Nonparametric Multivariate Method to Identify DE Gene Categories Across Two or More Treatments.

Russell Banks

Introduction

- Different technologies can provide “measures” of expression levels for a “large” number of genes.
- Differentially expressed genes (DE genes) across two or more treatments can be identified.
- However, researchers are often more interested in DE gene sets (categories) across treatments.
- Many methods exist to identify DE gene sets, and can be very different in their approach. Their common goal is to *test* whether gene categories consist of genes that are in some sense significantly more differentially expressed than all other genes. These categories are considered to be important on how treatments affect the transcriptional program of the organism under study. (Nettleton)

❑Null Hypotheses (Stevens)

- Competitive null
  - Compare DE of gene set (G) to a standard defined by the set's complement (G^c)
  - H_0^{comp}: The genes in G are at most often DE as the genes in G^c.
  - Less power, stronger statement.

- Self-contained null
  - Compare gene set to a fixed standard
  - H_0^{self}: No genes in G are DE.
  - More powerful to detect differences, more restrictive statement.

❑P-value Calculation Strategies (Stevens)

- Gene sampling
  - Genes are sampling units. The gene labels are permuted (column & row sums constant).
    Gene set: yes/no
    DE: significant/no.

- Subject sampling
  - Subjects are sampling units. The subject (array) labels are permuted.

MRPP Approach

Let T denote the number of treatments, n_i denote the number of independent replications of treatment i ($i = 1, \ldots, T$), and G denote the number of genes in a category of interest. For $i = 1, \ldots, T$ and $j = 1, \ldots, n_i$, let $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijG})'$ denote a vector of expression measurements, where Y_{ijk} represents the expression measurement associated with treatment i, replication j, and gene k. We assume that all $\mathbf{Y}_{ij}$ vectors are independent and that $\mathbf{Y}_{ij}$ has a continuous multivariate distribution F_i. We wish to test

$$H_0 : F_1 = \cdots = F_T \tag{1}$$

against all alternatives. When this null hypothesis is false, the multivariate distribution of the genes in the category of interest is not the same for all treatments. For a completely randomized experimental design, violation of H_0 implies that at least one treatment caused a change in the category's multivariate expression distribution. Thus, categories for which H_0 is false are of potential scientific interest.

Nettleton (2008)

MRPP Approach

To test H_0 against all alternatives, we propose to use the multiresponse permutation procedure (MRPP) described by Mielke and Berry (2001). The MRPP test statistic is given by

$$\bar{D} = \sum_{i=1}^{T} \frac{n_i}{N} D_i, \tag{2}$$

where $N = \sum_{i=1}^{T} n_i$ and D_i is the average of all the Euclidean distances between pairs of data vectors from the ith treatment group, i.e.

$$D_i = \frac{\sum_{j=1}^{n_i-1} \sum_{j'=j+1}^{n_i} ||Y_{ij} - Y_{ij'}||}{n_i(n_i-1)/2}. \tag{3}$$

The MRPP test uses a standard permutation approach to assess the significance of the observed value of $\bigcap_{c=1}^{C} H_0^{(c)}$. The permutation p-value is given by

$$\frac{1}{M} \sum_{m=1}^{M} \mathbb{1}(\bar{D} \geq \bar{D}_m), \tag{4}$$

where $\widehat{\mathrm{FDR}}$ denotes the value of the test statistic computed for the mth of $\bar{D}_m$ possible assignments of treatment labels to the observed data vectors, and $1(\dot{c})$ denotes the indicator function that takes the value 1 if its argument is satisfied and 0 otherwise. As with any permutation test, if the number of data permutations M is too large for timely computation, a randomly selected subset of permutations can be used to obtain an approximate permutation P-value.

Nettleton (2008)

MRPP Approach

- Can detect location shifts in the multivariate expression distribution caused by treatment effects.
- Can detect evidence against the null even when both treatment distributions have identical means, substantial overlap, and differences that are invisible to marginal approaches.
- Heterogeneity of variance among genes must be accounted (genes with larger variability can dominate the test).

Nettleton (2008)

Commensuration

- Different genes will have different levels of variation.
- The data from the kth gene are scaled by:

$$\left\{\sum_{i=1}^{T}\sum_{i'=i}^{T}\sum_{j=1}^{n_i}\sum_{j'=j}^{n_{i'}}(Y_{ijk}-Y_{i'j'k})^2\right\}^{-1/2} \qquad (5)$$

- Euclidean Commensuration as termed by Mielke and Berry (2001), will not remove heterogeneity of variance across treatments within a gene.

- If higher variation genes happen to contain the essential information about treatment differences, emphasis on high variation genes will be well placed, and power for detecting differences will be great with non-commensurated. The MRPP approach with non-commensurated data is very similar to the global test method Goeman *et.al* (2004)

Testing Across Multiple Gene Categories

- The gene sets can't be assumed to be independent because any given gene might be a member of more than one gene set.
- GSEA and SAFE are two testing methods that address the independent assumption violation when testing across multiple gene categories.
- Criticisms of GSEA and SAFE methods Nettleton (2009):
  - Each category statistic is obtained by comparing the gene statistics in a given category to the gene statistics for all genes outside of the category, the subset pivotality condition (controls error rates in multiple testing Yekutieli and Benjamini (1999)) is violated. (*Competitive*)
- MRPP approach calculates statistic for a given category with a function of only the genes within that category. (*Self-Contained Goeman and Buhlmann (2007)*)
- Generally competitive testing good for prominent categories while the self-contained allows better understanding of treatment impact.

ALL Data Example

R-code can be obtained by following the following link:

http://www.public.iastate.edu/~dnett/GeneCategoryAnalysis/GeneCategoryAnalysis.txt

Conclusions

- The strength of the method lies in its ability to test treatment differences within categories rather than difference in enrichment between categories Nettleton (2008)
- MRPP method in simulations to perform as well as other multivariate methods.
- MRPPc (uses standardized data) can be useful to detect difference for low-variability genes.(Genes with high-variability have large influence on GT test statistic Goeman *et al.* (2004))

References

- Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 2007; 23: 980-987.
- Miekle, P., Jr. and Berry, K. (2001) *Permutation methods: A Distance Function Approach.* Springer-Verlag, New York.
- Nettleton, D.,Recknor, J. and Reecy, J. (2008) http://bioinformatics.oxfordjournals.org/content/24/2/192.full?sid=33c6a37b-9f8e-4018-ae39-b4a2f917520a
- Yekutieli, D. and Benjamini, Y. (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J.Stat.Plan Inference,82, 171-196.*