An Overview of Gene Set Enrichment Analysis

CHELSEA JUI-TING JU, University of California, Los Angeles

Gene set enrichment analysis is a data mining approach designed to facilitate the biological interpretation of gene expression data. The main idea is to aggregate genes based on their commonalities, and assess the significant changes as a group. The framework for most of the current implementations can be divided into five components, including data collection and preprocessing, gene level statistics computation, gene set statistics computation, significance measurement, and multiple testing correction. Three softwares are reviewed and compared: GSEA, PAGE, and GSA. The comparison focuses on their statistical approaches in gene level statistics, gene set statistics, and significance measurement.

Categories and Subject Descriptors: G.3 [**Nonparametric statistics**] Kolmogorov-Smirnov statistic; J.3 [**LIFE AND MEDICAL SCIENCES**] Biology and genetics

General Terms: Statistic in Computational Biology

Additional Key Words and Phrases: Microarray, Genomic, Enrichment Analysis

ACM Reference Format:

Chelsea J.-T. Ju. 2014. An Overview of Gene Set Enrichment Analysis. *ACM Trans. Appl. Percept.* 0, 0, Article 0 (June 2014), 8 pages.

DOI:http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

The advent of "omic" studies has provided a high-throughput screening to quantify the changes in biological system. These studies include, but not limit to, genomics for gene expression profiling, proteomics for protein level quantification, and metabolomics for measuring metabolites abundance. In genomewide expression studies, the predominant technologies are DNA microarray and RNA Sequencing to monitor changes in expression of thousands of genes simultaneously. Similarly, liquid chromatography or gas chromatography followed by mass spectrometry (LC-MS or GS-MS) allow a large scale identification and quantification of proteins and metabolites in different biological systems. These -omics approaches often generate a large list of candidates, ranging from hundreds to thousands of genes, proteins, or metabolites. Consequently, mining through this large list of "interesting" candidates becomes a daunting task. As Subramanian et al. [2005] stated that the challenge no longer lies in obtaining molecular profiles, but rather in interpreting the results to gain insights into biological mechanism.

This work is part of the course curriculum of STATS M254 - Statistical Methods in Computational Biology, instructed by Dr. Jingyi Jessica Li at University of California, Los Angeles in Spring 2014.

DOI:http://dx.doi.org/10.1145/0000000.0000000

In order to facilitate the functional analysis across different phenotypes, many enrichment tools have been developed in the past ten years. The main idea is to aggregate genes, proteins or metabolties into a set that share a common theme, such as biological function, chromosomal location, protein interaction, regulation, or biochemical pathway. If the members of the predefined set are over-represented in the candidate list, then the list is said to be enriched by the predefined set [Hung et al. 2012]. The benefit of the enrichment analysis is two-fold. First, from a statistical point of view, grouping the candidates serves as a dimensionality reduction technique in data mining. The issues of false discovery rate in multiple testing are alleviated with a smaller number of objects undergoing statistical analysis. Second, from a biological point of view, placing the candidates into the context of biological understanding provides a more meaningful interpretation of the experimental results.

Since the genomic approach is the most mature field among all the -omic studies, most of the enrichment tools carry forward with the results from the microarray pipeline, and focus on identifying the significant gene sets that are enriched in the phenotype of interests. In this article, I will give an overview of the framework for these tools (Section 2), and discuss the statistical methods employed by three popular softwares, GSEA, PAGE, and GSA (Section 3). Although these softwares are designed for gene set analysis, the methods are applicable to identify the significant sets of proteins or metabolites.

2. OVERVIEW

Ackermann and Strimmer [2009] and Hung et al. [2012] summarized the framework of the Gene Set Enrichment procedures into five key components before making a statistical conclusion: data collection and preprocessing, gene level statistics for every single gene, gene set statistics for a predefined set, significance measurement, and multiple testing correction. This framework is illustrated in Figure 1.

2.1 Data Preprocessing

A typical gene expression profiling experiment involves comparing the expression patterns between two or more phenotypes. Data normalization is an essential step, which allow expression values from different experiments to be directly comparable [Irizarry et al. 2003]. In DNA microarray analysis, the expression values are represented by the color intensity of florescent dyes. The experimental artifacts are removed by normalization algorithms, such as RMA [Irizarry et al. 2003] for single color microarray, and print-tips loess [Smyth and Speed 2003] for two-color cDNA microarray. Alternatively in RNASeq data, the density of reads that map to a gene is normalized for the length of its transcript and for the sequencing depth of the experiment. The abundance of a gene is quantified by Reads Per Kilobase exon Model per million mapped reads (RPKM) [Mortazavi et al. 2008]. After normalization, log transformation of the expression values is commonly applied to avoid bias toward highly expressed genes.

