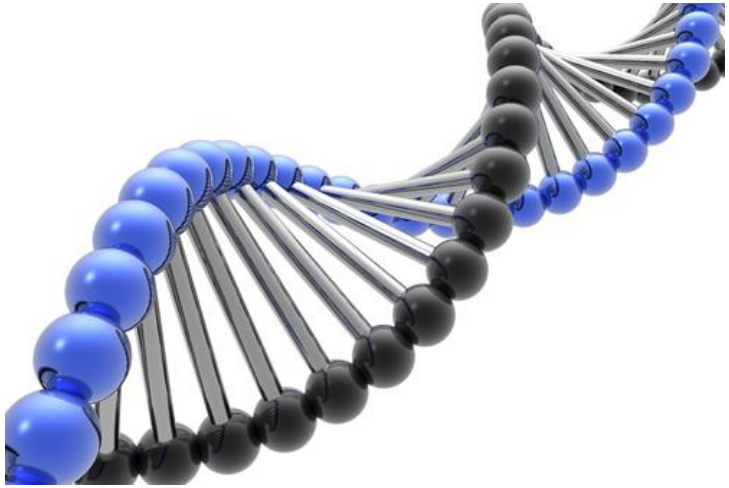
Fisher`s Method to Identify Enriched Gene Sets

Volha Tryputsen[2]

Joint work with Dhammika Amaratunga[1], Javier Cabrera[2] and An de Bondt[1]

[1]-Johnson & Johnson, [2]-Rutgers University



Outline

I. Comparative microarray experiments

II. MLP - p-value based method for identifying enriched gene sets

III. Methods of the approximation to the permutation p-values of MLP statistic

IV. Comparison of different approximation methods

V. Additional work (simulation study)

I. DNA Microarray. Gene expression data

- DNA microarrays are widely used in genomics research to monitor the expression levels of many thousands of genes simultaneously

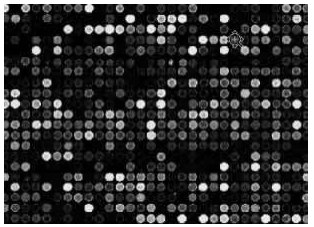
- In a typical microarray experiment, the data is of the form:
  $\{X_{gs}: g=1,\ldots,G; s=1,\ldots,N\}$, where
  g - indexes the genes (rows)
  s - indexes the samples (columns)
  X_{gs} - is a measure of gene expression for gene g in sample s

Example data set

Expression measures for *G* genes in *N* samples:

	C1	C2	C3	T1	T2	T3	…
G1	83	94	82	111	130	122	
G2	16	14	7	2	11	33	
G3	490	879	193	604	1031	962	
G4	46458	49268	74059	44849	42235	44611	
G5	32	70	185	20	25	19	
G6	1067	891	546	906	1038	1098	
G7	118	111	95	896	536	695	
G8	10	30	25	24	31	28	
G9	166	132	162	27	109	213	
G10	136	139	44	62	23	135	
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

45101 rows (genes) x 12 columns (samples)



Stage 1:
Assess quality & preprocess

Stage 2:
Analyze

Comparative microarray experiments

- It is of interest to determine which genes are differentially expressed across different groups of samples. (It doesn't have to be a parallel group design).

- This can be done by testing each gene (row) for a group difference (e.g., do *G t-tests* or a variation like *Ct* or *limma*).This generates a long list of *p*-values: $\{p_g\}$.

Gene set analysis

- Question: How to interpret these results?
  That's up to the biologist, but can statistician help?

- Solution: One way is to see whether we can identify "gene sets" that are "enriched".

- Mapping: Many of the *G* genes (about 50%) can be categorized into pre-defined "gene sets" based on their function or other characteristic.

- Test: whether a particular gene set is "*enriched*".
  A gene set is said to be *enriched* if the *p*-values of the genes that comprise it tend to be smaller than a typical random gene set of the same size.
  Enrichment could imply that the function associated with the gene set is operating differently in the two groups.

