GSEMT: A Gene Set Enrichment Analysis Method Based on Mantel Test

Na Yu*
School of Control Science and Engineering, Shandong University, Jinan 250061, China
*Email: yunayn@126.com

Abstract. Gene expression changes constantly with the occurrence and progression of diseases. The vast available gene expression data makes it possible for clinical researchers to understand the link between genotypes and phenotypes. However, it is still not an easy task because the information contained in the gene expression matrix is sparse. Gene set enrichment analysis is a powerful tool to meet the challenge of identifying complicated differential information underlying pathways. In this paper, we propose a method, called GSEMT, for gene set enrichment analysis by testing the correlation between a sample similarity matrix and a phenotype dissimilarity matrix. We implement experiments on knowledge-based gene sets and gene expression datasets for hepatocellular carcinoma. We justify the effectiveness and advantage of GSEMT by carrying out comparison studies. GSEMT outperforms GSEA and GSNCA in the classification performance on an experiment dataset and an independent validation dataset. The results show GSEMT is a useful alternative for gene set enrichment analysis.

1. Introduction
The development of next generation sequencing technique leads to the vast volume of gene expression data available to researchers in diverse fields, which triggers the boom of genomic analysis [1]. Gene expression data contains valuable information closely related to the occurrence and progression of disease, which provides great resources for clinical research and drug development. However, it is challenging to parse such big volume data and mine useful information. To understand which genes play significant roles in disease, researchers usually carry out differential expression analysis of hundreds of and thousands of genes between different phenotypes, e.g., control and disease. While in traditional differential expression analysis, genes are often evaluated in an isolated manner. In fact, they perform functions cooperatively and play in the form of pathway. To address this issue, gene set enrichment analysis methods have been built up to meet the requirement. By using gene set enrichment analysis method, we can obtain understandable information about the relevance between genotypes and phenotypes of interest. The statistical significance underlying individual gene sets indicates their functional implications to certain phenotypic states. Hepatocellular carcinoma (HCC) is one of the most common human malignancies and it is the third leading cause of cancer death worldwide [2]. Although liver transplantation and chemotherapy are valid in treating HCC, the 5-year overall survival rate is still only 7% [3]. Due to the condition that most HCCs cannot be diagnosed at their early stage, HCC patients’ recovery rate is quite low. It is generally recognized that HCC patients’ recovery rate can be improved by effective and accurate early diagnosis. Finding reliable biomarkers can be of great expectation to achieve this goal of early diagnosis for HCC [4].

[Creative Commons Attribution 3.0 licence]
In this paper, we propose a novel method for gene set enrichment analysis, called GSEMT, by calculating the correlation between the sample similarity matrix and the phenotype dissimilarity matrix. Specifically, we firstly define the correlation between a gene expression matrix (genotype) and a disease state indicators (phenotype). By employing a similar empirical process in Mantel test [5], we evaluate the statistical significance for the former correlation. We perform our experiments on knowledge-based gene sets and a public gene expression dataset for HCC. The results demonstrate the effectiveness of our proposed strategy. To show the advantage of GSEMT, we did the comparison studies with some existing methods, e.g., GSEA [6] and GSNCA [7], by evaluating the ability of distinguishing control samples from tumor ones of HCC via the top-ranked significant gene sets. GSEMT achieved better classification performance than GSEA and GSNCA.

2. Materials and Method

2.1 Gene Sets and Gene Expression Matrices

In this paper, we used two gene expression profiling datasets for HCC and a knowledge-based gene set dataset from Biocarta [8] to illustrate the effectiveness of GSEMT. The first microarray data of 243 non-tumor and HCC tissues were downloaded from GEO database with accession number GSE25097 [9]. TCGA HCC dataset was downloaded from https://xena.ucsc.edu/GDC TCGA Cancer (LIHC) [10]. It includes 50 control samples from tumor adjacent tissues and 374 tumor samples. We extracted 50 paired non-tumor and HCC samples from them and built up a new dataset named TCGApair. We used TCGApair as an independent dataset to validate the effectiveness of GSEMT in gene set enrichment between non-tumor and HCC conditions. And 289 gene set of Biocarta was obtained from MSigDB [6] curated gene sets.