2.2 Gene Level Statistics

The first step in a gene set enrichment analysis is to assess the amount of differential expression of the individual gene between two phenotypes. The values of differential expression can be represented by fold change, signal-to-noise ratio (mean to standard deviation ratio), regularized t-statistics, shrinkage correlation coefficient, coefficient of linear/logistic regression, and penalized log-likelihood ratio [Ackermann and Strimmer 2009]. Since most of the gene set enrichment approaches include the entire list of genes for downstream analysis, the choice of these methods is less critical. However, it is more suitable to use regularized version of test statistics due to the small sample size found in most instances of the experiment [Ackermann and Strimmer 2009]. In addition, the changes of gene expression can occur in opposite directions (either up-regulation or down-regulation), especially in a feedback mechanism.

If this type of mechanism is considered, it is recommended to eliminate the direction by taking the absolute or square of the gene statistics [Saxena et al. 2006; Hung et al. 2012].

2.3 Gene Set Statistics

To incorporate biological knowledge into the analysis, genes are combined into sets if they share a certain commonality. Gene Ontology [Ashburner et al. 2000] is the most commonly used knowledgebase as it provides a controlled vocabulary to describe gene roles in biological process, molecular function, and cellular component. Another common choice is the cascading pathways, where genes are grouped together if they are involved in the same pathway.

Gene set statistic provides a value to evaluate whether a gene set is significantly altered for a phenotype, and is defined by the properties of the genes in the set. This statistic can be computed by the sum or the mean or the median of the gene statistics, the modified Kolmogorov-Smirnov statistic, the maxmean statistic, the Wilcoxon rank sum test statistic, the sign test statistic or the conditional local FDR [Ackermann and Strimmer 2009]. The choice of these methods, together with the statistical assessment depend on the stated null hypothesis, which is addressed next.

2.4 Significance Measurement

There are two types of null hypothesis defined by Tian et al. [2005] and Ackermann and Strimmer [2009]. The first case, Q1, referred to as the "competitive null hypothesis". Given a phenotype, the test compares the phenotype association with genes in the set versus genes outside the set. This hypothesis considers all genes. In the other case, Q2, referred to as the "self-contained null hypothesis", which focuses only on the given gene set. It compares the gene set association with a phenotype versus a random phenotype. In general, Q2 is favored because it preserves the relationship of genes in a set, and directly address the question of enrichment [Hung et al. 2012].

The significance of the gene set statistic is evaluated by calculating the p-value from the null distribution. The null distribution can be generated in three different ways [Ackermann and Strimmer 2009], depending on the choice of null hypothesis. The competitive null hypothesis considers all genes, and therefore, the background distribution (null distribution) is obtained by shuffling genes in and out of the gene set. The self-contained null hypothesis compares the enrichment between phenotypes, so the background distribution can be simulated by randomly labeled the sample phenotypes. Both gene sampling and label permutation can be applied at the same time, generating another type of background distribution. The p-value is described by the fraction of gene set statistics in the re-sampled population that exceed (or fall below) the observed value.

2.5 Multiple Testing Correction

The multiple hypothesis testing problem arises when more than one gene sets are examined. A conservative approach is to use a sufficiently low corrected p-value, known as the Bonferroni correction [Shaffer 1995]. An alternative common approach is to control the false discovery rate using Benjamini-Hochberg procedure [Benjamini 1995].

3. IMPLEMENTATION

In this section, I choose three popular gene enrichment analysis softwares as examples, and focus on reviewing their methods for gene level statistics, gene set statistics, and significance measurement.

