

On multiple testing in bioinformatics and genetics

Ola Hössjer
Dept. of Mathematics
Division of Mathematical Statistics
Stockholm University

Multiple testing

Test large number m of null hypotheses H_1, \dots, H_m with

$$T_x \geq t \Rightarrow \text{reject } H_x.$$

	# not rejected	# rejected	
# true H_x	U	V	m_0
# false H_x	T	S	m_1
	W	R	m

How to choose threshold t ?

Either: **many false rejections** (V large)

or: **low power** (S small).

Note: More generally, t may depend x through the ordered $\{T_x\}$, as for step-down or step-up procedures (as will be seen below).

What to control?

Per-family error rate:	$\text{PFER} = E(V)$
Family-wise error rate:	$\text{FWER} = P(V > 0)$
False discovery rate:	$\text{FDR} = E\left(\frac{V}{R}1_{\{R>0\}}\right)$
Positive false discovery rate	$\text{pFDR} = E\left(\frac{V}{R} \mid R > 0\right)$
False discovery proportion:	$\text{FDP} = \frac{V}{R}1_{\{R>0\}}$

In general

$$\begin{aligned}\text{FDR} &\leq \text{FWER} \leq \text{PFER} \\ \text{FDR} &\leq \text{pFDR}\end{aligned}$$

Given a criterion C (= PFER, FWER, FDR or pFDR),

$$C \leq \alpha$$

means **controlling C at level α** .

Note: FDR and pFDR virtually identical when $P(R > 0) \approx 1$.

What kind of control?

Define $\Omega = \{1, \dots, m\}$ and

$$H_A = \bigcap_{x \in A} H_x$$

for any $A \subset \Omega$.

Weak control of C :

$$C \leq \alpha |H_\Omega$$

Strong control of C :

$$C \leq \alpha |H_A \text{ for all } A \subset \Omega.$$

Exact control of C :

$$C \leq \alpha |H_{A_0},$$

where A_0 is (unknown) set of true null hypotheses (hence $|A_0| = m_0$).

Applications in bioinformatics

Application	x	Control	C	Calculation of p -values
Microarray	gene	strong	FDR	simulation
Linkage	chromosome location	weak	FWER	simulation/ analytical
Association	chromosome location	weak/ strong	FWER/ FDR	simulation
Sequence alignment	local alignment	weak	FWER	simulation/ analytical

Application	H_x
Microarray	No differential expression of x between (e.g.) cases and controls
Linkage	Random (=Mendelian) segregation of alleles at x
Association	No association between phenotype and gene at x
Sequence	No similarity between aligned substrings according to x

Rules of Thumb

Type of control:

- **Weak:** Want to check if *any* H_x is false.
- **Strong:** Want to know exactly which H_x are false.

Type of criterion:

- **FWER:** $m_1 = 0$ likely
- **FDR, pFDR, FDP:** $m_1 = 0$ unlikely

p -values: Weak control

Raw p -values:

$$\begin{aligned} p_x &= \text{smallest } \alpha \text{ for which } H_x \text{ is rejected} \\ &\quad \text{based on } T_x \text{ only at level } \alpha \\ &= P(T_x \geq t_{x,\text{observed}} | H_x) \leq \alpha \end{aligned}$$

Adjusted p -value:

$$\tilde{p} = \text{smallest } \alpha \text{ for which } H_\Omega \text{ is rejected} \\ \text{at level } C \leq \alpha | H_\Omega$$

What is rejected?

$$H_\Omega \text{ is rejected if } \tilde{p} \leq \alpha.$$

Test statistics, weak control

Look at most significant test:

$$\max_x T_x \text{ or } \min_x p_x$$

Fisher:

$$-2 \sum_x \log(p_x)$$

Hybrid:

$$-2 \sum_{x; p_x \leq \gamma} \log(p_x)$$

for some threshold $0 < \gamma < 1$.

Note 1: When $\{p_x\}$ are independent under H_Ω , $\min p_x$ is distributed as minimum order statistic and Fisher's test statistic as $\chi^2(2m)$.

Note 2: When $\{p_x\}$ are dependent under H_Ω , distribution of test statistics often determined by permutation or resampling.

p -values: Strong control

Raw p -values:

$$\begin{aligned} p_x &= \text{smallest } \alpha \text{ for which } H_x \text{ is rejected} \\ &\quad \text{based on } T_x \text{ only at level } \alpha \\ &= P(T_x \geq t_{x,\text{observed}} | H_x) \leq \alpha \end{aligned}$$

Adjusted p -values for *individual tests* x :

$$\tilde{p}_x = \text{smallest } \alpha \text{ for which } H_A \text{ is rejected} \\ \text{at level } C \leq \alpha | H_A \text{ for all } A \ni x$$

What is rejected?

$$\{H_x; \tilde{p}_x \leq \alpha\}$$

Note: If $\min p_x$ test statistic is used for weak control, then $\tilde{p} \leq \min_x \tilde{p}_x$, i.e. weak control is more powerful than strong control under H_Ω , given the same criterion. However, often $\tilde{p} = \min_x \tilde{p}_x$, and then strong control is 'for free'.

