Single gene analysis of differential expression

Giorgio Valentini
valenti@disi.unige.it

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.

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}

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 $\Leftrightarrow \log_2 \frac{\overline{X}_{j(1)}}{\overline{X}_{j(2)}} \geq c$ or $\log_2 \frac{\overline{X}_{j(2)}}{\overline{X}_{j(1)}} \geq c$

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)

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).

Two sample t-test (1)

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

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

The sample means:

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

The sample variances:

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

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

Two sample t-test (3)

- Reject the null hypothesis (no difference in expression levels) at α significance level

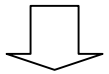
$$\Leftrightarrow \quad |t_j| > t_{a/2,d_j}$$

- **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 $t_{0.25,20}$=2.086
5. As $t_j > t_{0.25,20}$ then we reject the null hypothesis at α significance level.

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.

Variants of the t-test

- SAM, Significance Analysis if 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)

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 variance of 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

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

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

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

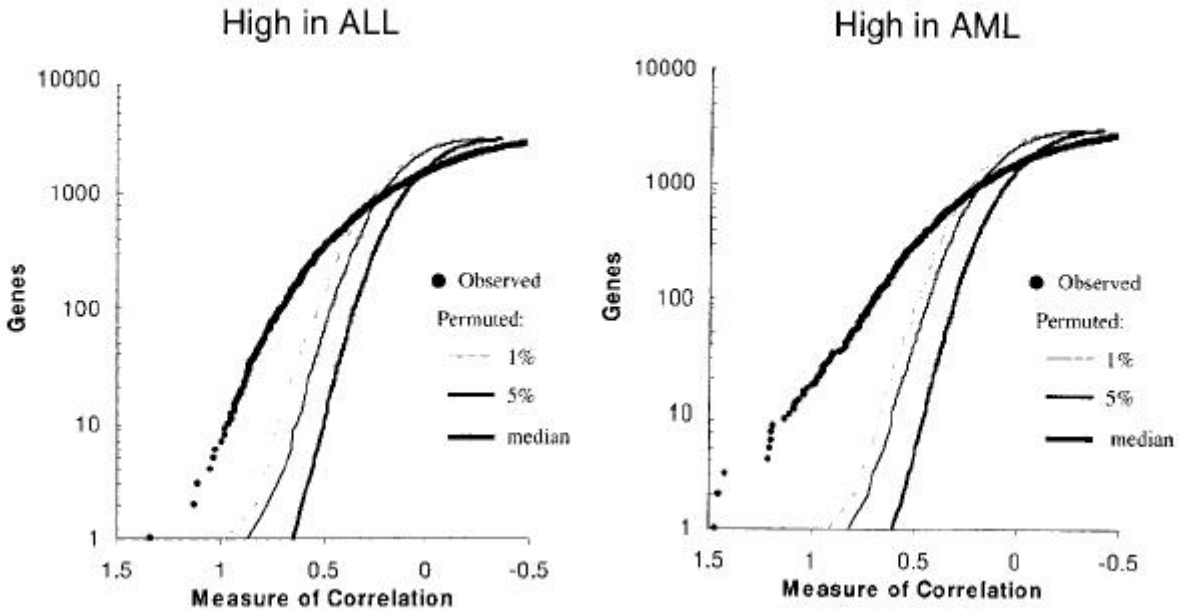
A non parametric permutation test (Golub, 1999) (2)

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_randj*.
   - Calculate the median, the 0.95 a_{95} and 0.99 a_{99} quantile of the *a_randj* empirical distribution for each j
4. If aj > 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:

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

Neighborood analysis

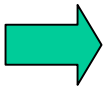


Gene-specific neighborhood analysis

- 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

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: $$c_i = \frac{(m_i^+ - m_i^-)}{(s_i^+ + s_i^-)}$$
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.