3.1 GSEA

Gene **S**et **E**nrichment **A**nalysis [Subramanian et al. 2005] is the pioneer tool for detecting enrichment. To date, it has more than six thousand citations according to Google Scholar. The software also provides

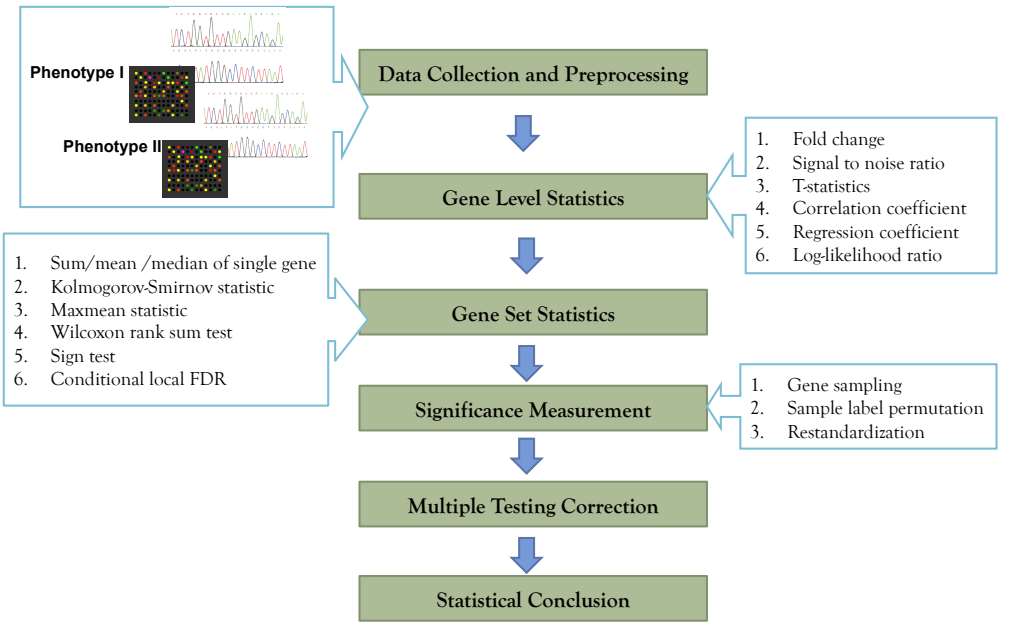

Fig. 1. The framework for gene set enrichment analysis.

a comprehensive molecular signatures database (MSigDB), which contains a wide range of collections of annotated functional gene sets. It takes the molecular profile data as an input, and allows users to select the gene set collection of interests.

Following the framework discussed in Section 2, GSEA uses values that can reflect the gene correlation with the phenotype of interests, such as the t-score or the signal-to-noise ratio, to represent the gene level statistics. The main idea of this approach follows the assumption that if the examined gene list is enriched by the member of a gene set, then these members tend to aggregate toward the top (or the bottom) of the list. Thus, the null hypothesis can be formally stated as

$$\text{Input : Gene List } L = \{l_1, l_2, \cdots, l_n\}, \text{Gene Set } S = \{s_1, s_2, \cdots, s_m\}$$
$$H_0 \text{ : m genes in a gene set S are randomly spread out in the list L among n genes}$$

Intuitively, the first step of the algorithm is to rank the genes in gene list L by sorting their gene level statistics. Here, the gene level statistics is denoted by r_j for gene l_j in gene list L. Using the ranked list, a weighted Kolmogorov-Smirnov statistic is computed, defined as the enrichment score (ES) of gene set S. The enrichment score is defined by the maximum deviation from zero of a running sum (running down the sorted list of genes). The score increases every time if a gene (l_j) in the list L is in the gene set S (Equation 1), and decreases otherwise (Equation 3). In addition, each gene set member (s_i) can be weighted by its absolute correlation with the phenotype, denoted as w_j. In another word, genes with high correlations to the phenotype contribute more to the enrichment score.

$$P_{hit}(S,i) = \sum_{l_j \in S; j \leq i} \frac{|r_j|^{w_j}}{N_R} \quad (1)$$
$$N_R = \sum_{l_j \in S} |r_j|^{w_j} \quad (2)$$

$$P_{miss}(S,i) = \sum_{l_j \notin S; j \leq i} \frac{1}{n-m} \quad (3)$$

$$ES(S) = max(P_{hit} - P_{miss}) \quad (4)$$

To estimate the significance of the enrichment score, sample labels permutation is used to generate the background distribution. ES_{NULL} is computed from the average enrichment score of one thousand permutations, where phenotype labels are randomly assigned to samples.

Although only the sample labels permutation is used to simulate the background distribution, the enrichment score considered the effect of genes outside the gene set S. Hence, Ackermann and Strimmer [2009] argue that this approach uses a hybrid of competitive null hypothesis (Q1) and self-contained null hypothesis (Q2).

