A False Discovery Rate approach to separate the score distributions of induced and non-induced genes

Stefanie Scheid

Max-Planck-Institute for Molecular Genetics
Computational Diagnostics
Ihnestrasse 63-73, D-14195 Berlin, Germany
stefanie.scheid@molgen.mpg.de
Microarray basics

DNA $\rightarrow$ mRNA $\rightarrow$ protein.

High demand for protein $\Rightarrow$ High corresponding gene expression.

Microarray measures gene expression for thousands of genes simultaneously.
Microarray Experiment

Two classes of disease status, set of patients for each class.

Question: Are there differences in gene expression between classes?

Genes showing differences are called differentially expressed, up/downregulated, induced.
How do we detect induced genes?

Statistically: Assign score for differential gene expression and test for significance.

Simplify: Test for difference in mean gene expression between classes.

Possible scores: t-test statistic, Wilcoxon ranksum score.
Overall score distribution is mixture of score distributions of induced and non-induced genes:
Overall score distribution is mixture of score distributions of induced and non-induced genes:
Overall score distribution is \textit{mixture} of score distributions of induced and non-induced genes:
Score distribution with twilight zone

What if score distributions overlap?
Score distribution with twilight zone

What if score distributions overlap?

Wilcoxon score

Density

700 800 900 1000 1100 1200

0.000 0.002 0.004 0.006

Wilcoxon score

Density

700 800 900 1000 1100 1200

0.000 0.002 0.004 0.006
Score distribution with twilight zone

What if score distributions overlap?

non-induced

?  non-induced
How can we reconstruct the mixture?

Significance testing causes multiplicity problem.

**False Discovery Rate (FDR):**

- Expected proportion of falsely called induced genes among all genes called induced

= Probability of genes in rejection area to be non-induced.
From extremal FDR to bin-wise FDR

Define rejection area by threshold values and estimate FDR:
From extremal FDR to bin-wise FDR

Define rejection area by threshold values and estimate FDR:

\[ \hat{FDR} = 4\% \]
From extremal FDR to bin-wise FDR

Define rejection area by threshold values and estimate FDR:
From extremal FDR to bin-wise FDR

Define rejection area by threshold values and estimate FDR:
From extremal FDR to bin-wise FDR

Define rejection area by threshold values and estimate FDR:
From extremal FDR to bin-wise FDR

Define rejection area by threshold values and estimate FDR:

Wilcoxon score
Density
700 800 900 1000 1100 1200
0.000 0.002 0.004 0.006

15% 96% 99% 35% 98% 13%
85% 4% 1% 65% 2% 87%

Wilcoxon score

Stefanie Scheid – FDR approach to separate score distributions – DSC 2003
Estimating bin-wise FDR

Define useful binning.

Estimate FDR as in Tusher et al. (2001) using class permutation:

\[ N_i = \text{Number of observed scores in bin } i \]

\[ N_i^p = \text{Number of scores of } p\text{th permutation in bin } i \]
For bin \( i \):

\[
FDR_i = \text{Prob}(\text{non-induced}|\text{"induced"})
\]

\[
= \frac{\text{Prob}(\text{"induced"}|\text{non-induced})}{\text{Prob}(\text{"induced"})} \text{Prob}(\text{non-induced})
\]
For bin $i$:

\[
FDR_i = \text{Prob}(\text{non-induced}|\text{“induced”}) \\
= \frac{\text{Prob}(\text{“induced”}|\text{non-induced})}{\text{Prob}(\text{“induced”})} \text{Prob}(\text{non-induced}) \\
\overline{FDR}_i = \frac{\text{median}(N^p_i)}{N_i} \text{Prob}(\text{non-induced})
\]
For bin $i$:

$$FDR_i = \text{Prob}(\text{non-induced}|\text{“induced”})$$

$$= \frac{\text{Prob}(\text{“induced”}|\text{non-induced})}{\text{Prob}(\text{“induced”})} \text{Prob}(\text{non-induced})$$

$$\widehat{FDR}_i = \frac{\text{median}(N^p_i)}{N_i} \text{Prob}(\text{non-induced})$$

requires prior knowledge or good estimator
Calculate lower and upper quartile of all permutation scores.

\[
\hat{Prob}(\text{non-induced}) = \frac{\text{Number of observed scores in } [q_{25}, q_{75}]}{0.5 \cdot \text{Total number of observed scores}}
\]
Calculate lower and upper quartile of all permutation scores.

\[ \hat{\text{Prob}}(\text{non-induced}) = \frac{\text{Number of observed scores in } [q_{.25}, q_{.75}]}{0.5 \cdot \text{Total number of observed scores}} \]

\[ \Rightarrow \hat{\text{FDR}}_i = \frac{\text{median}(N^p_i)}{N_i} \hat{\text{Prob}}(\text{non-induced}) \]
Simulation study

Two classes with 30 samples each, 1000 genes, 10 000 permutations, 20 bins.

Two features of (non-induced) gene expression:


2. Correlation due to pathways/coregulation, “clumpy dependence”: Add same standard normal error to blocks of 50 genes.
Induce a fraction $\pi$ of genes in one class with mean offset $\mu$ from $\mathcal{N}(\mu, \sigma = 0.2)$.

$\pi = 5, 15, 25, 50\%$

$\mu = 0.5, 0.7$
Induce a fraction $\pi$ of genes in one class with mean offset $\mu$ from $N(\mu, \sigma = 0.2)$.

$\pi = 5, 15, 25, 50\%$

$\mu = 0.5, 0.7$

Repeat each parameter combination 10 times $\Rightarrow$

Averaged $\overline{FDR}_i$ and averaged true proportion of non-induced genes.
\[ \pi_{\text{true}} = 15\% \quad \pi_{\text{est.}} = 11.98\% \quad \mu = 0.5 \]
\[ \pi_{\text{true}} = 15\% \quad \pi_{\text{est.}} = 11.98\% \quad \mu = 0.5 \]

Stefanie Scheid – FDR approach to separate score distributions – DSC 2003
\[ \pi_{\text{true}} = 25\% \quad \pi_{\text{est.}} = 20.54\% \quad \mu = 0.5 \]
\[ \pi_{\text{true}} = 50\% \quad \pi_{\text{est.}} = 48.8667\% \quad \mu = 0.7 \]
Further research

Simulation under several dependence structures, downregulation.

Improve $\widehat{Prob}(\text{non-induced})$.

Application to biological data.
FDR references

