BIOC 218 Final Project Dian Yang 06072012

A Critical Review of Gene Set Enrichment Analysis: Development and Improvement

Introduction:

Genome-wide expression analysis by microarrays has been one of the most widely used as measurement tools in biological research. Currently, the challenge of microarray is no longer to get enough gene expression profile data, but rather lies in analysis and interpretation of the results to understand the real biological meanings. Researchers kept working on extracting clear and coherent hypotheses from genome wide expression data (Goeman, 2007).

In early time, a common approach to analyze microarray data is just focusing on a handful of genes that are at either the top or the bottom of the list of genes. Attempting to understand each differentially expressed gene on a list of significant genes is laborious and demanding. It has also been shown that the gene list generated from a small number of samples can be highly variable (Pavlidis, 2003).

This kind of approach has some major limitations. A key limitation is that single gene analysis usually misses some important effects on pathways. Modest changes in all genes encoding members of a biological pathway may alter the pathway dramatically and might even be more important than a relative big change of a single gene. Also, because of the noise of microarray, modest biological differences cannot be detected after multiple hypothesis tests. It is also possible that a lot of genes are statistically significant but these genes don't have any unifying biological themes, thus it will be a big challenge to find meaningful biological meanings (Subramaniana, 2005). Lastly, the lists of significant genes from different groups working on the same biological system have very little overlaps (Fortunel, 2003).

In this paper, I will mainly discuss a power statistical tool gene set enrichment analysis, focusing on its development, mechanism, strength, weakness and improvement of this method.

2. The development of GSEA:

Due to the limitation of single gene analysis, scientists became interested in examining the association between known biological categories or pathways and outcomes.

Currently, there are two main types of method using gene sets to analyze differential expression data, the over-representation and the aggregate score

approaches. In both, gene categories or gene sets are generated before the statistical analysis. Most commonly, the gene sets are generated based on genes that are essential for a biological process, or have the same molecular function. In many cases, the gene sets are picked to specifically target the condition that is being studied. However, it is also more common to use category definitions directly from the Gene Ontology project (Lee, 2005). The Gene Ontology project sets a standard to describe gene and gene product attributes in any organism (The Gene Ontology Consortium 2000).

The Over-representation approach has a major limitation that it ignores all the genes that did not make the list of candidate genes. Therefore, the results mostly rely on the cutoff used in generating this list. In contrast, the aggregate score approach does not have this limitation. Basically, this approach is to assign scores to each gene set based on all the gene-specific scores for that gene set. There are various ways to calculate these aggregate scores (Pavlidis, 2002; Mootha et al. 2003).

Gene Set enrichment analysis (GSEA) is a method that evaluates microarray data at the level of gene sets that are defined based on prior biological knowledge. GSEA, developed by the Lander and Mesirov group, aims to determine whether members of a gene set tend to occur towards the top (or bottom) of the gene list. GSEA is mainly composed of the following four steps (Tian, 2005; Subramanian, 2005) :

(1) All genes are ranked by a signal-to-noise ratio.

(2) For each gene set, the distribution of gene ranks from the gene set is compared against the distribution for the rest of the genes by using the enrichment score (ES) based on a one-sided Kolmogorov–Smirnov statistic; The ES score reflects the degree to which a gene set is overrepresented at the top or bottoms of the entire list. The score is calculated in the following way. When we walk down the ranked list, when we encounter a gene in the gene set, the score is increased in a running sum statistical manner; on the contrary, the score decreases if we encounter a gene that is not in the gene sets. And the ES score is the maximum deviation from zero in the random walk, which correspond to a weighted Kolmogorov-Smirnov-like statistic.

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

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

Figrue 1. Calculation of Enrichment Score. (Subramaniana, 2005) S means gene sets, N is the total number of genes in the ranked list. To calculate the ES

value, we should rank the genes to generate a list {g1,...,gN} according to the correlation, r(gj)=rj, of their expression profiles with C. And then we should evaluate the fraction of genes in gene sets weighted by their correlation and the fraction of genes not in gene sets present up to a given position i in list. The ES is the maximum deviation from zero of P hit - P miss.