3.2 PAGE

GSEA uses a non-parametric approach to evaluate the enrichment, and Kim and Volsky [2005] argue that the method is not sensitive enough to detect the significantly altered gene sets. They propose a parametric approach, **P**arametric **A**nalysis of **G**ene Set **E**nrichment, which uses normal distribution for statistical inference. They claim that a normal distribution paradigm requires observations to be independent and identically distributed (iid). If the gene list L is enriched in the member of gene set S, then these members are interdependent (for example, they are co-regulated). Thus, the distribution of these interdependent genes deviates from the normal distribution. The null hypothesis can be formally stated as

$$\text{Input : Gene List } L = \{l_1, l_2, \cdots, l_n\}, \text{Gene Set } S = \{s_1, s_2, \cdots, s_m\}$$
$$H_0 : \text{all genes in gene list } L \text{ are independent of each other and identically distributed}$$

It uses Gene Ontology as the predefine gene sets, and the fold change values for each gene between two phenotypes as the gene level statistics. The gene set statistic is defined by a "normalized" average fold changes across m gene set members. The mean (μ) and standard deviation (σ) of total fold change of gene list L are calculated. The average fold change for a gene set S is denoted by μ_S. The Z score is calculated in Equation 5.

$$Z = \frac{(\mu_S - \mu) \times \sqrt{m}}{\sigma} \quad (5)$$

The *p*-value can be obtained directly for the Z score by comparing the observed distribution with the standard normal distribution.

3.3 GSA

Following the same assumption as in PAGE, Efron and Tibshirani [2007] propose several improvements over PAGE and GSEA. GSA uses t-score as the gene level statistics, and introduces a new test statistic, maxmean, to describe the gene set statistic. In addition, the test specifically evaluates both of the competitive null hypothesis (Q1) and the self-contained null hypothesis (Q2) as stated below:

$$\text{Input : Gene List } L = \{l_1, l_2, \cdots, l_n\}, \text{Gene Set } S = \{s_1, s_2, \cdots, s_m\}$$
$$H_0(Q1) : \text{Gene set S has been chosen by random selection}$$
$$H_0(Q2) : \text{Samples are independent and identically distributed among gene set S}$$

The first step of the algorithm is to transform the t-score of each gene l_j in gene list L to a z-value. Theoretically, the z-value has a standard normal distribution (Equation 6), and the transformation is defined in Equation 7, where Φ is the standard normal cumulative distribution function and F_{n-2} is the c.d.f for a t distribution having $n-2$ degree of freedom.

$$z_j \sim N(0,1) \tag{6}$$

$$z_j = \Phi^{-1}(F_{n-2}(t_j)) \tag{7}$$

The maxmean statistic of a gene set S containing m genes is defined as

$$T_{maxmean} = max(Score_s^+, Score_s^-) \tag{8}$$

$$Score_s^+ = \frac{1}{m}\sum_{l_j \in S} z_j^+ \tag{9}$$

$$Score_s^- = \frac{1}{m}\sum_{l_j \in S} z_j^- \tag{10}$$

Equation 9 and Equation 10 referred to the averages of the positive and negative parts of the scores. The separation allows detecting gene sets containing expression changes in both directions (up- and down- regulation), and is more sensitive in picking up gene sets with moderately large positive and negative z-values.

Taking two null hypotheses into consideration, the maxmean statistic is normalized by the means and standard deviations obtained from both gene sampling and phenotype permutation distributions. The procedure is referred to as “restandardization”, and is defined in Equation 11.

$$T^{\star\star} = \mu^\dagger + \frac{\sigma^\dagger}{\sigma^\star}\frac{T^\star - \mu^\star}{\sqrt{m}} \tag{11}$$

where $\mu^\dagger$ and $\sigma^\dagger$ are the mean and standard deviation of the distribution obtained by gene shuffling; $\mu^\star$ and $\sigma^\star$ are the mean and standard deviation obtained from the sample label permutation distribution, and $T^\star$ is the maxmean statistic computed from shuffling the phenotypes.