2.2 Method

GSEMT mainly consists of three steps. As is shown in figure 1, we use a gene expression matrix and one gene set in a pathway as the input. GSEMT generates a new expression matrix containing common genes shared by the input gene expression dataset and the gene set included in individual pathways. Then GSEMT creates two matrices from this new expression matrix. The two matrices are a sample-sample gene expression similarity (GES) matrix and a phenotype-phenotype dissimilarity (PPD) matrix whose elements are 1 and -1, respectively. Inspired by Mantel test [5], a method which can be used to calculate correlation between any two matrices, GSEMT carries out Mantel test on the GES.
matrix and PPD matrix. The null hypothesis we test in GSEMT is there is no correlation between matrices GES and PPD. The statistic of GSEMT is defined as expression (1). GSEMT executes N times random permutation of PPD matrix $P$ and calculate N new test statistic values. N can be set to any value that satisfies the research need. In this paper, we set $N = 10,000$. Then $p$-value can be calculated as equation (2). If the Benjamini-Hochberg adjusted $p$ value [11] is less than 0.05, we can determine the differential expression between two phenotypes in this gene set is statistically significant.

$$\text{Statistic of GSEMT: } CGEP = \sum_{u=1}^{s(2m-1)} s_u^* d_u^*$$

(1)

where $s$ represents upper triangle elements of GES, $d$ represents upper triangle elements of PPD. And $m$ is the number of control or disease samples. The statistic CGEP is the correlation of GES and PPD for each pair of samples.

$$p - value = \frac{\text{count}(CGEP \geq CGEP_0)}{N + 1}$$

(2)

where CGEP represents statistic value generated in permutation process. $CGEP_0$ is the initial statistic.

To show the advantage of GSEMT, we compare the enrichment results with those obtained by existing methods such as GSEA [6] and GSNCA [7]. We use fgsea [12] and GSAR packages for them respectively. All computations were implemented using R x64 3.6.2.

3. Results

3.1 Rank of Gene Sets

| Rank | GSEMT | GSEA | GSNCA |
|------|-------|------|-------|
|      | Statistics | p.adj | AUC | Statistics | p.adj | AUC | Statistics | p.adj | AUC |
| Top1 | 0.409 | 0.0001 | 0.990 ± 0.009 | 2.190 | 0.004 | 0.937 ± 0.039 | 0.495 | 0.0002 | 0.965 ± 0.019 |
| Top2 | 0.330 | 0.0001 | 0.989 ± 0.013 | 2.167 | 0.004 | 0.956 ± 0.016 | 0.458 | 0.0002 | 0.971 ± 0.021 |
| Top3 | 0.305 | 0.0001 | 0.984 ± 0.021 | 2.116 | 0.005 | 0.960 ± 0.027 | 0.443 | 0.0002 | 0.972 ± 0.014 |
| Top4 | 0.304 | 0.0001 | 0.984 ± 0.010 | 1.970 | 0.005 | 0.958 ± 0.038 | 0.441 | 0.0002 | 0.966 ± 0.031 |
| Top5 | 0.302 | 0.0001 | 0.985 ± 0.014 | 1.868 | 0.013 | 0.963 ± 0.013 | 0.423 | 0.0004 | 0.951 ± 0.031 |
| Mean ROC | 0.986 ± 0.002 | 0.965 ± 0.009 | 0.954 ± 0.007 |

Table 1. Top 5 gene sets and their performance in classification.

Firstly, we implemented GSEMT with gene expression data GSE25097 and gene sets documented in BioCarta. For comparison study, we carried out the experiments on GSEA and GSNCA similarly. We obtained three different outputs of ranks of these gene sets for each method. In each method, the top 5 ge
ne sets are shown in table 1. As shown, we ranked the gene sets by the value of each method’s statistic because the statistic characterizes the decision-making information for the difference between normal and tumor conditions. The statistic of GSEMT represents the correlation of GES matrix and PPD matrix. The statistic of GSEA represents whether genes in a specific gene set are enriched in the top or bottom of the list of differential genes in HCC. The statistic of GSNCA also represents the difference between normal and tumor conditions. A Benjamini-Hochberg adjusted p-value less than 0.05 is considered as statistically significant in our comparison study. From table 1, we also can get the metrics about the classification performance by each top-ranked gene set in the comparing method. For each method, the mean classification performance of the top 5 gene sets is also presented in table 1. The top 5 gene sets detected by GSEMT are "BIOCARTA_AKAP95_PATHWAY", "BIOCARTA_MPR_PATHWAY", "BIOCARTA_CDK5_PATHWAY", "BIOCARTA_STATHMIN_PATHWAY", and "BIOCARTA_AKAPCENTROSOME_PATHWAY". For GSEA, they are "BIOCARTA_MEMC_PATHWAY", "BIOCARTA_CELL_CYCLE_PATHWAY", "BIOCARTA_G2_PATHWAY", "BIOCARTA_ATRBRCA_PATHWAY" and "BIOCARTA_EFP_PATHWAY". And the top 5 gene sets for GSNCA are "BIOCARTA_G2_PATHWAY", "BIOCARTA_G1_PATHWAY", "BIOCARTA_EFP_PATHWAY", "BIOCARTA_ATRBRCA_PATHWAY" and "BIOCARTA_FREE_PATHWAY". "BIOCARTA_G2_PATHWAY", "BIOCARTA_ATRBRCA_PATHWAY" and "BIOCARTA_EFP_PATHWAY" are common to p gene sets detected by GSNCA and GSEA. Gene sets detected by GSEMT were unique.