(3) Class labels are permuted to generate a null distribution of ES. Nominal P value，reflecting the statistical significance， is estimated using an empirical phenotype-based permutation test procedure that preserves the complex correlation signature of gene expression data. In detail, we first permutate the phenotype labels and recompute the *ES* of the gene set for the permuted data, which generates a null distribution for the *ES.* The empirical, nominal *P* value of the observed *ES* is then calculated relative to this null distribution.

(4) The last step is to adjust the significance level to account for multiple hypothesis testing. To get a normalized ES for each gene set, we will adjust for variation in gene set size. And then the false discovery rate (FDR) is calculated by comparing the tails of the observed and null distributions for the NES. The FDR reflects the probability of false positive discoveries of the gene set with certain NES.

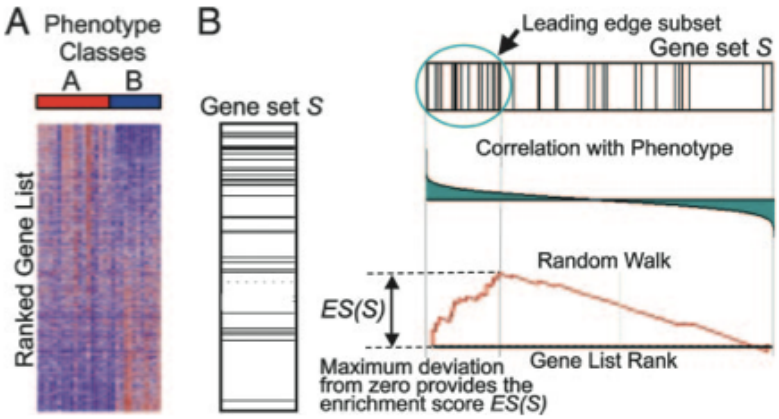


Figure 2. A GSEA overview (Subramanian et al. 2005) A. The expression data are sorted by correlation with phenotype. We can rank the genes in the order of differential expression. B. A Plot of the running sum for the gene sets in the ranked gene list.

3. Application of GSEA：

After its invention, GSEA has been widely used to analyze and interpret microarray as well RNA-Seq data. The original GSEA approach has also been refined into a more sensitive and robust analytical tool. GSEA eases the interpretation of a large scale experiment by indentifying pathways and process. Also, GSEA could be used to refine manually curated pathways and sets by indentifying the leading edge sets that are shared across diverse experimental data sets.

Until now, many researchers have utilized this powerful statistical approach to analyze genome wide transcription profile data, especially in a lot of disease study, in which they compare the expression data from normal people with patients with certain kind of disease, such as leukemia and lung cancer. And this powerful bioinformatics tool greatly foster new scientific discoveries.

To facilitate the use of GSEA, the Broad Institute launched the GSEA software freely available online. This software also supports R-programming and Java programming. The GSEA website is very user-friendly and the user guide includes very detailed instructions about the formatting issue, requirements and other useful tips for using GSEA and analyzing data.

