Comprehensive Comparison of Gene Set Analysis Tools

Zheng Liu[1], **Xuejun Li**[2], **Yate-Ching Yuan**[1], **and Xiwei Wu***[1]
[1]Bioinformatics Core, Department of Molecular Medicine, Beckman Research Institute, City of Hope National Medical Center, 1500 Duarte Rd, Duarte, CA 91010, USA
[2]Division of Biostatistics, Department of Information Sciences, Beckman Research Institute, City of Hope National Medical Center, 1500 Duarte Rd, Duarte, CA 91010, USA
*To whom correspondence should be addressed.

Abstract - *Gene set analysis has enhanced the microarray data analysis field with biological insights. The first introduced and widely used Over-representation analysis (ORA) method, has the limitation of the requirement of a predetermined differentially expressed genes list. To overcome this limitation, distribution based analysis (DBA) methods were developed with different analysis steps and null hypothesis. To understand the advantages and limitations of these methods, we present a comprehensive survey and evaluate the performance for nine commonly used gene set analysis tools. Methods testing self-contained hypothesis generally have better sensitivity and specificity than methods testing competitive hypothesis. But most of the methods have bias towards larger gene sets with self-contained methods more severe. Therefore, better sensitivity and specificity is obtained at the tradeoff of bigger bias in self-contained methods, and vice versa in competitive methods. We propose a combined performance plot to compare these methods, among which GSA demonstrated superiority over others.*

Keywords: Pathway analysis, microarray, gene set analysis.

1 Introduction

In the last decade, microarray technology largely expedited the biological discovery in basic, clinical and translational research. Initially, the analysis of microarray data was focused on differential expression analysis, where a list of genes that show statistically significant expression difference between conditions can be identified. However, biologists still face difficulties in correlating the target genes with biological significance, e.g. identification of signaling pathways that were differentially activated or repressed is often more interesting than a list of gene names. A gene set contains multiple genes sharing similar biological properties, e.g. gene ontology terms, signaling pathway, and chromosome location. The advantage of analyzing genes as a set is that it can detect coordinate changes that are usually moderate or weak at single gene level. To achieve a biologically relevant interpretation, the target gene list is usually compared to a reference gene list, which is typically all the genes on the microarray, for enrichment of certain gene ontology terms or biological pathways. We refer this method as over representation analysis (ORA). Because of the arbitrary selection of cutoff at the gene list identification step, important findings might be missed and the results are not stable. A number of cut-off free gene set analysis methods, which provide statistical methods to analyze multiple genes, were introduced later on to prevent any arbitrary cutoff. These tools are often denoted as distribution based analysis (DBA).

Recently, Nam et al. [1] thoroughly summarized and classified 26 gene analysis tools based on their null hypothesis and statistical methods. But the advantages and limitations of these methods are not completely understood. Tian et al. [2] suggested that tests based on both null hypotheses should be considered equally. Goeman et al. [3] further classified gene set analysis methods into three categories, self-contained, competitive and mixed. Dinu et al. [4] recently compared three self-contained analysis tools, SAM-GS [5], global test [6] and ANCOVA global test, and concluded that SAM-GS has slightly higher power. But none of them has conducted thorough performance comparison. We evaluated these tools, and systemically compared their performance using statistical simulation.

2 Methods

2.1 Analysis Tools

In the current study, we have compared GSEA [7] (both gene permutation and phenotype permutation), Tian/sigPathway (both gene permutation and phenotype permutation), ErmineJ [8] ORA, ErmineJ GSR, GSA [9], SAM-GS, SAFE [10], global test and PAGE [11]. Within these tools, there are 4 tools (Global_Test, SAFE, SAME-GS, and Tian_Pheno) testing the self-contained hypothesis, 5 tools (ErmineJ_GSR, ErmineJ_ORA, GSEA_Gene, PAGE, and Tian_Gene) testing the competitive hypothesis, and 2 tools (GSA, GSEA_Pheno) are mixed.