3.2 Classification Power of Selected Top Gene Sets

Then we did SVM-RFE [13] using R packages caret and kernlab for each top 5 gene set in the three different methods. For each gene set, we built up an SVM classifier based on these selected genes in the top 5 gene sets [14]. We tested the classification performance of each SVM classifier by 10-fold cross validation on the GEO GSE25097 dataset.

![Figure 2](image_url)

**Figure 2.** Mean ROC curves with standard deviations for GSEMT, GSEA and GSNCA. (a)-(d) refer to the results of GSE25097; (e)-(h) refer to the results of TCGA pair.

As illustrated in figure 2 (a) to (d), the top 5 gene sets detected by GSEMT achieved a mean AUC of 0.986 with standard deviation 0.002. Comparatively, the top 5 gene sets selected by GSEA achieved a mean AUC of 0.965 with standard deviation 0.009. And for GSNCA, the mean AUC of top 5 gene sets is 0.954 with standard deviation 0.007. Then we validated the classification performance on the ind
ependent dataset TCGApa. The results are shown in figure 2 (e) to (h). GSEMT again achieved a better classification performance than GSEA and GSNCA.

3.3 Functional Enrichment Analysis

To identify the enriched functional implications, we did GO analysis for each gene set from the top 5 gene sets selected by GSEMT, GSEA and GSNCA using g:Profiler [15]. The results are shown in table 2. Top 10 statistically significant GO terms for the top 5 gene sets in the three methods are listed simultaneously.

Table 2. Functional enrichment results for the top 5 gene sets in different methods.

| GSEMT                  | GSEA                       | GSNCA                      |
|------------------------|-----------------------------|----------------------------|
| GOterm_name            | -lg(p.adj)                  | GOterm_name                | -lg(p.adj) | GOterm_name                | -lg(p.adj) |
| Activation of protein kinase A activity | 13.270                      | DNA replication initiation | 32.497     | DNA repair                  | 23.917     |
| Cellular response to glucagon stimulus | 11.827                      | G1/S transition of mitotic cell cycle | 25.880 | Cellular response to DNA damage stimulus | 20.443     |
| Response to glucagon   | 10.766                      | Cell cycle G1/S phase transition | 25.299 | DNA metabolic process       | 19.798     |
| Renal water homeostasis | 10.592                      | DNA repair                  | 23.917     | Signal transduction involved in DNA damage checkpoint | 18.944 |
| Positive regulation of transferase activity | 9.949                      | DNA-dependent DNA replication | 23.814 | Signal transduction involved in DNA integrity checkpoint | 18.944 |
| Positive regulation of protein kinase activity | 9.699                      | Regulation of cyclin-dependent protein serine/threonine kinase activity | 21.001 | Signal transduction involved in cell cycle checkpoint | 18.874 |
| Regulation of transferase activity | 9.289                      | Regulation of cyclin-dependent protein kinase activity | 20.782 | Signal transduction in response to DNA damage | 18.287 |
| Regulation of protein kinase activity | 9.139                      | Mitotic cell cycle phase transition | 20.488 | Cell cycle checkpoint       | 17.896     |
| Positive regulation of kinase activity | 9.025                      | Cellular response to DNA damage stimulus | 20.443 | DNA damage checkpoint       | 17.678     |
| Activation of protein kinase activity | 8.644                      | Regulation of G1/S transition of mitotic cell cycle | 20.324 | DNA integrity checkpoint   | 17.330     |

As illustrated in table 2, the top 10 enriched GO functions of GSEMT are focused on PKA (protein kinase activity) [16], glucagon [17] and transferase [18]. These three biological processes are reported to be closely related with HCC. For GSEA, the top 10 statistically significant GO terms are mainly about DNA biological activity and cell cycle, which are common features shared by numerous tumor processes. Among the enriched GO functions of GSEA, cyclin-dependent protein kinase activity [16] is c
loosely related with HCC. For GSNCA, the enriched GO terms are about DNA biological activity and cell cycle, which are not specific for HCC because the changes of DNA biological activity and cell cycle are rather general.

