Review of *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles*

Grace Yi Chen

The two paper this week introduce a powerful analytical method called Gene Set Enrichment Analysis (GSEA), which utilizes prior biological knowledge to interpret gene expression data. Back then, the development of gene expression profiling allowed researchers to compare molecular characteristics of different groups of individuals. However, we could not understand the genes' biological and functional relationship if we analyze the genes individually. In the paper by Mootha et. al (2003), the authors develop a preliminary version of GSEA, which could get the biological insight into the identified genes. The authors apply GSEA when doing microarray data analysis in Type 2 diabetes mellitus (DM2) and identify a set of genes whose expression is correlated with insulin resistance and aerobic capacity. Later in the paper by Subramanian et. al (2005), the authors formalize a robust version of GSEA and generalize it for broader applicability. Specifically, the authors made additional adjustments like adding weights according to each gene's correlation with a phenotype when estimating the enrichment score (ES). Also, GSEA now use FDR instead of FWER to account for multiple hypotheses testing. Moreover, GSEA now could identify the leading-edge subset, which is the key part of high scoring gene sets contributing to ES. The authors apply GSEA in different cancer-related data sets such as cancer cell lines data, acute leukemias etc. and demonstrate its advantage.

These two papers are written well and easy for me to understand the motivation and the general idea of GSEA method, although I don't have much knowledge about the biological pathways. These two articles have a good motivation to develop GSEA. With the development of genomewide expression analysis, we want to know more about how we should interpret the identified genes collectively from the perspective of a biological mechanism. In addition, as is discussed in the paper, genomewide studies usually have limited sample sizes and high variability between individuals, which makes it difficult to distinguish true differences from noise. GSEA can increase the signal-to-noise ratio and detect modest gene expression changes. One thing we need to note is that, GSEA relies on predefined sets of genes, such as the Gene Ontology (GO) or KEGG pathway databases. The knowledge base may be updated so the GSEA results may lead to false positives or false negatives.

Question:

1. I am wondering if GSEA computationally intensive, especially for large datasets or complex gene sets as we need to use permutation test to test for statistical significance?