Statistical methods for single gene analysis of differential expression

Giorgio Valentini
DSI – Dipartimento di Scienze dell' Informazione
Università degli Studi di Milano
valentini@dsi.unimi.it

1

Comparing two conditions

- Each condition may be represented by one or more RNA samples.
- Using cDNA microarrays, samples can be compared:
  - directly (on the same microarray)
  - indirectly (by hybridizing each sample with a common reference sample)
- Null hypothesis: there is no difference in expression between the conditions
  - Direct comparison: expression ratio should be one
  - Indirect comparison: No difference between test sample and reference sample in the two conditions
- Similar approach with oligonucleotide microarrays.

2

Microarray data

- We assume that the expression levels have been suitably preprocessed ...

X_{jk} is the expression level of gene j in array k

We have N genes and $K = K_1 + K_2$ arrays

$$C_1 = \{X_{jk} \mid 1 \le k \le K_1, 1 \le j \le N\} \quad C_2 = \{X_{jk} \mid K_1 + 1 \le k \le K_1 + K_2, 1 \le j \le N\}$$

	Array1	Array2	...	ArrayK_1	ArrayK_1+1	...	Array$K$
Gene 1	X_{11}	X_{12}	...	X_{1K1}	X_{1K1+1}	...	X_{1K}
Gene 2	X_{21}	X_{22}	...	X_{2K1}	X_{2K1+1}	...	X_{2K}
...	...	...	...	...	...	...	...
Gene n	X_{N1}	X_{N2}	...	X_{NK1}	X_{NK1+1}	...	X_{NK}

3

Fold change

A gene "significantly" changes if its average ratio expression level varies most than a constant factor (De Risi et al., 1997):

The gene j is differentially expressed $\Leftrightarrow$ $\log_2 \frac{\overline{X}_{j(1)}}{\overline{X}_{j(2)}} \ge c$ or $\log_2 \frac{\overline{X}_{j(2)}}{\overline{X}_{j(1)}} \ge c$

where $\overline{X}_{j(1)} = \frac{\sum_{k=1}^{K_1} X_{jk}}{K_1}$ $\overline{X}_{j(2)} = \frac{\sum_{k=K_1+1}^{K_1+K_2} X_{jk}}{K_2}$

Usually c is set 1 (*two-fold* gene expression difference)

4

Fold change drawbacks

- It is not a statistical test (no level of confidence in the designation of genes as differentially expressed or not differentially expressed).
- It is subject to bias if the data have not been properly normalized:

  low-intensity genes may have a larger variance than high-intensity genes and small changes can result significant.
- Intensity-specific thresholds have been proposed as a remedy for this problem (Yang et al. 2002).

5

Two sample t-test (1)

Assumptions:

- two independent normal samples with unequal variances
- Having N genes and $K = K_1 + K_2$ arrays:

$$C_1 = \{X_{jk} \mid 1 \le k \le K_1, 1 \le j \le N\} \quad C_2 = \{X_{jk} \mid K_1 + 1 \le k \le K_1 + K_2, 1 \le j \le N\}$$

The sample means:

$$\overline{X}_{j(1)} = \frac{\sum_{k=1}^{K_1} X_{jk}}{K_1} \qquad \overline{X}_{j(2)} = \frac{\sum_{k=K_1+1}^{K_1+K_2} X_{jk}}{K_2}$$

The sample variances:

$$s^2_{j(1)} = \frac{\sum_{k=1}^{K_1} (X_{jk} - \overline{X}_{j(1)})^2}{K_1 - 1} \qquad s^2_{j(2)} = \frac{\sum_{k=K_1+1}^{K_1+K_2} (X_{jk} - \overline{X}_{j(2)})^2}{K_2 - 1}$$

6

Two sample t-test (2)

- *The t-statistic is*

$$t_j = \frac{\overline{X}_{j(1)} - \overline{X}_{j(2)}}{\sqrt{s^2_{j(1)} / K_1 + s^2_{j(2)} / K_2}}$$

- With d_j degrees of freedom: $d_j \approx K_1 + K_2 - 2$
- or, better: $d_j = \frac{(s^2_{j(1)} / K_1 + s^2_{j(2)} / K_2)^2}{(s^2_{j(1)} / K_1)^2 / (K_1 - 1) + (s^2_{j(2)} / K_2)^2 / (K_2 - 1)}$
- *The t-statistic* follows approximately a Student distribution

7

Two sample t-test (3)

- Reject the null hypothesis (no difference in expression levels) at α significance level $\Leftrightarrow \quad |t_j| > t_{\alpha/2, d_j}$

- **Example**. Test the null hypothesis "There is no difference in the expression level of a gene *j* in two different functional conditions":

1. Compute from the two samples extracted from the population the t-statistic t_j. E.g. t_j=2.785.
2. Compute the degrees of freedom d_j. E.g. d_j = 20.
3. Choose a significance level α. E.g. α = 0.05
4. From the tables of Student probability distribution look for $t_{0.025,20}$=2.086
5. As $t_j > t_{0.025,20}$ then we reject the null hypothesis at α significance level.

