False Discovery Rate-Controlled Test Decisions under Correlation in Gene Expression Studies

Michael G. Schimek^{1,2} Tomáš Pavlík²

¹Medical University of Graz, Institute for Medical Informatics, Statistics and Documentation, 8036 Graz, Austria

²Masaryk University, Department of Applied Mathematics and Institute of Biostatistics and Analyses, 66295 Brno, Czech Republic

Mathematical Biosciences Institute, Ohio State University, Statistical Genetics Journal Club Spring 2007

・ 同 ト ・ ヨ ト ・ ヨ ト

Introduction



The microarray technology

- It allows to study expression ('activity') in thousands of genes simultaneously
- We are interested in differences between experimental conditions

Motivation: For many research tasks involving classification and prediction it is necessary to preselect a set of differentially expressed genes Gain: Preselection helps improving the performance of the classifier or predictor Task: Selection of a set of differentially expressed genes

Challenges

- High-dimensionality (thousands of genes)
- Small sample sizes (due to limited availability of cases)
- Genes are co-regulated, hence differential expression can be substantially correlated (Klebanov et al. 2006)
- Insufficient biological background (pathways etc.)
- Stability of gene selection is an issue of increasing importance (Qiu et al. 2006)

When identifying differentially expressed genes via statistical tests we are confronted with a **multiple comparison problem**:

- Using the usual type I error rate α forces the number of false positives to grow enormously
- Need to control the type I error ⇒ the false discovery rate (*FDR*) approach due to Benjamini & Hochberg (1995) is most popular and quite useful

(B)

Parametric test statistics for gene i

Assume that *n* genes (i = 1, ..., n) have been measured over two experimental conditions (j = 1, 2) on K_1 arrays of condition 1 and K_2 arrays of condition 2 and $K_1 + K_2 = K$

 \bar{x}_{i1} and \bar{x}_{i2} mean gene expression for gene *i* under conditions 1 and 2

Standard test statistic

$$t_i = \frac{\bar{x}_{i2} - \bar{x}_{i1}}{s_i}$$

where s_i the pooled standard deviation for gene *i*

$$s_i = \sqrt{\left(\frac{1}{K_1} + \frac{1}{K_2}\right) \frac{\sum_{k_1=1}^{K_1} (x_{ik_1} - \bar{x}_{i1})^2 + \sum_{k_2=1}^{K_2} (x_{ik_2} - \bar{x}_{i2})^2}{K - 2}}$$

Modified test statistic

$$d_i = rac{ar{\mathbf{x}}_{i2} - ar{\mathbf{x}}_{i1}}{\mathbf{s}_i + \mathbf{s_0}},$$

where s_0 is a 'correcting' constant (also called 'fudge factor')

The correcting constant s_0

Motivation: It should make d_i approximately constant as a function of s_i

Detrimental effect: For a given confidence level the constant s_0 can dramatically affect the number of selected genes

- There is the following empirical evidence (Grant et al., 2005):
 - For $s_0 = 0$ (i.e. standard t_i) the d_i is large for a gene with small variance
 - For $s_0 > 0$ this effect is reduced
 - For *s*⁰ too large, expressed genes with small mean difference and/or small variance are obscured in the overall noise
- The effect of s₀ is unknown for co-regulated genes

When nonparametric alternatives are used (e.g. a rank-sum statistic) no s_0 specification needed, the results however less powerful (Schimek and Pavlik, 2006)

Goal

Identify as many differentially expressed genes as possible while incurring a relative low proportion of false positives

Let V be the number of false positives and R be the number of overall rejected hypotheses in a microarray experiment The FDR can be defined as

expectation of the ratio of V and R (have to account for possibility of R = 0)

$$FDR = \mathbf{E}\left(\frac{V}{R}\mathbf{1}_{\{R>0\}}\right).$$

However, it can be shown (Storey & Tibshirani, 2003) that $FDR = \mathbf{E} \left(\frac{V}{R} \right) \approx \frac{\mathbf{E}(V)}{\mathbf{E}(R)}$, which is easier to estimate and implement

(< 2)</p>

FDR estimation procedures: samr

R version of classical SAM procedure (Tusher et al., 2001) Let $t_{(1)} \leq t_{(2)} \ldots \leq t_{(g)}$ be the ordered observed test statistics The expected value for *i*th rank $\overline{t}_{(i)}$ is estimated via the set of *B* permutations of the data matrix Then (A arbitrary but fixed) goods satisfying