2.2 Simulation Method

Given the diversity of methods implemented in each tool, it is very interesting to examine whether their performance is also different. We developed a testing framework to systemically compare the performance using statistical simulation. We collected 464 signaling and metabolic pathways from KEGG and BioCarta, which are two commonly used canonical pathway databases. For testing purpose, we created 50 pseudo-pathways, each consisting of 20 pseudo-genes, which are differentially expressed between conditions. The major reason to include real pathways is to

generate some false positives so that we can assess the performance of each tool.

The simulation data were generated as a 20,000 x 20 matrix (20,000 genes, 10 normal and 10 treated samples) that follows a standard normal distribution. Differentially expressed (DE) genes were simulated by adding a small constant to the 10 treated samples. The magnitude of increase and the number of DE genes were carefully selected to mimic different scenarios in real experiments, as addressed in more detail in section 3. To prevent any biased results due to any particular simulated data set, one hundred independent simulation data sets were created for each scenario. These simulated data sets were then analyzed by using different analysis tools. Default or recommended parameters of each tool were used whenever possible. Receiver Operating Characteristics (ROC) curve was used to assess the performance of the tools based on the gene set ranks produced by each analysis tools. The mean and standard deviation of the AUCs from 100 simulations were obtained to represent the performance of each tool.

3 Results

3.1 Effects of number of DE gene

To examine the performance of the tools under different levels of differential expression in the gene sets, we generated 10%, 20%, 30%, 40% and 50% DE genes in each of the 50 pseudo-pathways. We also wanted to simulate the phenomenon in real microarray experiment that not all the DE genes belong to any gene sets, which likely to introduce additional level of noise to the data. Therefore, besides the DE genes within the pseudo-pathways, additional DE genes were created in each simulation data set to fix the number of DE genes at 2000. To determine how big the constant should be used to create DE genes, we tested 0.5, 1, and 2.5. Changes with 1 and 2.5 were so strong that all the tested analysis tools were able to achieve an AUC of almost 1. Therefore, we decided to use 0.5 as the constant and all the subsequent results were generated using this constant.

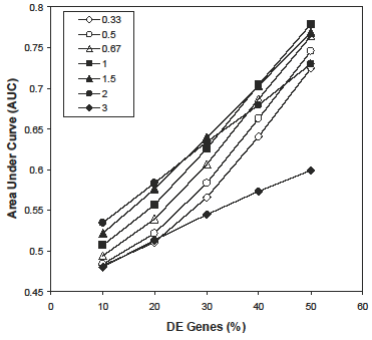

Figure 1. Performance of ORA method under different cutoffs

We found that the performance of ORA method was highly dependent on the selected cutoff. The performance of ORA method increases as the percentage of DE genes increases (Figure 1). Totally 5 cutoff values (0.33, 0.67, 1, 1.5, 2 and 3) were tested. The AUC values range from 0.55 with 10% DE genes to 0.85 with 50% DE genes with cutoff value of 1. More importantly, different cutoff values result in quite different performance. Cutoff of 1 has the largest AUC, followed by 0.67. Cutoff of 3 gives the lowest AUC, while cutoff of 0.33 and 2 resides in the middle. This result is expected because the theoretical t-statistics for DE genes in the simulated data set is close to 1.5.

3.1.1 Comparison of Gene Set Analysis Methods

To compare the ability to detect enriched gene sets for each analysis methods, 100 simulated data sets were generated and analyzed by each of the tools. To accurately estimate the false positive and false negative rate, gene sets reported as positively and negatively associated with the treatment phenotype were combined in all of the tools. The average AUCs across the 100 simulated data sets are shown in Figure 2 A-E. All the tools perform better with more DE genes in the gene sets. Global_Test and SAM_GS perform the best when the percentage of DE is low, and Sigpath_pheno and GSA perform the best when the percentage of DE is high. Note that even we used the best cutoff for ORA, it is almost the worst method and only better than PAGE and Tian_Gene. PAGE and Tian_Gene have almost identical performance. More importantly, we observed a general trend that phenotype resampling methods are better than gene set resampling methods. As a mixed hypothesis testing method, GSA seems to have consistently performance across different percentage of DE gene.