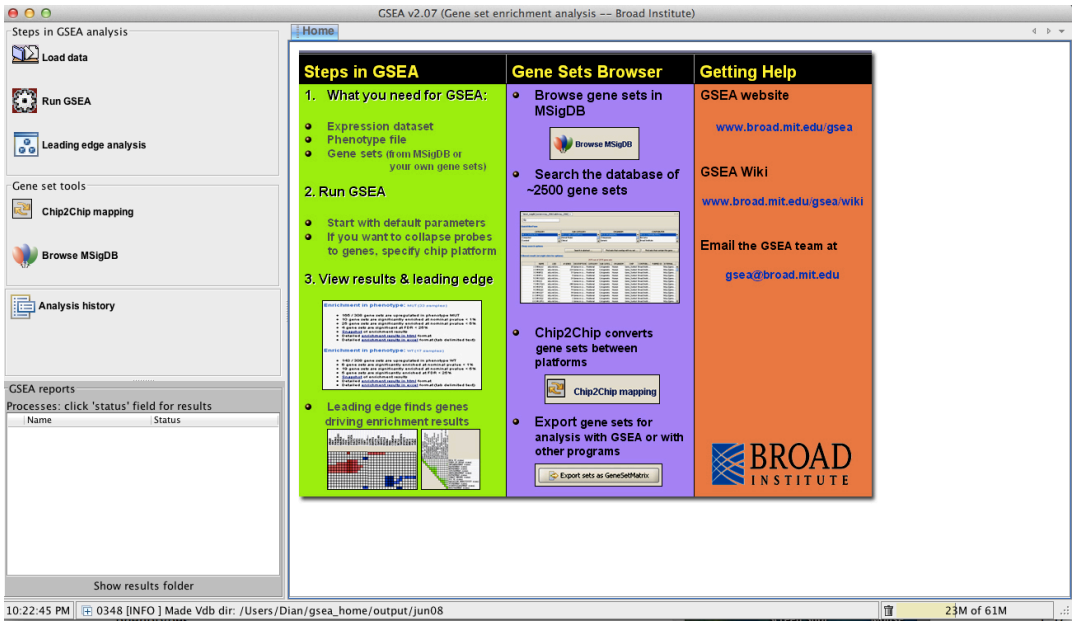


Figure 3. The GSEA software.

4. The Improvement of GSEA

Although GSEA is still currently the most population analysis for microarray and RNA-seq analysis, there are several shortcomings of this method. First, GSEA sometimes have very low power because, as mentioned in the user guide, the

recommended most suitable FDR threshold is 0.25. This low power might be because of the fact that the model and null hypothesis used to motivate the test statistic are different from those that are used for calculating p value (Yan, 2008).

Another problem of GSEA is that it loses the information on the degree of association between each gene and the binary phenotype by only using the relative ranking of genes rather than the absolute measurements (Dinu, 2007).

Third, as shown by Dinu et al, GSEA doesn't meet some simple requisite criteria for a gene-set enrichment analysis because in some cases, GSEA would frequently indentify gene sets as statistically significant when all of its genes have observed expressions completely uncorrelated with the phenotype (Dinu, 2007).

Another major problem of GSEA is that the enrichment score considers genes with the phenotype separately, even when they might have similar association with the phenotype. So GSEA is not powerful to detect a gene set with a mix of genes with positive and negative associations with the phenotype. For example, some feedback loops in the biological pathways involve several genes may cause a mix of genes with positive and negative relations with the certain phenotypes (Dinu, 2007).

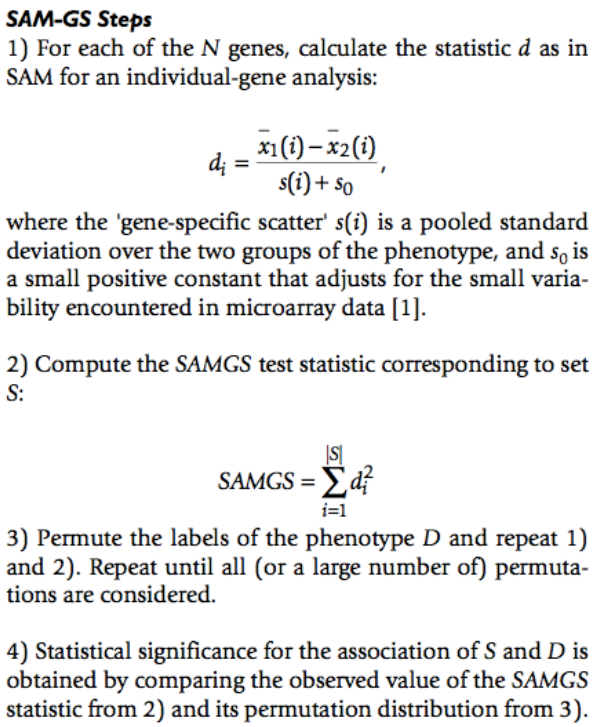
Taking the drawbacks into account, people think about improving the GSEA method. The Buhlmann group came up with an improved GSEA model in which they adapt the GSEA method to a self-contained null hypothesis and to calculate the P value using subject sampling (Goeman, 2007).