Then (Δ arbitrary but fixed) genes satisfying

$$t_{(i)} - \overline{t}_{(i)} \geq \Delta ext{ or } \overline{t}_{(i)} - t_{(i)} \geq \Delta$$

are called 'significant'

 $\widehat{FDR} = \widehat{\pi}_0 \frac{\text{median number of falsely called genes}}{\text{total number of genes called}},$ where $\widehat{\pi}_0 = \frac{\#\{t_i \in (q25, q75)\}}{g/2}$ is the estimated proportion of truly null hypotheses **Disadvantage:**

High memory requirements due to the storage of intermediate results

FDR estimation procedures: siggenes

A variant of the SAM procedure implemented in R (Schwender, Krause and Ickstadt, 2003) Major difference to original SAM: The estimation of the proportion of truly null hypotheses is based on spline smoothing (idea due to Storey & Tibshirani, 2003)

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1, \dots, g\}}{g(1 - \lambda)}, \ \lambda = 0.01, 0.02, \dots, 0.95$$

The final estimate of π_0 is set to $\hat{\pi}_0 = \hat{f}_{\lambda=1}, \hat{f}$ being a natural cubic spline with 3 degrees of freedom of $\hat{\pi}_0(\lambda)$ on λ Advantages:

- Feasible to use larger number of permutations without memory allocation problems
- The user can either decide for the median or the mean value of falsely significant genes obtained from the set of B permutation steps when estimating the FDR

FDR estimation procedures: Grant's procedure

 Idea is to estimate the *FDR* for an adequate set of values that covers the range of observed test statistics, finally picking the value which satisfies the pre-specified *α* level

Let *k* be an arbitrary but fixed value, G_k the set of genes *i* such that $t_i \ge k$, R_k be the size of G_k , and V_k be the number of truly null genes in G_k

 $E(R_k)$ we then estimate with R_k , $E(V_k)$ is estimated using the set $\{V_k^1, V_k^2, \ldots, V_k^B\}$ for a fixed kTaking $\hat{\mu}_k = 1/B \sum_{i=1}^B V_k^i$ for $E(V_k)$ would lead to overestimation; solution due to Grant et al. (2005): iterative algorithm

$$\hat{\mu}_{k}(1) = \frac{\hat{\mu}_{k}}{g}[g - (R_{k} - \hat{\mu}_{k})] \dots \hat{\mu}_{k}(i+1) = \frac{\hat{\mu}_{k}(1)}{g}[g - (R_{k} - \hat{\mu}_{k}(i))]$$
As the final estimate of $\mathbf{E}(V_{k})$ we are using $\hat{\mu}_{k}(n)$, where

 $\hat{\mu}_k(n) - \hat{\mu}_k(n-1) < 0.0001$

Comparison of the FDR estimation procedures

Procedure	Grant's	siggenes	samr	
Estimated formula	$\mathbf{E}(V)/\mathbf{E}(R)$	$\mathbf{E}(V)/\mathbf{E}(R)$	$\mathbf{E}(V)/\mathbf{E}(R)$	
Principle of V dist. estimation	permutations	permutations	permutations	
Type of test statistic	t or modified t	t or modified t	t or modified t	
Automatic s_0 calculation	no	yes	yes	
Proportion of truly null genes	not available	available	available	
Statistic for falsely called genes	mean	mean / median	median	

Questions of interest

- Are there differences in the obtained results (sets of selected genes)?
- Are there differences with respect to power and bias?
- Are there differences in computational costs?
- Can these permutation-based procedures cope well with correlated expression values?

We evaluated the procedures for the **ordinary** and for the **modified SAM** *t*-statistic with the following values of s_0 ('fudge factor'):

• 0, 0.5, 1, and 5, and \hat{s}_0 provided by siggenes and samr

Power, bias and stability of the number of correctly identified genes were studied for fixed *FDR* levels of $\alpha = 0.05$ and 0.1

We adopted the following setting:

- Grant's procedure with 10 000 permutation steps
- siggenes procedure applying the *mean* with 3000 permutation steps
- siggenes procedure applying the *median* with 3000 permutation steps
- samr procedure with 3 000 permutation steps

Simulation study outline continued

For the purpose of comparison an empirical Bayes thresholding (abb. EBT) procedure (no *FDR* control) was used (Johnstone and Silverman, 2004)

- Random thresholding assuming sparse signals (differential expression)
- Prior for each test statistic is mixture of an atom of probability at zero and a double exponential (heavy-tailed) probability
- Minimax squared error properties, hence related to FDR