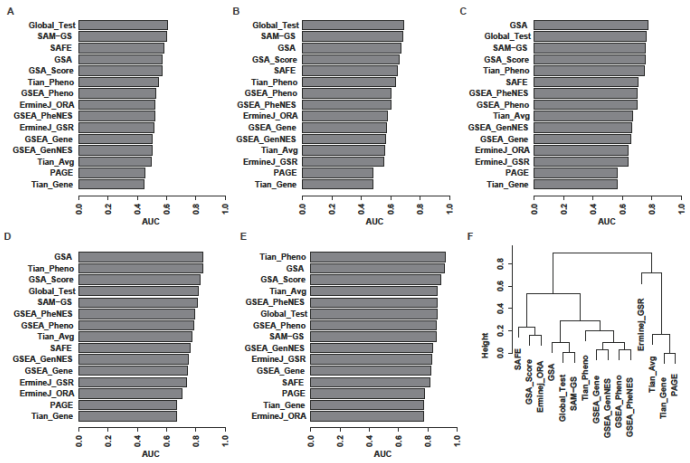

Figure 2. Performance comparison of gene set analysis tools. Mean and standard deviation of Area Under Curve (AUC) from 100 simulated data were calculated for each tool. Hierarchical Clustering of gene set analysis tools based on average ranks of gene sets in 100 simulated data. A~E shows the AUC with 10%, 20%, 30%, 40%, and 50% DE genes respectively. F. The plot shows the clustering result using 20% portion of DEG in a gene set. The color scale represents the similarity between each tool.

We next looked at how similar these tools are relative to each other. The similarity was determined by Euclidean distance between the average ranks of the gene sets across the 100 simulated data sets. We observed very similar results using different percentage of DE genes, and only the data with 30% DE genes are shown in Figure 2F. These tools can be classified into four groups using hierarchical clustering method. Global_Test, SAM-GS, GSA, Tian_Pheno and

GSEA form the biggest group. SAFE and ErmineJ_ORA the second group, while PAGE and Tian_Gene form the third. ErmineJ_GSR is the most distinct from all other methods. GSEA phenotype resampling and gene set resampling method form a subgroup, possibly due to its unique random walking algorithm. We also noted that the distance between PAGE and Tian_Gene is almost 0, which is not surprising because standardization based on large number of gene set resampling within Tian_Gene is equivalent to the standardization used in PAGE. It is unexpected though that ErmineJ_GSR is different from all other methods, because it is theorectically the same method as Tian_Gene. Its performance is also somewhat better than Tian_Gene (Fig. 2F).

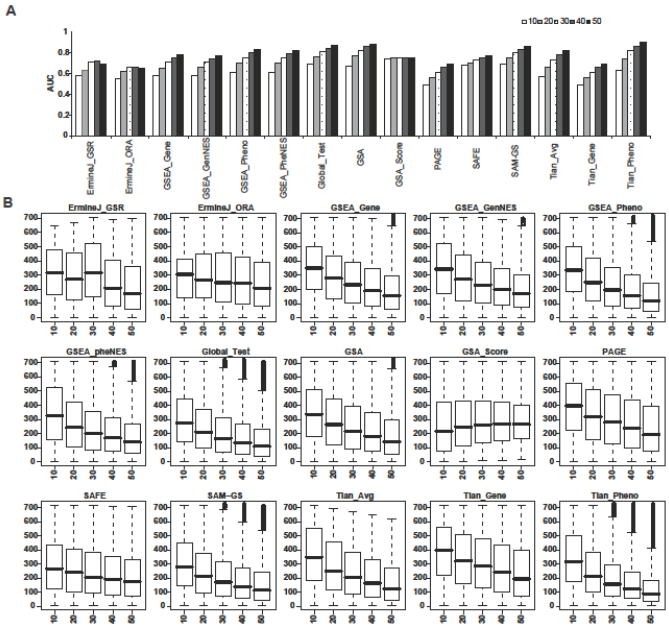

Figure 3. Effect of gene set size. A. The average AUC for different size of gene sets are plotted for each tools. B. The ranks for gene sets with different sizes. The x-axis is the size of gene sets, and the y-axis is the rank of the gene sets based on p-value.

