Some statistical methods for the identification of differentially expressed genes

Milan, May 2003

Anestis Antoniadis

Laboratoire IMAG-LMC
University Joseph Fourier
Grenoble, France

Introduction

- **Whats special about microarray data**

Introduction

- **Whats special about microarray data**
- thousands of variables (genes, p) and a very small number of (biological) replicates

Introduction

- **Whats special about microarray data**
- thousands of variables (genes, p) and a very small number of (biological) replicates
- variables (genes) are highly (?) correlated and data does not follow normal distribution (many outliers, heavy tails)

Introduction

- **Whats special about microarray data**
- thousands of variables (genes, p) and a very small number of (biological) replicates
- variables (genes) are highly (?) correlated and data does not follow normal distribution (many outliers, heavy tails)
- **selecting a subset**

Introduction

- **Whats special about microarray data**
- thousands of variables (genes, p) and a very small number of (biological) replicates
- variables (genes) are highly (?) correlated and data does not follow normal distribution (many outliers, heavy tails)
- **selecting a subset**
- select a statistic for “differential expression” and a cut off / quality measure for subset selection

Introduction

- **Whats special about microarray data**
- thousands of variables (genes, p) and a very small number of (biological) replicates
- variables (genes) are highly (?) correlated and data does not follow normal distribution (many outliers, heavy tails)
- **selecting a subset**
- select a statistic for “differential expression” and a cut off / quality measure for subset selection
- **Multiple Hypothesis Testing**

Introduction

- **Whats special about microarray data**
- thousands of variables (genes, p) and a very small number of (biological) replicates
- variables (genes) are highly (?) correlated and data does not follow normal distribution (many outliers, heavy tails)
- **selecting a subset**
- select a statistic for "differential expression" and a cut off / quality measure for subset selection
- **Multiple Hypothesis Testing**
- Empirical Bayes Thresholding, Bonferroni, FDR and adjusted p-values

Introduction

- **Whats special about microarray data**
- thousands of variables (genes, p) and a very small number of (biological) replicates
- variables (genes) are highly (?) correlated and data does not follow normal distribution (many outliers, heavy tails)
- **selecting a subset**
- select a statistic for “differential expression” and a cut off / quality measure for subset selection
- **Multiple Hypothesis Testing**
- Empirical Bayes Thresholding, Bonferroni, FDR and adjusted p-values
- Enhancing FDR by wavelets

DNA Microarrays

- Exciting new technology for measuring gene expression of thousands of genes simultaneously in a single sample of cells.

DNA Microarrays

- Exciting new technology for measuring gene expression of thousands of genes simultaneously in a single sample of cells.
- A multivariate quantitative way of measuring gene expression.

DNA Microarrays

- Exciting new technology for measuring gene expression of thousands of genes simultaneously in a single sample of cells.
- A multivariate quantitative way of measuring gene expression.
- The data can be organized into an $m \times n$ matrix Y, where m is the number of genes and n is the number of microarrays. The (i, j) entry of Y, say y_{ij} , is the expression level for the ith gene on the jth microarray.

DNA Microarrays

- Exciting new technology for measuring gene expression of thousands of genes simultaneously in a single sample of cells.
- A multivariate quantitative way of measuring gene expression.
- The data can be organized into an $m \times n$ matrix Y, where m is the number of genes and n is the number of microarrays. The (i, j) entry of Y, say y_{ij} , is the expression level for the ith gene on the jth microarray.
- Several technologies – cDNA array, oligonucleotide array, . . .

Notation

Basic measurement: one each array, for each spot, we have: y the log-ratio of the intensities of the (background corrected and normalized) red and green channels.

Number of measurements:

- $i = 1, \ldots, m$ genes
- $\ell = 1, \ldots, E$ experiments (conditions)
- $k = 1, \ldots, K$ slides per experiment (so $n = KE$)

Data: $y_{i,(k,\ell)}$ is log-ratio of intensities for gene i on slide k of the experiment ℓ.