II. MLP method

- Calculate MLP statistic for the gene set of a size *n*:

  MLP = mean(-log p)

- Assess significance by drawing a random gene set of a size *n* from the set of all p-value and calculate the value of MLP for the pseudo gene set (call it MLP*); repeat many times.

- The p-value for the gene set is the proportion of times that MLP* equals or exceeds the observed value MLP:

  $p = Pr[MLP^* \geq MLP]$

MLP significance assessment

Permutation

- Randomize genes, rather than samples since we are interested in assessing enrichment in a specific gene set compared to a random gene set
- Could be computationally burdensome if the number of genes is very large

Approximation to the permutation p-value

- Let $X_i = -\log(p_i)$
- MLP statistic = mean (X_i)
- Objective: to derive an approximation for the distribution of the mean of the gene set (MLP statistic)
- By drawing a random gene set of X_i from G we are actually sampling from a finite population of genes
- Could reduce or even avoid the computational effort

III. Approximation Methods

- Normal approximation
- Edgeworth approximation
- Saddlepoint approximation

Normal approximation

- From ***finite population sampling theory*** we can determine the mean and the variance of the Null distribution of the MLP

$$E(MLP) = \mu = \sum_{i=1}^{N} X_i / N \qquad Var(MLP) = \sigma^2 (1 - f)/(N - 1)f$$

where $\sigma^2 = \sum_{i=1}^{N} (X_i - \bar{X})^2 / N$ is the variance of $\{X_i\}$

and $f = n/N$ is the sampling fraction

- Then for gene sets of a size n (when n is large), based on the ***Central Limit Theorem*** Z_{MLP} is a standardized version of MLP and

$$Z_n = \sqrt{(N-1)}\, f\, (MLP-\mu)/(\sigma \sqrt{(1-f)}) \sim N(0,1)$$

and the p-value for the significance of a gene set is

$$p = 1 - P[Z_n \leq z_{obs.}] = 1 - \Phi(z_{obs.})$$

Normal approximation cont.

Advantages

- Reducing computational time by avoiding permutations

Disadvantages

- Since the distribution of {-log(p)} is very skewed and heavy tailed, convergence to normality will be slow
- Therefore, this approximation is most likely to be inadequate for small gene sets; since there are often many of them, we need a somewhat better approximation method

Edgeworth approximation

- Incorporates the ***skewness*** and ***kurtosis***

$$p = 1 - P[Z_n \leq z_{obs}]$$

$$= 1 - \Phi(z_{obs}) - \frac{p_1(z_{obs})\phi(z_{obs})}{\sqrt{n}} - \frac{p_2(z_{obs})\phi(z_{obs})}{n}$$

where

p_1 and p_2	are functions of $z_{obs.}$, corrected for skewness and kurtosis;
$\Phi(z_{obs.})$	standard normal distribution function;
$\varphi(z_{obs.})$	standard normal density function

Edgeworth approximation cont.

Advantages

- Adjusts for skewness and kurtosis
- Offers improved over Normal method accuracy of the results
- Reduces computational time by avoiding permutations

Disadvantages

- Slightly more complicated than Normal approximation
- The accuracy of approximation sometimes depends on gene set size

Saddlepoint approximation

$$p = 1 - G_n(x) = 1 - \Phi(w_x) - \phi(w_x)(w_x^{-1} - z_x^{-1})$$

where $w_x = [2n\{t_x x - R_n(t_x)\}]^{\frac{1}{2}} sgn(t_x)$

$z_x = t_x\{nR''(t_x)\}^{\frac{1}{2}}$

Φ and φ *are standard normal distribution function and density respectively*

Saddlepoint approximation cont.