3.1.2 Effects of Gene Set Size

To evaluate the effects of gene set size, we generated simulation data sets with gene set sizes of 10, 20, 30, 40 and 50 separately. Figure 3A shows the average AUCs of each tool across different gene set sizes with 100 simulations. Although to different extent, all of the methods have better performance to detect larger gene sets.

To further examine whether the analysis tools are biased to larger gene set size, we created 5 groups of pseudo gene sets, with size equal to 10, 20, 30, 40 and 50 genes respectively for each group. Each group contains 50 pseudo-gene sets. Therefore, there are 714 gene sets in the simulated data set, including 250 pseudo-gene sets and 464 real gene sets from KEGG and BioCarta. We created DE genes in the 10 treated samples, in 30% of genes for each of the 250 pseudo gene sets as well as randomly adding 0.5 in the genes not belonging to any gene sets to keep the overall number of DE genes being 2000 out of 20,000 total genes. The average ranks of gene sets with different sizes from 100 simulations were obtained. If the gene set size has no effects, the average ranks should be similar for gene sets with different sizes. However, as shown in Figure 3B, the gene sets with larger sizes rank better than those with smaller sizes, regardless of what tools are used. The bias is more severe in methods that included a standardization step based on the null distribution of ES, such as Tian/sigPathway and GSEA.

3.2 Performance plot

After the above simulation study, we conclude that both AUC and gene set size are critical factors to evaluate the performance of gene set analysis tools. Therefore, we present the AUC and gene set size effects together on the same plot so that the performance of each tool can be easily compared. The x-axis is the average AUC of simulated gene sets with 10% DE gene background, and y-axis is the slope of ranks among different sizes of gene sets. The best tool should have high AUC, which means better sensitivity and specificity, and low absolute slope, which means less bias to large gene sets. Therefore, the best tools should reside at the upper right corner of the plot. As shown in Figure 4, tools testing competitive hypothesis generally have less bias to gene set sizes, but also have lower AUC. In contrary, tools testing self-contained or mixed hypothesis have more severe bias to gene set sizes, but have better AUC.

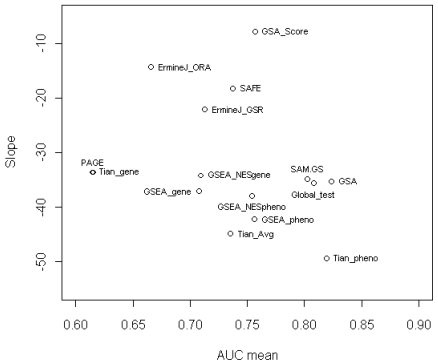

Figure 4. Performance evaluation of gene set analysis tools.

4 Methods performance on experimental data set

To confirm our simulation result with real-world scenario, we further compared the performance of the 11 gene set analysis methods by testing them on the p53 expression data on cancer cell lines. The dataset consisted of the transcriptional profiles from 17 p53+ and 33 p53 mutant cancer cell lines and was downloaded from the GSEA website. We utilize three p53 related pathways to measure the performance of pathway methods. More specifically, we roughly utilized the sum of the rank of the three p53 related pathways based on p-values or normalized enrichment scores assigned by each method to test whether these pathways appear as the top significant pathways.