Single Step Methods

Calculation of \tilde{p}_x is only based on p_x :

Bonferroni:

$$\tilde{p}_x = \min(mp_x, 1)$$

Sidak:

$$\tilde{p}_x = 1 - (1 - p_x)^m$$

minP:

$$\tilde{p}_x = P\left(\min_{1 \leq y \leq m} P_y \leq p_x | H_\Omega\right)$$

Note 1: All give strong control of FWER when $\min_{x \in A} p_x$ is test statistic for testing H_A .

Note 2: For minP, $\{P_y\}$ are random variables corresponding to raw p -values. In general, minP is implemented, using simulation, to account for dependency structure between different P_y .

Step-Down Procedures

Reject sequentially hypotheses $H_{(1)}, \dots, H_{(m)}$ based on ordered raw p -values $p_{(1)} \leq \dots \leq p_{(m)}$, with some stopping rule.

Holm:

$$\tilde{p}_{(x)} = \max_{y=1, \dots, x} \min((m - y + 1)p_{(y)}, 1)$$

Step-down Sidak:

$$\tilde{p}_{(x)} = \max_{y=1, \dots, x} \left(1 - (1 - p_{(y)})^{(m-y+1)} \right)$$

Step-down minP:

$$\tilde{p}_{(x)} = \max_{y=1, \dots, x} P\left(\min_{y \leq z \leq m} P_{(z)} \leq p_{(y)} \mid H_{\Omega} \right).$$

Note 1: All give strong control of FWER, but are more powerful than corresponding single step methods.

Note 2: For step-down minP: $P_{(1)} \leq \dots \leq P_{(m)}$ are the ordered $\{P_y\}$.

Step-Up Procedures

Accept sequentially hypotheses $H_{(m)}, \dots, H_{(1)}$ based on ordered raw p -values $p_{(1)} \leq \dots \leq p_{(m)}$, with some stopping rule.

BH procedure:

$$\tilde{p}_{(x)} = \min_{y=x, \dots, m} \min\left(\frac{m}{y} p_{(y)}, 1\right).$$

Note 1: Gives strong control of FDR.

Note 2: The procedure works for independent tests statistics under null hypothesis, but can be generalized to allow for certain kinds of dependency structures.

Bayesian Approach and Classification

Write $H_x = 0$ (1) when H_x is true (false).

Assume $\{H_x\}$ are identically distributed with

$$\pi_0 = P(H_x = 0),$$

so $\hat{\pi}_0 = m_0/m$ ideal estimate of π_0 . Assume

$$p_x \leq \gamma \Rightarrow \text{Reject } H_x$$

This is Bayesian classification problem with classification rule

$$\begin{aligned} p_x \in [0, \gamma] &\Rightarrow \text{Accept } H_x \\ p_x \in (\gamma, 1] &\Rightarrow \text{Reject } H_x \end{aligned}$$

Posteriors and pFDR

Assume

$$\begin{aligned} p_x | H_x = 0 &\sim F_0 = U(0, 1) \\ p_x | H_x = 1 &\sim F_1 \\ p_x &\sim F = \pi_0 F_0 + (1 - \pi_0) F_1 \end{aligned}$$

We have

$$\text{pFDR}(\gamma) = P(H_x = 0 | \text{Reject } H_x) = \frac{\pi_0 F_0(\gamma)}{F(\gamma)}$$

and posterior probability

$$\begin{aligned} P(H_x = 0 | p_x = u) &= \frac{\pi_0 f_0(u)}{f(u)} \\ &=: \text{FDR}(u) = \text{local FDR at } u, \end{aligned}$$

where $f_0 = F'_0$ and $f = F'$.

Bayesian Risk Functions

With Bayesian analysis, use e.g. risk function

$$\begin{aligned}\text{Risk}(\gamma) &= w \int_0^\gamma \text{FDR}(u) f(u) du \\ &\quad + (1 - w) \int_\gamma^1 (1 - \text{FDR}(u)) f(u) du \\ &= w P(p_x \leq \gamma) \text{pFDR}(\gamma) \\ &\quad + (1 - w) P(p_x > \gamma) \text{pFNR}(\gamma),\end{aligned}$$

where w is user-specified cost parameter and $\text{pFNR}(\gamma) = E(T/W | W > 0)$ is the positive *false non-discovery-rate*.