Original GSEA uses a competitive hypothesis rather than a self-contained null hypothesis. The competitive hypothesis suggests the gene in the ranked gene expression list are at most as often differentially expressed as the genes in gene sets, while the self-contained null hypothesis states that no genes in the gene list are differentially expressed. Statistically, the self-contained null hypothesis is more restrictive than the competitive hypothesis (Allison, 2006). In comparison of the two hypotheses, the self-contained one wins for several reasons. First, a test based on the self-contained null hypothesis often has more power than a test based on the competitive hypothesis because the restrictive nature of self-contained hypothesis. Second, the self-contained hypothesis has a desirable property that single gene testing and gene set testing are completely equivalent for singleton gene sets. However, competitive hypothesis doesn't treat a singleton gene set similarly to a single gene. Third, self-contained hypothesis allow us to look at the set of all genes on the chip, while this cannot be tested in a competitive way simply because there is no complement to test the gene set against. Thus, by adapting GSEA to a self-contained hypothesis by calculating the Kolmogorov–Smirnov statistic on the basis of the p value, we can improve the power of GSEA method (Goeman, 2007).

Alternatively, in recent years, with the development of bioinformatics, several new statistical analyses have been developed, such as Sub-GSE, SAM-GS and

GEGA. The Sub-GSE method was developed in 2008, which measures the enrichment of a predefined gene set or pathway, by testing its subsets. In real application, sub-GSE is shown to be more sensitive than GSEA in detecting gene sets assisted with a phenotype of interest, especially in cases where only a fraction of the genes in the set are associated with the phenotype. What's more, it is also shown that this sub-GSE method can detect more biologically meaningful gene sets than GSEA (Yan, 2008).

Another novel method is significance analysis of microarray to gene-set analysis (SAM-GS). Dinu introduced this SAM-GS method (Figure 4), which takes the same approach as SAM t-like statistic. Contrast to GSEA, SAM-GS tests a hypothesis that the mean vectors of expressions of genes in a gene set does not differ by the phenotype of interest. Also, SAM-GS takes into account of the absolute measures of each gene and requires measurement only of the expression the genes in the gene set to construct the test statistic. In their paper, they compare and contrast the power and sensitivity of both methods by their performance on 3 DNA microarray datasets. By comparison, we can see that SAM-GS has clear advantage than GSEA from both statistical and biological points (Figure 5, Dinu 2007).



SAM-GS Steps

1) For each of the N genes, calculate the statistic d as in SAM for an individual-gene analysis:

$$d_i = \frac{\bar{x}_1(i) - \bar{x}_2(i)}{s(i) + s_0},$$

where the 'gene-specific scatter' $s(i)$ is a pooled standard deviation over the two groups of the phenotype, and s_0 is a small positive constant that adjusts for the small variability encountered in microarray data [1].

2) Compute the *SAMGS* test statistic corresponding to set S:

$$SAMGS = \sum_{i=1}^{|S|} d_i^2$$

3) Permute the labels of the phenotype D and repeat 1) and 2). Repeat until all (or a large number of) permutations are considered.

4) Statistical significance for the association of S and D is obtained by comparing the observed value of the *SAMGS* statistic from 2) and its permutation distribution from 3).

Figure 4. A summary of SAM-GS method (Dinu, 2007). This figure described the key steps of SAM-GS.

Besides the methods mentioned above, another method called generally applicable gene set enrichment for pathway analysis (GAGE) also overcomes several limitations of GSEA. For example, GSEA is not appropriate for studies with under 8 gene chips per state and GSEA only consider transcription regulation in one direction. The GAGE method was developed in 2009, applying to databases with any number of samples and was based on a parametric gene randomization procedure. Contrary to general parametric analysis if gene set enrichment, it assumes a gene set comes from a different distribution than the background and uses two-sample t-test to account for the gene set specific variance and the background variance (Luo, 2009). In summary, the new GAGE method has the following advantages: 1. Better consistency across repeated studies and experiments. 2. Better sensitivity and specificity. 3. More biological relevance of the regulatory mechanisms.

Table 3: Results of the analyses of three datasets by GSEA and SAM-GS.