4. DISCUSSION

For the past ten years, the development of gene set enrichment tools has been greatly enhanced in addressing different statistical assumptions and incorporating various functional knowledgebase. Most of the variants of gene set enrichment procedures follow the same framework as described in Section 2. Three implementations are discussed in details. GSEA employs the weighted modified Kolmogorov-Smirnov statistic to represent the gene sets. This non-parametric statistic is distribution free, but may be less sensitive in detecting the changes. In addition, the computation bottleneck falls in the permutation step for background distribution simulation. PAGE uses a normal approximation approach, which requires less computational effort, and the mathematical intuition is relatively straightforward. It claims to detect more significantly altered gene sets than GSEA, but it fails to address the multiple testing correction. GSA defines a new statistic, maxmean, which allows the detection of gene sets that are moderately altered, and those contain both up- and down- regulated genes. The choice of gene level statistics is rather inconsequential [Hung et al. 2012]; however, the gene set statistics and significance measurement depends on the selection of null hypothesis, and can have different power in detecting the significantly altered gene sets.

APPENDIX

A.1 Kolmogorov-Smirnov Statistic

Kolmogorove-Smirnov Statistic is a non-parametric test. It facilitates the comparison between a sample and a reference probability distribution, or between two samples. The comparison is performed by quantifying the largest distance between the empirical distribution function of a sample and the cumulative distribution function of the reference distribution, or between two empirical distributions from samples. http://www.physics.csbsju.edu/stats/KS-test.html provides an excellent resource explaining the method. The procedures can be summarized into the following two steps.

(1) Compute the Empirical Distribution Function
The empirical distribution function F_n for n iid observations X_i is defined as

$$F_n(X) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

where $I(X_i \leq x)$ is an indicating function

$$I(X_i \leq x) = \begin{cases} 1 & if(X_i \leq x) \\ 0 & otherwise \end{cases}$$

(2) The Largest Distance between two Distributions
For a given cumulative distribution function $F(X)$, the statistic is described by the largest distance between two distributions

$$D_n = \sup_x |F_n(X) - F(X)|$$

REFERENCES

Marit Ackermann and Korbinian Strimmer. 2009. A general modular framework for gene set enrichment analysis. *BMC bioinformatics* 10 (Jan. 2009), 47. DOI:http://dx.doi.org/10.1186/1471-2105-10-47

M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 25, 1 (May 2000), 25–9. DOI:http://dx.doi.org/10.1038/75556

Bradley Efron and Robert Tibshirani. 2007. On testing the significance of sets of genes. *The Annals of Applied Statistics* 1, 1 (June 2007), 107–129. DOI:http://dx.doi.org/10.1214/07-AOAS101

Jui-Hung Hung, Tun-Hsiang Yang, Zhenjun Hu, Zhiping Weng, and Charles DeLisi. 2012. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in bioinformatics* 13, 3 (May 2012), 281–91. DOI:http://dx.doi.org/10.1093/bib/bbr049

Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)* 4, 2 (April 2003), 249–64. DOI:http://dx.doi.org/10.1093/biostatistics/4.2.249

Seon-Young Kim and David J Volsky. 2005. PAGE: parametric analysis of gene set enrichment. *BMC bioinformatics* 6 (Jan. 2005), 144. DOI:http://dx.doi.org/10.1186/1471-2105-6-144

Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* 5, 7 (July 2008), 621–8. DOI:http://dx.doi.org/10.1038/nmeth.1226

Vishal Saxena, Dennis Orgill, and Isaac Kohane. 2006. Absolute enrichment: gene set enrichment analysis for homeostatic systems. *Nucleic acids research* 34, 22 (Jan. 2006), e151. DOI:http://dx.doi.org/10.1093/nar/gkl766

J P Shaffer. 1995. Multiple Hypothesis Testing. *Annual Review of Psychology* 46, 1 (Jan. 1995), 561–584. DOI:http://dx.doi.org/10.1146/annurev.ps.46.020195.003021

Gordon K Smyth and Terry Speed. 2003. Normalization of cDNA microarray data. *Methods (San Diego, Calif.)* 31, 4 (Dec. 2003), 265–73. http://www.ncbi.nlm.nih.gov/pubmed/14597310

Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, and Benjamin L Ebert. 2005. Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide. (2005).

Lu Tian, Steven A Greenberg, Sek Won Kong, Josiah Altschuler, Isaac S Kohane, and Peter J Park. 2005. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America* 102, 38 (Sept. 2005), 13544–9. DOI:http://dx.doi.org/10.1073/pnas.0506577102