$

Then, the effective choice of $S = S(\boldsymbol{z}_{\mathcal{s}})$ depends on the individual scoring function $s_i = s(z_i)$. Let us consider the following two cases as before.

$$s^{(1)}(z) = z \quad \text{and} \quad s^{(2)}(z) = |z|$$

$s^{(1)}$ being *mean* has power against *shift* (*location*) in $z$values while $s^{(2)}$ being *absmean* (absolute value of the mean) has power against *scale* alternatives. A two-dimensional scoring function is also suggested as follows.

$$s(z) = \left(s^{(+)}(z),\ s^{(-)}(z)\right), \begin{cases} s^{(+)}(z) = \max(z, 0) \\ s^{(-)}(z) = -\min(z, 0) \end{cases}$$

and the *maxmean* statistic is defined to be

$$S_{\max} = \max(s_{\mathcal{s}}^{(+)}, s_{\mathcal{s}}^{(-)})$$

$S_{\max}$ is able to detect large z-values in either or both directions of departure. Our newly proposed summary statistics GSA.p are then

$$\boldsymbol{GSA.p} \begin{cases} S = \sum_1^m \dfrac{z_i^p}{m} & (mean.p) \\ S = \left\{\left|\sum_1^m \dfrac{z_i^{+p}}{m}\right|, \left|\sum_1^m \dfrac{z_i^{-p}}{m}\right|\right\} & (maxmean.p) \end{cases}$$

along with an exponentially transformed version given by

$$\boldsymbol{GSA.p.ett} \begin{cases} S = \sum_1^m \dfrac{sign(z_i^p) * e^{|z_i^p|-1}}{m} & (mean.p.ett) \\ S = \left\{\left|\sum_1^m \dfrac{sign(z_i^{+p}) * e^{|z_i^{+p}|-1}}{m}\right|, \left|\sum_1^m \dfrac{sign(z_i^{-p}) * e^{|z_i^{-p}|-1}}{m}\right|\right\} & (maxmean.p.ett) \end{cases}$$

The exponential transformation of the test scores of each gene is designed to amplify the significant difference between two or more groups of the given expression samples. In the next section, we study the performance of the newly proposed methods in comparison to the previously proposed methods through a simulation study. It was found that in some cases, the newly proposed methods are competitively better than the conventional methods in detecting significant gene-sets.

SIMULATION STUDY

We simulated 1000 gene expression values for 50 samples in each of two classes, control and treatment. Additionally, 50 gene-sets were also generated, each containing 20 genes. All measurements were standard normal random variates before the treatment effect was added under 5 different scenarios.

(1) All 20 genes of gene-set 1 are .2 units higher in class 2.
(2) The first 15 genes of gene-set 1 are .3 units higher in class 2.
(3) The first 10 genes of gene-set 1 are .4 units higher in class 2.
(4) The first 5 genes of gene-set 1 are .6 units higher in class 2.
(5) The first 10 genes of gene-set 1 are .4 units higher in class 2, and the second 10 genes of gene-set 1 are .4 units lower in class 2.

Table 1

The results of average p-values from the simulation study under 5 different scenarios using 200 permutations and 20 repetitions

	mean	absmean	maxmean	GSEA	GSEA.abs	mean.p (p = 2)	maxmean.p (p = 2)	mean.p (p = 3)	maxmean.p (p = 3)
(1)									
mean	.0028	.0590	.0008	.0320	.1920	.0380	.0020	.0113	.0058
sd	.0094	.0790	.0024	.0170	.0600	.0550	.0041	.0165	.0098
(2)									
mean	.0008	.0085	.0005	.0160	.0740	.0010	.0008	.0005	.0008
sd	.0024	.0182	.0022	.0080	.0340	.0021	.0018	.0015	.0024
(3)									
mean	.0005	.0055	.0008	.0310	.0570	.0005	.0263	.0023	.0025
sd	.0015	.0119	.0034	.0180	.0320	.0022	.1174	.0057	.0111
(4)									
mean	.0013	.0045	.0015	.0690	.0370	.0043	.0290	.0010	.0008
sd	.0036	.0089	.0024	.0380	.0140	.0190	.1297	.0045	.0034
(5)									
mean	.0178	.0000	.0003	.2330	.0110	.0000	.0005	.0623	.0000
sd	.1490	.0000	.0011	.0630	.0090	.0000	.0022	.0938	.0000

Note: The mean and standard deviation for GSEA and GSEA.abs were obtained from [1].
The results are based on 20 repeated simulations.

In each scenario, only the first gene-set was of interest. The results of the average *p*-values based on various summary statistics are tabulated in Table 1 above under 5 different scenarios using 200 permutations and 20 repetitions. The method that has consistently low *p*-values across all 5 different scenarios is considered the best. While *maxmean* is found to be such, our proposed method GSA.p in some scenarios is competitively better than GSA, GSEA, and GSEA.abs. In particular, under scenario (2), GSA.p has lower *p*-values than other previously proposed methods. The lower the *p*-value is, the more sensitive the method is in detecting significant gene-sets.

