Microarray Data Analysis (IV)

Multiple Testing

Hypotheses

- Define null hypothesis (H_0) and alternative hypothesis (H_1)

Example:

Are the expression levels of a gene the same

in two treatments?

H_0: the gene has same expression level.

H_1: the gene has different expression levels.

Steps of Hypothesis Testing

1. Determine the null and alternative hypothesis, using mathematical expressions if applicable.

2. Select a significance level (α).

3. Take a random sample from the population of interest.

4. Calculate a test statistic from the sample that provides information about the null hypothesis.

5. Decision
   - If the value of the statistic is consistent with the null hypothesis then do not reject H0.
   - If the value of the statistic is not consistent with the null hypothesis, then reject H0 and accept the alternative hypothesis.

		Test Conclusion	
		H_0	H_1
Reality	H_0	true negative	false positive (Type I error α)
	H_1	false negative (Type II error β)	true positive

H_0: the gene has same expression level.

H_1: the gene has different expression levels.

	Condition A				Condition B			
	rep1	rep2	rep3	rep4	rep1	rep2	rep3	rep4
g_1	x_{11}	x_{12}	x_{13}	x_{14}	y_{11}	y_{12}	y_{13}	y_{14}
g_2	x_{21}	x_{22}	x_{23}	x_{24}	y_{21}	y_{22}	y_{23}	y_{24}
…	…	…	…	…	…	…	…	…
g_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	y_{i1}	y_{i2}	y_{i3}	y_{i4}
…	…	…	…	…	…	…	…	…
g_n	x_{n1}	x_{n2}	x_{n3}	x_{n4}	y_{n1}	y_{n2}	y_{n3}	y_{n4}

Which genes are differentially expressed?

$H_0^{(1)}$: gene *1* has same expression level in both conditions
$H_0^{(2)}$: gene *2* has same expression level in both conditions
.....

$H_0^{(i)}$: gene *i* has same expression level in both conditions
.....
$H_0^{(n)}$: gene *n* has same expression level in both conditions

n = 6,000

Testing 6,000 gene-wise null hypotheses simultaneously!

Multiple Testing

- At a give significance level α,
  - For one test:

    Prob(making Type I error) = α

    Prob(Not making Type I error) = $1- \alpha$

  - For *n independent* tests:

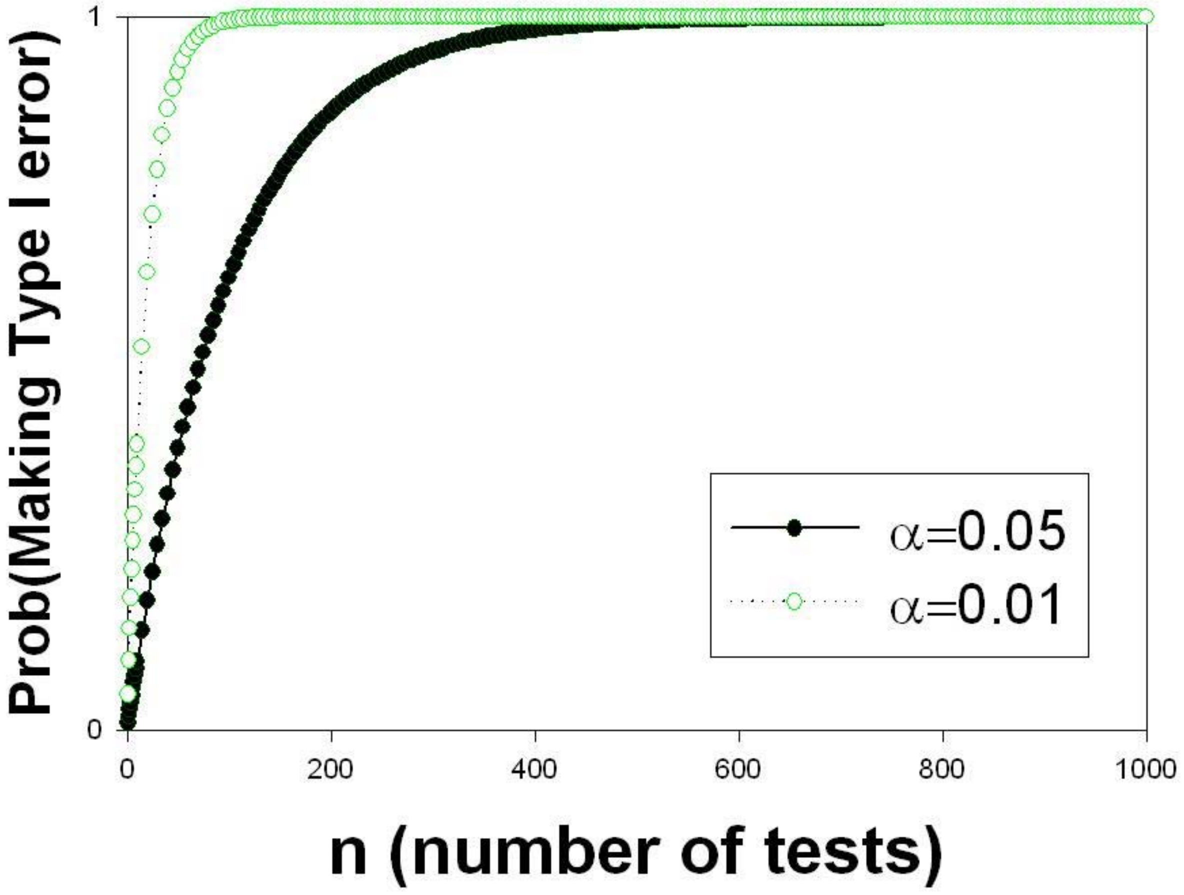
    Prob(Not making Type I error)

    = Prob(Not making Type I error for any test)

    = $(1- \alpha)^n$

    Prob(making Type I error for at least one test) = $1- (1- \alpha)^n$

problematic



α=0.01, n =100

Prob(making Type I error for at least one test) = 0.634 >> 0.01

Suppose out of the 6,000 genes, 100 are truly differentially expressed (i.e. they are true positives).

- α = 0.01, there are 6000 x 0.01 = 60 genes that are false positives, therefore, for the 160 reported genes that are differentially expressed in the two conditions, 37.5% are false positives.
- α = 0.05, 6000 x 0.05 = 300 false positive (75%).

The power of hypothesis testing is weakened/lost because too many tests are performed simultaneously.

➔ impose more stringent α values for individual tests so that the family-wise error rate (FWER) is about α

FWER (Family-wise Error Rate)

- Probability of making at least one Type I error when all null hypotheses are true. Let α represent this family-wise (Type I) error rate.
  - α is usually 0.01 or 0.05.
  - Each individual test uses more stringent Type I error rate.
- FWER methods:
  - Bonferroni correction (one-step)
  - Sidak correction
  - Holm's step-down version of Bonferroni correction
  - Other methods not covered (minP, maxT, etc)

- ❑ **Bonferroni correction (one-step)**
  - ➢ Individual tests use Type I error: α/n
- ❑ **Sidak correction**
  - ➢ Individual tests use Type I error: $1-\sqrt[n]{1-\alpha}$

If α=0.01, n=6000, α/n=1.667 x $10^{-6.}$ This means:

If we are testing the *n* hypotheses (*i*=1,2,...n)

$H_0^{(i)}$: gene *i* has same expression level in both conditions.

The probability that we make Type I error for any test is 1.667 x 10^{-6} and the expected number of false positive for all tests is 0.01. So it is extremely unlikely that a gene determined to be differentially expressed actually has the same expression level in the two conditions.

❑ Holm’s Step-down

- ➢ Use different Type I error rates for individual tests
- ➢ Less conservative, more powerful
- ➢ Use ordered *P*-values (hence genes are also ordered)

Step 1: Let $p_{(1)}, p_{(2)}, \ldots , p_{(n)}$ denote the *n* *p*-values ordered from smallest to largest.