Microarray Data

n arrays, m genes

	array 1	array 2	array 3	$\cdots$	array n
gene 1	1.23	-2.61	-3.87	$\cdots$	5.26
gene 2	3.89	-0.76	1.73	$\cdots$	-2.43
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
gene m	0.846	3.78	1.37	$\cdots$	-2.94

Question

Statistical inference: Suppose that we have n_1 microarrays taken from untreated cells and n_2 microarrays taken from treated cells (e.g., untreated=normal, treated=cancer); $n_1 + n_2 = n$. Which genes show a statistically significant difference in gene expression between these two types of cells? Answering this question helps to narrow down the search for genes involved in differentiating these cell types.

Notation:
θ_i indicates for every gene the change in expression between the two conditions.

Question: What are the values of θ_i? Which of them are different from 0?

Empirical Bayes thresholding models

The object of interest is a sequence of parameters θ_i on each of which we have a single observation Y_i subject to noise, so that $Y_i = \theta_i + \epsilon_i$ where the ϵ_i's are $N(0, \sigma)$ random variables.

Without some knowledge of the θ_i we are not going to be able to estimate them very efficiently. The method implemented in the package `EbayesThresh` developed by Johnstone and Silverman (2002) takes advantage of possible sparsity in the sequence.

A natural approach to this problem is thresholding: if the absolute value of a particular Y_i exceeds some threshold t then it is taken to correspond to a nonzero θ_i which is then estimated, most simply by Y_i itself. If $|Y_i| < t$ then the coefficient θ_i is estimated to be zero.

A Bayesian approach

Original motivation: function estimation via wavelets.

Within a Bayesian context:

sparsity $\Longleftrightarrow$ suitable prior distribution for the θ_i's.

Model: The θ_i's have independent prior distributions each given by the mixture

$$f_{\text{prior}}(\theta) = (1-w)\delta_0(\theta) + w\gamma(\theta).$$

The nonzero part of the prior, γ, is assumed to be a fixed unimodal symmetric density. Particular possibilities for the function are the *Laplace* or the *quasi–Cauchy* distribution, for which the procedures are entirely feasible computationally.

Thresholding idea

θ with the previous prior and $Y \sim N(\theta, \sigma)$. Find the posterior distribution of θ conditional on $Y = y$. Let $\hat{\theta}(y, w)$ be the median of this posterior distribution:

for any fixed w, the estimation rule $\hat{\theta}(y, w)$ will be a monotonic function of y with the thresholding property, i.e. there exists $t(w) > 0$ such that

$$\hat{\theta}(y, w) = 0 \text{ if and only if } |y| \leq t(w).$$

Once w has been specified, there are other possible estimation rules, for example the posterior mean $\tilde{\theta}(y, w)$ of θ given $Y = y$, or hard or soft thresholding with threshold $t(w)$.

Choice of w

Very important to make a good choice of mixing weight w, or equivalently of threshold $t(w)$.
JS approach is an Empirical Bayes: use the data once to obtain the estimate $\hat{w}$ by marginal maximum likelihood.
The same approach is used to estimate other parameters of the prior.

When the variance of the data is not known, then the package allows for its estimation from the median absolute deviation from zero. Provided the sequence θ_i is reasonably sparse, the median of the absolute deviations will not be affected by those observations that have nonzero means θ_i.

Multiple hypothesis testing

- We conduct a statistical test for each gene $g = 1, \ldots, m$ (t-test, Wilcoxon test, permutation test, $\ldots$).

Multiple hypothesis testing

- We conduct a statistical test for each gene $g = 1, \ldots, m$ (t-test, Wilcoxon test, permutation test, …).
- This yields test statistics T_g, the rejection regions and p-values p_g.