8

Esercizi

1. Si può affermare che un gene sia differenzialmente espresso a livello di significatività α=0.01, se nello stato funzionale 1 ha valore medio m_1=0.9 e varianza campionaria s_1=0.9 (6 campioni) mentre nello stato funzionale 2 ha valore medio m_2=0.3 e varianza campionaria s_2=0.3 (5 campioni) ?
2. I livelli di espressione di un gene sono misurati in 28 campioni di tessuti (16 malati e 12 sani). Il valore medio per i tessuti malati è m_m=1.3, per e in quelli sani m_s=0.8 con varianza campionaria rispettivamete s_m=0.3 ed s_s=0.4. Il gene è differenzialmente espresso? Se sì a quale livello di significatività?

9

Advantages and drawbacks of the t-test

- Advantages:
  - It takes into account the variance specific for each gene
  - We can get a p-value
- Disadvantages:
  - If *N* is small (e.g. *N*=4), we can underestimate the variance
  - Instability: if the variance of a gene is small by chance, the *t* value can be large even if the corresponding fold change is small.

⇩

Global t-test (variance pooled across different genes) if the variance is homogeneous between genes (Tanaka et al., 2000). This approach is biased if the assumption of homogeneous variance is violated.

10

Variants of the t-test

- SAM, Significance Analysis in Microarrays (Tusher, Tibshirani & Chu, 2001)
- Regularized t-test (Baldi & Long, 2001)
- B-statistic (Lonnsted and Speed, 2002)

Other approaches ...

- Normal mixture modeling (Pan, 2002)
- Regression modeling (Thomas et al., 2001)

11

SAM, Significance Analysis of Microarrays

- Applied to multiple hypothesis testing
- For binary outcomes it is similar to the t-test, with a correction c_0 for low expression levels:

$$m_j = \frac{\overline{X}_{j(1)} - \overline{X}_{j(2)}}{\sqrt{s^2_{j(1)} / K_1 + s^2_{j(2)} / K_2} + c_0}$$

- To compare m_j across all genes the distribution of m_j should be independent of the level of gene expression
- At low expression levels m_j can be high because of small values of s_j

⇨

- Adding a small value c_0 we could ensure that the variance of m_j is independent of the gene expression level.
- c_0 tries to minimize the coefficient of variation of m_j with respect to s_j

12

A non parametric permutation test (Golub, 1999) (1)

0. N genes and $K = K_1 + K_2$ arrays genes in two functional conditions:

$$C_1 = \{X_{jk} \mid 1 \le k \le K_1, 1 \le j \le N\} \quad C_2 = \{X_{jk} \mid K_1 + 1 \le k \le K_1 + K_2, 1 \le j \le N\}$$

1. For each gene g_j compute the following statistic:

$$a_j = \frac{\overline{X}_{j(1)} - \overline{X}_{j(2)}}{s_{j(1)} + s_{j(2)}}$$

2. Compute the Neighboroods $N_1(r)$ and $N_2(r)$ of radius r

$$N_1(r) = \{g_j \mid a_j > r\} \qquad N_2(r) = \{g_j \mid a_j < -r\}$$

$$-R \le r \le R, \quad R = \max |a_j|$$

13

A non parametric permutation test (Golub, 1999) (2)

3. Perform a permutation test to calculate whether the density of genes in a neighborood is significantly higher than expected:
   - Shuffle m times the class labels in a random way and each time calculate a_rand_j.
   - Calculate the median, the 0.95 a_{95} and 0.99 a_{99} quantile of the a_rand_j empirical distribution for each j

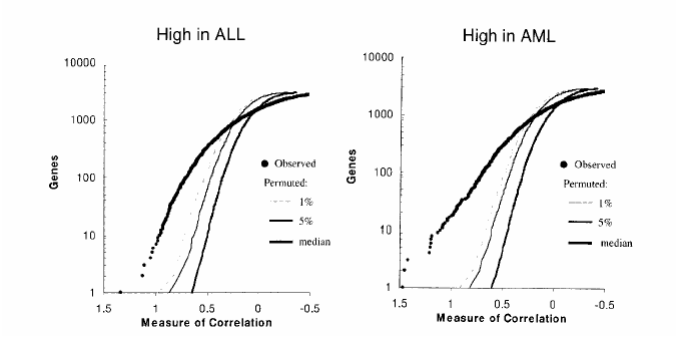
4. If $a_j > a_{95}$ then the difference between the two compared functional conditions of gene gj is significant at 0.05 level.

Hence the set $A_{0.05}$ of genes correlated to the functional condition 1 at 0.05 significance level are:

$$A_{0.05} = \{g_j \mid a_j > a_{95}\} \qquad \text{Analogously:} \qquad A_{0.01} = \{g_j \mid a_j > a_{99}\}$$

14

Neighborood analysis



15

Gene-specific neighborhood analysis

• It is a simple method $O(n \times d)$, n = number of examples, d = number of features (genes) to assess the correlation of genes with tumors.

• It estimates the significance of the matching of a given phenotype to a particular set of marker genes

• The permutation test is distribution independent: no assumptions about the functional form of the gene distribution.

