Reconstructing score distribution mixtures with Bin-wise False Discovery Rates

Stefanie Scheid and Rainer Spang

Max-Planck-Institute for Molecular Genetics, Berlin

- Computational Molecular Biology -

http://cmb.molgen.mpg.de

E-Mail: stefanie.scheid@molgen.mpg.de

Microarray experiments

Some basics:

- Measurement of abundance of mRNA transcripts in given tissue.
- Simultaneously, for up to 20000 genes.
- Question: Which genes are differentially expressed ("induced") if we compare two classes of tissue, disease, patients?

From a statistical point of view...

- Compare mean gene expression of classes.

⇒ Compute **Wilcoxon's ranksum score** for each gene and check for significance.

Score distributions

- Wilcoxon score distribution of only non-induced genes is unimodal and symmetrical.
- Induced genes show scores in distributional margins.
- Observed score distributions are **mixtures** of induced and non-induced scores.

Shown below are two score distributions where a certain fraction of genes is upregulated. Left side: Induced scores form a small hill easily separable by eye from the bulk of non-induced genes. Right side: Genes are slightly induced and thus are hidden in a **twilight zone** only recognizable as a subtle elevation.

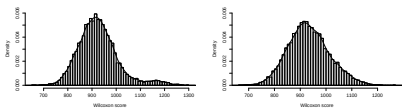

Figure 1: Two mixtures of induced and non-induced scores.

Reconstructing mixtures

- Twilight zones are informative for biological research.
- Not possible to detect induced genes hidden among non-induced, but...
- Possible to estimate the **percentage** of induced and non-induced genes in score intervals (bins) with theory of **False Discovery Rates (FDR)**.

⇒ Reconstruction of mixture in each bin.

⇒ Reconstruction of overall mixture using narrow bins.

FDR concepts

Classical FDR:

- Due to multiple testing situation in microarray experiments, see e.g. Benjamini and Hochberg (1995), Storey and Tibshirani (2001), Reiner et al. (2003).
- Given score threshold, genes with score exceeding threshold are called induced.
- FDR is defined as percentage of falsely called induced genes among all genes called induced, given extremal score threshold above.

⇒ FDR is proportion of non-induced genes and thus **reconstructs the mixture** in given rejection area.

Bin-wise FDR:

- Extension of FDR concept by dividing range of scores into bins.
- Define each bin separately as rejection area.

⇒ Bin-wise reconstruction of overall mixture (see Figure 2 as illustration).

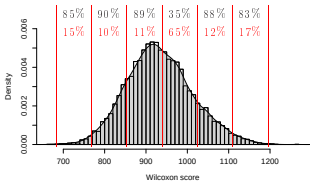

Figure 2: Reconstructed mixture with bin-wise FDR = percentage of non-induced genes (black) and resulting percentage of induced genes (red), illustration.

Simulation study

- Advantage of simulated data: Knowledge of **true** score distributions.
- Generate log expression data by using a lognormal model with normal errors and coregulation structure (clumpy dependence).
- Induce fraction π of genes in one class by adding a normally distributed offset with mean value μ.

Estimation:

- Use bins with approx. equal bin frequency as rejection areas, here 20 bins.
- Estimate bin-wise FDR by using class permutation and extending classical FDR estimate as given in Tusher et al. (2001)

Results

Shown below are stairplots of a simulation where 15% of genes were induced with mean offset 0.5 (slight induction) or 0.7 (moderate induction). The height of each stair denotes averaged percentage of non-induced genes after 10 simulations, the width corresponds to bin width.

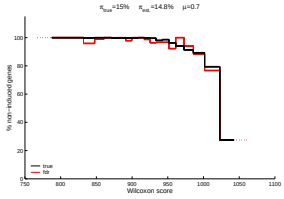

Figure 3: Bin-wise true (black) and estimated (red) percentages of non-induced genes for simulation with $\pi = 15\%$ and $\mu = 0.7$.

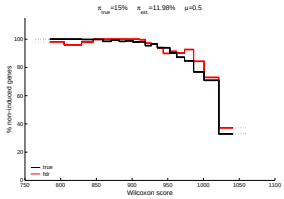

Figure 4: Bin-wise true (black) and estimated (red) percentages of non-induced genes for simulation with $\pi = 15\%$ and $\mu = 0.5$.

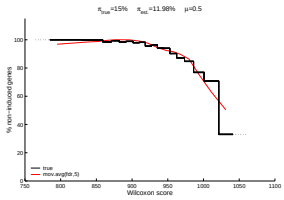

Figure 5: Bin-wise true percentages (black) of non-induced genes for simulation with $\pi = 15\%$ and $\mu = 0.5$. Red line denotes moving average of estimated percentages (window size 5).

References

Benjamini, Y., Hochberg, Y. 1995. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society B*, 57(1)289–300.

Reiner, A., Yekutieli, D., Benjamini, Y. 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–375.

Storey, J. D., Tibshirani, R. 2001. *Estimating False Discovery Rates Under Dependence, with Applications to DNA Microarrays*, Stanford University: Stanford Technical Report.

Tusher, V. G., Tibshirani, R., Chu, G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121.