Multiple hypothesis testing

- We conduct a statistical test for each gene $g = 1, \ldots, m$ (t-test, Wilcoxon test, permutation test, …).
- This yields test statistics T_g, the rejection regions and p-values p_g.
- p_g is the probability under the null hypothesis that the test statistic is at least as extreme as T_g. Under the null hypothesis,

$$\mathbb{P}(p_g < \alpha) = \alpha.$$

Statistical tests: Examples

- t-test: assumes homoscedastic normally distributed data in each class

Statistical tests: Examples

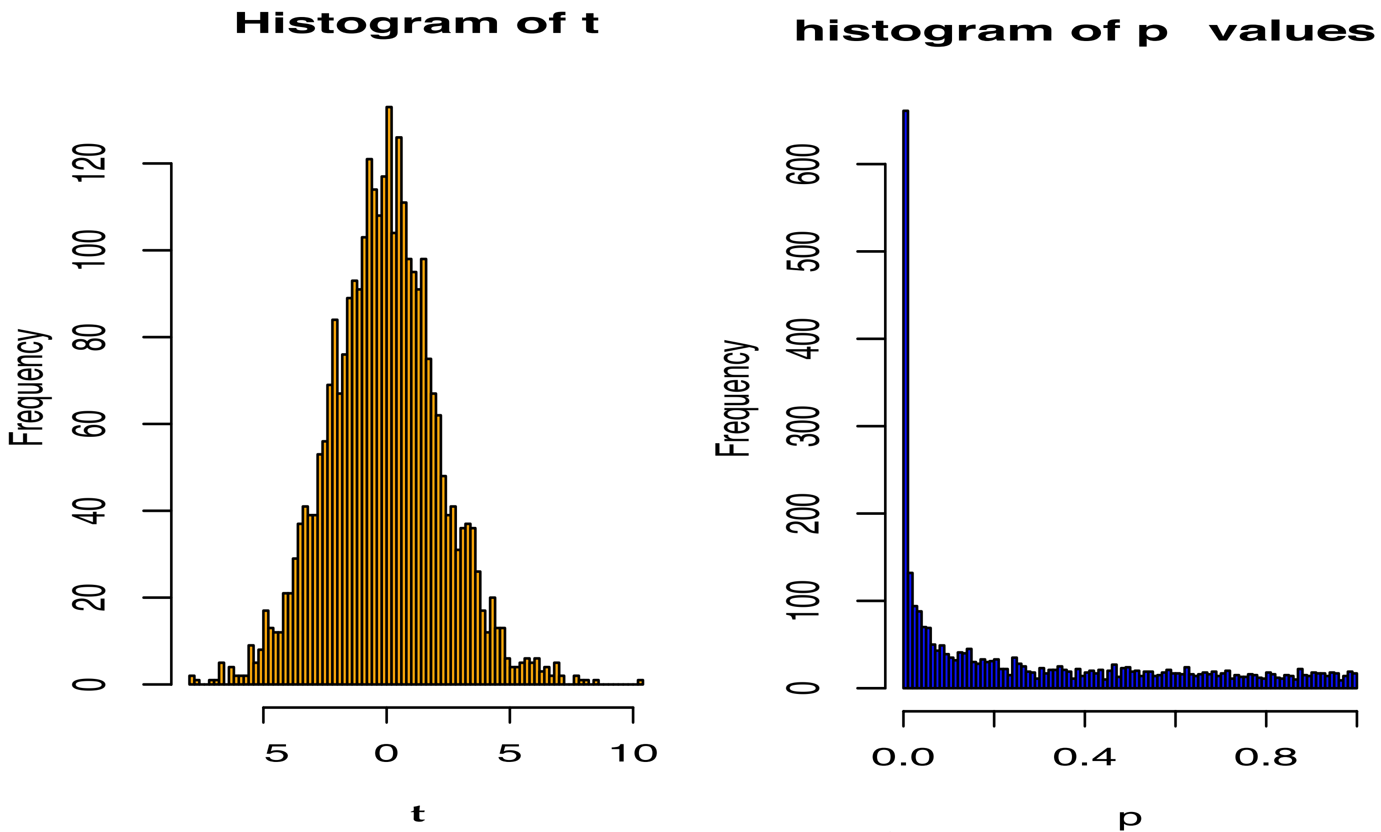
- t-test: assumes homoscedastic normally distributed data in each class
- Wilcoxon test: nonparametric, rank–based