Or, a more simplified version,

$$\text{Risk}(\gamma) = w \text{FDR}(\gamma) + (1 - w) \text{FNR}(\gamma).$$

Empirical Bayes and FDR Procedures

With pFDR or Bayesian criterion, we need to

1. Compute $\hat{\pi}_0$
2. Compute \hat{F} or \hat{f}

and plug into pFDR(γ) or Risk(γ). Choose either

$$\hat{\gamma} = \max\{\gamma; \widehat{\text{pFDR}}(\gamma) \leq \alpha\}$$

or

$$\hat{\gamma} = \arg \min_{\gamma} \widehat{\text{Risk}}(\gamma).$$

Note 1: When $\widehat{\text{pFDR}}$ is used $\hat{\pi}_0$ should be an upward conserv. estimate of π_0 .

Note 2: With hierarchical Bayesian approach, π_0 and (parameters of) f_1 are random, with certain hyperparameters.

Note 3: If $\{T_x\}$ are identically distributed, one may also use $G_0 = \mathcal{L}(T_x|H_x = 0)$ and $G_1 = \mathcal{L}(T_x|H_x = 1)$. Often G_0 and G_1 are easier to parametrize.

References:

General:

Benjamini and Hochberg. (1995). FDR is introduced. *J. Royal. Stat. Soc. B* **57**, 289-300

Genovese and Wasserman. (2002). Introduces FNR and analogies with Bayesian risk functions. *J. Royal. Stat. Soc. B* **64**(3), 499-517.

Holm. (1979). Seq. rejective Bonferroni. *Scand. J. Stat.* **6**, 65-70.

Korn et al. (2004). Introduce and control $C = \text{FDP}$ and generalize FWER to control $C = P(V > k)$ for any k . *J. Statist. Pl. Inf.* **124**, 379-398.

Marcus et al. (1976). Introduction of 'closed test principle' for strong control. *Biometrika* **63**(3), 655-660.

Storey. (2002). pFDR and q -values are introduced. *J. Royal. Stat. Soc. B* **64**(3), 479-498.

Westfall and Young. (1993). Book on *Resampling based multiple testing*, Wiley.

Zaykin et al. (2002). Hybrid of Fisher's and $\min p_x$ methods for weak control. *Genetic Epidemiology* **22**, 170-185.

Microarray:

Dudoit et al. (2002). Multiple testing for cDNA microarray experiments. *Statistica Sinica* **12**(1), 111-139.

Efron et al. (2001). Empirical Bayes analysis for Microarrays. *J. Amer. Statist. Assoc.* **96**, 1151-1160.

Ge et al. (2003). Overview of multiple testing and resampling-based algorithms with appl to microarrays (with discussion). *Test* **12**(1), 1-77.

Lönnstedt and Britton. (2004). (Full) Hierarchical Bayesian criterion for multiple testing (D -statistic). To appear in *Biostatistics*.

Lönnstedt and Speed. (2002). (Empirical) Bayesian criterion for multiple testing (B -statistic). *Statistica Sinica* **12**, 31-46.

Linkage:

Feingold et al. (1993). Applies extreme value theory for Gaussian processes for calculating FWER. *Am. J. Hum. Genet.* **53**, 234-251.

Lander and Kruglyak. (1995). Guidelines for reporting linkage results based on PFER. *Nature Genetics*, **11**, 241-247.

Ängquist and Hössjer. (2004). Extreme value theory for subordinated Gaussian processes for calculating FWER. To appear in *Biostatistics*.

Ängquist and Hössjer. (2004). Importance sampling for calculating FWER. *Stat. Appl. Gen. Mol. Biol.*, **3**(1), article 5.

Association:

Clayton. (2003). Invited talk at European Mathematical Genetics Meeting, Cambridge. <http://www-gene.cimr.cam.ac.uk/clayton/talks/>

Devlin and Roeder. (1999). Bayesian multiple testing for association studies. Genomewide scan to control for spurious association.

Biometrics **55**, 997-1004.

Lazzeroni and Lange. (1998). Conditional inference and adjusted p -values for TDT type tests. *Human Heredity* **48**, 67-81.

Nilsson. (2001). Use of FWER in association studies. PhD Thesis, Chalmers.

Westfall et al. (2002). Use of FWER in association studies. *Methods Mol. Biol.* **184**, 143-163.

Sequence Alignment:

Hansen. (2004). Markov controlled excursions (for alignment and structure prediction). PhD Thesis. Copenhagen University.

Hössjer. (2004). Short introduction/overview to multiple testing for sequence alignment (in Swedish). *Qvartilen*, **19**(2), 10-13.

Karlin and Dembo. (1992). Limit distribution of maximal alignment scores. *Adv. Appl. Prob.* **24**, 113-140.

Siegmund and Yakir. (2000). p -values for alignments with gaps. *Ann. Statist.* **28**(3), 657-680. Correction in **31:3**, 1027-1031.