Step 2: Find the largest integer *k* so that $p_{(i)} \leq \alpha/(n-i+1)$ for all $i = 1,\ldots,k$.

- – If no such *k* exists, set $c = 0$ (declare nothing significant).
- – Otherwise set $c = p_{(k)}$ (reject the nulls corresponding to the smallest *k* *p*-values).

Still, the expected number of false positive for all tests is α.

	Hypothesis	*P*-value (**ordered incrementally**)	Type I Error
g_1	$H_0^{(1)}$	p_1	α/n
g_2	$H_0^{(2)}$	p_2	$\alpha/(n\text{-}1)$
…	…	…	..
g_i	$H_0^{(i)}$	p_i	$\alpha/(n\text{-}i+1)$
…	…	…	…
g_n	$H_0^{(n)}$	p_n	α

An Example

- Suppose we conduct 5 tests and obtain the following *p*-values for tests 1 through 5.

Test	1	2	3	4	5
p-value	0.042	0.001	0.031	0.014	0.007

- Which tests' null hypotheses will you reject if you wish to control the FWER at level 0.05?

- Use both the Bonferroni method, Sidak method and the Holm method to answer this question.

Solution

Test	1	2	3	4	5
p-value	0.042	0.001	0.031	0.014	0.007

- The cutoff for significance is c = 0.05/5=0.01 using the Bonferroni method. Thus we would reject the null hypothesis for tests 2 and 5.

- The cutoff for significance is c = 0.0102 using the Sidak method. We would reject the null hypothesis for tests 2 and 5 as well.

$0.001 \leq 0.05/(5-1+1)=0.01$
$0.007 \leq 0.05/(5-2+1)=0.0125$
$0.014 \leq 0.05/(5-3+1)=0.0167$
$0.031 > 0.05/(5-4+1)=0.025$
$0.042 \leq 0.05/(5-5+1)=0.05$

- These calculations indicate that Holm’s method would reject null hypotheses for tests 2, 5, and 4.

Summary of FWER

- Focuses on the occurrence, not the number, of false positive.

  α = Probability of making *at least* one Type I error when all null hypotheses are true

- It does NOT consider the effect of the alternative hypothesis.

  If out of 100 genes identified to be differentially expressed, 50 are true positives, it is perfectly fine for experimentalists.

$\Rightarrow$ FWER is being replaced by False Discovery Rate (FDR) methods in very large datasets.

A Conceptual Description of FDR

- Suppose a scientist conducts 100 independent microarray experiments.
- For each experiment, the scientist produces a list of genes declared to be differentially expressed by testing a null hypothesis for each gene.
- For each list consider the ratio of the number of false positive results to the total number of genes on the list (set this ratio to 0 if the list contains no genes).
- The FDR is approximated by the average of the ratios described above.

False Discovery Rate (FDR)

	Not rejected hypothesis	Rejected hypothesis	Total
true hypothesis	U	V (false positive)	U+V
false hypothesis	T	S (true positive)	T+S
Total	U+T	R	n

Q = V/R is the ratio of genes falsely classified as differentially expressed.

Define: E(Q) = False Discovery Rate

$Q = 0 \quad$ (if $V=R=0$)

$Q = V/R \quad$ (if $R>0$)

FDR: expected proportion of false positive among the rejected hypotheses.

False Discovery Rate (FDR)

- FDR methods:
  - Benjamini-Hochberg step-up method
  - Benjamini-Yekutieli step-up method
  - Permutation methods (not covered)

- **Benjamini-Hochberg (BH) step-up method**
  - Specify false discovery rate r *(0<r<1, e.g. r=0.25)*
  - Assume the n tests are *independent* or there are positive regression dependence between tests.
  - Computes Q-value : $q_i = ir/n$

Let $p_{(1)}, p_{(2)}, \ldots , p_{(n)}$ denote the n p-values ordered from smallest to largest. Find the largest integer k so that

$$p_{(k)} \leq q_k = \mathrm{k}r/n.$$

- If no such k exists, set $c = 0$ (declare nothing significant).
- Otherwise set $c = p_{(k)}$ (reject the nulls corresponding to the smallest k p-values).