Statistical tests: Examples

- t-test: assumes homoscedastic normally distributed data in each class
- Wilcoxon test: nonparametric, rank–based
- permutation test: estimate the distribution of the test statistic (e.g., the t-statistic) under the null hypothesis by permutations of the sample labels:
  The p–value p_g is given as the fraction of permutations yielding a test statistic that is at least as extreme as the observed one.

Example

Golubdata, 27 ALLvs. 11 AMLsamples, 3,051 genes.



t-test: 1045 genes with$p < 0.05$.

Golub data

Multiple testing: the problem

Problem: thousands of hypotheses are simultaneously tested.

- Increased chance of false positives. E. g. suppose you have 10,000 genes on a chip and not a single one is differentially expressed. You would expect $10000 \times .01 = 100$ of them to have a p–value < 0.01.

Multiple testing: the problem

Problem: thousands of hypotheses are simultaneously tested.

- Increased chance of false positives. E. g. suppose you have 10,000 genes on a chip and not a single one is differentially expressed. You would expect $10000 \times .01 = 100$ of them to have a p–value < 0.01.
- Individual p–values of e.g. 0.01 no longer correspond to significant findings.

Multiple testing: the problem

Problem: thousands of hypotheses are simultaneously tested.

- Increased chance of false positives. E. g. suppose you have 10,000 genes on a chip and not a single one is differentially expressed. You would expect $10000 \times .01 = 100$ of them to have a p–value < 0.01.
- Individual p–values of e.g. 0.01 no longer correspond to significant findings.

Need to adjust for multiple testing when assessing the statistical significance of findings.

Multiple hypothesis testing

Outcomes when testing m hypotheses:

	Accept	Reject	Total
Null True	U	V	m_0
Alternative True	T	S	m_1
	W	R	m

Error measures

- FamilyWise Error Rate (FWER). The FWER is defined as the probability of at least one Type I error (false positive): $FWER = \mathbb{P}(V > 0)$.

Error measures

- FamilyWise Error Rate (FWER). The FWER is defined as the probability of at least one Type I error (false positive): $FWER = \mathbb{P}(V > 0)$.
- False discovery rate (FDR). The FDR (Benjamini & Hochberg 1995) is the expected proportion of Type I errors among the rejected hypotheses, including cases where no hypotheses are significant: $FDR = \mathbb{E}\{\frac{V}{R}|R > 0\}\mathbb{P}(R > 0)$.

Error measures

- FamilyWise Error Rate (FWER). The FWER is defined as the probability of at least one Type I error (false positive): $FWER = \mathbb{P}(V > 0)$.
- False discovery rate (FDR). The FDR (Benjamini & Hochberg 1995) is the expected proportion of Type I errors among the rejected hypotheses, including cases where no hypotheses are significant: $FDR = \mathbb{E}\{\frac{V}{R}|R > 0\}\mathbb{P}(R > 0)$.
- Positive false discovery rate (pFDR). The pFDR (Storey 2001) is the expected proportion of Type I errors among the true rejected hypotheses, considering only cases where at least one significant hypothesis is found: $pFDR = \mathbb{E}\{\frac{V}{R}|R > 0\}$.

Interpretation

- The FDR includes cases where no hypotheses are significant – the ”proportion” is set to zero.

Interpretation

- The FDR includes cases where no hypotheses are significant – the ”proportion” is set to zero.
- The pFDR only considers cases where at least one significant hypothesis is found.