Common features of artificial expression data

Sample size n = 3000 genes Unexpressed genes: simulated from N(0,1) Expressed genes:

- 100 up-regulated
- 200 down-regulated

in groups of 25 resp. 50

Correlated data generated from $x_{ij} = \sqrt{\rho} * a_j + \sqrt{(1-\rho)} * y_{ij}$, where i = 1, ..., 300, j = 1, ..., 25, $\rho = 0.4$ the assumed correlation, *a* a random vector for each group, and *y* the original vector of simulated values **Model C1** 'simple correlated'

- up-regulated from N(2,1)
- down-regulated from N(-2,1)

Model C2 'complex correlated'

- up-regulated from N(1,1), N(1,2), N(2,1), N(2,2) (25 genes each)
- down-regulated from N(-1,1), N(-1,2), N(-2,1), N(-2,2) (50 genes each)

通りメモトメモト

Simulation study outline continued

Model U1 'simple uncorrelated'

- up-regulated from N($\sqrt{0.4} * 2,1$)
- down-regulated from $N(\sqrt{0.4} * (-2), 1)$

Model U2 'complex uncorrelated'

- up-regulated from N($\sqrt{0.4}$, 1), N($\sqrt{0.4}$, 2), N($\sqrt{0.4}$ * 2, 1), and N($\sqrt{0.4}$ * 2, 2) (25 genes each)
- down-regulated from N($-\sqrt{0.4}$, 1), N($-\sqrt{0.4}$, 2), N($\sqrt{0.4}*(-2)$, 1), and N($\sqrt{0.4}*(-2)$, 2) (50 genes each)

Note that the mean is shifted for comparability with the correlated models

For each setting the sampling was replicated 10 times



Selected simulation results: Fudge factor





M. G. Schimek, T. Pavlík FDR-Controlled Test Decisions under Correlation

Selected simulation results: Fudge factor





ъ

-

2

M. G. Schimek, T. Pavlík FDR-Controlled Test Decisions under Correlation

Selected simulation results: Real *FDR* level for $\alpha = 0.1$



Figure 3

Procedure labels 'A': Grant, Liu and Stoeckert (2005), 'B': siggenes with mean,



'C': siggenes with median, 'D': samr, 'E': EBT

M. G. Schimek, T. Pavlík

FDR-Controlled Test Decisions under Correlation

Selec. sim. results: Correctly picked genes for $\alpha = 0.1$



Figure 4

Procedure labels 'A': Grant, Liu and Stoeckert (2005), 'B': siggenes with mean,



'C': siggenes with median, 'D': samr, 'E': EBT

M. G. Schimek, T. Pavlík

FDR-Controlled Test Decisions under Correlation

Selected simulation results: Model U1

Distribution of the correctly identified genes with respect to the overlap of the FDR procedures and EBT procedure evaluated for FDR levels 0.05 and 0.10 in the UNCORRELATED DATA MODEL 1

	Ov	erlap of the res according to	ults of FD the same	R and EBT pro s ₀ constant u	Overlap of the results of FDR procedures to the results of EBT procedure using s ₀ = 0						
		FDR procedu	res	EBT proced	ure		FDR procedu	ires	EBT proced	dure	
	C	% 50%	100% 0	% 50%	100%	0	% 50%	100% 0	% 50%	100%	
02	s ₀ = 0	100.0%		74.2%	25.8%	s ₀ = 0	100.0%	0.0~	74.2%	25.8%	
- -	$\mathbf{S}_0 \equiv \mathbf{\hat{S}}$	100.0%	0.0%	92.8%	72	$\mathbf{S}_0 \equiv \mathbf{\hat{S}}$	98.8%	1.5%	88.1%	1 <mark>1.9</mark> 9.	
le le	s ₀ = 0.5	100.0%	0.0%	86.1%	3.9%	s ₀ = 0.5	99.8%	0.2%	86.6%	13.49	
e e	s ₀ = 1	99.8%	0.2%	90.9%	9 <mark>.19</mark> .	s ₀ = 1	98.3%	1. %	89.0%	11.0%	
ē	s ₀ = 5	100.0%	0.0%	90.8%	121	s ₀ = 5	97.1%	2.0%	88.8%	11.2	
		FDR procedu	res	EBT proced	ure		FDR procedu	ires	EBT proced	lure	
	a	% 50%	100% 0	% 50%	100%	0	% 50%	100% 0	% 50%	100%	
2	s ₀ = 0	100.0%	0.0%	92.9%	719.	s ₀ = 0	100.0%	0.6%	92.9%	719.	
-	$\mathbf{S}_0 \equiv \mathbf{\hat{S}}$	97.0%	3.0%	99.3%	0.3%	s ₀ = ŝ	95.3%	4.7%	97.6%	2.	
e e	$s_0 = 0.5$	97.9%	2.	97.9%	2. %	s ₀ = 0.5	95.9%	4.1%	97.4%	2.	
e e	s ₀ = 1	98.2%	1.5%	99.1%	0.9%	s ₀ = 1	94.5%	5 <mark>5</mark> %	97.4%	2.	
e	s. = 5	97.7%	2.5%	99.8%	0.2%	s ₀ = 5	92.0%	80%	96.5%	3.	