Advantages

- Gives better than Edgeworth approximation, especially in the tails of the distribution
- Accuracy of approximation holds for small gene set sizes

Disadvantages

- Mathematically more complicated than Normal and Edgeworth methods
- Has some stability issues when the distribution of $-\log(p)$ is skewed or has long tails

IV. Comparison of different approximation methods

- To assess gene set significance permutation method was used, followed by Edgeworth and Saddlepoint approximations

- We studied the performance of the above methods on a few different datasets

Example: SLC17A5 (Day 18) data

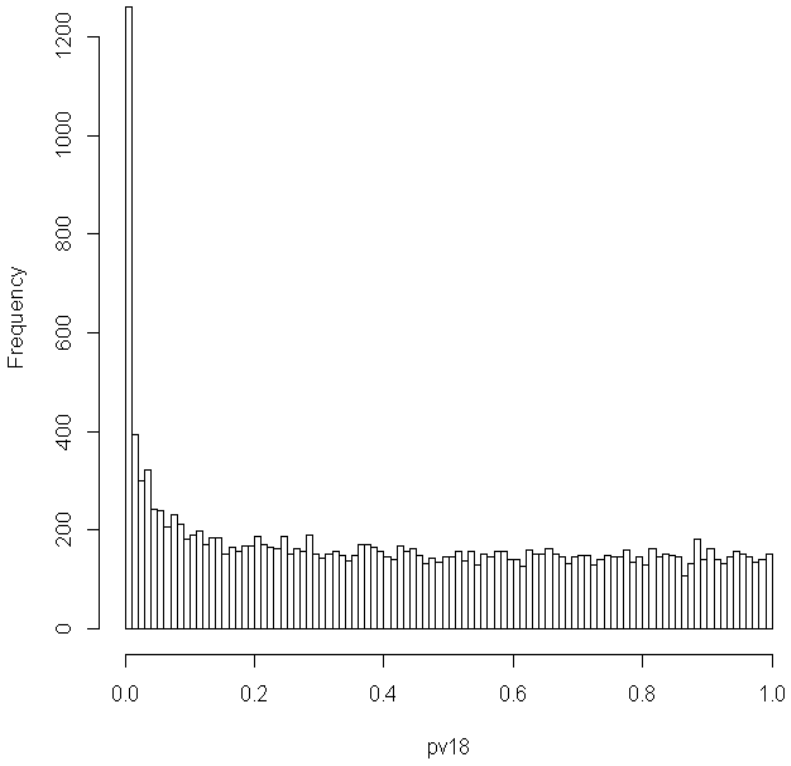
- Experiment: SLC17A5 “Day 18” data comes from an experiment that studied gene expression profiles of 2 groups of 18-day-old mice

- Groups and samples: 6 RNA samples were from the wild type (WT) mice and 6 were from mice whose Slc17A5 gene had been knocked out (KO)

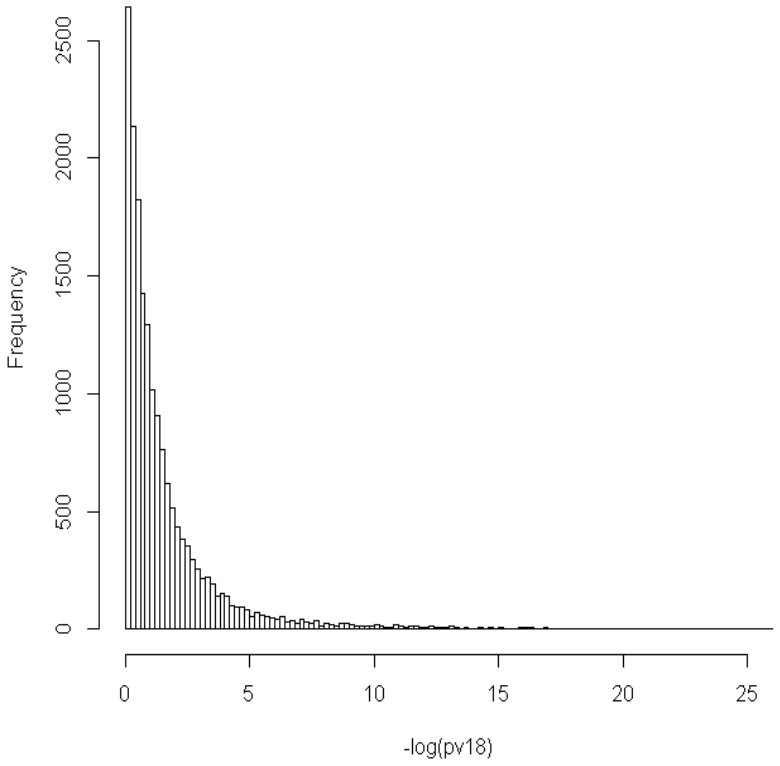
- Genes: 17370 genes were tested for differential expression using limma (linear models for microarray data). The result is a set of 17370 p-values