Interpretation

- The FDR includes cases where no hypotheses are significant – the ”proportion” is set to zero.
- The pFDR only considers cases where at least one significant hypothesis is found.
- If a procedure is applied to call hypotheses significant, then a pFDR of 5%, for example, says that on average the proportion of false positives among significant hypotheses is 5%.

Interpretation

- The FDR includes cases where no hypotheses are significant – the ”proportion” is set to zero.
- The pFDR only considers cases where at least one significant hypothesis is found.
- If a procedure is applied to call hypotheses significant, then a pFDR of 5%, for example, says that on average the proportion of false positives among significant hypotheses is 5%.
- Loosely . . . if we find 100 significant genes under some method with a pFDR of 5%, then we expect about 5 false positive genes.

Controlling a type I error rate

- Aim: For a given type I error rate α, use a procedure to select a set of ”significant” genes that guarantees a type I error rate α.

Controlling a type I error rate

- Aim: For a given type I error rate α, use a procedure to select a set of ”significant” genes that guarantees a type I error rate α.
- The type I error is defined with respect to a given configuration of true and false null hypotheses.

Controlling a type I error rate

- Aim: For a given type I error rate α, use a procedure to select a set of ”significant” genes that guarantees a type I error rate α.
- The type I error is defined with respect to a given configuration of true and false null hypotheses.
- Weak control of type I error: only under the assumption that all null hypotheses are true (complete null hypothesis, H_0).

Controlling a type I error rate

- Aim: For a given type I error rate α, use a procedure to select a set of ”significant” genes that guarantees a type I error rate α.
- The type I error is defined with respect to a given configuration of true and false null hypotheses.
- Weak control of type I error: only under the assumption that all null hypotheses are true (complete null hypothesis, H_0).
- Strong control of type I error: for all possible configurations of true and false null hypotheses.

FWER Procedures

Without loss of generality, we can assume the tests are performed with p–values $p_1, \ldots, p_m$ and rejection regions of the form $[0, t]$ for $0 < t \leq 1$.

- The Bonferroni correction: controls FWER; very conservative.

FWER Procedures

Without loss of generality, we can assume the tests are performed with p–values $p_1, \ldots, p_m$ and rejection regions of the form $[0, t]$ for $0 < t \leq 1$.

- The Bonferroni correction: controls FWER; very conservative.
- Westfall/Young procedures (Hochberg - Holm): Improves Bonferroni by resampling.

FWER Procedures

Without loss of generality, we can assume the tests are performed with p–values $p_1, \ldots, p_m$ and rejection regions of the form $[0, t]$ for $0 < t \leq 1$.

- The Bonferroni correction: controls FWER; very conservative.
- Westfall/Young procedures (Hochberg - Holm): Improves Bonferroni by resampling.
- Westfall/Young step–down multiple testing procedures via permutations of adjusted p–values (Dudoit et al. 2000).

FWER Procedures

Without loss of generality, we can assume the tests are performed with p–values $p_1, \ldots, p_m$ and rejection regions of the form $[0, t]$ for $0 < t \leq 1$.

- The Bonferroni correction: controls FWER; very conservative.
- Westfall/Young procedures (Hochberg - Holm): Improves Bonferroni by resampling.
- Westfall/Young step–down multiple testing procedures via permutations of adjusted p–values (Dudoit et al. 2000).

All methods are implemented in the Bioconductor package `multtest`, with fast algorithms.