Dataset	% of individual genes with FDR* ≤ 0.25	# of gene sets with FDR ≤ 0.01		# of gene sets with FDR ≤ 0.25		Sensitivity/Specificity (AUC†) of GSEA‡
		GSEA	SAM-GS	GSEA	SAM-GS	
Sex	0.1%	4	5	6	6	0.78/0.98 (0.94)
p53	0.3%	3	36	6	308	0.21/0.94 (0.68)
Leukemia	79.9%	0	182	5	182	0.06/NA§ (NA§)

* FDR = False discovery rate estimate
† AUC = Area under the ROC curve
‡ Taking SAM-GS p ≤ 0.05 as the target to be predicted
§ All gene sets in the leukemia dataset had SAM-GS p ≥ 0.05

Table 4: The 31 gene sets for which SAM-GS and GSEA strongly disagreed (SAM-GS FDR ≤ 0.01, GSEA FDR ≥ 0.49) in the *p53* analysis.

Gene Set	GSEA		SAM-GS		*p53* link
	FDR	p-value	FDR	p-value	
ATM Pathway	0.87	0.21	≤ 0.01	< 0.001	Pathway member
BAD Pathway	0.57	0.04	≤ 0.01	< 0.001	Apoptosis
Calcineurin Pathway	0.84	0.13	≤ 0.01	< 0.001	*p53*-induced proline oxidase mediates apoptosis via a calcineurin-dependent pathway (12)
Cell cycle regulator	0.90	0.29	≤ 0.01	< 0.001	Cell cycle
Mitochondria pathway	0.88	0.32	≤ 0.01	< 0.001	Apoptosis
***p53* signaling pathway**	0.51	0.01	≤ 0.01	< 0.001	Pathway member
Raccycd Pathway	0.83	0.56	≤ 0.01	< 0.001	Cell cycle
SA_TRKA_RECEPTOR	0.83	0.34	≤ 0.01	< 0.001	Integrated negative feedback loop between Akt and *p53* (11)
bcl2family and reg. network	0.83	0.42	≤ 0.01	0.001	Apoptosis
Cell cycle arrest	0.98	0.49	≤ 0.01	0.001	Cell cycle
Ceramide Pathway	0.88	0.30	≤ 0.01	0.001	Apoptosis
DNA DAMAGE SIGNALLING	0.85	0.23	≤ 0.01	0.002	Pathway member
SIG_IL4RECEPTOR IN_B_LYMPHOCYTES	0.93	0.27	≤ 0.01	0.002	Cytokines; JAK/STAT signaling
Cell cycle Pathway	0.89	0.72	≤ 0.01	0.003	Pathway member
G2 Pathway	0.81	0.50	≤ 0.01	0.003	Pathway member
Chemical Pathway	0.53	0.04	≤ 0.01	0.005	Pathway member
Drug resistance and metabolism	0.86	0.08	≤ 0.01	0.005	Pathway member
G1 Pathway	0.81	0.37	≤ 0.01	0.005	Pathway member
Breast cancer estrogen signaling	1.00	0.85	≤ 0.01	0.006	Pathway member
Ca_nf_at_signaling	0.78	0.08	≤ 0.01	0.007	Apoptosis (and cytokines)
Cytokine Pathway	0.53	0.05	≤ 0.01	0.007	Cytokines
ST_Interleukin_4_Pathway	0.84	0.07	≤ 0.01	0.007	Cytokines; JAK/STAT signaling
CR_DEATH	0.86	0.31	≤ 0.01	0.008	Pathway member
MAP00860: Porphyrin & chlorophyll metabolism	0.92	0.29	≤ 0.01	0.010	CPO regulated by *p53* (13)
Ck1 Pathway	0.49	0.02	≤ 0.01	0.011	Cdk5 phosphorylates *p53* (9)
Hivnef Pathway	0.95	0.48	≤ 0.01	0.011	Apoptosis
Ets Pathway	0.79	0.45	≤ 0.01	0.012	Ets1 required for *p53* transcriptional activation in UV-induced apoptosis (10)
ST_Wnt_Ca2_cyclic_GMP_Pathway	0.80	0.13	≤ 0.01	0.012	At least one known link between wnt and *p53* (14)
Chrebp Pathway	0.84	0.42	≤ 0.01	0.013	unknown
GPCRs_Class_A_Rhodopsin-like	0.60	0.04	≤ 0.01	0.013	unknown
ST_Fas_Signaling_Pathway	0.80	0.52	≤ 0.01	0.013	Pathway member

Figure 5. Comparison of GSEA and SAM-GS (Dinu, 2007). Table 3 shows the number of genes detected by both method. SAM-GS has much greater sensitivity than GSEA does. Table 4 compares the results from SAM-GS and GSEA of the most strongly disagreed genes. SAM is better at detecting genes and pathways while GSEA is not that sensitive.