Limits:

It assumes that the expression patterns of each gene are independent

It fails in detecting the role of coordinately expressed genes in carcinogenic processes

16

A filter approach to gene selection: Gene-specific neighborhood analysis

It is a method for gene selection applied before and independently of the induction algorithm (filter method).

It is an equivalent variant of the classic neighborhood analysis proposed by Golub et al. (1999)

1. For each gene the *S2N* ratio c_i is calculated: $c_i = \frac{(m_i^+ - m_i^-)}{(\sigma_i^+ + \sigma_i^-)}$
2. A gene-specific random permutation test is performed:
   i. Generate *n* random permutations of the class labels computing each time the *S2N* ratio for each gene.
   ii. Select a *p* significance level (e.g. $0<p<0.1$)
   iii. If the randomized *S2N* c_rand_i is larger than the actual S2N c_i in less than $p * n$ random permutations, select the i^{th} gene as significant for tumor discrimination at *p* significance level.

17

Hypothesis testing in microarray experiments

Null hypothesis H_g: No differential expression for gene *g*.

Example: Expression levels of gene *g* are not associated with a tumor class, or a treatment, drug, or a clinicl outcome.

α-level test for H_g: a test for which

$P(\text{Reject } H_g \mid H_g \text{ is true}) \leq \alpha \quad (0< \alpha <1)$

α is the *significance level* for the test

The *p-value* for the hypothesis H_g is the value at which one would just reject the null hypothesis:

$p_g = \inf\{\alpha : H_g \text{ is rejected at } \alpha \text{ significance level}\}$

18

Type I and type II error

- *Type I error* or false positive:
  state that a gene is differentially expressed when it is not: i.e. reject a true null hypothesis.

- *Type II error* or false negative:
  fail to identify a truly differentiate expressed gene, i.e. do not reject a false null hypothesis.

19

Multiple hypothesis testing in microarray experiments

- In microarray experiments many hypotheses are tested simultaneously: usually we test for differential expression of thousands of genes.
- In this case the chance of committing type I error (false positive) increases.

Example: The chance P of at least one *p-value* < α for *m* independent tests is:

$$P = 1-(1-\alpha)^m \qquad \text{and} \qquad \lim_{m\to\infty} 1-(1-\alpha)^m = 1$$

For *m* = 1000 and α = 0.01 this chance is 0.9999568

Individual *p-values* of 0.01 no longer correspond to significant findings: we need to *adjust for multiple testing*

20

Family Wise Error Rate (*FWER*)

Given *m* null hypotheses:

H_i={No differential expression for gene *i*}

then

FWER = P{at least one H_i is rejected|H_i is true, $1 \le i \le m$}

that is *FWER is the probability of at least one type I error*.

21

p-value adjustment

- In order to control the FWER, the adjusted *p-value* p_i for the hypothesis H_i is:

  $p_i = \inf\{\alpha : H_i \text{ is rejected at FWER } \alpha\}$

- Hypothesis H_i is rejected at FWER α if $p_i \le \alpha$

22

Adjusted *p-values* for *FWER*: single step procedures

Equal multiplicity adjustments are peformed for all hypotheses, regardless the unadjusted p-values: each hypothesis is evaluated using a critical value that is independent of the results of test of other hypotheses.

Example:

Bonferroni single-step adjusted p-values:

$$p_i = \min(mp_i, 1)$$

where m is the number of genes (hypotheses).

This test is highly conservative (the "true" adjusted p-value is in general much lower)

23

Adjusted *p-values* for *FWER*: stepwise procedures

In order to improve the power of the test, while preserving type I error rate control, *stepwise procedures* use the results of other hypotheses, that is rejection depends on the outcomes of the other procedures

Examples:

- *Holm (1979) step-down adjusted p-values*:

$$p_{r_i} = \max_{k=1,\ldots,i}\left(\min((m-k+1)p_{r_i}, 1)\right)$$

- *Westfall & Young (1993) permutation algorithm for step-down adjusted p-values.*

24

References

- J.Quackenbush Microarray data normalization and transformation, *Nature Genetics* vol.32 december 2002.
- V. Tusher et al. Siginificance analysis of microarrays applied to the ionizing radiation response, *Proc. of the National Academy of Science* vol.98 n. 9, 2001.
- De Risi et al. Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science* 278: 680-685, 1997.
- Golub et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* 286: 531-537, 1999
- Efron, B. et al. Microarrays and their use in a comparative experiment, Tech. Report, Dept. of Statistics, Stanford University, 2000
- Manduchi, E. et al. Generation of patterns from gene expression data by assigning confidence to differentially expressed genes, *Bioinformatics* 16: 685-698, 2000.
- Dudoit, S. et al. Statistical method for identifying differentially expressed genes in replicated cDNA microrarray experiments, *Statistica Sinica* 12(1): 111-139, 2002.
- Dudoit, S. et al. Multiple hypothesis testing in microarray experiments, Biostatistics W.P. Series, University of California, Berkeley, 2002.
- Lehmann E.L. *Testing Statistical Hypotheses*, Springer, New York, 1986

25