Bonferroni correction

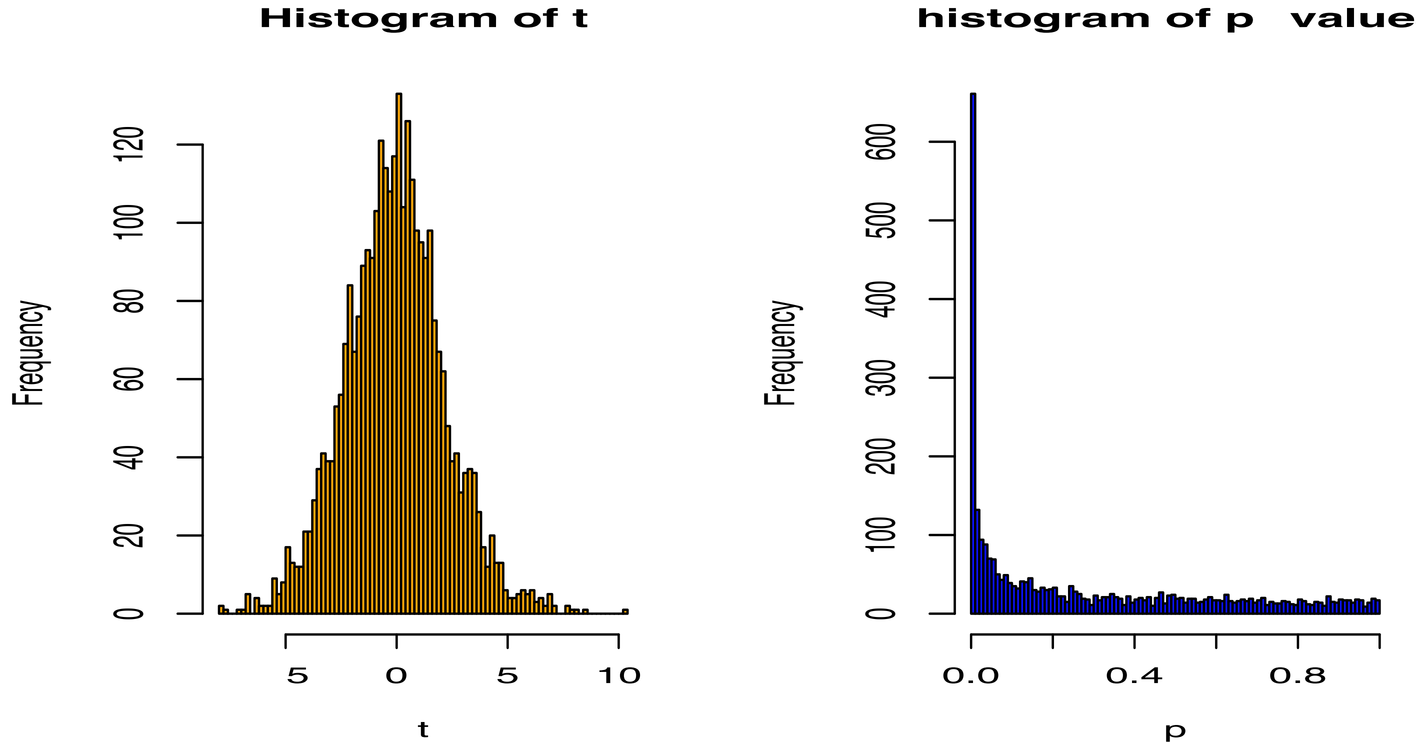
Use $\hat{p}_i \leq \alpha$ where $\hat{p}_i = \min(mp_i, 1)$

$$\begin{aligned} FWER = \mathbb{P}(V > 0) &= \mathbb{P}(\text{at least one } \hat{p}_i \leq \alpha | H_0) \\ &= \mathbb{P}(\text{at least one } p_i \leq \alpha/m | H_0) \\ &\leq \sum_{i=1}^{m} \mathbb{P}(p_i \leq \alpha/m | H_0) \\ &= m \cdot \alpha/m = \alpha. \end{aligned}$$

H_0 denotes the complete null hypothesis that no gene is differentially expressed.