4. Conclusion

In this paper, we propose a method called GSEMT to do gene set enrichment analysis. GSEMT calculates correlation between sample similarity matrix and phenotype dissimilarity matrix for each pair of samples as its test statistic. Based on Mantel test, GSEMT gets a p-value of statistical significant by empirically N times permutation test.

GSEMT outperforms GSEA and GSNCA in the classification of normal and tumor samples on two HCC datasets. The enriched GO functions show that the top gene sets detected by GSEMT have more specific relations with HCC than these gene sets detected by the other methods, e.g. GSEA and GSNCA. In addition, there is no need to do a differential expression analysis and generate a pre-ranked gene list in advance for GSEMT, which is much more convenient compared with GSEA. And GSEMT is applicable to datasets of any distribution because it uses a nonparametric test. Furthermore, GSEMT is more stable than GSNCA due to its inclusion of gene expression level across samples. GSEMT is expected to be a powerful tool to do gene set analysis and mine useful information from gene expression data. By using GSEMT, Researchers will gain valuable insights on the relationship between gene expression and disease progression.

Acknowledgments
Thank all lab colleagues for their kind encouragement and help. This work was partially supported by the National Natural Science Foundation of China (Grant Nos 61572287 and 61533011); Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project, 2019JZZY010423).

References

[1] Mardis ER: The impact of next-generation sequencing technology on genetics. Trends in genetics 2008, 24(3):133-141.
[2] Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A: Global cancer statistics, 2012. CA: a cancer journal for clinicians 2015, 65(2):87-108.
[3] Ilikhan SU, Bilici M, Sahin H, Akca ASD, Can M, Oz II, Guven B, Buyukuysal MC, Ustundag Y: Assessment of the correlation between serum prolidase and alphafetoprotein levels in patients with hepatocellular carcinoma. World journal of gastroenterology : WJG 2015, 21(22):6999-7007.
[4] Luo P, Wu S, Yu Y, Ming X, Li S, Zuo X, Tu J: Current Status and Perspective Biomarkers in AFP Negative HCC: Towards Screening for and Diagnosing Hepatocellular Carcinoma at an Earlier Stage. Pathology oncology research 2019, 26(2):599-603.
[5] Mantel N, Mantel N: The Detection of Disease Clustering and a Generalized Regression Approach. Cancer research (Chicago, Ill) 1967, 27(2):209-220.
[6] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES et al: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences - PNAS 2005, 102(43):15545-15550.
[7] Rahmatallah Y, Emmert-Streib F, Glazko G: Gene Sets Net Correlations Analysis (GSNCA): A multivariate differential coexpression test for gene sets. Bioinformatics 2014, 30(3):360-368.
[8] Nishimura D: BioCarta. Biotech Software & Internet Report 2001, 2(3):117-120.
[9] Lamb JR, Zhang C, Xie T, Wang K, Zhang B, Hao K, Chudin E, Fraser HB, Millstein J, Ferguson M et al: Predictive genes in adjacent normal tissue are preferentially altered by sCNV during tumorigenesis in liver cancer and may rate limiting. PloS one 2011, 6(7):e20090-e20090.

[10] Goldman MJ, Craft B, Hastie M, Repecka K, McDade F, Kamath A, Banerjee A, Luo Y, Rogers D, Brooks AN et al: Visualizing and interpreting cancer genomics data via the Xena platform. Nature biotechnology 2020, 38(6):675-678.

[11] Benjamini Y, Hochberg Y: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. 1995, 57(1):289-300.

[12] Sergushichev AA: An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. 2016:060012.

[13] Guyon I, Weston J, Barnhill S, Vapnik V: Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning 2002, 46(1):389-422.

[14] Chang C-C, Lin C-J: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2011, 2(3):1-27.

[15] Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, Vilo J: g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Research 2019, 47(W1):W191-W198.

[16] Massimi M, Ragusa F, Cardarelli S, Giorgi M: Targeting Cyclic AMP Signalling in Hepatocellular Carcinoma. Cells (Basel, Switzerland) 2019, 8(12):1511.

[17] Pun KK, Pun KK, Young RTT, Young RTT, Wang C, Wang C, Tam CF, Tam CF, Ho PWM, Ho PWM: The use of glucagon challenge tests in the diagnostic evaluation of hypoglycemia due to hepatoma and insulinoma. The journal of clinical endocrinology and metabolism 1988, 67(3):546-550.

[18] Sherman M, Campbell JAH, Titmuss SA, Kew MC, Kirsch RE: Glutathione S-transferase in human hepatocellular carcinoma. Hepatology (Baltimore, Md) 1983, 3(2):170-176.