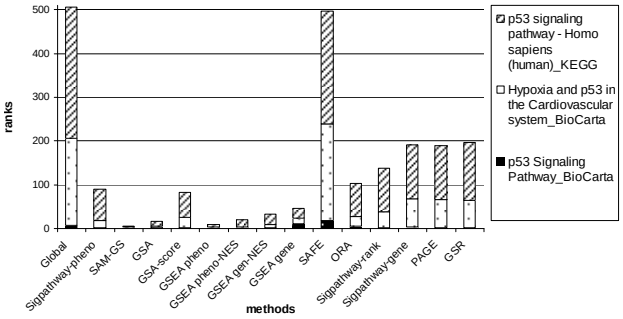

Figure 5. Performance of gene set analysis tools on p53 dataset.

In general, the phenotype-permutation based methods Sigpathway_Pheno, SAM-GS, GSA and GSEA identified the three pathways as relatively top pathways compared to the gene-permutation based methods Sigpathway_Gene, PAGE, and EmineJ_GSR (Figure 5). GSEA gene-set permutation retains its performance mainly due to the unique random-walk strategy.

5 Discussions

In this study, we have systemically compared 11 gene set analysis methods. To our knowledge, this is by far the most comprehensive comparison study. We confirmed that ORA method is highly sensitive to the selected cutoffs, which is likely to create very biased conclusion that is difficult to reproduce. Even when the best cutoff was used, methods based on ORA still have almost the worst sensitivity and specificity when compared to other analysis methods. The strength of ORA methods is that they have less bias to large gene sets. To some extent, we can consider that ORA methods are similar to gene set resampling methods, except that the latter is non-parametric.

We observed that the methods that are self-contained or mixed have better sensitivity and specificity than the methods that are purely competitive. A possible explanation is that gene resampling ignores the correlation structure in the gene sets, which might overestimate the variance in the null distribution of ES. This is also due to the fact that there are 10% DE genes in our simulated data sets, and this portion of genes results in a higher null ES value in gene resampling methods than in phenotype resampling methods. We feel that the chosen 10% DE genes is critical in the evaluation because it is quite common in real microarray experiments that there are significant portion of DE genes not belonging to any tested gene sets. Omitting the 10% DE genes in the simulated data sets will result in very similar performance between self-contained and competitive methods.

It is quite interesting to observe the bias towards large gene set size in most of the tools. This bias still exists even in the tools implementing a standardization step. The good performance of GSA scores suggests that a better scoring system without phenotype resampling can possibly overcome this limitation. As pointed out by Nam D and Kim SY, there are other factors, such as user friendly interface and species support, need to be considered when selecting the best analysis tools.

In summary, we have conducted systemic comparison of popular gene set analysis tools. Our results provide valuable information for researchers to understand the advantages and limitations of these tools.

6 References

[1] Nam D, Kim SY. “Gene-set approach for expression pattern analysis”. Briefings in Bioinformatics Jan 17 2008.

[2] Tian L, Greenberg SA, Kong SW, et al. Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci USA 2005;102:13544–9.

[3] Goeman JJ., Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 2007;23:980-7.

[4] Dinu I, Liu Q, Potter JD, et al. A biological evaluation of six gene set analysis methods for identification of differentially expressed pathways in microarray data. Cancer Informatics 2008;6:357-68.

[5] Dinu I, Potter JD, Mueller T, et al. Improving gene set analysis of microarray data by SAM-GS. BMC Bioinformatics 2007;8:242.

[6] Goeman JJ, van de Geer SA, de Kort F, et al. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 2004;20:93–9.

[7] Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005;102:15545–50.

[8] Lee HK, Braynen W, Keshav K, et al. ErmineJ: tool for functional analysis of gene expression data sets. BMC Bioinformatics 2005;6:269.

[9] Efron B, Tibshirani R. On testing the significance of sets of genes. Ann Appl Stat 2007;1:107–29.

[10] Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics 2005;21:1943-9.

[11] Kim SY, Volsky DJ. PAGE: parametric analysis of gene set enrichment. BMC Bioinformatics 2005;6:144.