	Hypothesis	*P*-value (**ordered incrementally**)	*Q*-value
g_1	$H_0^{(1)}$	p_1	$q_1 = r/n$
g_2	$H_0^{(2)}$	p_2	$q_2 = 2r/n$
…	…	…	..
g_i	$H_0^{(i)}$	p_i	$q_i = ir/n$
…	…	…	…
g_n	$H_0^{(n)}$	p_n	$q_n = r$

Our Example Revisited

- Suppose we conduct 5 tests and obtain the following *p*-values for tests 1 through 5.

Test	1	2	3	4	5
p-value	0.042	0.001	0.031	0.014	0.007

- Which tests’ null hypotheses will you reject if you wish to control the FDR at level 0.05?

- Use the Benjamini and Hochberg (1995) method to answer this question.

Solution

Test	1	2	3	4	5
p-value	0.042	0.001	0.031	0.014	0.007

$0.001 \leq 1*0.05/5=0.01$

$0.007 \leq 2*0.05/5=0.02$

$0.014 \leq 3*0.05/5=0.03$

$0.031 \leq 4*0.05/5=0.04$

$0.042 \leq 5*0.05/5=0.05$

The B&H method reject the null hypotheses for all 5 tests.

New Example (p_3 changed slightly)

- Suppose we conduct 5 tests and obtain the following *p*-values for tests 1 through 5.

Test	1	2	3	4	5
p-value	0.042	0.001	0.041	0.014	0.007

- Which tests' null hypotheses will you reject if you wish to control the FDR at level 0.05?

- Use the Benjamini and Hochberg (1995) method to answer this question.

Solution

Test	1	2	3	4	5
p-value	0.042	0.001	0.041	0.014	0.007

$0.001 \leq 1*0.05/5=0.01$
$0.007 \leq 2*0.05/5=0.02$
$0.014 \leq 3*0.05/5=0.03$
$0.041 > 4*0.05/5=0.04$
$0.042 \leq 5*0.05/5=0.05$

The B&H method would still reject the null hypotheses for all 5 tests even though 0.041>0.04.

❑ Benjamini-Yekutieli (BY) step-up method

- ➢Relax the assumption that the *n* tests are independent: arbitrary dependence between genes

- ➢Replace $q_i = ir/n$ by

$$q_i = ir/(n\Sigma(1/j)) \qquad j=1,2\ldots n$$

- ➢More conservative -- ($\Sigma(1/j)$ is a big number for large n)

The First Example

To control the FDR at level 0.05

Test	1	2	3	4	5
p-value	0.042	0.001	0.031	0.014	0.007

$0.001 \leq 0.004$

$0.007 \leq 0.009$

$0.014 > 0.013$

$0.031 > 0.018$

$0.042 > 0.022$

The B&Y method reject the null hypotheses for 2 and 5 tests.

Summary

- Multiple testing is now common in Genomics
- FWER is a framework to control of Type I error but it can be very conservative when there are very large number of tests.
- FDR gives more practical results for multiple testing such as microarray analysis and genome-wide genotyping data

R: multtest

- The multtest package contains a collection of functions for multiple hypothesis testing:
  - mt.teststat: compute test statistics for each row of a data frame.
  - mt.rawp2adjp: compute adjusted p-values from a vector of raw p-values
  - mt.reject: return the identity and number of rejected hypotheses

Related Papers

- S. Dudoit, J. P. Shaffer, and J. C. Boldrick. Multiple hypothesis testing in microarray http://www.bepress.com/ucbbiostat/paper110.
- J. P. Shaffer. Multiple hypothesis testing. Annu. Rev. Psychol., 46:561–584, 1995