

A Weighted Average Difference Gene-ranking for Gene Set Enrichment Analysis of Microarray Data

Jaeyoung Kim¹
widebrowboy@gmail.com

Hyungmin Lee²
lhm1103@knu.ac.kr

Miyoung Shin³
shinmy@knu.ac.kr

¹ Dept. of Information and Communication Engineering, Kyungpook National University, Daegu, South Korea

² Graduate School of Electrical Engineering and Computer Science, Kyungpook National University, Daegu, South Korea

³ School of Electrical Engineering & Computer Science, Kyungpook National University, Daegu, South Korea

Keywords: Gene set enrichment, differentially expressed genes, weighted average difference, gene ranking

1 Introduction

Recently, for the purpose of identifying *significant* gene sets showing differential expressions between two sample groups, the approach of gene set enrichment analysis [1,2] has taken many attentions which employs biological knowledge such as Gene-Ontology, KEGG Pathway, and Entrez along with microarray gene expression profiles in a unified analytical framework. The ultimate goal of the gene set enrichment analysis is to first generate candidate gene sets each of which is composed of the genes sharing a specific biological function in common and then identify significant ones out of the candidates showing *significantly* differential expression patterns between the two groups. Particularly the genes in a gene set needs to be ordered by a specific ranking method according to the difference in expression profiles between the two groups. For gene ranking, signal-to-noise ratio or fold change based methods are generally used. However, these ranking methods have the weakness that they do not actually consider the amount of signal intensities of gene expression for ranking. Thus, in the case of fold-change gene ranking [2], the genes showing relatively low signal intensities are likely to have larger statistic when they have larger variance, leading to the possible selection of false positive differentially expressed genes (DEGs). In the case of signal-to-noise ratio gene-ranking [1], the genes showing relatively low signal intensities are likely to have larger statistic if they have smaller denominator, i.e. the sum of standard deviations in two groups, even if the difference between signal intensities is insignificant. To handle this problem, we employ the weighted average difference (WAD) method[3] for gene ranking in the gene set enrichment analysis and evaluate its performance with a prostate tumor experiment data set.

2 Weighted Average Difference(WAD) method for Gene-ranking

As a kind of fold-change based gene ranking method, WAD[3] of a gene is computed by the combination of the average difference statistic and the weight term capturing relative average signal intensity. Thus, the WAD based gene ranking method tends to makes it that highly expressed genes are highly ranked on the average for the different conditions. Assuming that a gene set is composed of p genes and their corresponding expression profiles are given for two groups A and B, the WAD statistic for gene ranking is as follows:

$$WAD(i) = AD_i \times w_i = \left(\overline{x_i^B} - \overline{x_i^A} \right) \times \frac{\left((\overline{x_i^A} + \overline{x_i^B}) / 2 \right) - \min}{\max - \min}$$

Here $\overline{x_i^A}$ and $\overline{x_i^B}$ are the means of gene expression profiles in two groups A and B, respectively, for the i^{th} gene. The min and max are the maximum and minimum values in average expression values of p genes. As a gene has a larger absolute value of WAD statistic, it is considered more significant

3 Experiment and Results

For performance evaluation of gene set enrichment analysis by the WAD gene ranking, we used prostate tumor expression profiles produced by Singh et al[4] with Affymetrix gene chip. The prostate tumor experiment data set includes 12,533 gene expression profiles in 52 prostate tumor samples and 50 normal tissue samples. The gene expression data were transformed into log-scaled values and 190 KEGG pathways were generated as candidate gene sets for gene set enrichment analysis by selecting only the gene sets having at least 6 genes among KEGG pathways. Then, by applying two ranking methods of signal-to-noise ratio and WAD statistic for gene ranking, the most 40 significant gene sets were identified. These results were compared with the 13 prostate cancer-related pathways published by Huang D. et al[5]. In particular, out of these 13 pathways, the dorso-ventral axis formation pathway was excluded since it was not included any more in KEGG pathway database. Thus, 12 prostate cancer-related pathways were eventually used to evaluate our results, and F_1 measure[6], which is the harmonic mean of precision and recall, was employed as an evaluation metric. Table 1 is the results of gene set enrichment analysis on the prostate tumor data with signal-to noise ratio gene ranking and WAD gene ranking.

Table 1: Comparison of GSEA performance by signal-to-noise ratio gene ranking and WAD gene ranking

Gene Ranking methods	Precision	Recall	F_1 measure
Signal-to-Noise Ratio	0.075	0.25	0.115
WAD	0.2	0.667	0.308

4 Discussions

In this study we investigated the method of the gene set enrichment analysis using WAD gene ranking method. According to the experimental results, this approach seems to be superior in identifying significant pathways under certain conditions to the typical gene set enrichment analysis approach which uses signal-to noise ratio ranking method. As future works, further analyses with this method needs to be worthwhile in a variety of contexts.

Acknowledgement

This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MEST) (No. R01-2008-000-11089-0).

References

- [1] Subramanian A et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proc. Natl Acad Sci USA*, 102:15545-50, 2005.
- [2] Kim S. Y. et al., PAGE: parametric analysis of gene set enrichment., *BMC Bioinformatics*, 8;6:144, 2005.
- [3] Kadota K. et al., A weighted average difference method for detecting differentially expressed genes from microarray data., *Algorithms for Molecular Biology*, 26;3:8, 2008.
- [4] Singh D. et al., Gene expression correlates of clinical prostate cancer behavior., *Cancer Cell*, 1(2):203-9, 2002.
- [5] Huang D. et al., Identifying the biologically relevant gene categories based on gene expression and biological data: an example on prostate cancer., *Bioinformatics*, 15;23(12):1503-10, 2007.
- [6] Tan P. N. et al., *INTRODUCTION TO DATA MINING*, Addison Wesley, 2006