Bonferroni

Golub data, 27 ALL vs 11 AML samples, 3,051 genes

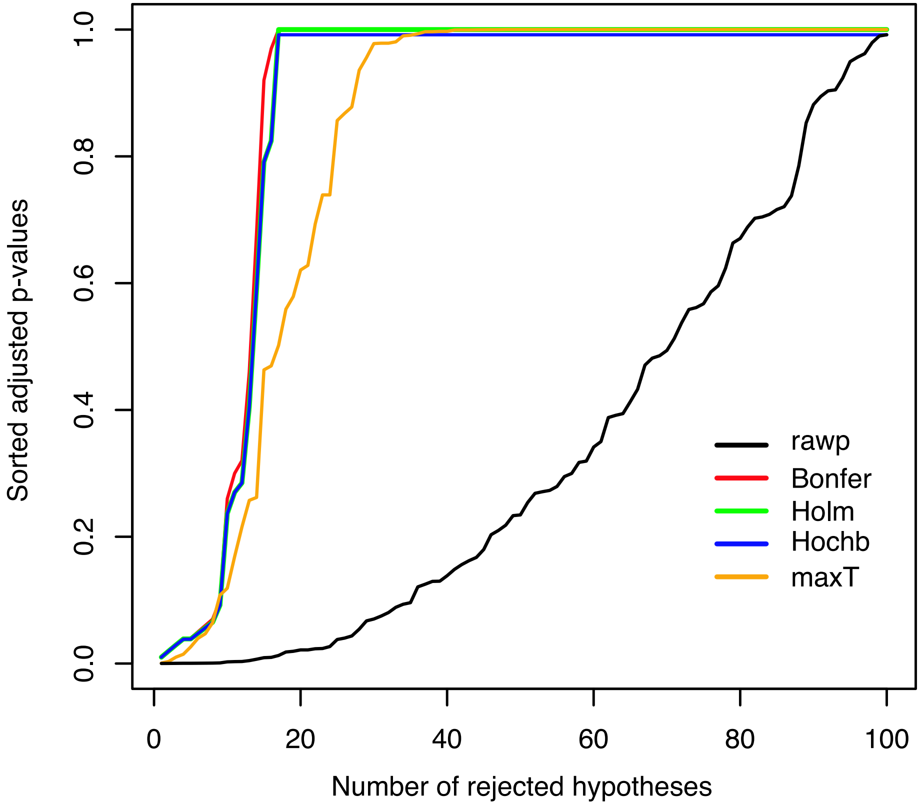


Golub data Bonferroni FWER

98 genes with Bonferroni $\hat{p}_i < 0.05 \leftrightarrow p_i < 0.000016$ (t-test)

FWER: Example

Golub data, 27 ALL vs. 11 AML samples, 3,051 genes.



From the multtest package in Bioconductor.

Comparisons of FWER on 100 p–values

FWER conservative (lack of power) : many interesting genes may be missed.

Controlling the FDR

Suppose that we know $FDR(t)$ for each rejection region $[0, t]$. How can we use these?

- Pick a rejection region$[0, t]$ a priori and note $FDR(t)$.

Controlling the FDR

Suppose that we know $FDR(t)$ for each rejection region $[0, t]$. How can we use these?

- Pick a rejection region$[0, t]$ a priori and note $FDR(t)$.
- A priori pick a level α at which to control the FDR. Take $t^{\alpha} = \max\{t : FDR(t) \leq \alpha\}$. Reject all $p_i \leq t^{\alpha}$. This controls FDR exactly at α.

Controlling the FDR

Suppose that we know $FDR(t)$ for each rejection region $[0, t]$. How can we use these?

- Pick a rejection region$[0, t]$ a priori and note $FDR(t)$.
- A priori pick a level α at which to control the FDR. Take $t^{\alpha} = \max\{t : FDR(t) \leq \alpha\}$. Reject all $p_i \leq t^{\alpha}$. This controls FDR exactly at α.
- Observe all rejection regions simultaneously, i.e. plot $FDR(t)$ versus t.

Controlling the FDR

Suppose that we know $FDR(t)$ for each rejection region $[0, t]$. How can we use these?