APPLICATION TO P53 DATA

p53 is a tumor protein and its gene codes for a protein that regulates the cell cycle and functions as a tumor suppressor. In principle, it is a cancer suppressor. The p53 signaling pathway activation is prompted by numerous cellular stress signals such as DNA damage, oxidative stress, and activated oncogenes [4]. For example, in normal cells, p53 protein level is low and stress signals may trigger the increase of p53 protein. Therefore, if a person inherits only one functional copy of the p53 gene, then that person is predisposed to cancer and will likely develop a variety of independent tumors. Here, the p53 protein is employed as a transcriptional activator of p53-regulated genes. This in turn gives three major outputs: cell cycle arrest, cellular senescence or apoptosis.

The p53 data containing the catalog of 522 gene-sets was obtained from [2], and Tables 2 to 5 below provide the lists of significant gene-sets found from p53 data by applying the methods mean.p and maxmean.p with FDR cutoff .10 and 200 permutations.

Table 2

The results of mean.p with p = 2; 25 significant gene-sets found from p53 data with FDR cutoff .10 and 200 permutations

1. p53 pathway *
2. p53 hypoxia *
3. hsp27 pathway *
4. p53 UP *
5. SA G1 and S phases *
6. radiation sensitivity *
7. MAP000251
8. rap down
9. glut down
10. atm pathway
11. bad pathway
12. bcl2family
13. CA NF at signaling
14. cell cycle regulator
15. ceramide pathway
16. DNA damage signal
17. drug resistance
18. G1 pathway
19. G2 pathway
20. P53 signaling
21. raccyc pathway
22. insulin signaling
23. SA TRKA receptor
24. calcineurin pathway
25. mitochondria pathway

* demonstrates significant gene-sets in [3].

Table 3

The results of mean.p with p = 3; 30 significant gene-sets found from p53 data with FDR cutoff .10 and 200 permutations

1. hsp27 pathway *
2. p53 signaling *
3. p53 hypoxia *
4. radiation sensitivity *
5. SA G1 and S phases *
6. p53 UP *
7. ccr3 pathway
8. atm pathway
9. bad pathway
10. bcl2 family
11. CA NF at signaling
12. calcineurin pathway
13. cell cycle arrest
14. cell cycle regulator
15. cell cycle pathway
16. ceramide pathway
17. chemical pathway
18. CR death
19. DNA damage signaling
20. drug resistance
21. G1 pathway
22. G2 pathway
23. mitochondria pathway
24. p53 pathway
25. raccycd pathway
26. SA TRKA receptor
27. SIG IL4 receptor
28. ST Fas signaling
29. breast cancer strong
30. pml pathway

* demonstrates significant gene-sets in [3].

Table 4

The results of maxmean.p with p = 2; 5 significant gene-sets found from p53 data with FDR cutoff .10 and 200 permutations

1. p53 hypoxia *
2. p53 pathway *
3. radiation sensitivity *
4. SA G1 and S phases *
5. p53 UP *

* demonstrates significant gene-sets in [3].

Table 5

The results of maxmean.p with p = 3; 10 significant gene-sets found from p53 data with FDR cutoff .10 and 200 permutations

1. fmlp pathway *
2. p53 hypoxia *
3. p53 pathway *
4. radiation sensitivity *
5. SA GA and S phases *
6. p53 UP
7. ccr3 pathway
8. atm pathway
9. ceramide pathway
10. p53 signaling

* demonstrates significant gene-sets in [3].

Table 6

The results of maxmean.p with p = 2 and exponential transformation; 5 significant gene-sets found from p53 data with FDR cutoff .10 and 200 permutations

1. radiation sensitivity *
2. p53 pathway *
3. cell cycle regulator
4. bad pathway

5. SA TRKA receptor

* demonstrates significant gene-sets in [3].

Table 7

The results of maxmean.p with p = 3 and exponential transformation; 5 significant gene-sets found from p53 data with FDR cutoff .10 and 200 permutations

1. radiation sensitivity *
2. p53 pathway *
3. p53 hypoxia
4. RAP UP
5. SA TRKA receptor

* demonstrates significant gene-sets in [3].

Tables 6 and 7 above provide the lists of significant gene-sets found from p53 data by applying the method maxmean.p with exponential transformation, again with FDR cutoff .10 and 200 permutations. The significant gene-sets detected by our proposed methods are indeed in agreement with the gene-sets detected by conventional GSEA, along with new gene-sets which were not discovered before. This demonstrates the stronger sensitivity of our proposed methods compared to the previously utilized methods.

SUMMARY & FUTURE STUDY

The proposed methods discover statistically significant gene-sets in microarray analysis. Through our foundation, new transformation functions and summary statistics are currently being explored to improve the sensitivity of uncovering significant gene-sets. The new approaches will be applied to various datasets including the *Molecular Signature Databases* to test their efficacy.

REFERENCES

[1] Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics*, **1:** 107–129.

[2] Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, **34:** 267–273.

[3] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpret-ing genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102:** 15545–15550.

[4] National Center for Biotechnology Information (US). The p53 tumor suppressor protein. *Genes and Disease*. https://www.ncbi.nlm.nih.gov/books/NBK22268/

[5] Scientist can study an organism's entire genome with microarray analysis. *Scitable by Nature Education*. https://www.nature.com/scitable/topicpage/scientists-can-study-an-organism-s-entire-6526266