Data specifics

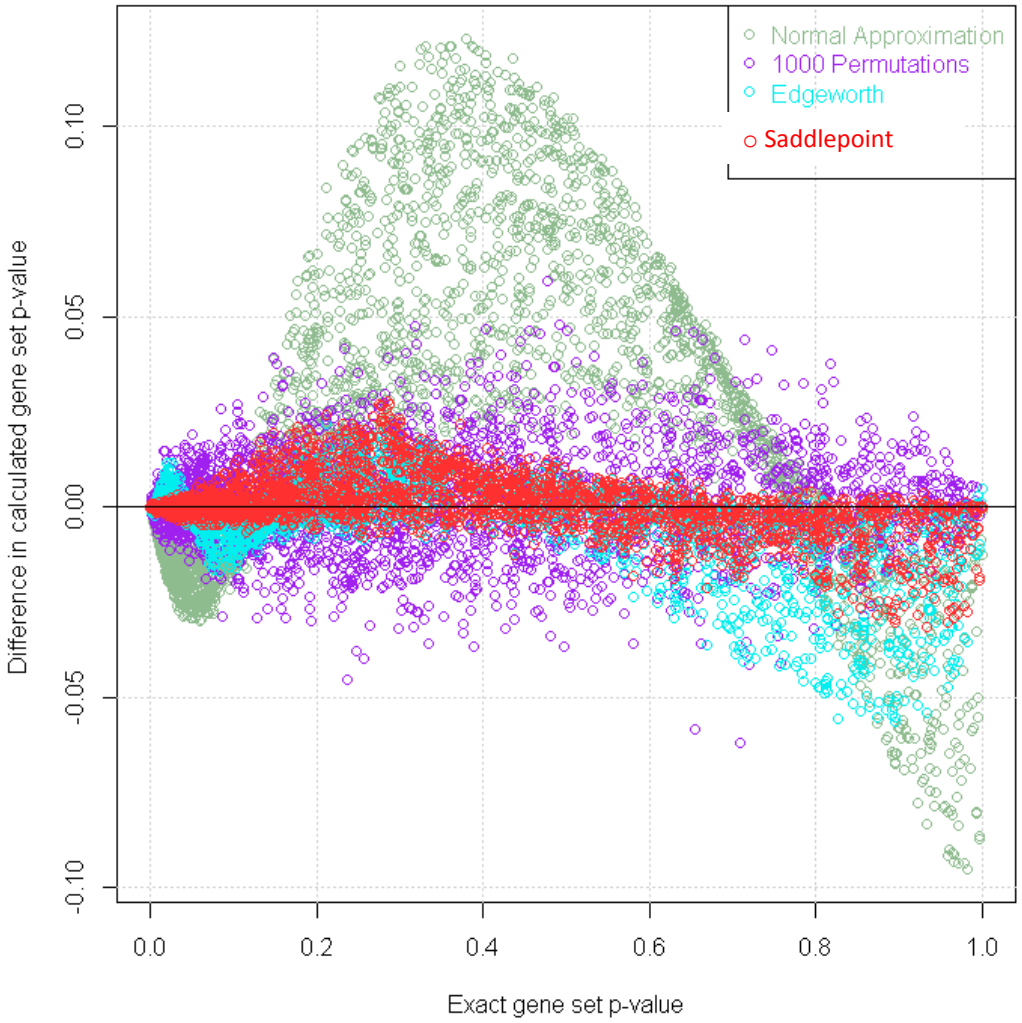
Gene p-values distribution



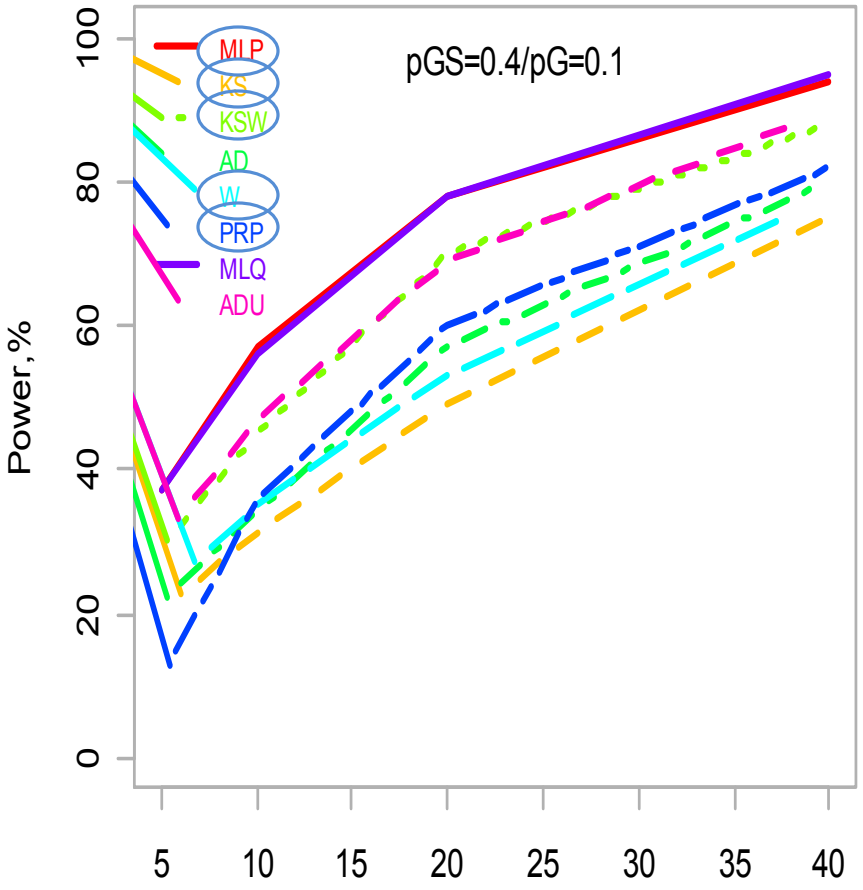
Gene -log(p) distribution



Slc17A5 “Day 18”



V. Additional work (simulation study)



- A simulation study was conducted to compare performance of some popular methods for gene set analysis (MLP, KS, KSW, W, PRP) in various settings
- We estimated a "detection power" – the ability of a GSA test to detect a signal and plotted it against gene set size

Summary

- The MLP statistic, MLP = mean(-log(p)), is essentially Fisher`s test statistic for pooling p-values
- Proposed approximation methods, based on Edgeworth and Saddlepoint approximations, are attractive alternatives to the computationally heavy permutation-based method for gene set enrichment assessment
- For gene set analysis, the MLP statistic has higher efficiency than either the modified KS statistic (which is the basis of GSEA) or Fisher`s exact test (which is the basis of many software packages for GSA)
- To determine significance, genes (rather than samples) are randomized, since we are interested in assessing enrichment in a specific gene set compared to a random gene set from the same system (rather than in assessing significance; i.e., the presence or absence of differential expression in that gene set)
- *Amaratunga, Cabrera, De Bondt and Tryputsen (in review, 2012)*

Thank you!