- Pick a rejection region$[0, t]$ a priori and note $FDR(t)$.
- A priori pick a level α at which to control the FDR. Take $t^{\alpha} = \max\{t : FDR(t) \leq \alpha\}$. Reject all $p_i \leq t^{\alpha}$. This controls FDR exactly at α.
- Observe all rejection regions simultaneously, i.e. plot $FDR(t)$ versus t.
- Or better, calculate the simultaneous controlling curves: $\alpha_{FDR}(t) = \inf_{s>t} FDR(s)$, which gives the minimum error rate attained when rejecting all p–values in $[0, t]$.

Estimation of the FDR (SAM)

Idea: Depending on the chosen cutoff-value for the p-value p_i of the test statistic T_i, estimate the expected proportion of false positives in the resulting gene list. For a threshold t, one may write

$$FDR(t) = \frac{\pi_0 \cdot t}{\mathbb{P}(p \leq t | R(t) > 0)}$$

where π_0 is the fraction of non-diff. genes among the m.

Estimates:

$$\begin{aligned} \hat{\mathbb{P}}(p \leq t | R(t) > 0) &= \frac{\max(\#\{p_i : p_i \leq t\}, 1)}{m} \\ \hat{\pi}_0 &= ??? \end{aligned}$$

Estimating π_0

We expect the p-values near 1 to be mostly nulls. The number of null p-values expected in $[\lambda, 1]$ is $(1-\lambda)\cdot m_0$. For some "well chosen" λ (automatic ways for that), estimate π_0 by:

$$\hat{\pi}_0 = \frac{\#\{p_i : p_i > \lambda\}}{(1-\lambda)m}$$

Adjusted p–values estimate the FDR SCC at the p-values :

$$\hat{\alpha}_{FDR,\lambda} = \min_{s \geq p_i} F\hat{D}R_\lambda(p_i)$$

Adjusted q–values estimate the FDR SCC at the p-values :

$$\hat{q}_\lambda = \min_{s \geq p_i} pF\hat{D}R_\lambda(p_i)$$

q-value = minimal FDR at which it appears significant.

Enhanced FDR

EFDR is based on controlling FDR, but differs through its reducing of the number of test statistics tested.

The number of hypotheses tested is decreased due to:

- the test-statistic signal $T_i, i = 1, \ldots, m$ is represented parsimoniously and decorrelated in the wavelet domain,

Enhanced FDR

EFDR is based on controlling FDR, but differs through its reducing of the number of test statistics tested.

The number of hypotheses tested is decreased due to:

- the test-statistic signal $T_i, i = 1, \ldots, m$ is represented parsimoniously and decorrelated in the wavelet domain,
- an optimal selection of m^* wavelets is made using an empirical Bayes thresholding procedure.

Enhanced FDR

EFDR is based on controlling FDR, but differs through its reducing of the number of test statistics tested.

The number of hypotheses tested is decreased due to:

- the test-statistic signal $T_i, i = 1, \ldots, m$ is represented parsimoniously and decorrelated in the wavelet domain,
- an optimal selection of m^* wavelets is made using an empirical Bayes thresholding procedure.

The test statistic is significant if $p^w_{(i)} \leq \frac{\alpha \cdot i}{m^*}$. Recall: Bonf. $p^w_{(i)} \leq \frac{\alpha}{m}$; FDR: $p^w_{(i)} \leq \frac{\alpha \cdot i}{m}$.

Some references

Y. Benjamini and Y. Hochberg (1995). Journal of the Royal Statistical Society B, Vol. 57, 289– 300.

S. Dudoit et al. (2002). Statistica Sinica, Vol. 12, 111–139.

J. D. Storey (2002). Journal of the Royal Statistical Society B, Vol. 60, 479–498.

V.G. Tusher et al. (2001). PNAS, Vol. 98, 5116–5121.

P.H. Westfall & S.S. Young (1993). Resampling-based multiple testing: examples and methods for p-value adjustment. Wiley.

Johnstone & Silverman (2001). Empirical Bayes Thresholding.
`http://www.statistics.bristol.ac.uk/`
`~ bernard/ebayesthresh`