10%	a	% 50%	100%	0%	50%	100%
4	s ₀ = 0	100.0%	0.6%		92.9%	711
%	$\mathbf{S}_0 \equiv \mathbf{\hat{S}}$	95.3%	4.7%		97.6%	2.
*	s ₀ = 0.5	95.9%	4.1%		97.4%	2
%	s ₀ ≡ 1	94.5%	5.5%		97.4%	2.0%
<u>-</u>	s ₀ = 5	92.0%	8.0%.		96.5%	3.

Percentage of correctly identified genes by the procedure being in the intersection of genes of FDR and EBT procedures Percentage of correctly identified genes by the procedure not being in the intersection of genes of FDR and EBT procedures



Selected simulation results: Model U2

Distribution of the correctly identified genes with respect to the overlap of the FDR procedures and EBT procedure evaluated for FDR levels 0.05 and 0.10 in the UNCORRELATED DATA MODEL 2

	Ove	erlap of the resi according to	ults of FD the same	R and EBT p s ₀ constant	Overlap of the results of FDR procedures to the results of EBT procedure using s ₀ = 0					
		FDR procedu	res	EBT proce	dure		FDR proced	lures	EBT proc	edure
	0	% 50%	100% 0	% 50%	100%	0	% 50%	100% (0% 50%	100
8	s ₀ = 0	100.0%	0.0%	50.3%	49.7%	s ₀ = 0	100.0%	0.0%	50.3%	49.7%
0	s ₀ ≡ ŝ	100.0%	0.0%	77.3%	22.7%	$\mathbf{S}_0 \equiv \mathbf{\hat{S}}$	98.5%	1. %	77.9%	22.1%
le le	s ₀ =0.5	99.5%	0.5%	82.7%	7.3%	s ₀ =0.5	91.0%	9 <mark>.0</mark> 9-	83.7%	16.3%
e.	s ₀ = 1	100.0%	0.0%	82.6%	17.4%	s ₀ = 1	84.5%	5.5%	83.8%	16.2%
ē	s ₀ = 5	100.0%	0.0%	70.8%	29.2%	s ₀ = 5	77.7%	22.3%	76.0%	24.0%
		FDR procedur	res	EBT proce	dure		FDR proced	lures	EBT proc	edure
	0	% 50%	100% 0	% 50%	100%	0	% 50%	100% 0	1% 50%	100
2	s ₀ = 0	100.0%	0.0%	74.3%	25.7%	s ₀ = 0	100.0%	0.0%	74.3%	25.7%
9	s _o = ŝ	95.9%	4	90.9%	9 <mark>.19</mark> .	s ₀ = ŝ	90.9%	9 <mark>.19</mark> .	91.5%	8 <mark>.5</mark> %
le l	s ₀ = 0.5	90.7%	9 <mark>.39</mark> .	93.8%	62	s ₀ =0.5	82.2%	17.8%	94.6%	54.
e e	s ₀ = 1	92.0%	8.09	96.7%	3.8%	s ₀ = 1	76.4%	23.6%	93.1%	69.
ē	s. = 5	99.6%	0.4%	90.8%	9.29	s. = 5	69.1%	30.9%	89.6%	10.4%