Summary:

Interpreting transcription profile data and discovering the biological meaning are challenges in the field. Traditional ways that focused on single genes missed a lot of information of the expression profiling data.

The development of the powerful analytical gene set enrichment analysis method derives its power by focusing on gene sets and highlighting pathway and process level. GSEA also has other adavantages, for example, it boosts the signal to noise ratio, make it possible to detect modest changes in individual genes and it also consider all of the genes without an arbitraty cutoff in terms of fold-change or significance. Because of all of the advantages, GSEA is now the most popular method for analyzing gene expression profile data.

However, we should also note that there are still several limitations of the current GSEA method. Focusing on those drawbacks and limitations, researchers have already developed more statistical methods in order to improve the sensitivity and power of GSEA, such as sub-GSE, SAM-GS and GAGE. With the development of these new methods, we can choose the most appropriate method for our specific purpose during research and discover new things that we might missed before.

Reference:

1. Aravind Subramaniana,b, Pablo Tamayoa,b, Vamsi K. Moothaa,c, Sayan Mukherjeed, Benjamin L. Eberta,e, Michael A. Gillettea,f, Amanda Paulovichg, Scott L. Pomeroyh, Todd R. Goluba,e, Eric S. Landera,c,i,j,k, and Jill P. Mesirova,k **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. The *Proceedings of the National Academy of Sciences USA 2005*

2. Weijun Luo, Michael S Friedman, Kerby Shedden, Kurt D Hankenson and Peter J Woolf **GAGE: generally applicable gene set enrichment for pathway analysis**. *BMC Bioinformatics 2009*

3. Fortunel, N. O., Otu, H. H., Ng, H. H., Chen, J., Mu, X., Chevassut, T., Li, X., Joseph,

M., Bailey, C., Hatzfeld, J. A., *et al.* **Comment on " 'Stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature".** (2003) *Science* 302, 393.

4. Irina Dinu, John D Potter, Thomas Mueller, Qi Liu, Adeniyi J Adewale, Gian S Jhangri, Gunilla Einecke, Konrad S Famulski, Philip Halloran and Yutaka Yasui **Improving gene set analysis of microarray data by SAM-GS** *BMC Bioinformatics* 2007, 8:242

5. M.F. Misman S. Deris S.Z.M. Hashim R. Jumali and M.S. Mohamad **Pathway-Based Microarray Analysis for Defining Statistical Significant Phenotype -Related Pathways: A Review of Common Approaches**. *International Conference on Information Management and Engineering 2009 IEEE*

6. Insuk Sohn, Kouros Owzar, Johan Lim, Stephen L George, Stephanie Mackey Cushman and Sin-Ho Jung **Multiple testing for gene sets from microarray experiments**. *BMC Bioinformatics 2011, 12:209*

7. Xiting Yan and Fengzhu Sun. **Testing gene set enrichment for subset of genes: Sub-GSE.** *BMC Bioinformatics* 2008, 9:362

8. Donna K. Slonim **From patterns to pathways: gene expression data analysis comes of age** *Nature Genetics* 2002 doi:10.1038/ng1033

9. Jelle J. Goeman, and Peter Bu¨ hlmann **Analyzing gene expression data in terms of gene sets: methodological issues** *Bioinformatics* Vol. 23 no. 8 2007, pages 980–987

10. Lu Tian, Steven A. Greenberg, Sek Won Kong, Josiah Altschuler, Isaac S. Kohane, and Peter J. Park. **Discovering statistically significant pathways in expression profiling studies.** The *Proceedings of the National Academy of Sciences USA 2005*