	FD	R proce		EBT procedure				
	0%	50%	100%	0%	50%	1005		
$s_0 \equiv 0$		100.0%	0.0%	t	74.3%	25.7%		
$\mathbf{S}_0 \equiv \mathbf{\hat{S}}$		90.9%	9 <mark>.1</mark> %		91.5%	8 <mark>5</mark> 5		
s ₀ = 0.5		82.2%	17.8%		94.6%	54		
s ₀ = 1		76.4%	23.6%		93.1%	69.		
s ₀ = 5	_	69.1%	30.9%		89.6%	10.4%		
	$\begin{tabular}{ c c c c c }\hline & s_0 \equiv 0 \\ \hline & s_0 \equiv \hat{s} \\ \hline & s_0 \equiv 0.5 \\ \hline & s_0 \equiv 1 \\ \hline & s_0 \equiv 5 \\ \hline \end{tabular}$	FD 0% $s_0 = 0$ $s_0 = 5$ $s_0 = 1$ $s_0 = 5$	FDR proce 0% 50% s ₀ = 0 100.0% s ₀ = \$ 90.9% s ₀ = 0.5 82.2% s ₀ = 1 76.4% s ₀ = 5 68.1%	FDR procedures 0% 50% 100% so = 0 100.0% 0.4% so = 0 50.0% 0.4% so = 0 50.0% 414 so = 0.5 52.2% 17.8% so = 1 78.4% 22.6% so = 5 68.1% 20.9%	$\begin{tabular}{lllllllllllllllllllllllllllllllllll$	FDR procedures EBT proced 0% 50% 100.0% 50% 50% 100.0% 40% 74.3% 50 50.3% 41% 74.3% 50 50.3% 41% 91.3% 50 50.3% 41% 91.3% 50 50.3% 12% 91.3% 50 50.3% 12% 91.5% 50 50.5% 12% 91.5% 50 50.5% 12% 91.5%		

Percentage of correctly identified genes by the procedure being in the intersection of genes of FDR and EBT procedures

Percentage of correctly identified genes by the procedure not being in the intersection of genes of FDR and EBT procedures

100%



M. G. Schimek, T. Pavlík FDR-Controlled Test Decisions under Correlation

Selected simulation results: Model C1

Distribution of the correctly identified genes with respect to the overlap of the FDR procedures and EBT procedure evaluated for FDR levels 0.05 and 0.10 in the CORRELATED DATA MODEL 1

	Ov	erlap of the res according to	ults of FDF the same	and EBT pro	Overlap of the results of FDR procedures to the results of EBT procedure using s ₀ = 0						
		FDR procedu	ires	EBT procedu	ıre		FDR proced	ures	EBT procedu	ure	
	a	ni 50%	100% 0%	50%	100%	0	% 50%	100% 0	% 50%	100%	
65	s ₀ = 0	100.0%		85.9%	4.19	s ₀ = 0	100.0%	0.0%	85.9%	4.19	
0	s _o ≡ŝ	99.8%	0.2%	91.9%	8.1%	$\mathbf{S}_0 \equiv \mathbf{\hat{S}}$	97.6%	2.8%	92.5%	755	
le l	s ₀ =0.5	100.0%	0.0%	91.8%	829-	s ₀ = 0.5	98.4%	1.	92.9%	711-	
÷	s ₀ = 1	100.0%	0.0%	91.9%	8.1%	s ₀ = 1	97.5%	2.	92.2%	787	
ē	s ₀ = 5	100.0%	0.0%	91.3%	8.7%	s ₀ = 5	95.6%	4.4%	90.5%	45%	
		FDR procedu	res	EBT procedu	ire		FDR procedu	ires	EBT procedu	Jre	
		ns 50%	100% 05	50%	100%		~ 50%	100% 01	6 50%	100%	
2	$s_0 \equiv 0$	99.8%	0.2%	98.2%	1.5%	$s_0 \equiv 0$	99.8%	0.2%	98.2%	1.5%	
°.	$\mathbf{S}_0 \equiv \mathbf{S}$	97.8%	2 %	99.2%	0.1%	$\mathbf{S}_0 \equiv \mathbf{\hat{S}}$	94.3%	5 <mark>7</mark> %	98.6%	1. %	
e e	s ₀ = 0.5	97.9%	2.	98.9%	1. %	s ₀ = 0.5	94.8%	5 <mark>2</mark> %	98.6%	1. %	
e e	s ₀ = 1	97.6%	2.	99.6%	0.9%	s ₀ = 1	94.1%	5 <mark>9</mark> 6	98.2%	1.	
ē	s _n = 5	98.9%	1. %	98.9%	1. %	s _n = 5	93.3%	6.7%	96.6%	3.	

dure		FDR procedu	res	EBT procedure
100%	0%	50%	100% 0%	50%
1.5%	s ₀ = 0	99.8%	0.5%	98.2%
0.1%	s ₀ ≡ ŝ	94.3%	5 <mark>7</mark> %	98.6%
1.5	s ₀ = 0.5	94.8%	52%	98.6%
0.4%	s ₀ = 1	94.1%	59%	98.2%
1. %	s ₀ = 5	93.3%	6.7%	96.6%

Percentage of correctly identified genes by the procedure being in the intersection of genes of FDR and EBT procedures Percentage of correctly identified genes by the procedure not being in the intersection of genes of FDR and EBT procedures



Selected simulation results: Model C2

Distribution of the correctly identified genes with respect to the overlap of the FDR procedures and EBT procedure evaluated for FDR levels 0.05 and 0.10 in the CORRELATED DATA MODEL 2

	00	according to	the same	s ₀ constant u	results of EBT procedure using s ₀ = 0						
		FDR procedu	res	EBT proce	procedure FDR			R procedures		EBT procedure	
	C	ni 50%	100% 0	% 50%	100%	0%	50%	100% 0	% 50%	10	
8	s ₀ = 0	100.0%	0.0%	66.2%	33.8%	s ₀ = 0	100.0%		66.2%	33.8%	
0	$s_0 \equiv \hat{s}$	100.0%	0.0%	80.9%	19.1%	s ₀ ≡ ŝ	93.7%	6 <mark>3</mark> %	82.6%	17,4%	
Ne l	s ₀ =0.5	99.5%	0.5%	83.0%	7.0%	s ₀ =0.5	92.1%	7 <mark>99</mark> 4	82.6%	17.4%	
÷	s ₀ = 1	99.6%	0.%	83.3%	16.7%	s ₀ = 1	89.0%	1 <mark>1.0</mark> %	81.7%	18.3%	
ē	s ₀ = 5	100.0%	0.0%	79.5%	20.5%	s ₀ = 5	82.5%	7.5%	75.7%	24.3%	
	c	FDR procedu	res 100% 0'	EBT proced	Jure 100%	0%	FDR procedu	100% 01	EBT proces	dure 10	
5	s _n = 0	100.0%	0.0%	90.8%	\$ 25	s _n = 0	100.0%		90.8%	42	
Ö	s ₀ = \$	91.5%	8.5%	96.7%	3.8%	S ₀ = S	84.9%	5.19	94.1%	59	
1	s _o = 0.5	89.8%	10.2%	97.5%	2.0%	s ₀ = 0.5	82.0%	18.0%	95.3%	47	
e e	s ₀ = 1	88.8%	1.24	98.8%	1.96	s ₀ = 1	79.4%	20.6%	93.0%	70	
ē	s, = 5	93.8%	62%	97.8%	2.5%	s. = 5	75.0%	25.0%	90.1%	1.91	

lure		FI	DR proced	ures	E	BT procedu	re
100%		0%	50%	100%	0%	50%	100%
\$ 2 5	$s_0 \equiv 0$		100.0%	0.0%		90.8%	423
3.8%	s _o ≡ŝ		84.9%	5.1%		94.1%	5 <mark>9</mark> %
2.	s ₀ =0.5		82.0%	18.0%		95.3%	42
1.3%	s ₀ = 1		79.4%	20.6%		93.0%	70
2.5%	s ₀ = 5	-	75.0%	25.0%		90.1%	999

Percentage of correctly identified genes by the procedure being in the intersection of genes of FDR and EBT procedures Percentage of correctly identified genes by the procedure not being in the intersection of genes of FDR and EBT procedures

the

100%



Summary of results and conclusions

- The behaviour of the SAM procedures with respect to power and bias is quite uniform
- An adequate choice of the correcting constant s₀ can improve the gene selection process, at least for the simple data models
- The automatic SAM choice of s₀ can be far from optimal
- The complexity of the data is definitely more relevant than the presence of correlation
- Empirical Bayes thresholding tends to outperform the SAM procedures at the cost of too large real *FDR* levels
- The behaviour of empirical Bayes thresholding can be further improved (bias reduction) for $s_0 > 0$

Summary of results and conclusions continued

- Grant's procedure requirers substantially more permutation steps compared to the other techniques and cannot be recommended
- The permutation-free **empirical Bayes thresholding** procedure is by far the **most efficient one** (recommended for huge data sets and screening purposes)
- The original SAM procedure performs reasonably well for the simple data model, even under correlation, but not for the complex data model
- The number of correctly identified genes interacts with the type of procedure and the specified *FDR* level ($\alpha = 0.1$ recommended, EBT approximates this value)

